JCTC Journal of Chemical Theory and Computation

Special Issue on Polarization

Computer simulations of organic and biomolecular systems consisting of thousands of explicitly represented atoms began in earnest in the 1970s. Molecular dynamics or Monte Carlo statistical mechanics were used to model, for example, liquid water, aqueous solutions of simple molecules and ions, organic liquids, and small proteins in vacuum. A key aspect of the work was the representation of the intra- and intermolecular energetics. In view of the size of the systems and available computer resources, the usual choice was classical force fields that had roots in ‘molecular mechanics’ studies of organic molecules going back to the 1950s. The simulation community was unified on this point, and the general force fields such as AMBER, CHARMM, and OPLS, which arose during the 1980s, adopted nearly identical functional forms. This included the representation of molecules as collections of atom-centered interaction sites with fixed partial charges. The electrostatic energy is then simply determined by the Coulombic interactions between the charged sites. Since the charges are fixed, there is no explicit treatment of electronic polarization, and intermolecular interactions are treated as pairwise additive. Though the impact of this approximation is diminished through the use of effective pair potentials with enhanced charges, the lack of explicit polarization is physically incorrect and is well-known to be problematic for interactions with charge concentrated ions, interactions of ions with π-electron systems, and even for less obvious cases such as polar solutes in low-dielectric media.

Consequently, there has been steady interest since the 1970s in the development and use of polarizable force fields with early work focusing on liquid water and ions in water. Nevertheless, after 30 years and universal agreement on the importance of the problem, generally accepted, broadly applicable polarizable force fields have not emerged, multiple treatments of polarizability (inducible dipoles, fluctuating charges, Drude oscillators, etc.) remain under consideration, and simulations of biomolecular systems with polarizable force fields are still uncommon. Though there is no denying that development and thorough testing of a polarizable force field are a large undertaking, overall, research in the area has taken a back seat to myriad applications of nonpolarizable force fields in modeling ever larger and more complex systems on longer timescales. Though the latter work allows contact with ongoing experiments in molecular biology, medicinal chemistry, and materials science, the impact and prospective capabilities of the simulation work are affected by the quality of the underlying description of molecular energetics. Quantum mechanical treatment of large systems and ab initio molecular dynamics have also advanced during this period and directly incorporate polarization effects; however, they do not provide a general solution as there will always be a class of problems for which use of more rigorous methods is not practical.

In this atmosphere, it was decided to have an issue of the *Journal of Chemical Theory and Computation* with a focus on current research on polarizability and polarizable force fields. Twenty-one articles are included that reflect the state-of-the-art and new developments. They provide a valuable platform for future advances on this important topic.

William L. Jorgensen
Editor-in-Chief, JCTC, Yale University

CT700252G

10.1021/ct700252g CCC: $37.00 © 2007 American Chemical Society
Published on Web 11/13/2007

JCTC Journal of Chemical Theory and Computation

Development of a Polarizable Intermolecular Potential Function (PIPF) for Liquid Amides and Alkanes

Wangshen Xie,[†] Jingzhi Pu,[†] Alexander D. MacKerell, Jr.,*[,‡] and Jiali Gao*[,†]

*Department of Chemistry and Supercomputing Institute, Digital Technology Center,
University of Minnesota, Minneapolis, Minnesota 55455, and Department of
Pharmaceutical Sciences, School of Pharmacy, University of Maryland,
Baltimore, Maryland 21201*

Received June 17, 2007

Abstract: A polarizable intermolecular potential function (PIPF) employing the Thole interacting dipole (TID) polarization model has been developed for liquid alkanes and amides. In connection with the internal bonding terms of the CHARMM22 force field, the present PIPF-CHARMM potential provides an adequate description of structural and thermodynamic properties for liquid alkanes and for liquid amides through molecular dynamics simulations. The computed heats of vaporization and liquid density are within 1.4% of experimental values. Polarization effects play a major role in liquid amides, which are reflected by an increase of 1.5−1.8 D in molecular dipole moment for primary and secondary amides. Furthermore, the computed polarization energies contribute to the total intermolecular interaction energy by 6−24%. The ability of the PIPF-CHARMM force field to treat protein backbone structures is tested by examining the potential energy surface of the amide bond rotation in *N*-methylacetamide and the Ramachandran surface for alanine dipeptide. The agreement with ab initio MP2 results and with the original CHARMM22 force field is encouraging, suggesting that the PIPF-CHARMM potential can be used as a starting point to construct a complete polarizable force field for proteins.

1. Introduction

Molecular mechanical force fields employing effective pairwise potential functions for electrostatic interactions are widely used and have been extremely successful in dynamics simulations of condensed-phase systems and biopolymers.[1,2] Undoubtedly, the most critical factor that determines the reliability of computational results is the accuracy of the potential energy functions. Consequently, there have been continuing efforts devoted to explicitly incorporate many-body polarization effects to further improve the accuracy of these force fields.[3−24] A straightforward approach for treating polarization effects in the current force fields is to include an induction term that depends on the instantaneous positions of the permanent charges and induction polarizations of the rest of the system.[25,26] Our goal is to incorporate explicit polarization terms into the CHARMM22 force field[27] by making adjustments to the nonbonded interaction terms and, at the same time, by minimizing the need for reparametrization of the internal bonding terms. We employ the same approach in the development of these force fields by first studying liquid properties of organic compounds representing different functional groups in proteins.[28] In this paper, we describe a polarizable intermolecular potential function (PIPF) for alkanes and amides.

One practical issue in developing a polarizable force field is that polarization effects are not uniquely described within the framework of classical force fields.[25,26] Thus, it is essential to first decide the functional form and the associated parameters to evaluate the polarization energy. Of course, molecular polarization is well-defined and can be

* Corresponding author e-mail: gao@chem.umn.edu.
[†] University of Minnesota.
[‡] University of Maryland.

10.1021/ct700146x CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/23/2007

properly treated by quantum mechanics (QM),[29,30] but its computational costs prevent it from applications to large molecular systems such as proteins and nucleic acids in aqueous solution.[31,32] Perhaps, the most widely used approach in molecular mechanics is based on the expression[25,26]

$$U_{\mathrm{pol}} = -\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{\mu}_i \cdot \mathbf{E}_i^o \tag{1}$$

where N is the number of interaction sites, $\mathbf{E}_i^o$ is the electric field at the ith atomic site due to the permanent charges of the system, and $\boldsymbol{\mu}_i$ is the induced dipole moment on the ith site. The associated parameters are the atomic polarizabilities which are given as a tensor. There is no rigorous way of defining these atomic polarizability tensors, and contributions due to higher order multipole moments are ignored or implicitly included by parametrization of eq 1. Despite these shortcomings, eq 1 provides a convenient approach to treat inductive polarization effectively as demonstrated by numerous studies in the past.[3−5,9−14,17,18,25,26,33−36] The present PIPF potential has been developed based on eq 1. In other studies, multipole moments have also been included in force field development.[18,19]

A closely related implementation is the Drude oscillator model (also called shell model),[20,37] which was originally introduced to treat dispersive interactions. In this approach, the partial charge on a polarizable site is redistributed among a set of off-center particles connected harmonically to the atomic site. The positions of these charge particles are determined self-consistently in response to the external field, and the charges and force constants are related to the atomic polarizability. Thus, the Drude oscillator model can be designed to yield the same results as the induced point dipole model.[20,37] Nevertheless, an additional choice must be made with respect to the number and distribution of these fictitious particles. The Drude model has been implemented into CHARMM, and efforts are being made to construct a complete force field for biopolymers.[20,21]

Both methods described above do not treat the charge-transfer effect, which has been suggested to be important for modeling proteins in aqueous solution.[38,] In recent years, the fluctuating charge model, which was derived on the basis of the principle of electronegativity equalization,[39−43] has been used by a number of groups to represent molecular polarization.[6,8,15,16,22−24,44] In this model, the values of the atomic charges are treated as dynamic variables, which can fluctuate subjected to the overall charge constraint and are dependent on the environmental electric field. In principle, the fluctuating charge model allows for charge transfer, although charge transfer between molecules in this model is often unphysical, and it is typically restricted within the same molecule to model charge polarization.[8,15,16,22] Kaminski et al. experimented with the combination of both fluctuating charges and point dipole induction.[15,16,44] It was concluded that the fluctuating charge model alone is inadequate in describing intermolecular interactions in a number of cases. The fluctuating charge model has also been implemented into CHARMM.[22−24]

A reasonable question that is often asked and was raised by an anonymous referee is whether or not there is a need for developing polarizable force field to study properties of systems that require a polarizable model. Indeed, most obvious thermodynamic quantities of simple liquids and solutions as well as biopolymers can be adequately described by effective potentials;[27,28] a convincing testimony is the widespread application and success of empirical force fields in biomolecular simulations. On one hand, it was thought that the seemingly anomalous behavior of alkylamine solvation was attributed to polarization effects,[33] but it was later shown that a reparametrization of the effective pairwise potential can indeed reproduce experimental results.[33e] On the other hand, the solvation of a chloride ion by a water sphere is dependent on the use of a pairwise potential or a polarizable model, and only the latter can produce results consistent with expectation.[34] Although little experimental information is available, undoubtedly, the simulation of protein folding will benefit from the use of a polarizable force field, because the charge distribution due to polarization for a fully solvent exposed peptide will be different than that when it is folded in the interior of the protein. Pairwise, effective potentials cannot capture these internal charge polarizations, whereas a carefully developed polarizable force field is more likely to be successful. The answers to these questions will continue to emerge as more tests and simulations are carried out using polarizable force fields.

In the following, we describe the results from molecular dynamics simulations of liquid alkanes and amides, making use of the PIPF potential for nonbonded interactions and the CHARMM22 force field for the remainder of the energy terms. We designate this combined force field as PIPF-CHARMM. Amides represent an important class of organic compounds as model systems for peptides, and it is essential to reproduce structural and energetic properties of these liquids in order to construct a force field for polypeptides. This is reflected in the development of a number of force fields, including the effective CHARMM[27] and OPLS force fields,[28,45] and the fluctuating charge model within CHARMM.[22] Although both alkanes and amides potentials have been developed, we focus our discussion on liquid amides, including formamide (primary amide), *N*-methylacetamide (NMA), and *N*-methylformamide (NMF, secondary amides), and *N*,*N*-dimethylformamide (DMF, tertiary amide). We first describe the polarization model that we use, followed by parametrization and computational details. In section 4, we present Results and Discussion. In section 5, we summarize the major findings of this work.

2. Theoretical Model

We employ the “standard” CHARMM force field plus a polarization term as follows[1,27]

$$U(\mathbf{R}) = \frac{1}{2}\sum_{\text{bonds}} K_b(b - b_0)^2 + \frac{1}{2}\sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \frac{1}{2}\sum_{\text{UB}} K_{\text{UB}}(S - S_0)^2 + \frac{1}{2}\sum_{\text{dihedrals}} K_\varphi(1 - \cos(n\varphi - \delta)) + \frac{1}{2}\sum_{\text{impropers}} K_\omega(\omega - \omega_0)^2 + \sum_{\text{nonbonded pairs}} \left\{ \epsilon_{ij}^{\text{min}} \left[\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \right\} - \frac{1}{2}\sum_{\text{atoms}} \boldsymbol{\mu}_i \cdot \mathbf{E}_i^o \quad (2)$$

where the potential energy, $U(\mathbf{R})$, is a sum over the internal and nonbonded terms as a function of the atomic coordinates $\mathbf{R}$. The internal terms include bond (b), valence angle (θ), Urey−Bradley (UB, S), dihedral angle (φ), and improper angle (ω) contributions, as shown in eq 2. The parameters K_b, K_θ, K_{UB}, K_φ, and K_ω are the respective force constants, and the variables with the subscript "0" are the corresponding equilibrium values. The nonbonded terms include Coulomb, induction (eq 1), and van der Waals interactions in the form of the Lennard-Jones potential. The variables in eq 2 have standard meanings,[1,27] and they are not explicitly described here in view of brevity.

We adopt the Thole interaction dipole (TID) model[46] in the present PIPF potential, which yields excellent results in the predicted molecular polarizabilities with a set of purely atomic isotropic polarizability parameters. Furthermore, it has been shown that these parameters are remarkably transferable and can provide a reasonable estimate of the anisotropy in molecular polarizability even though atomic isotropic parameters are used.[46,47] In the TID model, the induced dipole at the ith interaction site due to the homogeneous external electric field $\mathbf{E}_i^o$ is given by

$$\boldsymbol{\mu}_i = \alpha_i(\mathbf{E}_i^o - \sum_{j \neq i}^{N} \mathbf{T}_{ij} \cdot \boldsymbol{\mu}_j) \quad (3)$$

where N is the number of polarizable sites, α_i is the atomic polarizability tensor, and $\mathbf{T}_{ij}$ is the dipole field tensor defined by

$$\mathbf{T}_{ij} = \frac{1}{r_{ij}^{\;3}}\mathbf{I} - \frac{3}{r_{ij}^{\;5}}\begin{bmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{bmatrix} \quad (4)$$

where $\mathbf{I}$ is the identity matrix, and x, y, and z are the Cartesian components along the vector between atoms i and j at a distance r_{ij}. In principle, all atomic interactions can be included within the same molecule, although short-range interactions (1−2, 1−3 and 1−4 terms) are excluded in the present study.

To avoid infinite polarization at a distance shorter than $(4\alpha_i\alpha_j)^{1/6}$ between two interacting induced dipoles, a phenomenon called the "polarization catastrophe", Thole[46] introduced a damping scheme in which the dipole field tensors can be derived from the first-order elements

$$(\mathbf{T}_{ij})_p^D = -(1 - e^{-au^3})\,\frac{(\mathbf{r}_{ij})_p}{r_{ij}^{\;3}} \quad (5)$$

where the subscript p is a Cartesian component of the vector $\mathbf{r}_{ij}$, and the superscript D denotes a damped interaction tensor. The damping scheme is equivalent to considering a smeared charge distribution between two interacting sites whose charge distribution is given as follows[46,47]

$$\rho(u_{ij}) = \frac{3a}{4\pi}\, e^{-au_{ij}^3} \quad (6)$$

where $u_{ij} = r_{ij}/(\alpha_i\alpha_j)^{1/6}$ is the effective distance between sites i and j. The factor a is a dimensionless width parameter of the smeared charge distribution which controls the strength of damping. The damping factor used in the PIPF is a = 0.572.

The modified higher-order $\mathbf{T}$ matrix elements can be obtained successively by taking the derivative of the preceding lower rank elements

$$(\mathbf{T}_{ij})_{pq}^D = \nabla_p(\mathbf{T}_{ij})_q^D = \lambda_5\,\frac{3(\mathbf{r}_{ij})_p(\mathbf{r}_{ij})_q}{r_{ij}^{\;5}} - \lambda_3\,\frac{\delta_{pq}}{r_{ij}^{\;3}} \quad (7)$$

$$(\mathbf{T}_{ij})_{pqr}^D = \nabla_p(\mathbf{T}_{ij})_{qr}^D = -\lambda_7\,\frac{15(\mathbf{r}_{ij})_p(\mathbf{r}_{ij})_q(\mathbf{r}_{ij})_r}{r_{ij}^{\;7}} + \lambda_5\,\frac{3[(\mathbf{r}_{ij})_p\delta_{qr} + (\mathbf{r}_{ij})_q\delta_{pr} + (\mathbf{r}_{ij})_r\delta_{pq}]}{r_{ij}^{\;5}} \quad (8)$$

where the parameters λ_i are given as follows

$$\lambda_3 = 1 - \exp(-au^3) \quad (9)$$

$$\lambda_5 = 1 - (1 + au^3)\exp(-au^3) \quad (10)$$

$$\lambda_7 = 1 - \left(1 + au^3 + \frac{3}{5}a^2u^6\right)\exp(-au^3) \quad (11)$$

As in the procedure used by Ren and Ponder[17,18] interactions are damped only between interacting induced dipoles, while the electric field due to the permanent charges is not affected.

Equation 3 shows that each induced dipole depends on the polarization of all other dipoles. Thus, it must be solved self-consistently. A standard iterative procedure is often used such that an initial guess of the induced dipoles (or simply set to 0) is first made to estimate a set of induced dipoles, which are inserted into eq 3 to yield a newer set of induced dipoles. These induced dipoles are then used in the next iteration, and the process continues until a predefined convergence criterion is satisfied.[10−14] Typically, a few iteration steps are sufficient to achieve the required accuracy to ensure energy conservation, and the iterative procedure provides the most practical, converged results in molecular dynamics simulations.[10−14] Alternatively, eq 3 can be rearranged such that the induced dipole moments are determined exactly by inverting the dipole interaction matrix[10]

$$\boldsymbol{\mu} = \mathbf{A}^{-1}\mathbf{E}^o \quad (12)$$

where $\boldsymbol{\mu}$ is the column induced dipole vector, $\mathbf{E}^o$ is the electric field matrix due to atomic partial charges, and the interaction matrix **A** is defined as

$$\mathbf{A} = \begin{bmatrix} \alpha_1^{-1} & \mathbf{T}_{12} & \cdots & \mathbf{T}_{1N} \\ \mathbf{T}_{21} & \alpha_2^{-1} & \cdots & \mathbf{T}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}_{N1} & \mathbf{T}_{N2} & \cdots & \alpha_N^{-1} \end{bmatrix} \tag{13}$$

Although eq 12 yields the exact results, which is useful for validating the results from the iterative procedure, the shortcoming is the unbearable computation costs for large systems, which scales as 27 × N^3 to invert the **A** matrix (3N × 3N-dimension).[9]

If one treats the induced dipoles as independent dynamic variables, the nuclear dynamics and molecular polarization can be propagated simultaneously. Sprik and Klein,[37] among others, have used this approach to perform molecular dynamics simulations with an induced polarizable dipole force field. In this case, the Lagrangian of the system is extended with a fictitious kinetic term associated with these extra variables, and the dynamics of the induced dipole moment is governed by the following equation of motion (in matrix form)[48]

$$\mathbf{m}_\mu \ddot{\boldsymbol{\mu}} = \mathbf{E} - \frac{\boldsymbol{\mu}}{\alpha} \tag{14}$$

where $\mathbf{m}_\mu$ is an "inertial factor" associated with the extra dynamical variables whose dimensions are those of mass × charge^{-2}, **E** is the total electric field, and $\ddot{\boldsymbol{\mu}}$ is the second time derivative of the induced dipole. In comparison with solving the self-consistent equations, this approach is a very efficient way of computing induced dipoles with an increase in computer time by a factor of about 2.[48] As pointed out by Van Belle et al.,[48] the induced dipoles will fluctuate about its average orientation during the dynamics simulations, although the converged induced dipoles are always oriented along the direction of the local electric field.

All three methods have been implemented into the program CHARMM (c33a1) for the TID model, and a critical evaluation of their performance and convergence in computed thermodynamic and structural properties has been carried out (to be published).

3. Computational Details

3.1. Parametrization. The parametrization of the PIPF-CHARMM force field follows the procedure employed previously in the development of the original CHARMM22 force field.[27] For convenience, the internal terms and nonbonded terms are optimized iteratively. Experimental structural data and bulk liquid properties for different organic functionalities are used as the primary targets of parameter optimization. Potential energy surfaces, relative energies of different conformations, and vibrational spectra calculated from high level QM methods are used as Supporting Information where experimental results are not available.

3.1.1. Nonbonded Terms. Nonbonded terms include van der Waals interactions, which are modeled by the Lennard-Jones form in CHARMM, Coulomb interactions among fixed (permanent) atomic partial charges, and polarization interactions, which include both charge-induced dipole and induced dipole−induced dipole contributions. We used the Lennard-Jones parameters from the CHARMM22 force field as an initial input for the same types of interaction site,[27] whereas the partial atomic charges are scaled to yield the correct dipole moments in the gas phase, using the TID model, for the model compounds selected in the present study. The original set of isotropic atomic polarizabilities was fitted for H, C, N, and O to a set of 16 molecular polarizabilities,[46] and later, van Duijnen and Swart extended the optimization set to 52 molecules with halogen and sulfur atoms.[47] Although different optimization schemes were used, they found that the original set was very similar to those from the new optimization.[47] We have also tried to reoptimize these atomic polarizabilities and reached the same conclusion. Thus, we decided to directly use the atomic polarizabilities from Thole's work.[46]

Then, condensed-phase simulations were carried out by slightly readjusting the Lennard-Jones parameters and partial charges to reproduce the experimental heats of vaporization and liquid densities. Consequently, the finals set of charges and atomic polarizabilities do not yield the exact, although very close, gas-phase dipole moments for these amides. The optimized parameters are given in Table S1, Supporting Information.

3.1.2. Internal Terms. Having optimized an initial set of the nonbonded parameters for alkanes and amides, we further examined the internal energy terms, including bond stretching, angle bending, out-of-plane bending, and torsion of dihedral angles, using the original values in the CHARMM22 force field.[27] We focused on the torsional potential energy surface and vibrational frequencies of NMA and the conformational energies and the Ramachandran map of alanine dipeptide obtained from QM calculations at the LMP2/cc-pVQZ(-g)//MP2/6-31G(d) level of theory.[49] Then, we returned to liquid simulations to further optimize the nonbonded energy terms until both liquid-phase results and internal energy terms are satisfactory. Finally, to match the ab initio Ramachandran map, an energy correction map (CMAP)[49] is also made to the φ, ψ dihedrals using the PIPF-CHARMM force field. For additional details of the CMAP procedure, readers are directed to the original paper.[49] The final force field is given as Supporting Information, which can be download as the parameter file for CHARMM.

3.2. Simulation Details. Nonbonded parameters were optimized through liquid simulations of four alkanes and six amides. In each case, molecular dynamics were executed with the CHARMM program using the isothermal−isobaric (NPT) ensemble at 1 atm and a temperature indicated below. The temperature is maintained by the Nose-Hoover thermostat,[50] while the pressure is controlled via the Langevin piston method.[51] The velocity Verlet algorithm was used to integrate the equations of motion with a time step of 1 fs.[52] In the present PIPF potential employing the TID model for molecular polarizations, intramolecular interactions between atom pairs that form a covalent bond (1−2), a bond angle (1−3), and a dihehdral angle (1−4) are excluded from the

Table 1. Optimized Parameters for Alkanes and Amides in the Nonbonded Energy Terms of the Present PIPF Potential along with the Original Values in the CHARMM Force Field

	$R_{\min}/2$ (Å)		ϵ (kcal/mol)		*q* (e)		α (Å^3)
atom type	CHARMM	PIPF	CHARMM	PIPF	CHARMM	PIPF	PIPF
C(R−CH3)	2.040	2.020	−0.078	−0.080	−0.27	−0.09	1.334
C(R2−CH2)	2.175	2.120	−0.055	−0.060	−0.18	−0.06	1.334
C(R3−CH)	2.275	2.200	−0.027	−0.035	−0.09	−0.03	1.334
C(O═C−CH3)	2.060	2.020	−0.080	−0.080	−0.27	−0.09	1.334
C(N2°−CH3)	2.060	2.020	−0.080	−0.080	−0.11	0.02	1.334
C(N3°−CH3)	2.060	2.020	−0.080	−0.080		0.05	1.334
C(O═C)	2.000	1.960	−0.110	−0.110	0.51	0.45	1.334
O(C═O)	1.700	1.730	−0.120	−0.120	−0.51	−0.45	0.837
N(N1°)	1.850	1.850	−0.200	−0.200	−0.64	−0.68	1.073
N(N2°)	1.850	1.850	−0.200	−0.200	−0.47	−0.49	1.073
N(N3°)	1.850	1.850	−0.200	−0.200		−0.46	1.073
H(N1°)	0.2245	0.7577	−0.046	−0.015	0.32	0.34	0.496
H(N2°)	0.2245	0.7577	−0.046	−0.015	0.31	0.29	0.496
H(C═O)	1.320	1.340	−0.022	−0.025	0.08	0.00	0.496
H(R−CH3)	1.320	1.340	−0.022	−0.025	0.09	0.03	0.496
H(N−CH3)	1.320	1.340	−0.022	−0.025	0.09	0.06	0.496
H(O═CCH3)	1.320	1.340	−0.022	−0.025	0.09	0.03	0.496

dipole interaction tensor. We have also examined the possibility to include 1−4 interactions in molecular polarization, but we found that the best results are obtained without these short-range terms. We employed the iterative procedure to converge the induced dipoles with a criterion of less than 0.0001 Debye per atom. In all simulations, a spherical cutoff was used to generate a nonbonded list for all pairs within 12.5 Å, and the interaction forces are smoothed to 0 by a switching function between 11.0 and 12.0 Å for Coulomb interactons and a shifted potential for Lennard-Jones interactions.[53c] Although the use of a spherical cutoff for nonbonded interactions may introduce some errors in the computed thermodynamic properties, the development of the OPLS and the CHARMM22 force fields as well as the SPC, the TIP3P, and the TIP4P models utilized even shorter truncation distances at that time. Yet, numerous applications suggest that these force fields still perform exceptionally well when long-range electrostatics is explicitly included. Certainly, models that are specially derived for Ewald calculations have shown improved properties, especially in computed dielectric constants.[53] We are currently implementing the PME-based method for the present model to further validate the present parameters. Spherical truncation was made at 12.5 Å for interactions involving induced dipoles. All the bonds connecting to a hydrogen atom are fixed by the SHAKE algorithm.[54]

In each case, a cubic box was used, consisting of 256 molecules with periodic boundary conditions. The box size varied from about 26 × 26 × 26 Å^3 for formamide to as large as 35 × 35 × 35 Å^3 for isobutane. The simulations were run for ethane, propane, and butane at their boiling points, which are 184, 231, and 273 K, respectively; for isobutane, formamide (FORM), and *N*-methylformamide (NMF) at 298 K; for acetamide (ACEM) at 373 K and its boiling point (494 K); for *N*-methylacetamide (NMA) and *N*,*N*-dimethylacetamide (DMA) at 373 K; and for *N*,*N*-dimethylformamide (DMF) at 298 and 373 K. Each system was first equilibrated by at least 1 ns, followed by another 1 ns for averaging. Tests suggest that the computed results are sufficiently converged and show little variations at much longer simulation time for these simple liquid systems. Statistical uncertainties ($\pm 1\sigma$) for the computed properties reported here are determined through averages of batches of 50 ps simulations.

The average energy of the gas-phase molecule was calculated by Monte Carlo (MC) simulations of a single molecule at the same temperature as in the corresponding MD simulations. Standard Metropolis sampling[2] was used that included Cartesian moves for all atoms. Each Monte Carlo simulation consisted of at least 1 × 10^6 configurations of equilibration followed by 5 × 10^6 configurations of averaging. The energy convergence for a single molecule in the gas phase is much better achieved employing Monte Carlo simulations than using molecular dynamics where large fluctuations in temperature complicate potential energy convergence.

4. Results and Discussion

4.1. Polarization. Optimized nonbonded parameters are listed in Table 1. In general, partial charges are smaller than those in the CHARMM22 force field. This is expected since the fixed charge force field mimics many-body polarization effects in an average way such that the molecular dipoles are greater than those in the gas phase. Except for the hydrogen atom on nitrogen, the Lennard-Jones parameters in the present PIPF potential are very similar to the original values in CHARMM.[27] The “polar” hydrogen radius was increased from 0.2245 to 0.7577 Å (for type H only in the CHARMM force field definition in the present set of compounds). Atomic polarizabilities for each element are directly taken from the TID model,[46] which were fitted to experimental *anisotropic* molecular polarizabilities for a small set of molecules in the gas phase,[47] but they require

Table 2. Computed and Experimental Dipole Moment and Molecular Polarizability for Amides[a]

	μ_g (gas phase)		μ_{tot} (liquid)		μ_{ind} (liquid)		α (Å^3)	
liquid	exp[b]	PIPF	PIPF	QM/MM[d]	PIPF	QM/MM[d]	exp[b]	PIPF
formamide	3.73	3.70	5.3	4.9	1.6	1.2	4.08	4.08
acetamide	3.68	3.59	5.4, 5.1[c]		1.8, 1.5		5.67	5.91
NMA	3.72	3.31	5.0	4.7	1.7	0.9	7.82	7.97
NMF	3.83	3.35	4.9	4.4	1.5	1.2	5.91	5.91
DMA	3.70	3.31	4.4		1.1		9.63	9.24
DMF	3.82	3.48	4.4, 4.3[c]	4.6	0.9 ,0.8	0.5	7.81	7.81

[a] Dipole moment in Debye, polarizability in Å^3. [b] Experimental data are from ref 42. [c] Corresponding to simulations at the second (higher) temperature (see text). [d] Computed using the AM1 model for each amide in its liquid treated by the polarizable model in ref 14.

Table 3. Computed Energetic Results (kcal/mol) for Alkanes and Amides at Specified Temperatures[a]

species	$-\Delta E_i$	$-\Delta E_{elec}$	$-\Delta E_{pol}$	ΔE_{intra}	$-\Delta E_{tot}$	ΔH_v(exp)[b]	ΔH_v(calc)
ethane (184)	3.12	0.00	0.01	−0.03	3.10	3.52	3.46 ± 0.02
propane (231)	4.03	−0.01	0.01	−0.04	3.99	4.49	4.45 ± 0.04
butane (273)	4.90	0.02	0.01	−0.05	4.85	5.35	5.39 ± 0.05
isobutane (298)	4.09	0.01	0.01	−0.02	4.06	4.57	4.66 ± 0.06
FORM (298)	14.41	7.84	3.51	−0.01	14.40	14.7	14.99 ± 0.03
ACEM (373)	14.35	6.58	3.41	0.06	14.41		15.16 ± 0.04
ACEM (494)	12.10	5.38	2.84	0.12	12.22	13.4	13.20 ± 0.05
NMA (373)	12.83	4.46	2.11	−0.09	12.73	13.3[c]	13.48 ± 0.05
NMF (298)	12.93	5.21	2.37	0.00	12.93	13.52	13.55 ± 0.04
DMA (298)	12.24	2.73	0.77	0.21	12.46	12.7	13.05 ± 0.03
DMF (298)	10.95	2.89	0.65	0.10	11.05	12.00	11.79 ± 0.03
DMF (373)	9.72	2.54	0.61	−0.04	9.68	10.4	10.42 ± 0.04

[a] Temperatues are indicated in parentheses in Kelvin. [b] Experimental data are from refs 42 and 49. See text for discussion. [c] Reference 62. Experimental heat of vaporization for *N*-methylacetamide has been reported at 14.2 kcal/mol (ref 57).

no further modification in liquid simulations. Importantly, only a single parameter (isotropic atomic polarizability) is needed for each element for all interaction types. The remarkable transferability of the TID model has been thoroughly examined by van Duijnen and Swart.[47] The transferability is a major advantage of the TID model, for which few other polarizable models exhibit such a good behavior.

Table 2 depicts the computed molecular dipole moments in the gas phase and in the liquid phase along with the average molecular polarizabilities for amides, for which all intramolecular interactions are included in the calculation. The computed molecular polarizabilities are in excellent agreement with experimental results. Gas-phase dipole moments are generally underestimated in the present TID model, with a mean unsigned error of less than 8% in comparison with the experimental data.[55−57] This is in contrast to the effective pairwise potentials, in which the molecular dipoles are typically overestimated by 10−20% to account for polarization effects.[1,27,28] In the present study, we decided not to strictly enforce the requirement that gas-phase dipole moments exactly reproduce the corresponding experimental data. Our experience from early studies[13,14] shows that the increased flexibility allows for condensed-phase properties to be better described, and similar observations have been made in recent applications.[22] Clearly, the dipole moments are greatly enhanced in going into the liquid phase, although the quantitative accuracy of the average dipole moments in the fluid phase is difficult to assess because there are no experimental data for comparison. As expected, the enhancement in dipole moment for the primary and secondary amides is more significant than that for the tertiary amides due to hydrogen bonding interactions in the former, which are absent in the latter systems. Previously, a similar trend was observed from Monte Carlo simulations of formamide, acetamide, *N*-methylformamide, and *N*-methylacetamide using a polarizable intermolecular potential function (PIPF-A)[14] with a set of atomic polarizabilities similar to the Applequist[58] values without considering intramolecular polarization interactions. However, the present TID model (Table 2) yields induced dipoles twice as large as the previous PIPF-A potential.[14] The present results are in better agreement with combined QM/MM simulations in which one solute is represented by the semiempirical AM1 method embedded in a solution of the same amide.[14] For the primary and secondary amides the computed induced dipoles are 0.9−1.2 D from the QM/MM simulations.[14]

4.2. Liquid Properties. The computed energetic results are summarized in Table 3. The heat of vaporization is related to the total intermolecular interaction energy of the liquid, E_i(l), the intramolecular energies in the liquid, E_{intra}(l), and in the gas phase, E_{intra}(g), and the work term, which is RT for 1 mol of ideal gas.

$$\begin{aligned}\Delta H_v(T) &= -E_i(\mathrm{l}) - E_{intra}(\mathrm{l}) + E_{intra}(\mathrm{g}) + RT \\ &= -E_{tot}(\mathrm{l}) + E_{intra}(\mathrm{g}) + RT \qquad (15)\end{aligned}$$

In computing the intermolecular interaction energy for the liquid, we have included a correction for long-range van der Waals interactions beyond the cutoff distance by assuming the distribution function is uniform.[14,59] The correction to

Table 4. Computed and Experimental Molecular Volume, Diffusion Constants, and Dielectric Constant for Simulated Liquids[a]

	V (Å^3)		D (10^{-9} m^2/s)		ϵ	
species	exp	calc	exp	calc	exp	calc[b]
ethane (184)	91.5	91.3 ± 0.4	4.27 (182)	4.66 ± 0.02		
propane (231)	126.0	127.4 ± 1.2	5.13 (243)	4.86 ± 0.05		
butane (273)	160.3	162.1 ± 1.6		5.00 ± 0.02		
isobutane (298)	175.1	178.1 ± 2.6		6.89 ± 0.09		
FORM (298)	66.3	65.0 ± 0.2		0.41 ± 0.01	109.5	138
ACET (373)	99.9	96.8 ± 0.3		0.75 ± 0.004		154
ACET (494)		108.0 ± 0.5		2.90 ± 0.03		145
NMA (373)	135.8	132.4 ± 0.6	1.47 (373)	1.46 ± 0.02	101	152
NMF (298)	98.3	97.8 ± 0.3	0.85 (298)	0.72 ± 0.01	186	200
DMA (298)	154.5	153.1 ± 0.3	1.37 (298)	0.57 ± 0.004	38.9	116
DMF (298)	128.5	128.9 ± 0.4	1.64 (298)	1.08 ± 0.01		
DMF (373)	139.0	139.7 ± 0.7	3.93 (373)	2.67 ± 0.02		165

[a] Temperatues are indicated in parentheses in Kelvin. Experimental data are from refs 42, 49, and 50. [b] Since the simulation is relatively short for computing the slowly converging total molecular dipole moment, it is likely that the computed dielectric constants have not fully converged. Further simulations are being carried out.

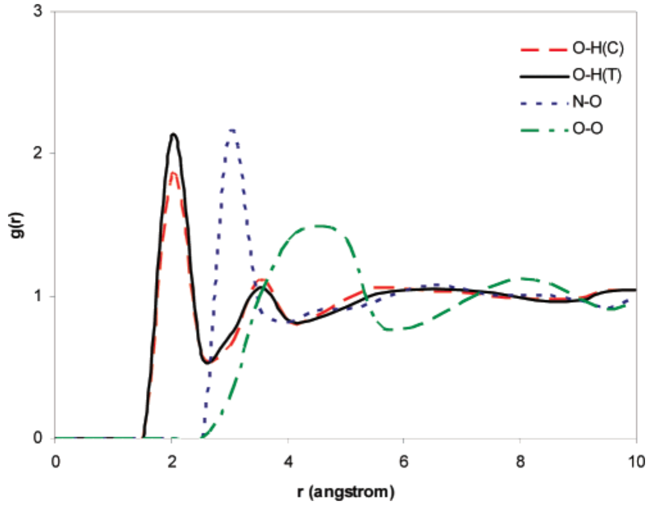


Figure 1. Computed O−H(C), O−H(T), N−O, and O−O radial distribution functions for liquid formamide at 25 °C. H(C) and H(T) specify the amino hydrogen atoms cis and trans to the carbonyl group. Distances are in angstroms.

the computed heat of vaporization due to departure from ideal behavior of the vapor was found to be negligible for amides, and thus they are not included.[60,61] The mean unsigned error in the calculated heats of vaporization for the four alkanes and six amides (two performed at two termperatures) is 1.4% compared to the experimental data.[55−57,62−66] For comparison, the average error reported by Jorgensen et al. is about 2% for hydrocarbons and amides for the OPLS-AA potentials.[28] Simulation of liquid alkanes and amides using CHARMM fixed charge force field yields an error ranging from 0 to 6%.[22] In a separate study employing the PIPF-A potential,[14] the average error was about 2% for four amides, formamide, NMF, NMA, and DMF.[14] Thus, the present energetic results are comparable to or slightly better than the performance of earlier force fields. It should be noted that we have used 100% of the trans configuration for NMA and NMF, and they remained in that configuration, whereas there are about 2−3% of the cis population in natural abundance in experiments.[67,68]

Table 3 also lists the average polarization energies in these liquids. Obviously, there is little contribution from molecular polarization in liquid alkanes, whereas polarization effects are significant in liquid amides. The largest polarization contributions are found in primary and secondary amides, amounting to about 24% and 18% of total intermolecular interaction energy, respectively. For comparison, the PIPF-A model has polarization effects between 12 and 14% of the total energy for primary and secondary amides.[14] In that model, the gas-phase dipole moments for these amides are slightly greater than the corresponding experimental value except NMF,[14] whereas they are smaller in the present case.

Table 4 lists the computed liquid density (volume), self-diffusion constants, and dielectric constants at various temperatures used in the dynamics simulations. The mean unsigned error in the computed molecular volume and liquid density is about 1.3% in comparison with experimental data (Table 4). Overall, the TID model shows excellent results for these organic liquids. We note that during the parametrization process, DMF was only considered at 373 K. Interestingly, when these parameters are used to perform simulations of DMF at 298 K, we obtain a liquid density and ΔH_v within 1% and 2% from the corresponding experimental values. The dielectric constants have only been averaged for 1 ns of simulation time, and they are almost certain not yet converged. Extended simulations with particle-mesh Ewald treatment of long-range electrostatics are being carried out. The calculated self-diffusion coefficients are somewhat underestimated for NMF, DMF, and DMA, while it is in excellent agreement with experiment for NMA.[69,70] For comparison, the CHARMM-FQ model yields a value of $D = 1.93 \times 10^{-9}$ m^2/s for liquid NMA. The effective CHARMM force field produced a value of 2.04×10^{-9} m^2/s.[27]

4.3. Radial Distribution Functions. The structure of the liquids are characterized by radial distribution functions (rdfs), $g_{xy}(r)$, which specifies the probability of finding an atom y at a distance r from atom x. All rdfs are normalized to the bulk density. To simplify our discussion, we focus on rdfs involving hydrogen bonding interactions. Errors associated with data collection are about half of the width of the bin size, which is 0.05 Å.

4.3.1. Formamide and Acetamide. Figure 1 shows the radial distribution functions for liquid formamide, in which the hydrogen atom trans to the carbonyl group is denoted by H(T) and cis to the carbonyl group by H(C). The strong first peaks of the carbonyl oxygen and amide hydrogen pairs, O−H(C) and O−H(T), centered at 1.95 Å are due to hydrogen bonding interactions. The results are in excellent agreement with the peak at 1.9 Å assigned to O−H contacts from diffraction experiments.[71−74] In an early study using the PIPF potential, the first O−H peak in liquid formamide occurs at 1.85 Å.[14] The agreement with the OPLS[28] and CHARMM[27] potentials is also good with the first peak occurring at 1.9 Å. Integration to the minima of the first peaks gives 0.9 and 1.0 nearest neighbors around H(C) and H(T). For comparison, other studies using PIPF potential give

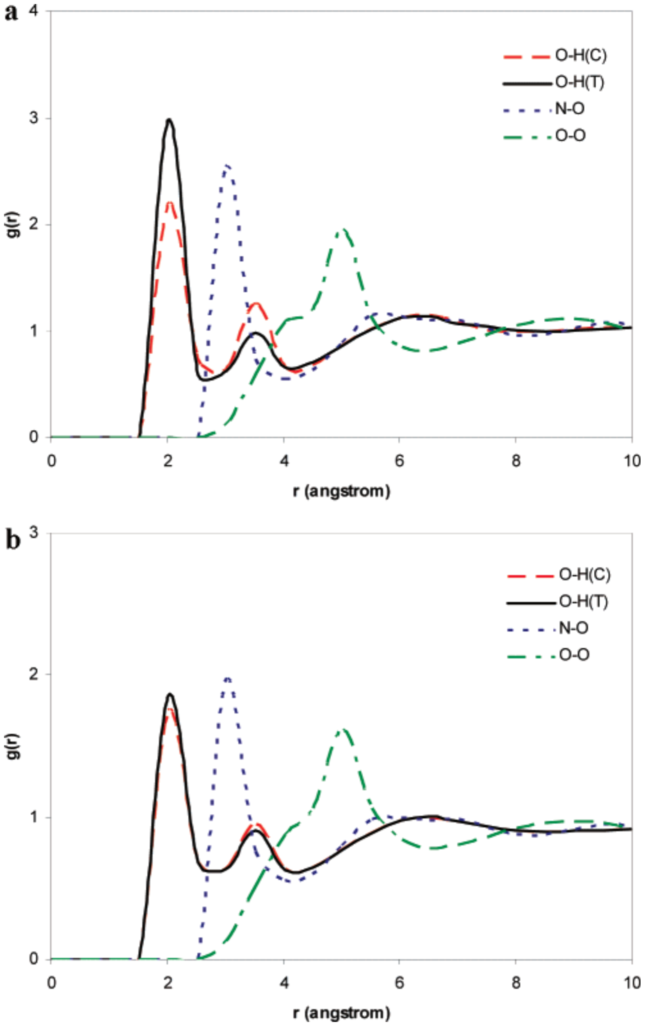


Figure 2. Computed O−H(C), O−H(T), N−O, and O−O radial distribution functions for liquid acetamide at (a) 373 K and (b) 494 K.

0.9 and 1.1 nearest neighbors for these two hydrogen atoms, respectively.[14] Hydrogen bonding interactions are also reflected by the heavy atom distributions, in particular the N−O rdf, which has a strong first peak at 2.93 Å. For comparison, the previous PIPF potential has a peak at 2.8 Å in the N−O distribution.[14] Excellent agreement has been observed with OPLS potential which gives a peak at 2.9 Å. Integration of the first peak to the minimum at 4.02 Å yields 3.41 contacts. This is greater than the finding from previous PIPF and OPLS potentials, which yield a value of 2.5−2.7. Different diffraction studies produced values of 2.9, 3.03, and 3.05 Å,[71−74] reflecting the uncertainty range from experiments. We note that the computed contact number depends strongly on the minimum position in the rdf, which is often not precisely defined due to overlap between the tails from the first and second solvation layers.

The computed rdfs for acetamide are displayed in Figure 2 at two simulation temperatures, corresponding to 373 and 494 K. The first peaks in the N−O distribution function are found at 2.90 and 2.93 Å at 373 and 494 K, respectively, which are slightly shorter than a distance of 3.03 Å from a recent X-ray diffraction experiment of liquid acetamide at 346 K.[75] An important qualitative feature is that the heights of the two O−H peaks decrease noticeably as the temperature increases, suggesting greater thermal fluctuations and less structured interactions. Figure 2a also reveals that the trans hydrogen, H(T), dominates hydrogen bonding interactions due to favorable dipolar orientations, which is also indicated in Figure 1. However, at higher temperature, the contributions from both hydrogens in hydrogen bonding interactions become similar. Integration of the first peaks in O−H(C), O−H(T), and N−O rdfs to the first minima results in 0.8, 0.9, and 2.4 nearest neighbors at 373 K. At higher temperature, the number of the first contacts decreases to 0.7, 0.7, and 2.0.

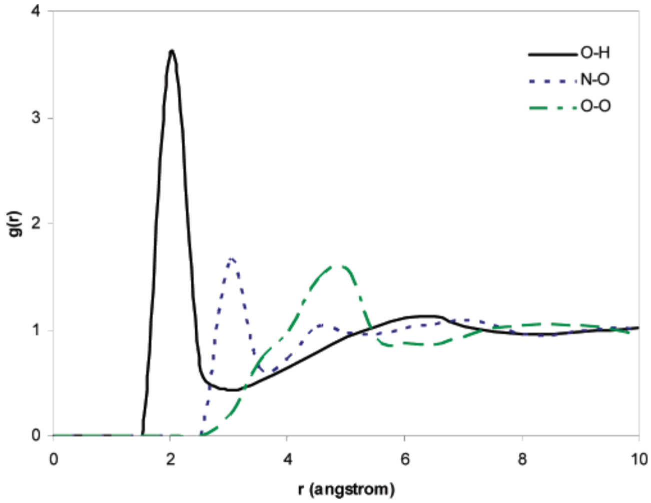


Figure 3. O−H, N−O, and O−O radial distribution functions for liquid *N*-methylformamide at 25 °C.

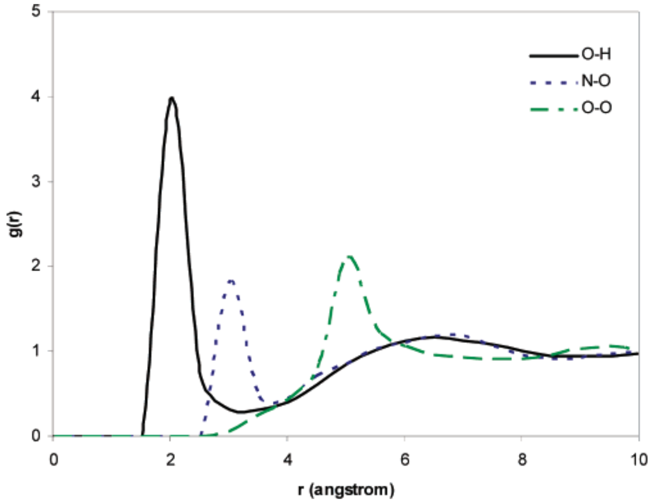


Figure 4. O−H, N−O, and O−O radial distribution functions for liquid *N*-methylacetamide at 100 °C.

4.3.2. N-Methylformamide and N-Methylacetamide. The computed rdfs for NMF and NMA are displayed in Figures 3 and 4. Strong hydrogen bonding in these two liquids is clearly indicated by the first striking peaks in the O−H and N−O rdfs. The computed rdf for O−H shows a strong peak at 1.95 Å for both NMF and NMA, in good agreement with the neutron diffraction experiment with fully deuterated NMF which gives the O−H contact at 1.89 Å.[76,77] Integration to the first minima gives 1.0 and 1.1 nearest neighbors in liquid NMA and NMF, respectively. The other peak extensively

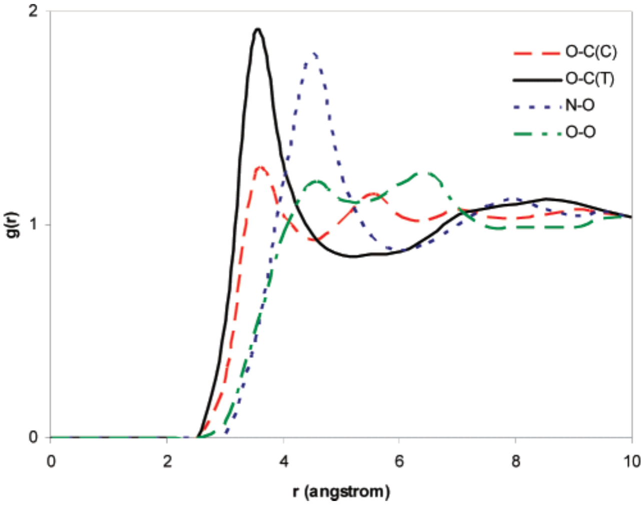


Figure 5. O−C(C), O−C(T), N−O, and O−O radial distribution functions for liquid *N,N*-dimethylformamide at 25 °C. C(C) and C(T) specify the methyl carbon atoms cis and trans to the carbonyl group.

studied by diffraction experiments is the peak in N−O rdf. The N−O interaction due to hydrogen bonding in liquid NMF and NMA are estimated to be 3.02 Å for NMF in X-ray diffraction[76,77] and 3.03 Å for NMA from both diffraction experiment and DFT calculations.[78] Our calculation gives a value of 2.90 and 2.93 Å in liquid NMF and NMA, respectively. Integration to the first minima yields 1.1 and 1.2 for NMA and NMF, respectively. In contrast, MC simulations from previous study using PIPF-A force field give the first peak at 1.85 and 2.75 Å for O−H and N−O contact, respectively.[14]

We note that the present PIPF potential employing the TID model for polarization yields N−O peaks for the primary and secondary amides in close agreement (2.90−2.95 Å from simulations vs 3.03 Å from diffraction experiments), which may be compared with previous simulations (2.75−2.8 Å) employing a polarizable potential and the Applequist-like polarizabilities.[14] Furthermore, it has often been suggested that it is necessary to have shorter hydrogen-bonding peaks in liquid simulations with effective pair potentials to account for polarization effects by making stronger and shorter hydrogen bonds. For example, the OPLS force field has a distance of 2.9 Å at the first peaks in the N−O rdfs for formamide and NMA.[79] We also notice that the height of the first peak and shape in the N−O rdf for NMA, including a shoulder at about 4.5 Å and a broad peak at about 7 Å, are found to be in good agreement between the present PIPF and OPLS force field.

4.3.3. N,N-Dimethylformamide and N,N-Dimethylacetamide. Despite the fact that there is no hydrogen bond donor in DMF and DMA, the computed rdfs displayed in Figures 5 and 6 show significant local order due to the interaction between methyl groups and carbonyl oxygen in the view of the peaks in O−C(C) and the O−C(T) rdfs centered at 3.50 Å and the N−O rdf at 4.55 Å. This is consistent with the conclusions of Jorgensen and Swenson using the OPLS-UA potential[79] and our PIPF-A model.[14] Interestingly, dipolar interactions favor closer contacts between the trans methyl group and the carbonyl oxygen in both the present and early PIPF potentials. On the other hand, the OPLS-UA potential does not seem to distinguish between the two methyl groups in DMF.[79] In contrast, X-ray diffraction experiments suggest no significant local order in liquid DMF and DMA;[80,81] however, it is difficult to specifically resolve the total diffraction pattern into specific pair interactions without significant hydrogen bonding interactions and isotope replacement.

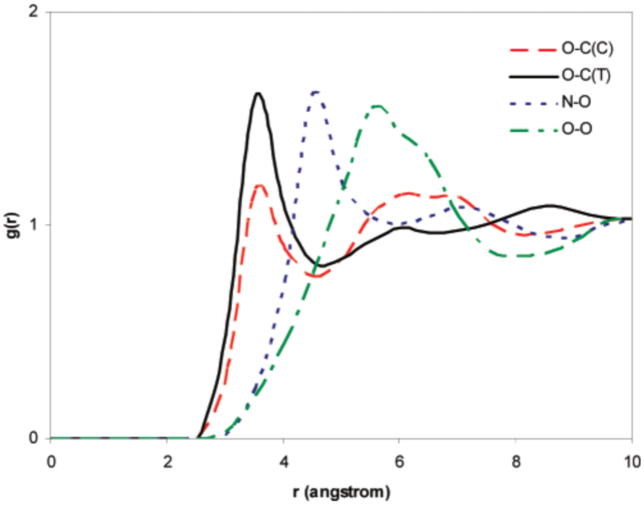


Figure 6. O−C(C), O−C(T), N−O, and O−O radial distribution functions for liquid *N,N*-dimethylacetamide at 25 °C. C(C) and C(T) specify the methyl carbon atoms cis and trans to the carbonyl group.

Table 5. Torsional Terms and Parameters That Have Been Adjusted from the Original CHARMM22 Force Field for use in Connection with the Present PIPF Potential for Amides

dihedral type	K_φ	n	δ
CT3−C−NH1−CT3	1.2	1	0
CT3−CT1−NH1−C	1.6	1	0
HA−CT3−NH1−H	0.11	3	0
O−C−CT3−HA	0.04	3	180

Table 6. Relative Energies (in kcal/mol) and Conformations (in deg) of the Alanine Dipeptide C7$_{eq}$, C7$_{ax}$, and C5 Minima from Ab Initio (LMP2/cc-pVQZ(-g)//MP2/6-31G(p,d)) and Force Field Calculations (ref 49)

	C7$_{eq}$	C7$_{ax}$	C5
QM	0	2.20	1.01
PIPF	0	1.97	1.53
		Conformations (ϕ,ψ)	
QM	−83, 78	74, −64	−158, 162
PIPF	−79, 77	69, −73	−152, 156

4.4. Internal Parameters. The internal bonded parameters are reoptimized for amides functional groups making use of NMA and alanine “dipeptide” as the model compounds. We began with the original CHARMM22 force field for all bonding terms, and it was found that all parameters associated with bonds and angles and with the improper dihedral angle terms can be kept without alteration. The only

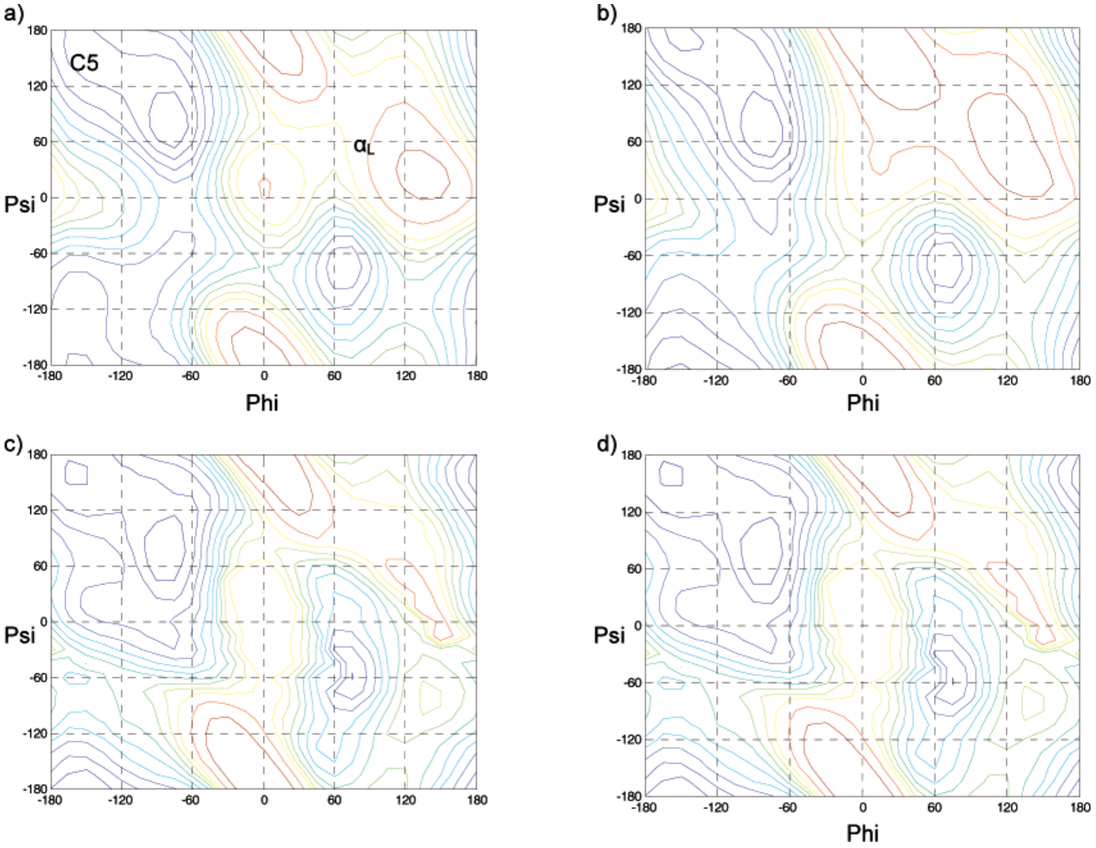


Figure 7. Adiabatic alanine dipeptide potential energy surface calculated from (a) PIPF, (b) CHARMM22, (c) PIPF with CMAP, and (d) QM. Contour represents 1−10 kcal/mol with 1 kcal/mol interval, 12 kcal/mol, and 15 kcal/mol.

parameters that were further modified are some torsional terms, which are listed in Table 5. These values are optimized to reproduce ab initio conformation energies of NMA and the potential energy surface (Ramachandran map) of alanine dipeptide.

With these minor readjustments of the CHARMM22 parameters, we found that it is possible to obtain the relative conformational energies for NMA and the alanine dipeptide using the present polarizable nonbonded interaction terms. In particular, we found that cis conformer of NMA is 2.44 kcal/mol higher in energy than the trans configuration, and the rotation barrier about the peptide bond is 21.1 kcal/mol. For comparison, MP2/cc-pVTZ//MP2/6-31G(d) calculations yield values of 2.39 and 20.5 kcal/mol, respectively.[27] NMR studies revealed a rotational barrier of 19.8 ± 1.8 kcal/mol.[82] The relative conformational energies for the alanine dipeptide are given in Table 6, which are compared with QM calculations at LMP2/cc-pVQZ(-g)//MP2/6-31G(d) level. The relative energy of $C7_{ax}$ calculated by the present force field is underestimated by 0.2 kcal/mol, while the energy of C5 is overestimated by 0.5 kcal/mol. The largest deviation was found in the Ψ angle for the $C7_{ax}$ conformer, which is overestimated by nearly 20° at −73° compared to −64° from LMP2/cc-pVQZ(-g)//MP2/6-31+G(d) calculations.

Ramachandran plot (phi−psi map) computed using the present CHARMM-PIPF potential, the original CHARMM22 force field, and the ab initio MP2/TZVP//6-31G(d,p) are shown in Figure 7, along with a fully corrected energy contour by the CMAP (see below) procedure. In comparison with the MP2 results, the CHARMM22 force field shows a steep surface at the C5 region, and the energy in the α_L region is overestimated. The PIPF-CHARMM force field yields somewhat improved features in these two regions. As noted elsewhere,[49] artifacts exist to overly sample the π-helical populations using empirical potentials, whereas π-helices are rarely observed experimentally. This problem was traced to subtle deviations between the empirical and ab initio Ramachandran maps. MacKerell et al.[49] made a bold proposal by introducing a spline correction map (CMAP) to reproduce almost exactly the MP2 results. Now, the CMAP is a standard option in the CHARMM22 force field, which significantly improved conformational sampling of low population structures. A CMAP has also been constructed for the present CHARMM-PIPF potential, which gives a mean deviation from the MP2 results by merely 0.0002 kcal/mol.

As in the standard CHARMM force field, we have two options that the users may choose from, with and without the inclusion of the CMAP for the PIPF-CHARMM potential. The use of CMAP slightly increases computational time. If this is not a major concern, it is recommended to include the CMAP procedure because it significantly reduces the tendency to form a π-helix.

The transferability of the bond and angle parameters from the CHARMM22 force field is further tested by vibrational spectral analysis of NMA and three conformers of the alanine dipeptide. Calculated frequencies and key characteristic components for each mode have been determined for each molecule; these results are given as Supporting Information in Tables S1−S4. Comparison with experimental and ab initio results at the LMP2/cc-pVQZ(-g)//MP2/6-31G(d) level of theory suggest that only small differences exist for lower frequency modes between the polarizable force field and the CHARMM22 force field, and the agreement with ab initio force field analyses is are also of similar quality both in computed vibrational frequencies and contributing vibrational motions.

5. Conclusions

A polarizable intermolecular potential function (PIPF) employing the Thole interacting dipole (TID) polarization model has been developed for liquid alkanes and amides. In connection with the internal force field terms of the CHARMM22 force field, with minor modifications of several torsional terms only, the present PIPF-CHARMM potential provides an adequate description of structural and thermodynamic properties for liquid alkanes and for liquid amides. The computed heats of vaporization and liquid density are within 1.4% of experimental values. Although polarization effects are negligible in liquid alkanes, they make major contributions to the potential energy of liquid amides. The average molecular dipole moments are enhanced by 1.5−1.8 D for primary and secondary amides, from gas-phase values of about 3.3−3.7 D to condensed-phase values of 5−5.4 D. This represents as large as a 50% increase from induction polarizations and is reflected by the computed polarization contributions, ranging from 6 to 24% of total potential energies. The average induced dipoles are nearly twice as large as a previous set of polarizable potentials, making use of Applequist-like atomic polarizabilities without intramolecular interactions, but they are in closer agreement with combined QM/MM simulations in which one molecule of amide was treated quantum-mechanically in a liquid of the same amides represented classically.[14] The ability of the PIPF-CHARMM force field to treat protein backbone structures is tested by examining the potential energy surface of the amide bond rotation in *N*-methylacetamide and the Ramachandran surface for alanine dipeptide. The agreement with ab initio MP2 results and with the original CHARMM22 force field is encouraging, suggesting the PIPF-CHARMM potential can be used as a starting point to construct a complete polarizable force field for proteins.

Acknowledgment. This work was partially supported by the National Institutes of Health.

Supporting Information Available: Four tables containing results from vibrational force field analyses using the present PIPF-CHARMM potential, the CHARMM22 force field, and ab initio LMP2/cc-pVQZ(-g)//MP2/6-31G-(p,d) method plus a full listing of the PIPF-CHARMM force field parameters for alkanes and amides. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) MacKerell, A. D., Jr. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Computat. Chem.* **2004**, *25*, 1584–1604.

(2) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford University Press: London, 1987.

(3) Stillinger, F. H.; Weber, T. A.; David, C. W. *J. Chem. Phys.* **1982**, *76*, 3131.

(4) Van Belle, D.; Couplet, I.; Prevost, M.; Wodak, S. J. *J. Mol. Biol.* **1987**, *198*, 721.

(5) Niesar, U.; Corongiu, G.; Clementi, E.; Kneller, G. R.; Bhattacharya, D. K. *J. Phys. Chem.* **1990**, *94*, 7949.

(6) Sprik, M.; Klein, M. L.; Watanabe, K. *J. Phys. Chem.* **1990**, *94*, 6483.

(7) Wallqvist, A.; Berne, B. J. *J. Phys. Chem.* **1993**, *97*, 13841.

(8) Rick, S.; Stuart, S;, Berne, B. *J. Chem. Phys.* **1994**, *101*, 6141.

(9) Bernardo, D. N.; Ding, Y.; Krogh-Jespersen, K.; Levy, R. M. *J. Comput. Chem.* **1995**, *16*, 1141.

(10) Ding, Y.; Bernardo, D. N.; Krogh-Jespersen, K.; Levy, R. M. *J. Phys. Chem.* **1995**, *99*, 11575.

(11) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.

(12) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048.

(13) Gao, J.; Habibollazadeh, D.; Shao, L. *J. Phys. Chem.* **1995**, *99*, 16460.

(14) Gao, J.; Pavelites, J. J.; Habibollazadeh, D. *J. Phys. Chem.* **1996**, *100*, 2689.

(15) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730.

(16) Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694.

(17) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.

(18) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

(19) Rasmussen, T. D.; Ren, P.; Ponder, J. W.; Jensen, F. *Int. J. Quantum Chem.* **2007**, *107*, 1390.

(20) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

(21) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2005**, *109*, 18988.

(22) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1.

(23) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1504.

(24) Patel, S.; Brooks, C. L., III *J. Chem. Phys.* **2005**, *123*, 164502.

(25) Ponder, J. W.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27.

(26) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.

(27) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, S.; Fischer, M. J.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(28) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(29) Gao, J. *J. Phys. Chem. B* **1997**, *101*, 657.

(30) Gao, J. *J. Chem. Phys.* **1998**, *109*, 2346.

(31) (a) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471. (b) Chen, B.; Ivanov, I.; Klein, M. C.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *91*, 215503.

(32) (a) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300. (b) Kuo, I.-F. W.; Mundy, C. J. *Science* **2004**, *303*, 658.

(33) (a) Morgantini, P.-Y.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 6057. (b) Menge, E. C.; Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 2367. (c) Ding, Y.; Bernardo, D. N.; Krogh-Jespersen, K.; Levy, R. M. *J. Phys. Chem.* **1995**, *99*, 11575. (d) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775. (e) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827.

(34) (a) Stuart, S. J.; Berne, B. J. *J. Phys. Chem.* **1996**, *100*, 11934. (b) Stuart, S. J.; Berne, B. J. *J. Phys. Chem. A* **1999**, *103*, 10300.

(35) Dang, L. X.; Chang, T.-M. *J. Chem. Phys.* **1997**, *106*, 8149.

(36) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89.

(37) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1998**, *89*, 7556.

(38) Nadig, G.; Van, Zant, L. C.; Dixon, S. L.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1998**, *120*, 5593.

(39) Sanderson, R. T. *Science* **1951**, *114*, 670.

(40) Parr, R. G.; Yang, W. *Density-functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, 1989.

(41) Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315.

(42) Rappé, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.

(43) York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *104*, 159.

(44) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741.

(45) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(46) Thole. B. T. *Chem. Phys.* **1981**, *59*, 341.

(47) van Duijnen, P. T.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399.

(48) Van Belle, D.; Froeyen, M.; Lippens, G.; Wodak, S. *Mol. Phys.* **1992**, *77*, 239.

(49) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400.

(50) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(51) Feller, S. E.; Zhang Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613.

(52) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117.

(53) (a) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665. (b) Horn, H. W.; Swope, W. C.; Pitera, J. W. *J. Chem. Phys.* **2005**, *123*, 194504/1. (c) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.

(54) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(55) Lide, D. R. *CRC Handbook of Chemistry and Physics*; CRC Press: U.S.A., 2006.

(56) Yaws, C. L. *Chemical Properties Handbook*; McGraw-Hill, 1999.

(57) Riddick, J. A.; Bunger, W. B. *Organic Solvents*, 3rd ed.; Wiley-Interscience: New York, 1970.

(58) Applequist, J.; Carl. J. R.; Fung, K. *J. Am. Chem. Soc.* **1972**, *94*, 2952.

(59) Jorgensen, W. L; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638.

(60) Daubert, T. E.; Bartakovits, R. *Ind. Eng. Chem. Res.* **1989**, *28*, 641.

(61) Reid, R. C.; Praunitz, J. M; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987; pp 42, 97, 136−143, 209.

(62) Lemire, R. J.; Sears, P. G. *Top. Curr. Chem.* **1978**, *74*, 45.

(63) Covington, A. K.; Dickinson, *T. Physical Chemistry of Organic Solvent Systems*; Plenum: London, 1973.

(64) Gopal, R.; Rigzi, **S.** A. *J. Ind. Chem. Soc.* **1966**, *43*, 179.

(65) Geller, B. Zegers, H. C.; Somsen, *G. J. Chem. Thermodyn.* **1984**, *16*, 225.

(66) Somsen, G.; Coops, J. *Rec. Trau. Chim.* **1965**, *84*, 985.

(67) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664.

(68) Drakenberg, T.; Dahlqvist, K.-I.; Forsen, S. *J. Chem. Phys.* **1972**, *76*, 2178.

(69) Greiner-Schmid, A.; Wappmann, S.; Has, M.; Lüdemann, H.-D. *Chem. Phys.* **1991**, *94*, 5643.

(70) Chen, L.; Gross, T.; Lüdemann, H.-D. *Z. Phys. Chem. (Muenchen)* **2000**, *214*, 239.

(71) Kalman, E.; Serke, I.; Palinkas, G.; Zeidler, M. D.; Wiesmann, F. J.; Bertagnolli, H.; Chieuz, P. *Z. Naturforsch., A: Phys. Sci.* **1983**, *38A*, 231.

(72) DeSando, R. J.; Brown, G. H. *J. Phys. Chem.* **1968**, *72*, 1088.

(73) Ohtaki, H.; Funaki, A.; Rode, B. M.; Reibnegger, G. J. *Bull. Chem. Soc. Jpn.* **1983**, *56*, 2116.

(74) Ohtaki, H.; Katayama, N.; Ozutsumi, K.; Radnai, T. *J. Mol. Liq.* **2000**, *88*, 109.

(75) Nasr, S.; Mounir, Ghédira, M.; Cortès, R. *J. Chem. Phys.* **1999**, *110*, 10487.

(76) Hammami, F.; Nasr, S.; Bellissent-Funel, M.; Oumezzine, M. *J. Phys. Chem. B* **2005**, *109*, 16169.

(77) Hammami, F.; Nasr, S.; Oumezzine, M.; Cortès, R. *Biomol. Eng.* **2002**, *19*, 201.

(78) Trabelsi, S.; Bahri, M.; Nasr, S. *J. Chem. Phys.* **2005**, *122*, 024502.

(79) Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 569.

(80) Borrmann, H.; Persson, I.; Sandström, M.; Stålhandske, C. M. V. *J. Chem. Soc., Perkin Trans.* **2000**, *2*, 393.

(81) Takamuku, T.; Matsuo, D.; Tabata, M.; Yamaguchi, T.; Nishi, N. *J. Phys. Chem. B* **2003**, *107*, 6070.

(82) Drakenberg, T.; Forsén, S. *Chem. Commun.* **1971**, *21*, 1404.

CT700146X

JCTC Journal of Chemical Theory and Computation

Design of a Next Generation Force Field: The X-POL Potential

Wangshen Xie[†] and Jiali Gao*[,†,‡]

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, and Centro Nacional de Supercomputación, Programa Biología Computacional, C/ Jordi Girona 29, 08034 Barcelona, Spain

Received July 4, 2007

Abstract: An electronic structure-based polarization method, called the X-POL potential, has been described for the purpose of constructing an empirical force field for modeling polypeptides. The X-POL potential is a quantum mechanical model, in which the internal, bonded interactions are fully represented by an electronic structure theory augmented with some empirical torsional terms. Nonbonded interactions are modeled by an iterative, combined quantum mechanical and molecular mechanical method, in which the molecular mechanical partial charges are derived from the molecular wave functions of the individual fragments. In this paper, the feasibility of such and electronic structure based force field is illustrated by small model compounds. A method has been developed for separating a polypeptide chain into peptide units, and its parametrization procedure in the X-POL potential is documented and tested on glycine dipeptide. We envision that the next generation of force fields for biomolecular polymer simulations will be developed based on electronic structure theory, which can adequately define and treat many-body polarization and charge delocalization effects.

1. Introduction

Molecular mechanics or force fields, employing empirical potential energy functions,[1] play a central role in computer simulations, which ultimately determine the accuracy of computational results.[2] Although remarkable success has been achieved in molecular dynamics and Monte Carlo simulations of liquids, solutions, and biopolymers, thanks to the development and careful validation of several force fields,[3−8] there are also a number of deficiencies in the current generation of force fields. First, the choice of the energy terms and the associated degrees of freedom are arbitrary in defining a force field.[9] As a result, different force fields often have seemingly very different parameters (for example, partial charges), but the computed dynamic and thermodynamic properties can be similar. Second, by and large, most force fields make use of the harmonic approximation to bond stretch and angle bending, and there is a lack of consideration of the coupling among different energy terms. Of course, a number of force fields do include anharmonicity and cross terms,[7,8,10,11] but there is no systematic way of improving the functional form and its performance because little is known about their effects in biomolecular simulations. Empirical force fields can be developed by parametrizing against observed vibrational frequencies.[1,9,11] Third, the treatment of many-body polarization effects suffers from difficulties in choosing the functional form and empirical parameters.[8,12−21,29] Just as atomic partial charges, there is no rigorous definition of atomic (or group) polarizabilities, nor is it measurable experimentally. There are many ways of describing molecular polarization, which of course is a well-defined quantity. Fourth, charge-transfer effects are ignored in all empirical force fields, and there is no obvious way of including these effects.[22−25] Although the energy contributions are typically small in most applications, charge transfer can be important in certain situations.[23,24] Finally, the form of the empirical potentials is not appropriate for modeling chemical reactions involving bond formation and bond breaking or regions significantly away from the adiabatic ground state. Although specialized potentials[26a] or general functional forms[26b] can be developed to treat specific

* Corresponding author e-mail: gao@chem.umn.edu.

[†] University of Minnesota.

[‡] Centro Nacional de Supercomputación.

10.1021/ct700167b CCC: $37.00
Published on Web 08/23/2007

processes, they are restricted to applications to that particular case only. In view of these difficulties, it is of interest to consider alternative approaches to design a force field for biomolecular simulation and modeling that takes these effects into account. It is our hope that the present paper may contribute to this goal.

Combined quantum mechanical and molecular mechanical (QM/MM) potentials,[27−30] in principle, provide a reasonable solution to all the deficiencies mentioned above, at least for the region that is explicitly treated by quantum mechanics. The internal energy terms used in the force field as well as electrostatic interactions are described by the quantum chemical method used to represent the "QM" region. There is no ambiguity in the functional form, nor in the selection of internal degrees of freedom. Molecular polarization and charge transfer are naturally represented by electronic structure theory. And, of course, such a method can be used to model chemical reactions.[31] Ten years ago, we described a method for treating many-body polarization effects in fluid simulations by combined QM/MM techniques.[32] This method takes a different approach to treat polarization and electronic effects which do not have the problems in the classical polarizable force fields noted above.[8,12−21,29] We had envisioned developing an electronic structure based polarization force field, which hereafter is called the X-POL potential, for biomolecular simulations that naturally includes molecular polarization. In paper 1, we outlined the general principles and formalisms of this approach,[32] and in the second paper, we demonstrated the feasibility of carrying out fluid simulations with such a polarizable potential function for liquid water.[33] This approach was also applied to liquid HF.[34] In this paper, we develop the theory for constructing a force field to treat polypeptides on the basis of the X-POL potential.

We emphasize that the goal here is the development of a molecular-orbital based force field, which is empirical and contains parameters; these parameters shall be optimized to reproduce the properties of a set of target molecular systems following the same philosophy in developing and validating molecular mechanics force fields. Yet, the fundamental elements including the bonded and electrostatic "terms" are determined by electronic structure theory. We do not aim at developing a linear scaling method for a fully solvated protein system,[35−37] although such a treatment is esthetically appealing and the X-POL potential can be developed into a quantum model for biopolymers. This indeed is the approach undertaken in ab initio molecular dynamics and Car−Parrinello molecular dynamics which have been successfully used in numerous applications.[38−40] However, these studies are limited to small systems with relatively short simulations. It would be difficult in the near future to extend the method to much larger systems such as a solvated protein or nucleic acid. Semiempirical models overcome the problem of computational costs,[41−44] but they are considered inaccurate for general applications. Some time ago, Head-Gordon estimated that a Hartree−Fock (HF) or density functional theory (DFT) calculation of a protein of about 10 000 atoms would only be remotely feasible with the assumption of 100-fold increase in computer speed and a true linear scaling in molecular size.[37] Such calculations in a single energy evaluation are interesting, which still represent a daunting task today. However, it is necessary to perform statistical mechanical simulations to obtain ensemble averages in order to compare the computed results with experiments. The need to repeat energy and gradients evaluations for millions of times is the computation bottleneck in electronic structure methods for condensed phase systems.

Clearly, the only possibility that HF or DFT methods can be used as a reliable force field in biomolecular simulations is to introduce approximations, keeping in mind that it is necessary to be able to evaluate the energy and gradients for a given configuration within a few seconds of time. To this end, an X-POL potential, which was also called the molecular-orbital derived empirical potential for liquids (MODEL),[32] has been constructed for simulations of liquid water and hydrogen fluoride,[33,34] in which three approximations were made: (a) the wave function of the entire system is constructed as a Hartree product of the antisymmetric wave functions of individual molecules, (b) the interactions between any pair of residues are evaluated by combined QM/MM techniques, and (c) the electronic structure of individual molecules are treated by a semiempirical HF model. In this method, the electronic structure of each solvent molecule or amino acid residue in a polypeptide is influenced and polarized by the electrostatic field generated from the rest of the system, which in turn affects the wave functions of other molecules. Consequently, the total energy of the system is determined self-consistently. Clearly, many-body polarization effects are naturally described by electronic structure theory. The main advantage of this approach is that the treatment of molecular polarization as well as other energy components can be systematically improved by using ab initio HF, DFT, or advanced techniques such as perturbation, multiconfiguration, and couple cluster theories.

The X-POL potential also differs from combined QM/MM approaches that treat the induced dipoles classically in the MM region.[28,29,45−50] The difficulties and uncertainty of representing polarization in classical force field[8,12−21,29] still exists in these coupled QM/MM-pol methods, whereas the X-POL potential treats the entire system equally. Furthermore, charge-transfer effects can easily be included in the X-POL potential, which would be exceedingly difficult in the classical treatment. The X-POL force field is designed as a quantum mechanical model for biomolecular simulations.

In what follows, we first review the X-POL potential for treatment of liquid systems without covalent-bond connection between molecules. Then, in section 3, we introduce the new theory for treating polypeptides in which covalent bonding connections between residues must be separated. Section 4 presents the algorithm and computational details, and section 5 highlights the results and parametrization process. Finally, we summarize the major findings of this study and future perspectives.

2. Theoretical Background

For completeness, this section reviews the method presented in refs 32 and 33. The next section contains the new methodology developed in the present work. We first consider a system consisting of *N* molecules that are not covalently connected in a primary unit cell with periodic boundary conditions along with nearest image convention. For the sake of brevity, we assume this is a simple liquid system with identical solvent molecules such as liquid water; obviously there is no restriction for solutions and mixed solvents in the method presented below.[33] To focus our discussion, we assume that the readers are familiar with combined QM/MM methods. We make the first assumption that the wave function of the liquid system (Φ) is represented by a Hartree product of the antisymmetric wave functions of the individual molecules, $\{\Psi_I; I = 1, \cdots, N\}$

$$\Phi = \prod_{I=1}^{N} \Psi_I \tag{1}$$

where the individual molecular wave function is written as a Slater determinant of *M* doubly occupied molecular orbitals (MOs), $\{\phi_i(I)\}$ with 2*M* electrons in each molecule. As usual, the MOs are linear combinations of an atomic orbital basis set, $\{\chi_\mu\}$, spanning over the entire molecule, which are subjected to the orthonormal constraint

$$\Lambda_{ij}(I) = \sum_{\mu\nu} c_{i\mu}(I) c_{j\nu}(I) S_{\mu\nu}(I) - \delta_{ij} = 0 \tag{2}$$

where $S_{\mu\nu}(I)$ is the overlap integral between atomic orbitals χ_μ and χ_ν in molecule *I*.

The assumption made in eq 1 neglects the exchange correlation interactions between molecules, thus, the entire system does not satisfy the Pauli exclusion principle, but this approximation is quite reasonable in the spirit of a force field development. To account for the short-range exchange repulsion and the long-range dispersion interactions, we use an empirical function to parametrically model these effects, and we adopt the popular Lennard-Jones potential

$$E_{IJ}^{\mathrm{vdW}} = \sum_{\alpha=1}^{A} \sum_{\beta=1}^{B} 4\epsilon_{\alpha\beta} \left[\left(\frac{\sigma_{\alpha\beta}}{R_{\alpha\beta}} \right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{R_{\alpha\beta}} \right)^{6} \right] \tag{3}$$

where *A* and *B* are the number of atoms in molecules *I* and *J*, which are the same in this discussion, and the parameters $\epsilon_{\alpha\beta}$ and $\sigma_{\alpha\beta}$ can be derived using the combining rules such that $\epsilon_{\alpha\beta} = (\epsilon_\alpha \epsilon_\beta)^{1/2}$ and $\sigma_{\alpha\beta} = (\sigma_\alpha + \sigma_\beta)/2$. ϵ and σ are atomic empirical parameters that are considered to have the same meaning and treatment as in a typical MM force field, and they depend on the specific functional type.

The Hamitonian of the system can be written as

$$\hat{H} = \sum_{I=1}^{N} \hat{H}_I^o + \frac{1}{2} \sum_{I=1}^{N} \sum_{J \neq I}^{N} \hat{H}_{IJ} \tag{4}$$

where $\hat{H}_I^o$ is the electronic Hamiltonian of an isolated molecule in the gas phase, and $\hat{H}_{IJ}$ describes the interactions between molecules *I* and *J*. The interaction Hamitonian can be expressed by eq 5:

$$\hat{H}_{IJ}(\Psi_J) = -\sum_{i=1}^{2M} V_i(\Psi_J) + \sum_{\alpha=1}^{A} Z_\alpha(I) V_\alpha(\Psi_J) + E_{IJ}^{\mathrm{vdW}} \tag{5}$$

Here, $Z_\alpha(I)$ is the nucleus charge of atom α in molecule *I*, and $V_t(\Psi_J)$ is the electrostatic potential of molecular *J* at either the electronic ($t = i$) or nuclear ($t = \alpha$) positions of molecule *I*. The electrostatic potential due to molecule *J* is defined as follows

$$V_t(\Psi_J) = -\int \frac{d\mathbf{r}\rho_J(\mathbf{r})}{|\mathbf{r}_t - \mathbf{r}|} + \sum_{\beta=1}^{B} \frac{Z_\beta(J)}{|\mathbf{r}_t - \mathbf{R}_\beta(J)|} \tag{6}$$

where $\rho_J(\mathbf{r})$ is the electron density of molecule *J*, derived from the molecular wave function, $\rho_J(\mathbf{r}) = |\Psi_J(\mathbf{r})|^2$.

The total potential energy of the system is

$$E_{\mathrm{tot}} = <\Phi|\hat{H}|\Phi> - \sum_{I=1}^{N} E_I^o \tag{7}$$

where $E_I^o = <\Psi_I^o|\hat{H}_I^o|\Psi_I^o>$ is the energy of molecule *I* in the gas phase with the wave function Ψ_I^o, which has a constant value and is used here purely for setting the zero of energy of the condensed phase system corresponding to that of infinitely separated or noninteracting species.

In principle, eq 7 can be determined by standard HF theory with or without optimization of the instantaneous molecular wave function Ψ_I in the presence of all other molecules. Of course, the method above is not restricted to HF theory and can be equivalently written in the form of DFT or any other electronic structure methods. For exampled, Wesolowski and co-workers have used a frozen density functional method for large systems without optimization of the electron density of individual fragments.[51,52] The fragmental molecular orbital method developed by Kitaura and co-workers allows for full optimization of the wave function.[53]

Without further approximation, it is necessary to compute the two-electron integrals arising from different molecules, which would be too expensive for a force field. Fortunately, this problem can be adequately treated by a combined QM/MM approach, which is the second assumption of the X-POL potential. Here, the electronic integral in eq 6 is expressed as a multipole expansion on molecule *J*. The two-electron two-center Coulomb integrals can also be evaluated in exactly the same way as that described by Dewar and Thiel in semiempirical NDDO methods.[54] Alternatively, if we only use the monopole term, i.e., partial atomic charges, the interaction Hamiltonian can be simplified to

$$\hat{H}_{IJ}(\Psi_J) = -\sum_{i=1}^{2M} \sum_{\beta=1}^{B} \frac{e \cdot q_\beta(\Psi_J)}{|\mathbf{r}_i - \mathbf{R}_\beta(J)|} + \sum_{\alpha=1}^{A} \sum_{\beta=1}^{B} \frac{Z_\alpha(I) q_\beta(\Psi_J)}{R_{\alpha\beta}} + E_{IJ}^{\mathrm{vdW}} \tag{8}$$

where $q_\beta(\Psi_J)$ is the partial atomic charge on atom β of molecule *J*, fitted to reproduce the electrostatic potential of eq 6 from the wave function Ψ_J, and $R_{\alpha\beta}$ is the distance between two atoms. Previously, we have shown that intermolecular interactions can be adequately described simply by scaling the Mulliken population charges in the simulation of liquid water.[33] In developing an X-POL force field for biopolymers, it is desirable to include at least the dipolar expansion terms.

With this treatment, the potential energy of eq 7 is consistently optimized to obtain the ground-state potential energy of the system.

Recently, Gascon et al.[55] described a self-consistent space-domain decomposition method for computing electrostatic potentials of proteins. The method reported in that work appears to be the same as that described above except that Morokuma's ONIOM and the 6-31G(d) basis set were used.[56,57] Surprisingly, these authors do not appear to be aware of the method reported in refs 32−34. Soon after the publication of ref 32, Field also described a similar implementation making use of both the AM1 and HF/STO-3G method.[58]

3. The Electronic-Structure Polarization Force Field for Proteins

For biomolecular systems such as proteins and nucleic acids, the division of the entire system into individual molecular fragments is not obvious because each residue is covalently connected to its neighbors.[29,59−64] In this case, it is necessary to decide the basic unit for the "molecular partition" and to consider the effects of charge delocalization between neighboring fragments. In this section, we present a novel procedure for constructing a force field based on molecular orbital theory, for energy minimization and dynamics simulations of proteins.

3.1. Quantum Mechanical Model. Before we begin constructing the X-POL force field, a critical decision must be made on the choice of a specific quantum mechanical model to represent the system. Of course, it would be ideal to use an accurate electronic structure theory such as CCSD-(T) or a well-tested DFT model along with a large basis set. However, these methods are not practical in the foreseeable future, and it is not clear if DFT methods can yield accurate results without rigorous treatment of dispersion interactions. The most practical choice is semiempirical quantum mechanical models coupled with proper parametrizations, such as the self-consistent extended Huckel theory (SC-EHT),[66,67] and the formalisms based on the neglect diatomic differential overlap (NDDO) approximation.[41] The recent parametrization of the self-consistent charge density functional tight-binding (SCC-DFTB) model[68,69] yielded promising results; however, the procedure and the use of tabulated electronic integrals make it difficult for force field development. It would be of interest to improve the atomic parameters in the SC-EHT method. We find that the general formalisms used in the MNDO,[70] AM1,[42] and PM3[43] models are most appealing because the theory is well-defined and has been extensively tested. The semiempirical formalisms contain atomic parameters, and the total number of parameters are no more than those associated with a given atom type in the current empirical force fields.

We anticipate that the semiemipirical parameters will be fully optimized for each functional group in the amino acid residues, keeping in mind that we are interested in developing a *force field* rather than a "general" QM model. The parametrization will necessarily include optimizations of the molecular geometries (including radial distribution functions), energies (such as heats of formation), electronic structural properties (such as molecular dipoles, electron affinities, and ionization potentials), spectroscopic data (for example, vibrational frequencies, NMR chemical shifts), and condensed phase properties (such as heats of vaporization, density, diffusion constants, relaxation time, and solvation free energies) among others for a set of selected compounds representing different functionalities. Of course, the X-POL force field can be systematically improved by increasing the level of the QM theory employed. It is expected that the parametrization process will become less dependent on fitting against experimental data (or high-level QM results) as the level of the QM model increases. In this paper, our focus is on the theory for constructing a polypeptide chain represented by a QM model. We adopt the Austin Model 1 (AM1)[42] method to demonstrate the method without further optimization for specific functional groups, except the boundary atoms discussed below.

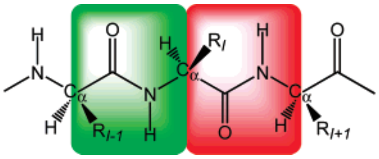


Figure 1. Definition of peptide units and the division of the C_α boundary atom. Two quantum mechanical fragments are highlighted in green and red, corresponding to residues *I-1* and *I*.

3.2. System Partition and Boundary Definition. We consider a system of polypeptide of N residues, which is divided into N subunits or fragments. The interactions among different subunits are determined through a combined QM/MM algorithm. It would be natural to use the formal chemical structure of each amino acid residue as the "QM" subunit; however, it is more appropriate to keep intact the resonance delocalization in a peptide bond in electronic structure calculations. Thus, we adopt the "peptide unit" convention defined in the IUPAC nomenclature (*Pure Appl. Chem.* **1974**, *40*, 291−308. http://www.chem.qmul.ac.uk/iupac/misc/ppep1.html) which consists of the −CHR−CO−NH− atoms. For our computational purpose, which will become clear below, we make the sequence separation across the C_α atoms of adjacent residues, as illustrated in Figure 1. Thus, the Ith peptide unit contains the atoms $-C_\alpha^I R^I-CO-NH-C_\alpha^{I+1}H-$, in which $NH-C_\alpha^{I+1}H$ belongs to the (I+1)th amino acid residue. In our definition, the C_α atoms are equally *shared* by adjacent peptide units. In the following, we follow the IUPAC recommendation to simply refer the "peptide unit" as a "residue" when no ambiguity arises.

In this partition scheme, each residue (peptide unit) shares two C_α atoms with the neighboring residues, except the N- and C-termini. We call these atoms the boundary atoms (Figure 1). With the use of a semiempirical QM model, the boundary carbon atom has the standard valence s and p orbitals and four valence electrons. Adopting the generalized hybrid orbital (GHO) approach for the treatment of a QM-MM boundary in combined QM/MM calculations,[59,60] we make the same transformation of the sp atomic orbitals (AOs)

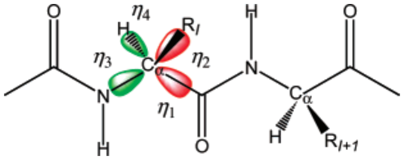


Figure 2. Assignment and sequence of hybrid orbitals on the boundary atom. Hybrid orbitals in red and green belong to different QM fragments.

on the boundary atom into a set of four orthonormal hybrid orbitals (HOs), on the basis of the local geometry about the C$_\alpha$ atom

$$\begin{pmatrix} \eta_1^I \\ \eta_2^I \\ \eta_3^I \\ \eta_4^I \end{pmatrix} = [\mathbf{T}_\mathrm{B}(I)]^{-1} \begin{pmatrix} s_1^I \\ p_x^I \\ p_y^I \\ p_z^I \end{pmatrix} \tag{9}$$

where the superscript I specifies that the orbitals are located on atom C$_\alpha$(I), and the subscript B indicates that the transformation matrix has dimensions of (4 × 4) for the boundary atom. The transformation matrix in eq 9 is defined by the geometry of the four atoms bounded to the boundary atom and their local coordinates, and its expression has been given in ref 60. We note that the hybrid orbitals are orthonormal by construction, and they are also orthogonal to all other AO basis functions due to the NDDO approximation. In ab initio HF theory, the HOs need to be orthogonalized to the rest of the basis functions, and procedures have been described previously.[61,63]

In defining the HOs in eq 9, the transformation matrix, $\mathbf{T}_\mathrm{B}(I)$, is constructed in such a way that the orientations of the four HOs are pointing sequentially toward the carbonyl carbon, the C$_\beta$(I) (or H$_{\alpha 2}$ for glycine) atom of the side chain, the amino nitrogen of the (I-1)th peptide unit, and the H$_\alpha$ atom (Figure 2). Therefore, there are two boundary atoms and four boundary hybrid orbitals in the Ith residue in the present QM partition: the first two hybrid orbitals, η_1^I and η_2^I, on the C$_\alpha$(I) atom and the third and the fourth hybrid orbitals, η_3^{I+1} and η_4^{I+1}, on the C$_\alpha$(I + 1) atom. Assuming that there are N_I atomic orbital basis functions on all other atoms in residue I, we have a total of $N_I + 4$ basis functions, called active orbitals, that are mixed atomic and hybrid orbitals to form the MOs of the subsystem. We note that, similar to the GHO method,[59−62] the remaining four hybrid orbitals, two from C$_\alpha$(I) and two from C$_\alpha$(I + 1), will also be used in constructing the Fock matrix, but they are not variationally optimized in self-consistent field (SCF) calculations of residue I.

3.3. Potential Energy Surface and Procedure. In optimizing the antisymmetric wave function Ψ_I, of residue I, we note that this QM subunit is embedded in the electric field of classical partial charges (or multipoles if higher order of density expansion is used) of the rest of the system, although they are also treated quantum-mechanically. Thus, in principle, the computational procedure is identical to that used in the GHO method for combined QM/MM systems,[59,60] except that two "active" hybrid orbitals from each boundary atom participate in the SCF optimization. Specifically, if the density matrix for residue I is $\mathbf{P}^\mathrm{H}(I)$, which has dimensions of $[N_I + 4] \times [N_I + 4]$, the total *interaction energy* between residue I and the rest of the system is

$$E_I = \sum_{\mu\nu} P^\mathrm{H}_{\mu\nu}(I) H^\mathrm{H}_{\mu\nu}(I) + 1/2 \sum_{\mu\nu}\sum_{\lambda\sigma} P^\mathrm{H}_{\mu\nu}(I) P^\mathrm{H}_{\mu\nu}(I) W^\mathrm{H}(\mu\nu,\lambda\sigma) + E^\mathrm{nuc}_{IP} + E^\mathrm{nuc}_{I} - E^o_I \tag{10}$$

where the superscript H indicates that all matrix elements are given in the mixed AO and HO basis, $H^\mathrm{H}_{\mu\nu}(I)$ is an element of the "effective" one-electron integral matrix that includes the interaction Hamiltonian of eq 8, W^H $(\mu\nu,\lambda\sigma)$ is the usual two-electron integrals including both Coulomb $(\mu\nu,\lambda\sigma)$ and exchange $(\mu\lambda,\nu\sigma)$ terms, and E^nuc_I and E^nuc_{IP} are the nuclear Coulombic energies within residue I and that with the rest of the protein system, respectively. The effective Hamiltonian matrix element $H^\mathrm{H}_{\mu\nu}(I)$ is given below[59,60]

$$H^\mathrm{H}_{\mu\nu}(I) = H^\mathrm{o,H}_{\mu\nu}(I) + J^\mathrm{H}_{\mu\nu}(I) + \frac{1}{2}\sum_{i,j=3}^{4} P^\mathrm{H}_{\eta_i\eta_j}(I) W^\mathrm{H}(\mu\nu,\eta_i^I\ \eta_j^I) + \frac{1}{2}\sum_{i,j=1}^{2} P^\mathrm{H}_{\eta_i\eta_j}(I+1) W^\mathrm{H}(\mu\nu,\eta_i^{I+1}\ \eta_j^{I+1}) \tag{11}$$

where η_i specifies the boundary auxiliary orbitals from residues (I) and (I +1) that are *not* optimized in the SCF for residue I, $H^\mathrm{o,H}_{\mu\nu}(I)$ is the standard one-electron matrix for residue I, $J^\mathrm{H}_{\mu\nu}(I)$ is the "QM/MM" one-electron integral due to the first term of eq 8 summed over all other residues other than I, and $P^\mathrm{H}_{\eta_i\eta_j}(I)$ and $P^\mathrm{H}_{\eta_i\eta_j}(I+1)$ are the populations of the auxiliary hybrid orbitals specified by eq 11.

The total potential energy of the entire system is

$$E_\mathrm{tot} = \frac{1}{2}\sum_{I=1}^{N} E_I \tag{12}$$

We further define the interaction energy between residues I and J by[32,33]

$$E_{IJ} = \frac{1}{2}[<\Psi_I|\hat{H}_{IJ}|\Psi_I> + <\Psi_J|\hat{H}_{JI}|\Psi_J>] \tag{13}$$

This is necessary to ensure that $E_{IJ} = E_{JI}$ because in combined QM/MM calculations the two integrals in the bracket parentheses may not be identical.

Although the wave function of eq 1 can be variationally optimized for all residues simultaneous in each SCF cycle, it is more convenient to sequentially optimize the wave function of each residue, by keeping the partial charges (derived from the corresponding wave functions) of the rest of the system fixed. Thus, we have a double iterative SCF procedure: (1) the SCF optimization of the wave function of each residue and (2) the SCF optimization of the mutual polarization of the entire system. Specifically, after the individual wave functions for all residues are converged, which constitute one iterative cycle in the "system" SCF, we check the convergence of the total energy of the system in eq 11. This is repeated until the total energy is converged to a given tolerance. This double-SCF procedure has been used previously in our treatment of liquid water in Monte

Carlo simulations; typically less than 5 system iterations are sufficient to achieve convergence.[33]

4. Computational Details

The computational procedure of the X-POL force field follows roughly the same approaches outlined in refs 32 and 33. However, there are several new aspects that need to be defined here. It is clear that eq 11 requires the density matrix elements, $\{P^{H}_{\eta_i\eta_j}(I);\ I = 1, \cdots, N\}$, for the auxiliary orbitals on the two boundary atoms in each residue during SCF optimizations. In the original GHO method developed for QM/MM calculations, the density is obtained by transferring the partial atomic charge on the boundary carbon from the MM force field, to the three auxiliary orbitals, plus the density of one electron.[59] In the X-POL potential, there are two hybrid orbitals from each boundary atom, and the nuclear charges are not adjusted as in QM/MM calculations; however, the major difference here is that these auxiliary orbitals are also active orbitals in the neighboring residues, which are fully optimized in SCF calculations. Consequently, the "auxiliary" densities are no longer invariant, but they are dynamically changing due to the change of molecular geometry and instantaneous charge polarizations. Furthermore, these optimized densities provide the necessary input for the auxiliary orbitals in the neighboring residues. At convergence, the chemical potential of these hybrid orbitals (active and auxiliary and vice versa) are fully equalized.

The next critical issue is to define an appropriate procedure for determining the partial charges for all other residues in eq 8 when the wave function of residue I is being optimized. A number of possibilities are available, including Mulliken population charges,[71] electrostatic potential fitted charges,[72] and the class IV (CM2) charges proposed by Cramer and Truhlar.[73−76] A good charge mapping procedure will ensure that "QM/MM" interactions be accurately determined in comparison with experimental data—through parametrization of the force field. However, special care must be taken in any charge mapping procedure such that the charge density to be used as the auxiliary density is appropriately neutralized by the atomic charges in the subunit where these "auxiliary orbitals" reside and are "active". This can be done by imposing a charge constraint if an electrostatic potential fitting procedure is used. In the present study, which is aimed at demonstrating the feasibility of the X-POL force field, we adopt the Mulliken population analysis, which properly divides the charge population between the "auxiliary orbitals" (note that they are "active orbitals" in the fragment where they are determined) and the rest of the QM subunit. Clearly, in future development of a reliable force field, both the "classical" representation of the electrostatic potential from a given wave function in terms of multipole expansions and the specific method for obtaining these multipoles shall be a primary focus of study.

Third, in the X-POL treatment, short-range exchange repulsion and long-range dispersion forces are represented by the traditional Lennard-Jones potential. Obviously, these are empirical terms that must be properly optimized against experimental and high-level ab initio results on bimolecular interactions and liquid properties including density. We have demonstrated in Monte Carlo simulations of liquid water using the X-POL approach that the van der Waals parameters in the Lennard-Jones potential can be similarly adjusted as in the development of empirical potentials, e.g., the TIP3P and TIP4P models. In the present study, we employ the corresponding parameters in the CHARMM22 force field without further modification. Note that it might be tempting to use "pure" electronic structure methods to determine the repulsive and dispersive energies; however, this would be futile in force field development if computational speed is taken into consideration.

We present an algorithm for optimization of the individual molecular wave function with the GHO boundary treatment.[59] The convergence of the entire system is achieved by an iterative SCF procedure outline below.

(a) Determine the transformation matrices $\{\mathbf{T}_{\mathbf{t}}(I);\ I = 1, \cdots, N\}$ for the interconversion between the AOs and a set of mixed AOs and HOs, $\mathbf{C}^{HO}_{I} = \mathbf{T}_{\mathbf{t}}(I)^{-1}\mathbf{C}^{AO}_{I}$, where the subscript "**t**" specifies that the matrix has dimensions of $[N_I + 8] \times [N_I + 8]$. Compute the $[N_I + 4] \times [N_I + 4]$-dimensional density matrix, $\{\mathbf{P}^{H}(I);\ I = 1, \cdots, N\}$ using the active HOs for each residue.

(b) For each residue, perform SCF optimization sequentially, beginning from residue $I = 1$.

(c) For residue I, expand $\mathbf{P}^{H}(I)$ into full dimension by adding the auxiliary density matrix elements and transform it into the AO basis $\mathbf{P}^{AO}_{\mathbf{t}}(I) = [\mathbf{T}_{\mathbf{t}}(I)^{-1}]^{+}\ \mathbf{P}^{HO}_{\mathbf{t}}(I)[\mathbf{T}_{\mathbf{t}}(I)^{-1}]$. Construct the full Fock matrix $\mathbf{F}^{AO}_{\mathbf{t}}(I)$ in AO basis, including QM/MM interaction terms. Transform $\mathbf{F}^{AO}_{\mathbf{t}}(I)$ into HO basis, $\mathbf{F}^{HO}_{\mathbf{t}}(I) = [\mathbf{T}_{\mathbf{t}}(I)]^{+}\ \mathbf{F}^{AO}_{\mathbf{t}}(I)[\mathbf{T}_{\mathbf{t}}(I)]$. Remove the columns and rows corresponding to the auxiliary hybrid orbitals to yield the "active" Fock matrix $\mathbf{F}^{AO}(I)$, which is of $[N_I + 4] \times [N_I + 4]$-dimension. Diagonalize $\mathbf{F}^{AO}(I)$ and compute the energy and new density matrix $\mathbf{P}^{H}(I)$.

(d) Test convergence. If not satisfied, go to step **(c)**. If convergence is met, compute new partial charges from the optimized wave function and set the densities of the active HOs as auxiliary densities for other SCF optimizations. Incrementing I by one until $I = N$, and then, go to step **(c)**.

(e) Compute the total energy and test convergence. If not satisfied, go to step **(b)**.

We point out that the matrices transformations in step (c) are quite simple because it only involves orbitals on the boundary atoms (a total of 8 orbitals). Thus, it takes a negligible amount of computer time. As can be seen from the algorithm above, the Fock matrix construction and diagonalization are performed for each individual residue, and there are a total of N separate SCF calculations of the size of each residue in each system iteraction. In our experience on the simulation of liquid water, the total number of system iterations does not increase significantly, perhaps by one or two cycles, with increased system size. Thus, the total computational time is linear scaling by $S \times N \times O(N^{3}_{max})$, where S is the number of iterations in system SCF, and $O(N^{3}_{max})$ is the computing efforts for the largest residue. The difference beween electronic structure calculations for a molecule of the size of $\sum_{I}^{N} N_I$ orbitals and that of N separate calculations of the size of N_{max} is obvious because the former would scale as $O([\sum_{I}^{N} N_I]^3)$ due at least to diagonalization.

Table 1. Modified Parameters for the Carbon Boundary Atom in the Present X-POL Force Field along with the Original AM1 Values and Those Used in the GHO Model[a]

parameters	AM1	GHO	X-POL
β_s	−15.715783	−5.500524	−12.85205
β_p	−7.719283	−14.666638	−5.680080
U_{ss}	−52.028658	−52.028658	−49.774256
U_{pp}	−39.614239	−38.703112	−39.573436

[a] All values are given in eV.

5. Results and Discussion

To illustrate the feasibility of the X-POL force field, we present test cases to demonstrate the procedure for optimizing parameters associated with the boundary atoms and an application to a tetrapeptide model interacting with a single water molecule. Here, we have assumed that the semiempirical AM1 model is adequate for treating the individual residues; obviously, the AM1 model itself is not satisfactory for constructing a force field for protein simulations. However, there is little doubt that they can be parametrized to accurately treat specific functional groups and interactions. The parametrization of the NDDO-based semiempirical QM model for different functional groups and atom types shall be left for future exploration.

5.1. Parametrization of Boundary Atoms. To parametrize the semiempirical force field for the boundary carbon atom and to assess its performance, we consider three model compounds: propane, 2-methylpropane (isobutene), and 2,2-dimethylpropane (neopentane). In each case, a single boundary atom is defined at the C_2 atom position, and two fragmental QM subunits are treated. We aim at the selection of a minimum number of model compounds in the parametrization process to achieve transferability by satisfying key quantum chemical requirements.[59] These include (1) the balance of electronegativity, (2) the properties of chemical bonding, and (3) the conformational potential energy profiles involving the boundary atom. Our experience in the development of the GHO methods, at the semiempirical level,[59,60] semiempirical SCC-DFTB treatment,[62] ab initio HF level,[61a,62] and DFT method,[61b] shows that if the electron-withdrawing power and the formation of the chemical bonds are adequately balanced with the QM model that the boundary atom mimics, the empirical parameters for the boundary atoms are fully transferable, just as all other standard semiempirical parameters or basis sets.

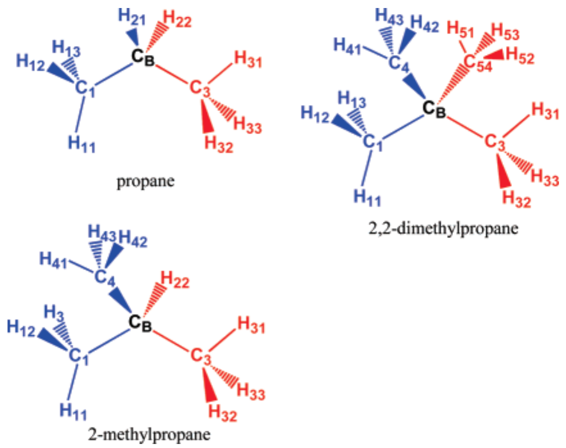
Taking the three criteria listed above into consideration, we found that we only need to modestly modify the parameters of the original AM1 Hamiltonian for carbon. We focused on the one-center core integrals U_{ss} and U_{pp} and the resonance integrals β_s and β_p, the latter of which are closely related to chemical bonding (see below). We slightly decreased these values to obtain the best overall results, but it is closer to that of the original AM1 value. Note that in the full parametrization process of the X-POL potential, the balance with all other atoms will also be consistently considered. The parameters for the boundary carbon atom are listed in Table 1 along with the standard AM1 values and those used in the GHO model.[59]

Table 2. Optimized Bond Lengths (Å) and Bond Angles (deg) Using the X-POL Potential and the Full AM1 Method

	propane		2-methylpropane		2,2-dimentylpropane	
bond	AM1	X-POL	AM1	X-POL	AM1	X-POL
C_B−C1	1.516	1.523	1.523	1.524	1.527	1.523
C_B−C3	1.516	1.526	1.523	1.524	1.527	1.526
C_B−C4(H)	1.118	1.123	1.523	1.524	1.527	1.523
C_B−C5(H)	1.118	1.120	1.121	1.118	1.527	1.519
H11−C1	1.114	1.115	1.114	1.115	1.114	1.115
H12−C1	1.114	1.113	1.114	1.114	1.114	1.114
H13−C1	1.114	1.113	1.114	1.113	1.114	1.114
H31−C3	1.114	1.115	1.114	1.114	1.114	1.114
H32−C3	1.114	1.113	1.114	1.115	1.114	1.115
H33−C3	1.114	1.113	1.114	1.114	1.114	1.114
H41−C2			1.114	1.113	1.114	1.114
H42−C4			1.114	1.115	1.114	1.115
H43−C4			1.114	1.113	1.114	1.114
H51−C5					1.114	1.115
H52−C5					1.114	1.114
H53−C5					1.114	1.114

	propane		2-methylpropane		2,2-dimentylpropane	
angle	AM1	X-POL	AM1	X-POL	AM1	X-POL
C1−C_B−C3	111.87	107.25	110.78	114.75	109.5	114.6
C1−C_B−C4			110.96	107.34	109.5	107.2
C1−C_B−C5					109.5	106.5
C3−C_B−C4			110.78	107.25	109.5	107.3
C3−C_B−C5					109.5	106.7
C4−C_B−C5					109.5	114.7

Scheme 1. Atom Numbers Assigned to the Three Alkanes Which Are Separated into Two Quantum Mechanical Fragments Across a Boundary Atom C_B



The semiempirical parameters for the resonance integrals, β_s and β_p, are most directly responsible for chemical bonding and molecular geometry. The optimized parameters in the table show that the X-POL potential has very similar values compared with the original AM1 parameters. This is in contrast to the GHO method,[59] which does not have the double self-consistent field treatment to optimize the auxiliary hybrid orbital densities. The bond lengths and selected bond angles of propane, 2-methylpropane, and 2,2-dimethylpropane, optimized using the X-POL potential and the AM1 method, are given in Table 2 (see Scheme 1 for atom assignment). The key parameters to be examined are the bond

Table 3. Mulliken Population Charges (au) Obtained Using the X-POL Potential and the Full AM1 Method. Values in Parentheses Are Sums Over Hydrogens

	propane		2-methylpropane		2,2-dimentylpropane	
bond	AM1	X-POL	AM1	X-POL	AM1	X-POL
C_B	−0.160	−0.159	−0.111	−0.107	−0.060	−0.054
C1	−0.210 (0.004)	−0.230 (0.003)	−0.206 (0.010)	−0.223 (0.012)	−0.202 (0.015)	−0.230 (0.013)
H11	0.071	0.072	0.072	0.072	0.072	0.072
H12	0.071	0.084	0.07203	0.085	0.072	0.086
H13	0.071	0.077	0.072	0.078	0.072	0.086
C3	−0.210 (0.004)	−0.235 (−0.003)	−0.206 (0.010)	−0.223 (−0.001)	−0.202 (0.015)	−0.230 (0.007)
H31	0.071	0.072	0.072	0.085	0.072	0.085
H32	0.071	0.083	0.072	0.073	0.072	0.072
H33	0.071	0.077	0.072	0.077	0.072	0.086
C4 (H)	0.076	0.076	−0.206 (0.010)	−0.242 (0.012)	−0.202 (0.015)	−0.236 (0.013)
H41			0.072	0.084	0.072	0.085
H42			0.072	0.073	0.072	0.073
H43			0.072	0.084	0.072	0.085
C5 (H)	0.076	0.083	0.081	0.084	−0.202 (0.015)	−0.225 (0.020)
H51					0.072	0.072
H52					0.072	0.086
H53					0.072	0.086

lengths and bond angles associated with the boundary atom, C_B, which is placed at the C2 position in all three cases. The present comparison is best made with the values optimized using the AM1 model, rather than the experimental or high-level ab initio results, because the main purpose is to evaluate the possibility to parametrize the boundary atom to reproduce the results from the QM model used to describe the QM subunits. For the nine C_B−C bonds and three C_B−H bonds in these three model compounds, the average unsigned error is 0.004 and 0.003 Å, respectively. The agreement with the AM1 results is good. Bond angles see somewhat greater variations mainly because the way that the hybrid orbitals are defined. In the original GHO approach, the hybrid orbital pointing toward the QM fragment is defined based on the local (instantaneous) geometry of the other (MM) three bonds connected to the boundary atom.[59] The remaining three auxiliary hybrid orbitals are created by using Schmidt orthorgonalization and equal hybridization.[60] In the present application, we have adopted the same definition and partitioning scheme for the hybrid orbitals, but, of course, all hybrid orbitals should be treated equally based on the respective local geometry followed by a Lowdin-type orthorgonalization. The latter approach would resolve the slight imbalance caused by the hybridization method. Nevertheless, the optimized bond angles are still reasonable for the present test purposes.

On the other hand, the fundamental criterion necessary to ensure transferability of these atomic parameters for boundary atoms is the balance of the electron-withdrawing power of the boundary atom with that of the QM model that it mimics (AM1 in the present case). Therefore, the boundary atom must have the same electronegativity as that of an "AM1 carbon" atom so that there is no charge transfer between two identical groups. The most relevant parameters in the semiempirical theory are the one-center U_{ss} and U_{pp} terms, which are optimized in connection with the resonance integral parameters (as they are not independent in energy calculations). We found that it is possible to achieve this goal by only optimizing these four parameters listed in Table 1. This is illustrated by the computed Mulliken population charges for the three alkane model compounds. It perhaps should be emphasized here that the Mulliken population[71] is in fact the best charge analysis to examine the balance of the electron-withdrawing abilities of different elements or between atoms of the same type, but they are treated differently (AM1 vs GHO).

Ideally, there is no net charge transfer between two neighboring groups across the boundary atom in the X-POL potential, although a slight variation is inevitable since the boundary atom is, after all, an approximation to the original QM method. Propane is used as the primary target in the parametrization, and the goal is to have as little charge transfer as possible between two fragments via C_B. Table 3 shows that the sum of the total charges of the two QM fragments are nearly the same (0.080 au) both from AM1 and the X-POL calculations, and the partial charge on the boundary carbon only differs by 0.001 au (Scheme 1). To remove the effect of the hydrogen atoms on the C_B atom, which have not been reparameterized, we examine the second symmetric system, neopentane. The unrestrained AM1 calculation yields a total net charge of 0.030 au for two methyl groups, which may be compared to values of 0.026 and 0.027 au from the X-POL potential. Importantly, the two fragments are reasonably balanced. The small difference among the individual methyl groups is again due to the definition of the hybrid orbitals for the boundary atom, which are not completely equivalent. For isobutene, the difference between the two QM fragments is 0.060 au more positive for one methyl group and an H_B than two methyl groups from full AM1 calculations, whereas the difference is 0.059 au. The agreement between AM1 and X-POL partial charges on the C_B atom is also good, which shows the absolute amount of charge imbalance between the boundary atom and the standard carbon.

The torsional potential energy profiles for isobutene about a C−C dihedral are illustrated in Figure 3. The torsional

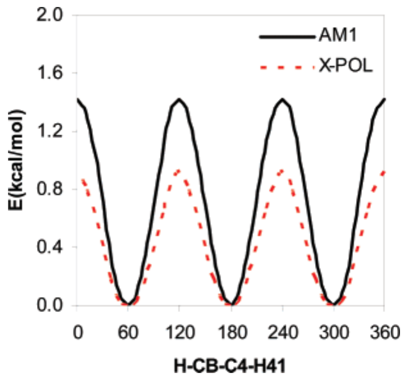


Figure 3. Torsional potential energy profiles for 2-methylpropane (isobutene) about the H22−CB−C4−H41 dihedral angle from AM1 and X-POL optimizations. Energies are in kcal/mol and dihedral angles are in degrees.

Table 4. Mulliken Population Charges (au) for Glycine Obtained Using the X-POL Potential and the Full AM1 Method

		X-POL	
atom	AM1	C_α−C	C_α−N
N	−0.349	−0.350	−0.348
HT1	0.157	0.128	0.126
HT2	0.151	0.180	0.160
H_α1	0.122	0.108	0.120
C_α	−0.047	−0.056	−0.067
H_α2	0.085	0.126	0.120
C	0.302	0.266	0.283
OT1	−0.342	−0.374	−0.355
OT2	−0.314	−0.282	−0.291
HO2	0.236	0.254	0.252

energy from the X-POL potential has contributions from “pure” QM, QM/MM, and “pure” MM (van der Waals) terms, and the computed barrier height is about 0.5 kcal/mol lower than the full AM1 energy. This trend has been observed in the GHO boundary approach,[59] and this difference is easily corrected by adjusting the semiempirical method or by including a classical torsional term. The latter scenario is probably a simple choice for constructing an empirical force field, and this would be the only internal bonding terms required in the X-POL potential.

5.2. Glycine and Glycine-Dipeptide. To examine the possibility that the X-POL potential can be applied to proteins, we optimized the structures of glycine and glycine dipeptide, which are compared with the full original AM1 calculations. Note that the main goal here is to show that the boundary parameters optimized above are transferable to polypeptides and that the development of the X-POL potential will involve full optimization of the QM model itself for different functional groups. Listed in Table 4 are computed partial atomic charges for glycine when the boundary atom is placed on the C_α atom with the first hybrid orbital pointed either toward the carbonyl carbon (C_α−C) or toward the amino nitrogen (C_α−N), as the current definition of the hybrid orbitals still does not yield exactly equivalent hybridizations. The results are compared with the AM1 values, while geometrical parameters are presented in Figure 4.

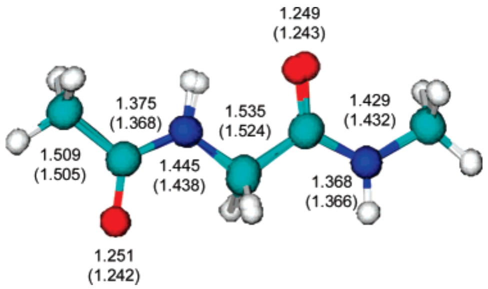


Figure 4. Superposition of the optimized structures using the AM1 and the X-POL potentials for glycine dipeptide. Bond lengths in angstroms are given for the X-POL potential and for the AM1 model in parentheses.

Table 5. Mulliken Population Charges (au) for Glycine Dipeptide Obtained Using the X-POL Potential and the Full AM1 Method

		X-POL	
atom/group	AM1	C_α−C	C_α−N
CH_3	0.060	0.060	0.058
CO	−0.067	−0.091	−0.094
NH	−0.130	−0.132	−0.142
H_α1	0.111	0.096	0.108
C_B	−0.039	−0.002	−0.013
H_α2	0.111	0.128	0.121
CO	−0.085	−0.133	−0.111
NH	−0.147	−0.123	−0.124
CH_3	0.185	0.195	0.197

Overall the partial charges show reasonable transferability. The amino group has a total net partial charge of −0.041 au from AM1, whereas it is −0.042 and −0.062 au from the X-POL potential when the hybrid orbitals are defined based on the C_α−C vector and the C_α−N vector, respectively. For the carboxyl group, the total net charges are −0.136, −0.111, and −0.118 au from AM1, X-POL (C_α−C bond), and X-POL (C_α−N bond), respectively. The difference is only about 0.02 au for such a strong electron-withdrawing group. Overall, there is little charge transfer between the two QM fragments for both partition schemes in comparison with the AM1 partial charges.

The optimized structure of glycine dipeptide for the extended conformation is illustrated in Figure 4 along with some selected bond lengths from the X-POL and AM1 potentials. Overall, the structural agreement is excellent with a root-mean-square difference of 0.007 Å for bonds shown in Figure 4. The combined group charges are given in Table 5; the trend of charge polarization is adequately retained in the X-POL potential, and the charge delocalization across the boundary atom is also good.

6. Summary

An electronic structure-based polarization potential, which is called the X-POL potential, has been described for the purpose of constructing an empirical force field for modeling polypeptides. The X-POL potential takes an entirely different philosophical approach toward the development of a force field and the treatment of electronic polarization and charge delocalization. The internal, bonded interactions are fully represented by an electronic structure theory augmented with some empirical torsional terms. Nonbonded interactions are modeled by an iterative, combined quantum mechanical and molecular mechanical method, in which the molecular mechanical partial charges are derived from the molecular wave functions of the individual fragments. In this paper, the X-POL potential is illustrated by making use of the neglect of diatomic differential overlap (NDDO) approximation and the AM1 model as the quantum mechanical method, without further parametrization for specific functional groups. The main purpose of this study is to demonstrate the feasibility of such an electronic structure force field and to develop a practical and well-defined method for separating a polypeptide chain into peptide units. The boundary is treated following the ideas of the generalized hybrid orbital (GHO) technique developed for the treatment of QM and MM boundaries and extended to bridge two QM regions in the X-POL potential. The parametrization procedure and philosophy for the boundary treatment between QM fragments in the X-POL potential is documented and tested by a number of simple compounds.

The X-POL model presented here is an empirical *force field*, although it is based on quantum mechanical formalisms. If one finds that a particular QM model used or the approximations made are inadequate to treat certain properties, for example, the torsional potential energy profile about a single bond rotation, one can include a purely empirical energy term such as the sine and cosine function series in the current force fields. Although this might be deemed ad hoc, the method is nevertheless systematic in that one can always seek for a better, more accurate QM representation of the individual residues such that these empirical functional terms can be eliminated.

In two forthcoming papers, we describe the analytical energy gradient techniques for the X-POL potential and an application to a solvated protein. The exact treatment and construction of individual force fields in the future may differ from the method presented here, but the general direction seems to be clear. We envision that the next generation of force fields for biomolecular polymer simulations will be developed based on electronic structure theory, which is one way to properly define and treat many-body polarization and charge delocalization effects.

Acknowledgment. The work is supported in part by the National Institutes of Health, the Minnesota Supercomputing Institute, and the Office of Naval Research (ONR) under grant number N00012-05-01-0538. J.G. thanks the Ministerio de Educación y Ciencia, España, for support and Professor Modesto Orozco for his hospitality during his sabbatical leave.

References

(1) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.

(2) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.

(3) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(4) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(5) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(6) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. *Eur. Biophys. J.* **2005**, *34*, 273.

(7) Maple, J. R.; Hwang, M. J.; Jalkanen, K. J.; Stockfisch, T. P.; Hagler, A. T. *J. Comput. Chem.* **1998**, *19*, 430.

(8) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.

(9) Dinur, U.; Hagler, A. T. *Rev. Comput. Chem.* **1991**, *2*, 99.

(10) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730.

(11) Warshel, A. *J. Chem. Phys.* **1994**, *101*, 6141.

(12) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.

(13) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.

(14) Gao, J.; Habibollazadeh, D.; Shao, L. *J. Phys. Chem.* **1995**, *99*, 16460.

(15) Gao, J.; Pavelites, J. J.; Habibollazadeh, D. *J. Phys. Chem.* **1996**, *100*, 2689.

(16) (a) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141. (b) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741.

(17) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.

(18) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504.

(19) Patel, S.; Brooks, C. L., III. *Mol. Simul.* **2006**, *32*, 231.

(20) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.

(21) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587.

(22) Lee, T. S.; York, D. M.; Yang, W. *J. Chem. Phys.* **1995**, *102*, 7549.

(23) Nadig, G.; Van Zant, L. C.; Dixon, S. L.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1998**, *120*, 5593.

(24) Van der Vaart, A.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1999**, *121*, 9182.

(25) Mo, Y.; Gao, J. *J. Phys. l Chem. B* **2006**, *110*, 2976.

(26) (a) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*, 3049. (b) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218.

(27) Gao, J.; Xia, X. *Science* **1992**, *258*, 631.

(28) Gao, J. Methods and applications of combined quantum mechanical and molecular mechanical potentials. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995; Vol. 7, pp 119.

(29) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(30) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

(31) Gao, J. *Acc. Chem. Res.* **1996**, *29*, 298.

(32) Gao, J. *J. Phys. Chem. B* **1997**, *101*, 657.

(33) Gao, J. *J. Chem. Phys.* **1998**, *109*, 2346.

(34) Wierzchowski, S. J.; Kofke, D. A.; Gao, J. *J. Chem. Phys.* **2003**, *119*, 7365.

(35) Lee, T.-S.; York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *105*, 2744.

(36) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1996**, *104*, 6643.

(37) Head-Gordon, M. *J. Phys. Chem.* **1996**, *100*, 13213.

(38) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(39) Rothlisberger, U.; Carloni, P.; Doclo, K.; Parrinello, M. *J. Biol. Inorg. Chem.* **2000**, *5*, 236.

(40) Tuckerman, M. E.; Marx, D.; Parrinello, M. *Nature* **2002**, *417*, 925.

(41) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.

(42) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(43) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(44) Zerner, M. C. *Rev. Comput. Chem.* **1991**, *2*, 313.

(45) Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374.

(46) Thompson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492.

(47) Gao, J. *J. Comput. Chem.* **1997**, *18*, 1062.

(48) Gao, J.; Byun, K. *Theor. Chem. Acc.* **1997**, *96*, 151.

(49) Lin, Y.-l.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1484.

(50) Poulsen, T. D.; Ogilby, P. R.; Mikkelsen, K. V. *J. Chem. Phys.* **2002**, *116*, 3730.

(51) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050.

(52) Wesolowski, T.; Muller, R. P.; Warshel, A. *J. Phys. Chem.* **1996**, *100*, 15444.

(53) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(54) Dewar, M. J. S.; Thiel, W. *Theor. Chim. Acta* **1977**, *46*, 89.

(55) Gascon, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. *J. Chem. Theory Comput.* **2006**, *2*, 175.

(56) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.

(57) Dapprich, S.; Komiromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *Theochem* **1999**, *461−462*, 1.

(58) Field, M. J. *Mol. Phys.* **1997**, *91*, 835.

(59) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.

(60) Amara, P.; Field, M. J.; Alhambra, C.; Gao, J. *Theor. Chem. Acc.* **2000**, *104*, 336.

(61) (a) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 632. (b) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Phys. Chem.* **2005**, *6*, 1853.

(62) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 5454.

(63) (a) Ferré, N.; Assfeld, X.; Rivail, J.-L. *J. Comput. Chem.* **2002**, *23*, 610. (b) Thery, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269. (c) Assfeld, X.; Ferré, N.; Rivail, J.-L. *ACS Symp. Ser.;* Gao, J., Thompson, M. A., Eds.; 1998; Vol. 712, p 234.

(64) Antes, I.; Thiel, W. *ACS Symp. Ser.;* Gao, J., Thompson, M. A., Eds.; 1998; Vol. 712, p 50.

(65) Calzaferri, G.; Forss, L.; Kamber, I. *J. Phys. Chem.* **1989**, *93*, 5366.

(66) Carbo, R.; Fornos, J. M.; Hernandez, J. A.; Sanz, F. *Int. J. Quantum Chem.* **1977**, *11*, 271.

(67) Anderson, A. B.; Hoffmann, R. *J. Chem. Phys.* **1974**, *60*, 4271.

(68) Elstner, M.; Porezag, D.; Juugnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Sukai, S.; Seifect, G. *Phys. Rev. B* **1998**, *58*, 7260.

(69) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.

(70) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.

(71) Mulliken, R. S. *J. Chem. Phys.* **1964**, *61*, 20.

(72) Momany, F. A. *J. Phys. Chem.* **1978**, *82*, 592.

(73) Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385.

(74) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.

(75) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291.

(76) Zhu, T.; Li, J.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 5503.

CT700167B

JCTC Journal of Chemical Theory and Computation

Derivation of Distributed Models of Atomic Polarizability for Molecular Simulations

Ignacio Soteras,[†] Carles Curutchet,[†] Axel Bidon-Chanal,[†] François Dehez,[‡]
János G. Ángyán,[*,§] Modesto Orozco,[*,||,⊥] Christophe Chipot,[*,‡] and F. Javier Luque[*,†]

Departament de Fisicoquímica and Institut de Biomedicina, Facultat de Farmàcia, Universitat de Barcelona, Avgda, Diagonal 643, Barcelona 08028, Spain, Équipe de Dynamique des Assemblages Membranaires, Unité Mixte de Recherche CNRS/UHP 7565 and Équipe Modélisation Quantique et Cristallographique, LCM3B UMR 7036, Nancy Université, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France, Departament de Bioquímica i Biología Molecular, Facultat de Química, Universitat de Barcelona, c/. Martí i Franqués 1, 08028, Barcelona, and Unitat de Modelització Molecular i Bioinformàtica, Institut de Recerca Biomèdica, Parc Científic de Barcelona, c/. Josep Samitier 1, 08028 Barcelona, Spain

Received May 14, 2007

Abstract: The main thrust of this investigation is the development of models of distributed atomic polarizabilities for the treatment of induction effects in molecular mechanics simulations. The models are obtained within the framework of the induced dipole theory by fitting the induction energies computed via a fast but accurate MP2/Sadlej-adjusted perturbational approach in a grid of points surrounding the molecule. Particular care is paid in the examination of the atomic quantities obtained from models of implicitly and explicitly interacting polarizabilities. Appropriateness and accuracy of the distributed models are assessed by comparing the molecular polarizabilities recovered from the models and those obtained experimentally and from MP2/Sadlej calculations. The behavior of the models is further explored by computing the polarization energy for aromatic compounds in the context of cation-π interactions and for selected neutral compounds in a TIP3P aqueous environment. The present results suggest that the computational strategy described here constitutes a very effective tool for the development of distributed models of atomic polarizabilities and can be used in the generation of new polarizable force fields.

Introduction

The assumption that induction effects can be treated in an average sense by means of an appropriate parametrization justifies the success of pairwise, additive potential energy functions for cost-effective statistical simulations of organic and biomolecular systems. An important ingredient in the development of such force fields consists in increasing artificially the polarity of the participating molecules to mimic intermolecular induction phenomena.[1] A popular approach for implicit polarization is based upon the observation that, compared to the experimental gas-phase quantities, molecular dipole moments computed at the Hartree−Fock (HF) level with a split-valence 6-31G(d) basis set are systematically exaggerated.[2] Thus, it has thus become customary to use net atomic charges derived by fitting the HF/6-31G(d) electrostatic potential or suitably scaled HF/

[*] Corresponding author e-mail: fjluque@ub.edu (F.J.L.), Christophe.Chipot@edam.uhp-nancy.fr (C.C.), modesto@mmb.pcb.ub.es (M.O.), Janos.Angyan@lcm3b.uhp-nancy.fr (J.G.A.).

[†] Facultat de Farmàcia, Universitat de Barcelona.

[‡] Unité Mixte de Recherche CNRS/UHP 7565, Nancy Université.

[§] LCM3B UMR 7036, Nancy Université.

[||] Facultat de Química, Universitat de Barcelona.

[⊥] Institut de Recerca Biomèdica.

10.1021/ct7001122 CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/29/2007

6-31G(d) interaction energies in the development of nonpolarizable force fields.[3−11]

The effective polarization implicit to two-body force fields is not equivalent to a rigorous, atomic-level description of the molecular response to a nonuniform, external electric field. This has been illustrated by numerous studies that have examined structural[12−24] and energetic[25−28] properties on a variety of chemical and biochemical systems. The growing body of evidence that induction forces can play a pivotal role in the fine description of the structural and energetic features of highly polarizable systems, in conjunction with the enhanced computational capabilities witnessed in recent years, has stimulated the development of strategies to treat polarization explicitly in classical simulations. Much effort has, therefore, been invested to explore the suitability of such induction schemes such as fluctuating charge[29−33] or induced point dipole[34−45] models, Drude oscillators,[46−48] modified sets of atomic charges,[49−51] or schemes that combine the above.[52−54]

The implementation of explicit polarization schemes is associated with the availability of models of distributed atomic polarizabilities, which are neither uniquely defined nor physically measurable. The partitioning scheme put forward by Stone,[55,56] which merges an earlier formulation of the susceptibility function of the charge density[57] with the distributed multipole analysis method,[58] provides distributed polarizabilities from quantum mechanical (QM) calculations of the response of an isolated molecule to an external perturbation. A closely related procedure[59] relies on a topological partitioning of the molecular space into atomic regions, according to the *atoms-in-molecules* (AIM) theory.[60] These models yield polarizabilities that reproduce the induced moments due to a local electrostatic potential and its successive derivatives experienced at another site. Nevertheless, they include a plethora of terms that rapidly become cumbersome to handle, thus limiting their usefulness in force field simulations.

Applequist derived a heuristic approach wherein atomic polarizabilities were derived by minimizing the deviation between calculated and experimental molecular polarizabilities.[34] This strategy, which was originally devised in the framework of the induced dipole model, was subsequently refined through the introduction of a modified polarizability tensor to smear out the dipole interaction[61] or by extending the model to monopole and dipole polarizabilities.[62] Alternative schemes rely upon atomic hybrid, bond, or group polarizabilities[63] or have resorted to the fitting of molecular polarizabilities determined from QM calculations with large basis sets.[64,65] An inherent feature of these approaches is a substantial component of arbitrariness in the parametrization of the model and the assumption of transferability of the atomic polarizabilities.

In the spirit of the electrostatic potential-fitted (ESP) charge approach,[66−68] alternative schemes targeted at the construction of models of distributed polarizabilities based on a least-squares fitting to the induction energy have been devised.[69−72] Their strength resides in the possibility of generating rather easily compact, flexible sets of polarizability parameters at any given order. Their main limitation stems from the enormous cost associated with the computation of induction energies, since in its most straightforward formulation it requires N_p distinct QM calculations to evaluate the induction energy due to the presence of a nonpolarizable point charge placed at any of the N_p points used to discretize the space around the molecule. This shortcoming has been recently circumvented in two computational strategies that rely essentially upon a single QM calculation. The first scheme is based on second-order perturbation theory (PT) and, upon suitably chosen scaling factors, reproduces the variational induction energy from one QM calculation at the HF level.[73−75] The second strategy consists of mapping grids of induction energies from a single high-level QM calculation, followed by a topological partitioning of the electron density (TPED) response into atomic regions.[59,76] For neutral molecules, induction energies obtained from PT and TPED schemes agree closely.[75]

In the present study, the PT strategy is used to derive distributed models of atomic polarizabilities in the framework of the induced dipole theory for a set of neutral molecules which includes prototypical organic compounds. The atomic polarizabilities are derived by considering their implicit and explicit interaction. The quality of the models is assessed in a comparison of the molecular polarizabilities recovered from the models with both experimental and MP2/Sadlej values. In addition, the ability of the atomic polarizabilities to reflect the induction energy determined for selected chemical interactions (i.e., cation-π complexes and neutral polar solutes in aqueous solution) is examined. The results are discussed in the light of the potential implementation of the distributed models in classical force fields.

Methods

Induction Energies. The variational calculation of the induction energy ($U_{\rm ind}$) due to a nonpolarizable point charge (q_k) placed at point $\mathbf{r}_k$ in the space surrounding a molecule is expressed as

$$U_{\rm ind} = E_{\rm total,k} - E^{\circ} - q_k V(\mathbf{r}_k) \qquad (1)$$

where $E_{\rm total,k}$ is the total energy of the molecule in the presence of the point charge at $\mathbf{r}_k$, E° is the energy of the isolated molecule, and $V(\mathbf{r}_k)$ is the electrostatic potential created by the isolated molecule at $\mathbf{r}_k$.

Within the PT approach, the induction energy appears at the second order of perturbation energy and is given by

$$U_{\rm ind} = \left\langle \psi^{(o)} \middle| \frac{q_k}{|\mathbf{r}_k - \mathbf{r}|} \middle| \psi^{(1)} \right\rangle \qquad (2)$$

where $\psi^{(o)}$ and $\psi^{(1)}$ denote the wave function of the isolated molecule and its first-order correction term, respectively.

In the framework of the Hartree−Fock (HF) theory, eq 2 can be approximated as[73]

$$U_{\rm ind} = \sum_{a}^{\rm occ} \sum_{r}^{\rm vir} \frac{1}{\epsilon_a - \epsilon_r} \left[\sum_{\mu} \sum_{\nu} c^{*}_{\mu a} c_{\nu r} \left\langle \varphi_{\mu} \middle| \frac{q_k}{|r_k - r|} \middle| \varphi_{\nu} \right\rangle \right]^2 \qquad (3)$$

where ϵ_a and ϵ_r stand for the energy of occupied and virtual molecular orbitals, respectively, and ϕ_μ and ϕ_ν correspond to atomic orbitals in the occupied and virtual molecular orbitals.

The strength of the PT scheme embodied in by eq 3 lies in its reduced computational cost, as only one single QM calculation at the HF level is required to estimate the density matrix. The induction energies, however, tend to be underestimated in absolute value relative to the variational ones, as expected from the fact that eq 3 is based upon an uncoupled form of the HF equations. Such a deviation can, nevertheless, be corrected by an appropriate scaling of the PT induction energies.[75,78] In particular, we have used here the distance-dependent scaling factor defined by eq 4,[75] which was derived by comparing the exact (MP2/Sadlej) and PT induction energies for a series of small neutral compounds

$$\zeta(r) = \frac{\alpha_{\text{exact}}}{\alpha_{\text{UCHF}}}\left(a_o + \frac{a_1}{r} + \frac{a_2}{r^2}\right) \tag{4}$$

where α_{exact} and α_{UCHF} are the exact (i.e., the derivative of the energy with respect to the electric field) and the uncoupled HF estimates of the molecular polarizability, and a_x ($x = 0, 1, 2$) are adjustable parameters.

Atomic Polarizabilities. For all intents and purposes, the distributed models of atomic polarizabilities have been derived in the framework of the induced dipole theory, though an extension to other models is straightforward. For the sake of simplicity, in all cases off-diagonal components of the polarizability tensor of the constituent atoms have been neglected in all cases. The atomic dipole components have been imposed to be isotropic.

Two different procedures have been considered to account for the coupling between distributed units.[79] In the model of explicitly interacting distributed polarizabilities, the many-body nature of the polarizability response experienced by the molecule due to the presence of the nonpolarizable point charge (q_k; see above) is accounted for by an explicit interaction between the different units, which in the induced dipole model is summarized in

$$U_{\text{ind}} = -\frac{1}{2}\sum_i \mu_i E_i^\circ \tag{5}$$

where μ_i is the induced dipole created at atom i, and E_i° is the local external electric field applied to the molecule.

The induced dipole (see eq 6) depends on the point atomic dipole polarizabilities, α_i°, and the total local electric field, E_i, which consists of the permanent electric field and that generated other induced dipoles (eq 7)

$$\mu_i = \alpha_i^\circ E_i \tag{6}$$

$$E_i = E_i^\circ - \sum_{j\neq i} T_{ij}\mu_j \tag{7}$$

where T_{ij} is the ijth element of the dipole field tensor.

Equations 5−7 are solved iteratively during the fitting to the PT induction energies until self-consistency in the induced dipoles is achieved. The induction energy is then determined as

$$U_{\text{ind}} = -\frac{1}{2}\sum_{\text{i}} \alpha_i^\circ E_i E_i^\circ \tag{8}$$

In the model of implicitly interacting distributed polarizabilities, the coupling of the polarizability response of the different subunits is omitted during the fitting to the PT induction energies, so that the induction energy is expressed as

$$U_{\text{ind}} = -\frac{1}{2}\sum_{\text{i}} \alpha_i^{\text{eff}} E^\circ E_{ii}^\circ \tag{9}$$

Equation 9 simplifies the calculation of the induction energies, while assuming that the fitting of atomic polarizabilities is able to capture the coupling between the induced dipoles located at the different polarizable sites of the molecule. It is, therefore, worth stressing that the atomic polarizabilities appearing in eqs 8 and 9 are different, as the many-body nature of the induction energy is taken into account explicitly (α_i°) during the derivation of the distributed models used in eq 8. In contrast, these effects are assumed to be incorporated implicitly into the effective atomic polarizabilites (α_i^{eff}) used in eq 9.

Computational Details. The PT induction energies of eq 3 were determined for a variety of neutral compounds, which include small molecules (CH_4, CO, H_2O, H_2S, HF, CO_2, NH_3), organic compounds containing prototypical functional groups (C_2H_6, C_2H_4, C_2H_2, C_6H_6, CH_3OH, CH_3NH_2, CH_3F, CH_3CN, CH_3OCH_3, $HCOCH_3$, CH_3COCH_3, HCOOH, CH_3CONH_2, $CH_3CH_2NO_2$, CH_3COOCH_3), heterocyclic rings (pyridine, pyrrole, furan, imidazole, and indol), and a series of aromatic derivatives (fluorobenzene, chlorobenzene, phenol, aniline, benzonitrile, 1,4-difluorobenzene, and 1,3,5-trifluorobenzene).

The computation of the molecular polarizabilities, the value of which largely depends on the level of theory,[80−84] was performed at the MP2 level using the Sadlej basis set.[83] This protocol has proven to offer a good compromise between accuracy and computational investment.[75] The geometry optimizations were performed at the MP2/6-31G-(d,p) level, and the molecular polarizabilities were estimated subsequently at the MP2/Sadlej level using the Gaussian03 suite of programs.[84]

The grids of PT induction energies were determined from the HF/Sadlej wave function using the MOPETE[85] program. The grids were constructed by using the OPEP program[86,87] and fixing the multiplicative factor (ξ) for the atomic van der Waals radii[88] of atoms to 5. The number of points (N_p) was adjusted by varying the grid step (see ref 86 for details). Using this procedure, the dependence of the polarizability models on the density of points was examined by varying N_p from ca. 500 to ca. 5000 points.

The isotropic atomic polarizabilities used in eqs 8 and 9 (models A and B, respectively) were restrained to be positive during the fitting procedure in order to avoid physically unrealistic values due to the excessive simplicity of the model. Moreover, in the model of explicitly interaction polarizabilities (eq 8; model A), the coupling between induced dipoles borne by contiguous atoms (1−2 and 1−3 interactions) was neglected, since their interaction is assumed to be described appropriately by the bonded terms implemented in classical force fields. In addition, Thole's damping function[61] was used whenever necessary to couple the induced dipoles located at non-neighboring (1−4 interactions or greater) sites. Following Thole's approach, the scaling distance used to smear out the dipole interaction was built up from the atomic polarizabilities of the interacting sites. The fitting of the PT induction energies was performed using the FITPOL program.[89] Finally, the effect of eliminating the restraint that forces atomic polarizabilities to be positive was investigated for the model of implicitly interacting polarizabilities (eq 9; model C). In this case, the fitting was performed using the OPEP program.

Table 1. Dependence of Atomic Dipolar Polarizabilities (in au^3) Obtained from Models A−C on the Number of Points Used for Mapping of Induction Energies for Methanol, Methylamine, Acetone, and Acetamide[a]

	A (eq 8, $\alpha > 0$)			B (eq 9, $\alpha > 0$)			C (eq 9, no restraint)		
atom	500	1500	5000	500	1500	5000	500	1500	5000
					CH_3OH				
O	7.7	7.5	7.7	7.6	7.4	7.6	10.9	10.4	10.6
H(O)	1.7	1.8	1.7	1.8	1.9	1.8	2.3	2.3	2.1
C	0.1	0.1	0.1	0.1	0.1	0.1	−13.3	−10.6	−10.1
H(C)	3.7	3.8	3.8	3.7	3.8	3.8	7.2	6.6	6.5
					CH_3NH_2				
N	9.4	11.2	11.2	8.8	11.1	10.9	11.6	14.6	13.7
H(N)	2.1	1.6	1.6	2.3	1.6	1.7	2.7	2.0	2.0
C	0.1	0.1	0.1	0.1	0.1	0.1	−8.8	−10.5	−7.6
H(C)	3.9	3.8	3.8	4.0	3.8	3.8	6.4	6.6	6.0
					CH_3COCH_3				
O	11.1	10.9	10.6	11.3	11.3	11.1	13.5	13.3	13.2
C(O)	4.2	4.4	5.2	2.9	2.7	3.2	5.3	4.7	5.6
C	0.1	0.1	0.1	0.1	0.0	0.0	−7.9	−5.7	−6.5
H(C)	3.9	4.0	3.9	4.1	4.1	4.1	6.4	5.8	6.0
					CH_3CONH_2				
O	13.6	12.2	12.2	12.7	11.5	11.5	13.6	14.1	13.9
C(O)	0.1	0.1	0.1	1.0	0.9	0.1	0.9	−0.9	−1.0
N	10.9	8.3	8.6	11.9	9.2	9.6	11.8	10.9	12.8
H(N)	2.2	2.2	2.1	1.8	1.9	1.9	2.1	2.3	1.8
C	0.2	0.2	0.1	0.2	0.1	0.1	−9.2	−5.2	−5.5
H(C)	4.5	4.1	4.1	4.6	4.1	4.2	6.7	6.0	6.0

[a] Only values obtained for around 500, 1500, and 5000 points are shown.

Results and Discussion

Effect of the Grid. Based on a previous series of experiments on the influence of the grid on QM electrostatic potential derived atomic point charges, the definition of the grid over which the induction energies are mapped can be a critical factor on the models of distributed atomic polarizabilities. Contrary to the grids used to fit point charges, it has been noted that mapping of induction energies must sample regions of space far enough from the nuclei to warrant an appropriate reproduction of molecular polarizabilities.[75] Accordingly, grids have been constructed in such a way that only those points located between envelopes corresponding to 2 and 5 times the atomic van der Waals radii are considered. Here, our attention is focused mainly on the density of points defined between these envelopes.

The dependence on the grid density of the atomic and the molecular polarizabilities obtained for the three models is illustrated in Table 1 and Figure 1 for four representative molecules, viz. methanol, methylamine, acetone, and acetamide. The results clearly demonstrate that both atomic and molecular polarizabilities remain only marginally affected by the density of points, except for those grids involving less than ca. 1000 points. From a practical point of view, it can be concluded that for the set of small and medium sized molecules examined here models of atomic dipolar polarizabilities can be derived from grids containing ca. 1500 points, which corresponds to a grid step of about 1.3 Å. In the following the discussion will be limited to the results obtained using above definition of the grid.

Atomic Dipole Polarizabilities. Table 2 shows the atomic dipole polarizabilities obtained for models A−C for the whole set of neutral molecules considered in this investigation.

The analysis of the results shown in Table 2 reveals qualitative differential trends in the dipole polarizabilities between certain atom types. For instance, the larger polarizability of third-row atoms is reflected in the comparison of the values obtained for S (ca. 22 au^3) and Cl (ca. 21 au^3) relative to O (ca. 8 au^3) and F (ca. 5 au^3). Likewise, polar hydrogen atoms, i.e., bonded to N, O, and F, bear polarizabilities (ca. 1.7 au^3) lower than those found for hydrogen atoms in benzene (ca. 3.3 au^3). The polarizabilities of the nitrogen atom in methylamine and aniline are rather similar (ca. 12 au^3), just like for the nitrogen atom in acetonitrile and benzonitrile (ca. 15 au^3), the fluorine atom in methyl fluorine and fluorobenzene, 1,4-difluorobenzene, and 1,3,5-trifluorobenzene (ca. 6 au^3), and the oxygen atom in methanol and phenol (ca. 8 au^3), which in turn differs from the polarizability of the oxygen atom in those compounds featuring a carbonyl moiety (ca. 12 au^3). The different polarizability borne by the two nitrogen atoms in imidazole is also worth underlining, amounting to about 7 and 13 au^3 for NH and N, respectively, and hence resembling the values obtained for the nitrogen atom in pyrrole (ca. 6.5 au^3) and pyridine (ca. 14 au^3). Moreover, the polarizability of the carbon atom in benzene (ca. 8.5 au^3) lies close to the average value determined for the carbon atoms in the monosubstituted benzene derivatives (excluding the carbon atom bearing the substituent). Yet, the results also show that the atomic polarizabilities depend on the nature of the substituent and the position of the carbon atom relative to that substituent.

The preceding trends support the generally accepted assumption of a certain degree of transferability for the atomic polarizability of atom types in specific chemical groups. Great care must, however, be taken in the physical interpretation of the distributed models due to anomalous atomic polarizabilities generally found for occluded atoms, such as the carbon atom in methyl or carbonyl groups and the carbon atom bearing the substituent in benzene derivatives. In these cases model C generally yields negative atomic polarizabilities, which are shown to be close to zero in models A and B, where positivity restraints are enforced. This behavior stems mainly from the statistical fitting

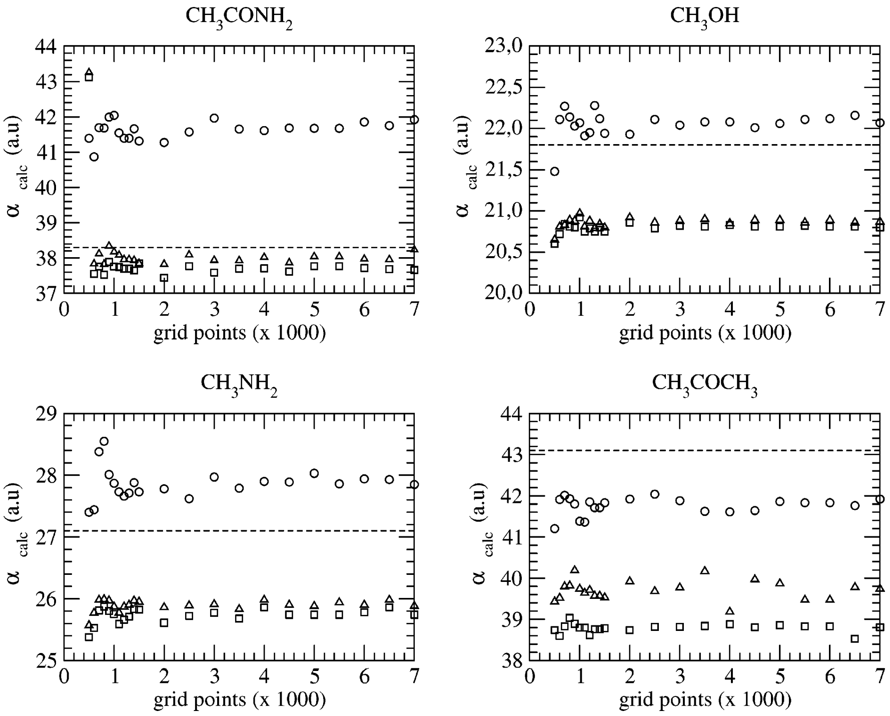


Figure 1. Dependence of the molecular polarizability (in au^3) derived from atomic polarizabilities obtained for models A (eq 8, $\alpha > 0$; triangle), B (eq 9, $\alpha > 0$; square), and C (eq 9, no restraint; circle) on the number of points used for mapping of induction energies for methanol, methylamine, acetone, and acetamide.

performed to minimize the difference between the reference and the calculated induction energies. Thus, even though the distributed models are mathematically consistent, their physical meaning can be affected by the anomalous values assigned to atoms buried in the interior of the molecule. To alleviate this effect, one possibility might consist in enforcing suitable restraints to the polarizabilities of those eclipsed atoms during the fitting procedure. Alternatively, it is reasonable to expect that more elaborate models of distributed polarizabilities, including for instance charge-flow and quadrupole polarizabilities, or where ill-defined components are eliminated,[71,72] should yield more realistic models.

Finally, it is also worth noting that upon exclusion of those compounds with less than four atoms or with negative atomic polarizabilities, there is in general a close similarity between the atomic polarizabilities derived for models A−C (see the Supporting Information), as noted by the scaling coefficient (c) of the regression equations α(model A) = c α(model B) (c = 1.05, r = 0.98, F = 675.8), α(model A) = c α(model C) (c = 0.92, r = 0.98, F = 885.7), and α(model B) = c α(model C) (c = 0.88, r = 0.99, F = 3530.0; in the preceding equations r is the Pearson's correlation coefficient, and F is the Snedeckor's distribution parameter). Within the specific conditions imposed here to the mathematical models used in eqs 8 and 9 (see Methods), the similar results obtained for distributed models of explicitly or implicitly interacting atomic polarizabilities mainly reflects the large screening effect introduced in model A by neglecting the coupling between induced dipoles borne by contiguous atoms (see above).

Molecular Polarizabilities. The reliability of the models of distributed atomic polarizabilities can be checked from their ability to reproduce the molecular polarizability. Table 3 reports the molecular polarizabilities determined from MP2/Sadlej computations and from experimental measurements[90] for the series of compounds and the corresponding values obtained from models A−C.

The root-mean-square deviation (rmsd) between the experimental and the calculated molecular polarizabilities is comparable in the three models, ranging from 2.2 to 3.3 au.^{3} Yet, whereas model C tends to overestimate slightly the molecular polarizability with a mean deviation of 1.7 au,^{3} the reverse trend is found for models A and B, which underestimate the molecular polarizability by 1.6 and 2.5 au,^{3} respectively. Figure 2 shows the regression equations obtained for the comparison of the experimental and the calculated values. In all cases there is a close agreement between the calculated and the experimental polarizabilities, as noted in the scaling coefficients of equation α_M(exptal) = c α_M(model), which amount to 1.04 (r = 1.00, F = 10746), 1.06 (r = 1.00, F = 9635.2), 1.07 (r = 1.00, F = 8329), and 0.96 (r = 1.00, F = 4291) for models A−C, respectively.

Induction Energies. The suitability of the distributed models can be further checked by examining the induction energies for cation-π interactions, where polarization plays a critical contribution to the total stabilization energy of the complex (see for instance ref 74).

The induction energy profiles determined for the approach of a positively charged particle toward benzene were determined from MP2/Sadlej computations and using the three distributed models. The profiles were computed for three possible orientations of the approaching particle (see Figure 3): (i) along the middle of a C−C bond (x-direction), (ii) along the C−H bond (y-direction), and (iii) perpendicular

Table 2. Atomic Dipolar Polarizabilities (in au^3) Obtained from Models A−C for the Series of Neutral Compounds

atom	A (eq 8, α>0)	B (eq 9, α>0)	C (eq 9, no restraint)	atom	A (eq 8, α>0)	B (eq 9, α>0)	C (eq 9, no restraint)	atom	A (eq 8, α>0)	B (eq 9, α>0)	C (eq 9, no restraint)
CH_4											
C	0.0		−8.0	H(C)	4.0		6.3	rmsd[a]	0.02		0.01
NH_3											
N	9.6		10.9	H	1.9		2.3	rmsd	0.07		0.09
H_2O											
O	7.4		8.5	H	1.6		1.8	rmsd	0.06		0.07
HF											
F	4.9		5.1	H	1.2		1.8	rmsd	0.04		0.05
H_2S											
S	20.9		23.6	H	2.4		2.8	rmsd	0.05		0.06
CO											
C	8.1		9.0	O	5.7		6.6	rmsd	0.06		0.07
CO_2											
C	0.0		−11.6	O	9.4		15.2	rmsd	0.20		0.04
C_2H_6											
C	4.5	0.0	−3.2	H	3.2	4.4	5.9	rmsd	0.04	0.02	0.03
C_2H_4											
C	9.0	9.4	10.6	H	2.3	2.1	2.4	rmsd	0.07	0.07	0.08
C_2H_2											
C	6.9	6.8	7.7	H	4.1	4.2	4.7	rmsd	0.04	0.04	0.05
C_6H_6											
C	8.9	7.5	8.6	H	2.3	3.2	3.5	rmsd	0.06	0.07	0.08
CH_3OH											
C	0.1	0.1	−10.6	H(C)	3.8	3.8	6.6	rmsd	0.03	0.04	0.03
O	7.5	7.4	10.4	H(O)	1.8	1.9	2.3				
CH_3NH_2											
C	0.1	0.1	−10.5	H(C)	3.8	3.8	6.6	rmsd	0.05	0.05	0.05
N	11.2	11.1	14.6	H(N)	1.6	1.6	2.0				
CH_3F											
C	0.0		−11.1	H(C)	3.6		6.6	rmsd	0.03		0.01
F	5.8		8.3								
CH_3CN											
C(N)	0.1	0.1	−1.7	N	14.4	14.5	17.5	rmsd	0.04	0.04	0.04
C(H)	0.1	0.1	−1.3	H	4.1	4.1	5.1				
CH_3OCH_3											
C	0.1	0.1	−11.3	H	3.8	3.9	6.8	rmsd	0.06	0.04	0.02
O	9.8	8.6	15.0								
$HCOCH_3$											
C(O)	0.1	0.1	−3.5	$C(CH_3)$	0.2	0.2	0.2	$H(CH_3)$	4.3	4.3	4.8
O	11.0	11.0	14.7	H(C═O)	4.3	4.3	5.7	rmsd	0.06	0.05	0.06
CH_3COCH_3											
C(═O)	4.4	2.7	4.7	$C(CH_3)$	0.1	0.0	−5.7	rmsd	0.04	0.03	0.03
O	10.9	11.3	13.3	H	4.0	4.1	5.8				
HCOOH											
C	0.1	0.2	−8.6	O(H)	8.2	8.5	11.6	H(O)	1.2	0.9	1.2
O(═C)	9.8	9.9	14.0	H(C)	3.4	3.3	6.3	rmsd	0.06	0.06	0.05
CH_3CONH_2											
C(═O)	0.1	0.9	−0.9	N	8.3	9.2	10.9	H(N)	2.2	1.9	2.3
C(H)	0.2	0.1	−5.2	H(C)	4.1	4.1	6.0	rmsd	0.07	0.04	0.05
O	12.2	11.5	14.1								
$CH_3CH_2NO_2$											
$C(CH_3)$	0.1	0.1	−12.2	O	9.2	9.3	16.3	$H(CH_2)$	1.8	1.7	2.1
$C(CH_2)$	11.1	10.7	19.5	$H(CH_3)$	3.2	3.2	6.6	rmsd	0.13	0.12	0.08
N	0.1	0.1	−19.8								

Table 2 (Continued)

atom	A (eq 8, $\alpha>0$)	B (eq 9, $\alpha>0$)	C (eq 9, no restraint)	atom	A (eq 8, $\alpha>0$)	B (eq 9, $\alpha>0$)	C (eq 9, no restraint)	atom	A (eq 8, $\alpha>0$)	B (eq 9, $\alpha>0$)	C (eq 9, no restraint)
CH_3COOCH_3											
C(−COO)	0.1	0.1	−6.6	O(═C)	11.9	11.3	13.3	H(CH_3O)	2.9	3.3	5.8
C(═O)	0.5	0.6	1.0	O(−CH_3)	11.9	9.0	14.0	rmsd	0.04	0.03	0.03
C(−OCO)	0.2	0.2	−10.0	H(CH_3C)	3.7	4.0	5.9				
Pyridine											
C_α	0.1	0.1	−3.3	N	13.9	13.6	17.4	H(C_γ)	3.6	4.0	4.7
C_β	14.8	13.6	19.1	H(C_α)	4.0	4.5	5.4	rmsd	0.06	0.07	0.08
C_γ	0.1	0.1	−2.4	H(C_β)	2.3	2.6	2.2				
Pyrrole											
C_α	5.5	5.9	6.7	H(C_α)	3.5	3.8	4.1	H(N)	1.9	2.5	2.8
C_β	11.3	10.4	11.8	H(C_β)	1.8	2.2	2.4	rmsd	0.06	0.06	0.08
N	7.7	5.7	6.5								
Furan											
C_α	3.6	4.5	4.2	O	8.1	7.1	8.7	H(C_β)	2.0	2.4	2.4
C_β	10.2	9.2	11.0	H(C_α)	3.7	3.8	4.3	rmsd	0.05	0.06	0.07
Imidazole											
C_2	0.1	0.1	0.1	N_3	13.4	13.5	15.2	H(C_5)	2.6	3.1	3.3
C_4	4.7	4.5	5.3	H(C_2)	4.4	4.6	5.1	H(N_1)	2.0	2.4	2.9
C_5	9.5	8.6	10.2	H(C_4)	3.0	3.3	3.6	rmsd	0.06	0.06	0.07
N_1	7.5	6.5	6.9								
Indole											
C_2	0.1	0.1	−2.7	C_8	0.1	0.0	−13.7	H(C_4)	3.4	4.6	3.4
C_3	22.0	22.8	29.1	C_9	0.1	0.1	−4.2	H(C_5)	2.7	3.2	3.0
C_4	9.3	5.8	13.5	N	17.0	16.4	24.6	H(C_6)	1.8	2.3	2.3
C_5	8.7	7.9	10.2	H(N)	0.4	0.5	0.5	H(C_7)	3.7	3.8	3.9
C_6	12.6	14.3	14.7	H(C_2)	3.2	3.4	4.8	rmsd	0.07	0.09	0.09
C_7	8.0	6.0	12.4	H(C_3)	0.1	0.1	−0.6				
Fluorobenzene											
C(F)	1.4	1.1	1.4	C_{para}	8.7	7.5	9.2	H(C_{meta})	2.2	3.2	3.3
C_{ortho}	10.8	9.5	10.3	F	5.7	6.0	6.7	H(C_{para})	2.5	3.2	3.1
C_{meta}	8.6	7.2	8.5	H(C_{ortho})	2.1	3.1	3.6	rmsd	0.06	0.06	0.08
Chlorobenzene											
C(Cl)	0.1	0.1	−4.7	C_{para}	9.0	8.6	9.6	H(C_{meta})	2.2	3.4	3.3
C_{ortho}	12.1	8.8	12.6	Cl	18.5	19.4	22.7	H(C_{para})	2.6	2.9	3.2
C_{meta}	8.3	7.3	8.5	H(C_{ortho})	1.9	3.3	3.2	rmsd	0.05	0.06	0.07
Phenol											
C(OH)	0.1	0.1	−1.3	O	7.5	7.8	9.2	H(C_{meta}/C'_{meta})	2.8/2.3	3.5/2.5	4.0/3.2
C_{ortho}/C'_{ortho}	12.2/12.3	9.9/10.1	12.2/12.9	H(O)	1.5	1.7	1.6	H(C_{para})	3.0	3.6	3.4
C_{meta}/C'_{meta}	6.7/7.3	6.9/9.2	6.1/7.9	H(C_{ortho}/$C'_{\mathbf{ortho}}$)	2.3/2.7	3.3/3.2	3.6/3.6	rmsd	0.06	0.07	0.08
C_{para}	9.7	7.3	10.9								
Aniline											
C(NH_2)	0.1	0.1	−2.4	N	12.0	11.1	13.4	H(C_{meta})	2.8	3.5	3.7
C_{ortho}	13.2	10.1	13.1	H(N)	0.9	1.5	1.5	H(C_{para})	1.7	2.3	2.9
C_{meta}	4.6	5.4	6.1	H(C_{ortho})	2.4	3.5	3.6	rmsd	0.07	0.07	0.09
C_{para}	15.5	13.5	14.2								
Benzonitrile											
C(≡N)	0.1	0.1	−4.5	C_{para}	8.1	8.8	10.6	H(C_{meta})	2.5	3.3	3.7
C(CN)	11.1	8.5	15.5	N	13.8	15.0	17.9	H(C_{para})	2.4	3.2	3.1
C_{ortho}	7.9	6.7	6.3	H(C_{ortho})	2.7	3.4	4.2	rmsd	0.07	0.08	0.09
C_{meta}	8.9	7.5	8.0								
1,4−Difluorobenzene											
C(F)	0.2	0.1	−0.3	F	5.8	6.1	6.7	rmsd	0.05	0.06	0.07
C	11.8	9.8	11.5	H	1.6	2.9	3.1				
1,3,5−Trifluorobenzene											
C(F)	0.1	0.1	-4.5	F	5.7	6.1	7.1	rmsd	0.05	0.06	0.07
C	13.8	12.4	18.8	H	1.9	2.5	2.2				

[a] Root-mean square deviation (in kcal/mol) between PT induction energies and the values recovered by the distributed models.

Table 3. Molecular Polarizabilities (in au^3) Determined from Atomic Dipolar Polarizabilities Obtained from Models A−C and MP2/Sadlej Computations and Measured Experimentally

molecule	A (eq 8, $\alpha>0$)	B (eq 9, $\alpha>0$)	C (eq 9, no restraint)	MP2/ Sadlej	exptl
CH_4	16.1	16.1	17.1	16.5	17.5
CO	13.8	13.8	15.6	13.3	13.2
H_2O	10.6	10.6	12.2	9.8	9.8
H_2S	25.7	25.7	29.1	24.6	25.5
HF	6.1	6.1	6.9	5.7	5.4
CO_2	18.8	18.8	18.8	18.6	19.6
NH_3	15.5	15.5	17.7	14.4	15.3
C_2H_6	28.2	26.4	28.7	28.3	30.2
C_2H_4	27.3	27.4	30.9	27.3	28.7
C_2H_2	21.9	21.9	24.8	22.8	22.5
C_6H_6	67.3	64.6	72.6	69.4	69.6
CH_3OH	20.8	20.8	21.9	21.1	21.8
CH_3NH_2	26.0	25.8	27.7	25.7	27.1
CH_3F	16.5	16.5	17.0	16.8	20.0
CH_3CN	27.2	26.9	29.9	30.2	30.2
CH_3OCH_3	32.8	32.3	33.0	33.4	34.8
$HCOCH_3$	28.6	28.3	31.4	30.2	31.0
CH_3COCH_3	39.5	38.8	41.9	41.9	43.1
HCOOH	22.7	22.7	24.5	22.9	22.9
CH_3CONH_2	37.9	37.9	41.3	39.8	38.3
$CH_3CH_2NO_2$	43.1	42.4	44.0	44.5	47.2
CH_3COOCH_3	44.8	43.3	46.7	46.1	46.0
pyridine	60.2	59.2	66.5	64.1	61.9
pyrrole	53.8	52.7	59.4	55.0	53.6
furan	47.0	46.6	52.5	48.6	48.8
imidazole	47.3	46.6	52.5	49.3	48.5
indole	94.5	91.3	101.2	102.8	NA
fluorobenzene	66.4	63.8	71.8	69.4	69.5
chlorobenzene	79.7	76.6	85.9	84.0	83.0
phenol	71.0	68.9	77.2	75.1	74.9
aniline	77.8	75.0	84.1	81.4	81.6
benzonitrile	80.4	77.4	86.8	86.4	84.3
1,4−difluorobenzene	65.9	63.2	71.0	69.3	66.1
1,3,5−trifluorobenzene	64.9	63.2	70.7	69.8	65.7
msd[a]	−1.6	−2.5	1.7	−0.1	
rmsd[b]	2.2	3.3	2.8	1.5	

[a] Mean signed deviation (in au^3) of calculated values relative to the experimental ones. [b] Root-mean-square deviation (in au^3) relative to the experimental values.

to the center of the aromatic ring (z-direction). As noted in Figure 4, in all cases there is a rather promising agreement between the induction energies determined from the three distributed models and from MP2/Sadlej computations. This is particularly true in the range of distances corresponding to noncovalent interactions, especially for the approach of the nonpolarizable point charge along the z-axis, which corresponds to the cation-π interaction.

Figure 5 shows the variation in the magnitude of the total dipole moment induced in the benzene ring by the approach of the nonpolarizable positive point charge along the three different directions depicted in Figure 3. Keeping in mind the simplicity of the distributed models investigated here, which rely on isotropic atomic dipole polarizabilities, it is not surprising to find deviations between the induced dipole determined from MP2/Sadlej computations (carried out for the benzene in the presence and absence of the positive point charge) and from models A−C at those distances where the point charge penetrates the van der Waals region, where higher order polarization effects should be considered. At greater separations there is, however, qualitative agreement between the distance-dependent profiles obtained for the induced dipole moment determined from models A−C, which reproduce the trends witnessed at the quantum mechanical level.

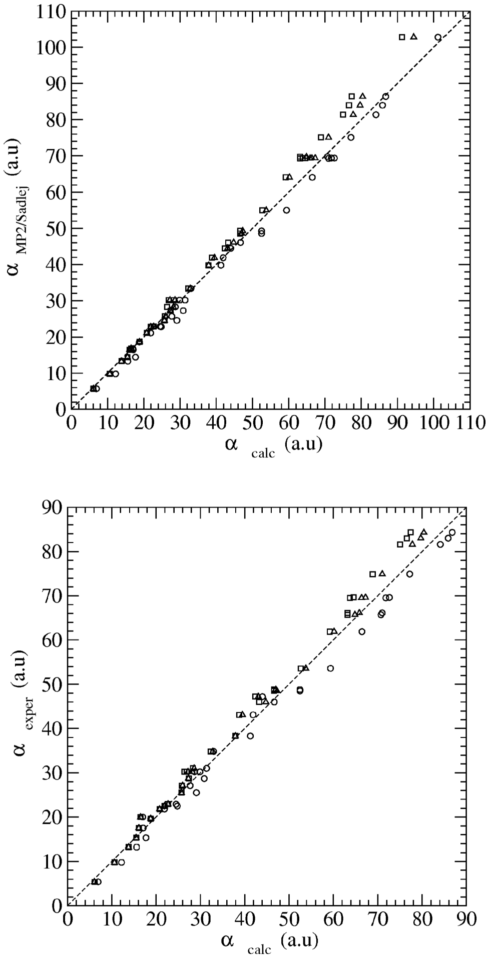


Figure 2. Comparison of the molecular polarizability (in au^3) determined at the MP2/Sadlej level (*top*) and experimentally (*bottom*) in front of the values obtained from distributed models A (triangle), B (square), and C (circle).

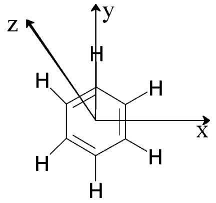


Figure 3. Orientations considered for the approach of a nonpolarizable point charge to benzene. The x-, y-, and z-axis corresponds to the approach along the axis passing (i) through the midpoint of the C−C bond, (ii) the C−H bond, and (iii) the center of the ring along the normal to the molecular plane.

Table 4 shows the polarization energies obtained variationally for a series of cation-π complexes constructed by placing a positive unit point charge at 2.5 Å along the normal axis passing through the center of the ring. The compounds include benzene and all its substituted derivatives, pyridine,

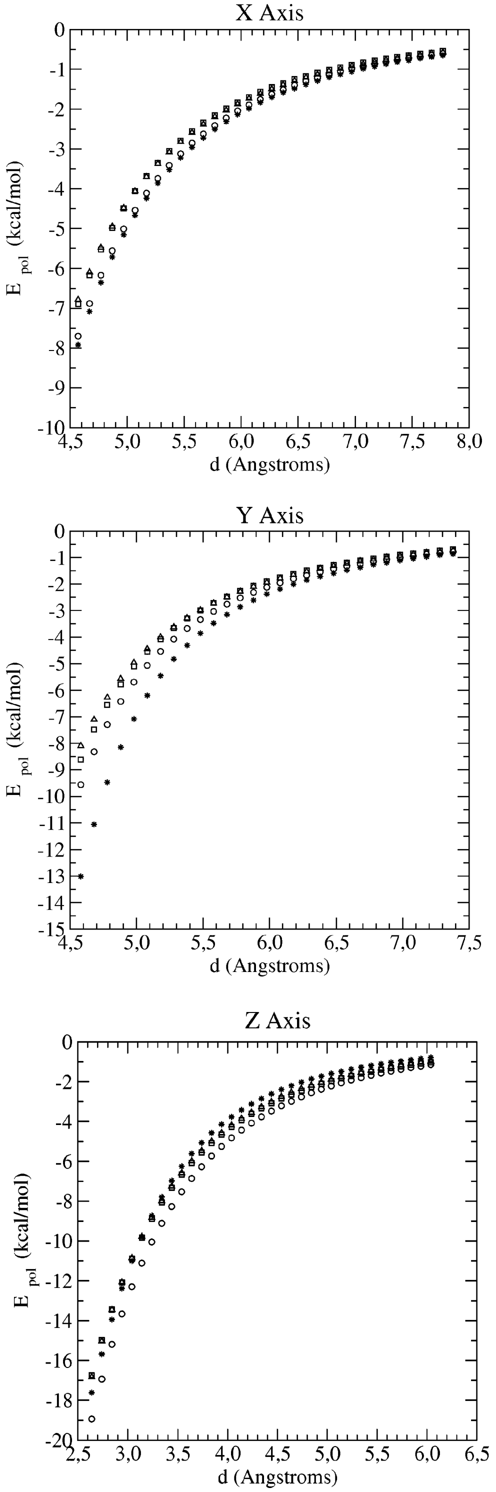


Figure 4. Induction energy profiles (in kcal/mol) for the approach of a positively charged particle to benzene determined from MP2/Sadlej calculations (black dots) and from distributed models A (triangle), B (square), and C (circle). Distances (in Å) are taken from the center of the benzene ring.

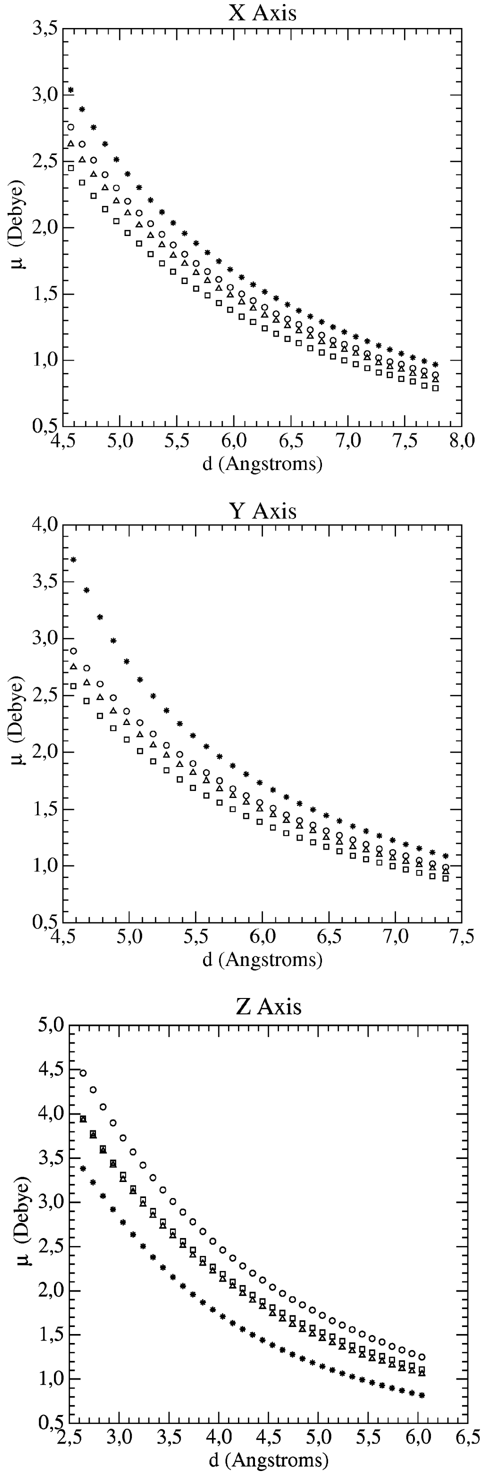


Figure 5. Induced dipole moment (in Debye) in the benzene molecule by a positively charged particle placed at different distances from the center of the ring determined from MP2/Sadlej calculations (black dots) and from distributed models A (triangle), B (square), and C (circle). Distances (in Å) are taken from the center of the benzene ring.

pyrrole, furan, imidazole, and indole (in the latter case, the positive charge was placed above the six-membered ring). For aniline and indole deviations between the MP2/Sadlej and classical induction energies of 4−5 kcal/mol are found, which suggests that at short intermolecular distances the simple models of distributed isotropic polarizabilities considered here might not be adequate to account properly for the induction effects felt by the polarizable sites in certain complexes. Nevertheless, in spite of the short distance separating the nonpolarizable point charge from the center of the ring, the polarization energies generally reproduce satisfactorily the MP2/Sadlej variational values, as the

Table 4. Induction Energies (in kcal/mol) Determined from Atomic Dipolar Polarizabilities Obtained from Models A−C and MP2/Sadlej Computations for Selected Cation-π Complexes

compound	A (eq 8, α>0)	B (eq 9, α>0)	C (eq 9, no restraint)	MP2/ Sadlej
benzene	−20.5	−20.3	−23.0	−21.4
indole	−19.4	−20.4	−21.2	−25.1
fluorobenzene	−19.4	−19.3	−21.9	−21.2
phenol	−19.5	−20.0	−22.3	−22.4
aniline	−19.8	−19.6	−21.9	−24.3
chlorobenzene	−20.5	−20.0	−22.6	−22.3
benzonitrile	−20.9	−20.8	−24.1	−22.2
1,4−difluorobenzene	−18.4	−18.5	−20.9	−21.0
1,3,5−trifluorobenzene	−18.3	−17.7	−19.9	−21.0
pyrrole	−19.9	−19.6	−22.2	−22.4
furan	−17.7	−17.8	−20.1	−20.5
imidazole	−15.9	−15.7	−17.7	−18.9
pyridine	−18.8	−18.7	−21.2	−19.8
msd[a]	2.6	2.6	0.3	
rmsd[b]	2.9	2.9	1.6	

[a] Mean signed deviation (in kcal/mol) of calculated values relative to the MP2/Sadlej ones. [b] Root-mean-square deviation (in kcal/mol) relative to the MP2/Sadlej induction energies.

deviations amount on average to 2.9 kcal/mol for models A and B and to 1.6 kcal/mol for model C.

As a final test to evaluate the goodness of the distributed models, we determined the induction energy for acetamide and pyridine in an aqueous environment. To this end, we selected five distinct configurations from Monte Carlo classical discrete simulations of those compounds solvated in an aqueous solution (TIP3P[91] water molecules). For each configuration, the induction energy created by the TIP3P water molecules on the solute was determined from MP2/ Sadlej computations (see eq 1) and at the classical level using the atomic polarizabilities obtained from models A−C (eqs 8 and 9). In the two cases the water molecules were treated as an assembly of point particles located at the position of O and H atoms bearing the standard partial charges defined in the TIP3P model. The average induction energies determined by including the water molecules placed at 2.5, 4.5, 6.5, and 8.5 Å (around 6, 33, 77, and 150 waters, respectively) from the solute are shown in Figure 6. The results indicate that the induction energies estimated from models A−C reproduce satisfactorily the QM values, as noted in deviations between QM and classical polarization energies around 0.2 kcal/mol. In all cases the dependence of the induction energy on the distance of the water molecules from the solute is well captured by models A−C.

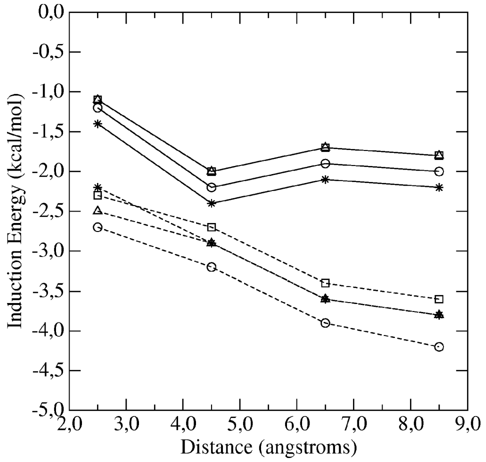


Figure 6. Induction energy (in kcal/mol) determined for acetamide (dashed) and pyridine (solid) in aqueous solution. The plot shows the average value of the induction energy determined from MP2/Sadlej calculations (star) and from distributed models A (triangle), B (square), and C (circle) for five distinct snapshots. Computations were performed by considering the solute at the QM level or classical levels and the water molecules (treated by using the TIP3P model) having any atom at a distance of 2.5, 4.5, 6.5, and 8.5 Å from any atom of the solute.

Conclusion

We have presented within the framework of the induced dipole model a computational strategy relying on the numerical fitting of atomic polarizabilities to induction energies determined from a perturbational scheme. Models of explicitly and implicitly interacting distributed polarizabilities have been considered. For a series of small, neutral organic compounds, our results indicate that the models reproduce rather nicely the molecular polarizability, as reflected in the RMSDs of about 3 au^3 for a series of compounds with a range of molecular polarizabilities close to 100 au^3. In addition, they predict in a satisfactory fashion the polarization energy determined variationally for a series of representative cation-π complexes, where induction effects have proven to contribute significantly to the stabilization energy of the complexes. They are also capable of reproducing the induction energy determined for acetamide and pyridine in aqueous environments.

For all intents and purposes, the present results suggest that the computational strategy outlined here can be a useful, effective tool to derive distributed models of atomic polarizabilities. Clearly, additional detailed studies are required to check the suitability of the models of both explicitly and implicitly interacting atomic polarizabilities in the framework of classical, discrete molecular simulations. At this point, it is worth noting that the distributed polarizability models considered here are rather simple, and they can be ameliorated in several ways, by including for instance charge-flow and quadrupole polarizabilities. In addition, the reliability of the distributed models must be supported by the accuracy in reproducing the induction energy determined at a high level of QM theory as well as in providing a correct description of anisotropy and nonadditivity features of induction forces for a variety of molecular complexes. Finally, the implementation of the distributed polarizabilities into a given force field must be accompanied by an extensive calibration of the different energy contributions and by appropriate corrections in order to maintain the subtle balance between the different energy contributions.[92] Even though the case examples tackled here were limited to models of isotropic atomic polarizabilities within the induced dipole theory, extension of the computational strategy presented here to more elaborate models is expected to be rather straightforward.

Acknowledgment. This work was partially supported by grants from the Spanish Ministerio de Educación y Ciencia (grant CTQ2005-08797-C02-01/BQU). The Centre de Supercomputació de Catalunya (CESCA) is kindly acknowledged for computational facilities.

Supporting Information Available: Numbering of atoms in pyridine, pyrrole, furan, imidazole, and indole in Table 2 (Figure S1) and plots for the comparison of the atomic polarizabilities derived from models A−C (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Computer Simulation of Chemical and Biomolecular Systems. Beveridge, D. L., Jorgensen, W. L., Eds.; *Ann. N.Y. Acad. Sci.* **1986**, *482*, 1.

(2) Hehre, W. J.; Radom, L.; Schleyer, P. V. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley-Interscience: New York, 1986.

(3) MacKerell, A. D., Jr.; Karplus, M. Importance of Attractive van der Waals Contribution in Empirical Energy Function Models for the Heat of Vaporization of Polar Liquids. *J. Phys. Chem*. **1991**, *95*, 10559.

(4) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. OPLS Potential Functions for Nucleotide Bases. Relative Association Constants of Hydrogen-Bonded Base Pairs in Chloroform. *J. Am. Chem. Soc*. **1991**, *113*, 2810.

(5) Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. Accuracy of Free Energies of Hydration for Organic Molecules from 6-31G*-Derived Partial Charges. *J. Comput. Chem*. **1993**, *14*, 1240.

(6) Orozco, M.; Jorgensen, W. L.; Luque, F. J. Comparison of 6-31G*-Based MST/SCRF and FEP Evaluations of the Free Energies of Hydration for Small Neutral Molecules. *J. Comput. Chem.* **1993**, *14*, 1498.

(7) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem*. **1993**, *97*, 10269.

(8) MacKerell, A. D., Jr., Wiórkiewicz-Kuczera, J.; Karplus, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *J. Am. Chem. Soc*. **1995**, *117*, 11946.

(9) Fox, T.; Kollman, P. A. Application of the RESP Methodology in the Parametrization of Organic Solvents. *J. Phys. Chem. B* **1998**, *102*, 8070.

(10) McDonald, N. A.; Jorgensen, W. L. Development of an All-Atom Force Field for Heterocycles. Properties of Liquid Pyrrole, Furan, Diazoles, and Oxazoles. *J. Phys. Chem. B* **1998**, *102*, 8049.

(11) Foloppe, N.; MacKerell, A. D., Jr. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Macromolecular Target Data. *J. Comput. Chem*. **2000**, *21*, 86.

(12) Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. Gas-Phase and Liquid-State Properties of Esters, Nitriles, and Nitro Compounds with the OPLS-AA Force Field. *J. Comput. Chem*. **2001**, *22*, 1340.

(13) Chipot, C.; Ángyán, J. G.; Maigret, B.; Scheraga, H. A. Modeling Amino Acid Side XChains. 3. Influence of Intra- and Intermolecular Environment on Point Charges. *J. Phys. Chem.* **1993**, *97*, 9797.

(14) New, M. H.; Berne, B. J. Molecular Dynamics Calculation of the Effect of Solvent Polarizability on the Hydrophobic Interaction. *J. Am. Chem. Soc*. **1995**, *117*, 7172.

(15) Chipot, C.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. Molecular Dynamics Potential of Mean Force Calculations: A Study of the Toluene-Ammonium π-Cation Interactions. *J. Am. Chem. Soc*. **1996**, *118*, 2998.

(16) Cieplak, P.; Caldwell, J.; Kollman, P. A. Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases. *J. Comput. Chem*. **2001**, *22*, 1048.

(17) Allen, T. W.; Bastug, T.; Kuyucak, S.; Chung, S.-H. Gramicidin A Channel as a test Ground for Molecular Dynamics Force Fields. *Biophys. J*. **2003**, *84*, 2159.

(18) Grossfield, A.; Ren, P.; Ponder, J. W. Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc*. **2003**, *125*, 15671.

(19) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III CHARMM Fluctuating Charge Force Field for Proteins: II Protein/Solvent Properties from Molecular Dynamics Simulations Using a Nonadditive Electrostatic Model. *J. Comput. Chem*. **2004**, *25*, 1504.

(20) Yan, T.; Burnham, C. J.; Del Pópolo, M. G.; Voth, G. A. Molecular Dynamics Simulation of Ionic Liquids: The Effect of Electronic Polarizability. *J. Phys. Chem. B* **2004**, *108*, 11877.

(21) Allen, T. W.; Andersen, O. S.; Roux, B. Energetics of Ion Conduction through the Gramicidin Channel. *Proc. Nat. Acad. Sci U.S.A.* **2004**, *101*, 117.

(22) Kim, B.; Young, T.; Harder, E.; Friesner, R. A.; Berne, B. J. Structure and Dynamics of the Solvation of Bovine Pancreatic Trypsin Inhibitor in Explicit Water: A Comparative Study of the Effects of Solvent and Protein Polarizability. *J. Phys. Chem. B* **2005**, *109*, 16529.

(23) Ishida, T. Polarizable Solute in Polarizable and Flexible Solvents: Simulation Study of Electron transfer Reaction Systems. *J. Phys. Chem. B* **2005**, *109*, 18558.

(24) Sakharov, D. V.; Lim, C. Zn Protein Simulations Including Charge Transfer and Local Polarization Effects. *J. Am. Chem. Soc*. **2005**, *127*, 4921.

(25) Guo, H.; Gresh, N.; Roques, B. P.; Salahub, D. R. Many-Body Effects in Systems of Peptide Hydrogen-Bonded Networks and Their Contributions to Ligand Binding: A Comparison of the Performances of DFT and Polarizable Molecular Mechanics. *J. Phys. Chem. B* **2000**, *104*, 9746.

(26) Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W. Parallel ab Initio and Molecular Mechanics Investigation of Polyccordinated Zn(II) Complexes with Model Hard and Soft Ligands : Variations of Binding Energy and of its Components with Number and Charges of Ligands. *J. Comput. Chem.* **2000**, *21*, 1011.

(27) Gresh, N.; Piquemal, J.-P.; Krauss, M. Representation of Zn(II) Complexes in Polarizable Molecular Mechanics. Further Refinements of the Electrostatic and Short-Range Contributions. Comparisons with Parallel ab Initio Computations. *J. Comput. Chem.* **2005**, *26*, 1113.

(28) Gresh, N.; Sponer, J. Complexes of Pentahydrated Zn^{2+} with Guanine, Adenine, and the Guanine-Cytosine and Adenine-Thymine Base Pairs. Structures and Energies Characterized by Polarizable Molecular Mechanics and ab Initio Calculations. *J. Phys. Chem. B* **1999**, *103*, 11415.

(29) Rappé, A. K.; Goddard, W. A., III. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem*. **1991**, *95*, 3358.

(30) Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *J. Chem. Phys*. **1994**, *101*, 6141.

(31) Field, M. J. Hybrid Quantum Mechanical/Molecular Mechanical Fluctuating Charge Models for Condensed Phase Simulations. *Mol. Phys*. **1997**, *91*, 835.

(32) Banks, J. L.; Kaminsky, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys*. **1999**, *110*, 741.

(33) Bret, C.; Field, M. J.; Hemmingsen, L. A Chemical Potential Equalization Model for Treating Polarization in Molecular Mechanical Force Fields. *Mol. Phys*. **2000**, *98*, 751.

(34) Applequist, J.; Carl, J. R.; Fung, K.-K. An Atom Dipole Interaction Model for Molecular Polarizability. Application to Polyatomic Molecules and Determination of Atom Polarizabilities. *J. Am. Chem. Soc*. **1972**, *94*, 2952.

(35) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol*. **1976**, *103*, 227.

(36) Lybrand, T. P.; Kollman, P. A. Water-Water and Water-Ion Potential Functions Including Terms for Many Body Effects. *J. Chem. Phys*. **1985**, *83*, 2923.

(37) Caldwell, J.; Dang, L. X.; Kollman, P. A. Implementation of Nonadditive Intermolecular Potentials by Use of Molecular Dynamics: Development of a Water-Water Potential and Water-Ion Cluster Interactions. *J. Am. Chem. Soc*. **1990**, *112*, 9144.

(38) Voisin, C.; Cartier, A. Determination of Distributed Polarizabilities to be Used for Peptide Modeling. *J. Mol. Struct. (Theochem)* **1993**, *286*, 35.

(39) Meng, E. C.; Kollman, P. A. Molecular Dynamics Studies of the Properties of Water around Simple Organic Solutes. *J. Phys. Chem*. **1996**, *100*, 11460.

(40) Meng, E.; Caldwell, J. W.; Kollman, P. A. Investigating the Anomalous Solvation Free Energies of Amines with a Polarizable Potential. *J. Phys. Chem*. **1996**, *100*, 2367.

(41) Kamisnki, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. Development of a Polarizable Force Field for Proteins via ab Initio Quantum Chemistry: First Generation Model and Gas Phase Tests. *J. Comput. Chem*. **2002**, *23*, 1515.

(42) Ren, P.; Ponder, J. W. Consistent Treatment of Inter- and Intramolecular Polarization in Molecular Mechanics Calculations. *J. Comput. Chem*. **2002**, *23*, 1497.

(43) Borodin, O.; Smith, G. D. Development of Quantum Chemistry-Based Force Fields for Poly(ethylene oxide) with Many-Body Polarization Interactions. *J. Phys. Chem. B* **2003**, *107*, 6801.

(44) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from ab Initio Quantum Chemistry. *J. Phys. Chem. A* **2004**, *108*, 621.

(45) Borodin, O.; Smith, G. D. Development of Many-Body Polarizable Force Fields for Li-Battery Components: 1. Ether, Alkane, and Carbonate Solvents. *J. Phys. Chem. B* **2006**, *110*, 6279.

(46) Cao, J.; Berne, B. J. Theory and Simulation of Polar and Nonpolar Polarizable Fluids. *J. Chem. Phys*. **1993**, *99*, 6998.

(47) Lamoureux, G.; Roux, B. Modeling Induced Polarization with Classical Drude Oscillators: Theory and Molecular Dynamics Simulation Algorithm. *J. Chem. Phys*. **2003**, *119*, 3025.

(48) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys*. **2003**, *119*, 5185.

(49) Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. Towards Improved Force Fields: III. Polarization thorugh Modified Atomic Charges. *J. Comput. Chem*. **1999**, *20*, 704.

(50) Ferenczy, G. G.; Reynolds, C. A. Molecular Polarization through Induced Atomic Charges. *J. Phys. Chem. A* **2001**, *105*, 11470.

(51) Curutchet, C.; Bofill, J. M.; Hernández, B.; Orozco, M.; Luque, F. J. Energy decomposition in Molecular Complexes: Implications for the Treatment of Polarization in Molecular Simulations. *J. Comput. Chem*. **2003**, *24*, 1263.

(52) Stern, H. A.; Kaminsky, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. Fluctuating Charge, Polarizable Dipole, and Combined Models: Parametrization from ab Initio Quantum Chemistry. *J. Phys. Chem. B* **1999**, *103*, 4730.

(53) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. Combined Fluctuating Charge and Polarizable Dipole Models: Application to a Five-Site Water Potential Function. *J. Chem. Phys*. **2001**, *115*, 2237.

(54) Masia, M.; Probst, M.; Rey, R. On the Peformance of Molecular Polarization Methods. I. Water and Carbon Tetrachloride Close to a Point Charge. *J. Chem. Phys*. **2004**, *121*, 7362.

(55) Stone, A. J. Distributed Polarizabilities. *Mol. Phys.* **1985**, *56*, 1065.

(56) Le Sueur, C. R.; Stone, A. J. Localization Methods for Distributed Polarizabilities. *Mol. Phys.* **1994**, *83*, 293.

(57) Maaskant, W. J. A.; Oosterhof, L. J. Theory of Optical Rotatory Power. *Mol. Phys.* **1964**, *8*, 319.

(58) Stone, A. J. Distributed Multipole Analysis, or How to describe a Molecular Charge Distribution. *Chem. Phys. Lett.* **1981**, *83*, 233.

(59) Ángyán, J. G.; Jansen, G.; Loos, M.; Hättig, C.; Hess. B. A. Distributed Polarizabilities Using the Topological Theory of Atoms in Molecules. *Chem. Phys. Lett.* **1994**, *219*, 267.

(60) Bader, R. F. W. *Atoms in Molecules* − *A Quantum Theory*; Oxford University Press: London, 1990.

(61) Thole, B. T. Molecular Polarizabilities Calculated with a Modified Dipole Interaction. *Chem. Phys.* **1981**, *59*, 341.

(62) Applequist, J. Atom Charge Transfer in Molecular Polarizabilities. Application of the Olson-Sundberg Model to Aliphatic and Aromatic Hydrocarbons. *J. Phys. Chem*. **1993**, *97*, 6016.

(63) Miller, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc*. **1990**, *112*, 8533.

(64) Stout, J. M.; Dykstra, C. E. Static Dipole Polarizabilities of Organic Molecules. Ab Initio Calculations and a Predictive Model. *J. Am. Chem. Soc*. **1995**, *117*, 5127.

(65) Zhou, T.; Dykstra, C. E. Additivity and Transferability of Atomic Contributions to Molecular second Dipole Hyperpolarizabilities. *J. Phys. Chem. A* **2000**, *104*, 2204.

(66) Bonaccorsi, R.; Petrongolo, C.; Scrocco, E.; Tomasi, J. Theoretical Investigations on the Solvation Process. *Theor. Chim. Acta* **1971**, *20*, 331.

(67) Momany, F. A. Determination of Partial Atomic Charges from ab Initio Molecular Electrostatic potentials. Application to Formamide, Methanol, and Formic Acid. *J. Phys. Chem.* **1978**, *82*, 592.

(68) Cox, S. R.; Williams, D. E. Representation of the Molecular Electrostatic potential by a Net Atomic Charge Model. *J. Comput. Chem.* **1981**, *2*, 304.

(69) Nakagawa, S.; Kosugi, N. Polarized One-Electron Potentials Fitted by Multicenter Polarizabilities and Hyperpolarizabilities. Ab Initio SCF-CI Calculation of Water. *Chem. Phys. Lett.* **1993**, *210*, 180.

(70) Alkorta, I.; Bachs, M.; Perez, J. J. The Induced Polarization of the Water Molecule. *Chem. Phys. Lett.* **1994**, *224*, 160.

(71) Celebi, N.; Ángyán, J. G.; Dehez, F.; Millot, C.; Chipot, C. Distributed Polarizabilities Derived from Induction Energies: A Finite Perturbation Approach. *J. Chem. Phys.* **2000**, *112*, 2709.

(72) Dehez, F.; Soetens, J. C.; Chipot, C.; Ángyán, J. G.; Millot, C. Determination of Distributed Poalrizabilities from a Statistical Analysis of Induction Energies. *J. Phys. Chem. A* **2000**, *104*, 1293.

(73) Luque, F. J.; Orozco, M. Polarization Effects in Generalized Molecular Interaction Potential: New Hamiltonian for Reactivity Studies and Mixed QM/MM Calculations. *J. Comput. Chem.* **1998**, *19*, 866.

(74) Cubero, E.; Luque, F. J.; Orozco, M. Is Polarization Important in Cation-Pi Interactions? *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5976.

(75) Chipot, C.; Dehez, F.; Ángyán, J.; Millot, C.; Orozco, M.; Luque, F. J. Alternative Approaches for the calculations of Induction Energies: Characterization, Effectiveness, and Pitfalls. *J. Phys. Chem. A* **2001**, *105*, 11505.

(76) Dehez, F.; Chipot, C.; Millot, C.; Ángyán, J. G. *Chem. Phys. Lett.* **2001**, *338*, 180.

(77) Francl, M. M. Polarization Corrections to Electrostatic Potentials. *J. Phys. Chem*. **1985**, *89*, 428.

(78) Chipot, C.; Luque, F. J. Fast Evaluation of Induction Energies: A Second-Order Perturbation Theory Approach. *Chem. Phys. Lett.* **2000**, *332*, 190.

(79) Chipot, C.; Ángyán, J. G. Continuing Challenges in the Parametrization of Intermolecular Force Fields. Towards and Accurate Description of Electrostatic and Induction Terms. *New. J. Chem.* **2005**, *29*, 411.

(80) Voisin, C.; Cartier, C.; Rivail, J. L. Computation of Accurate Electronic Molecular Polarizabilities. *J. Phys. Chem*. **1992**, *96*, 7966.

(81) Liu, S. Y.; Dykstra, C. E. Multipole Polarizabilities and Hyperpolarizabilities of AHn and A2Hn Molecules from Derivative Hartree-Fock Theory. *J. Phys. Chem*. **1987**, *91*, 1749.

(82) Spackman, M. A. A Simple Quantitative Model of Hydrogen Bonding. *J. Chem. Phys*. **1986**, *85*, 6587.

(83) Sadlej, A. J. Medium-Size Polarized Basis Sets for High-Level Correlated Calculations of Molecular Electric Properties. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995.

(84) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; D. K. Malick, Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03, Revision B.04*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(85) Curutchet, C.; Alhambra, C.; Orozco, M.; Luque, F. J. *MOPETE*; University of Barcelona: Barcelona, 2003.

(86) Ángyán, J. G.; Chipot, C.; Dehez, F.; Hättig, C.; Jansen, G.; Millot, C. OPEP: A Tool for the Optimal Partitioning of Electric Properties. *J. Comput. Chem.* **2003**, *24*, 997.

(87) Ángyán, J. G.; Chipot, C.; Dehez, F.; Hättig, C.; Jansen, G.; Millot, C. *OPEP*; Université Henri Poincaré: Nancy, 2002.

(88) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem*. **1964**, *68*, 441.

(89) Soteras, I.; Orozco, M.; Luque, F. J. *FITPOL*; University of Barcelona: Barcelona, 2006.

(90) Atomic and Molecular Polarizabilities. In *CRC Handbook of Chemistry and Physics*, Internet Version 2007 (87th ed.); Lide, D. R., Ed.; Taylor and Francis: Boca Raton, FL.

(91) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.

(92) Dehez, F.; Angyán, J. G.; Soteras Gutiérrez, I.; Luque, F. J.; Shulten, K.; Chipot, C. Modeling Induction Phenomena in Intermolecular Interactions with an ab Initio Force Field. *J. Chem. Theor. Comput*. **2007**, *3*, 1914−1926.

CT7001122

JCTC Journal of Chemical Theory and Computation

Modeling Induction Phenomena in Intermolecular Interactions with an Ab Initio Force Field

François Dehez,[†] János G. Ángyán,*[,†] Ignacio Soteras Gutiérrez,[‡] F. Javier Luque,*[,‡] Klaus Schulten,[§] and Christophe Chipot*[,†]

Equipe de dynamique des assemblages membranaires, UMR 7565 and Equipe de modélisation quantique et cristallographique, LCM3B, UMR 7036, Nancy Université, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France, Departament de Fisicoquímica and Institut de Biomedicina, Facultat de Farmàcia, Universitat de Barcelona, Avgda, Diagonal 643, Barcelona 08028, Spain, and Theoretical and Computational Biophysics Group, Beckman Institute, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801

Received June 26, 2007

Abstract: One possible road toward the development of a polarizable potential energy function relies on the use of distributed polarizabilities derived from the induction energy mapped around the molecule. Whereas such polarizable models are expected to reproduce the signature induction energy with an appreciable accuracy, it is far from clear whether they will perform equally well in the context of intermolecular interactions. To address this issue, while pursuing the ultimate goal of a “plug-and-play’’-like approach, polarizability models determined quantum mechanically and consisting of atomic isotropic dipole plus charge-flow polarizabilities were combined with the classical, nonpolarizable Charmm force field. Performance of the models was probed in the challenging test cases of cation-π binding and the association of a divalent calcium ion with water, where induction effects are envisioned to be considerable. Since brute force comparison of the binding energies estimated from the polarizable and the classical Charmm potential energy functions is not justified, the individual electrostatic and induction contributions of the force field were confronted to the corresponding terms of a symmetry-adapted perturbation theory (SAPT) expansion carried out with the 6-311++G(*d*,*p*) basis set. While the quantum-mechanical and the molecular-mechanical electrostatic and damped induction contributions agree reasonably well, overall reproduction of the binding energies is plagued by an underestimated repulsion that underlines the necessity of de novo parametrization of the classical 6-12 form of the van der Waals potential. Based on the SAPT expansion, new Lennard-Jones parameters were optimized, which, combined with the remainder of the polarizable force field, yield an improved reproduction of the target binding energies.

Introduction

One of the keys to the success of pairwise additive macromolecular force fields resides in the assumption that in numerical simulations, polarization phenomena can be accounted for in an average sense by means of an appropriately parametrized electrostatic term. Such effective potential energy functions compensate for missing through-space induction effects by inflating artificially the polarity of the constituent molecules.[1] A popular implicit polarization scheme, which has pervaded over the past 20 years, relies upon the observation that at the Hartree−Fock (HF) level of approximation, the split-valence 6-31G(*d*) basis set exhibits a conspicuous tendency to overestimate systematically gas-phase molecular dipole moments.[2] In a vast number of instances where explicit polarization phenomena can be

* Corresponding author e-mail: Christophe.Chipot@edam.uhp-nancy.fr (C.C.), Janos.Angyan@1cm3b.uhp-nancy.fr (J.G.A.), fjluque@ub.edu (F.J.L.).

[†] Nancy Université.

[‡] Universitat de Barcelona.

[§] University of Illinois at Urbana−Champaign.

10.1021/ct700156a CCC: $37.00
Published on Web 10/06/2007

ignored, additive force fields, in which the electrostatic term consists of point charges either derived from HF/6-31G(d) electrostatic potentials or optimized to reproduce the interaction with surrounding water molecules, have proven to describe reasonably well the underlying physical properties of the molecular assemblies.[3,4] An obvious advantage for turning to an implicit polarization approach is the cost-effectiveness of the numerical simulation, obviating the crucial need for an accurate evaluation of the induced dipole moments, which generally represents an appreciable overhead in the calculation of the potential energy. Given the success of implicit polarization schemes, it is legitimate to call into question the necessity to craft new nonadditive potential energy functions, when additive ones seemingly perform just as well.[5]

Unfortunately, in numerous examples, an exaggerated polarity becomes clearly insufficient to describe adequately the response of the molecular charge distribution to a nonuniform, external electric field—chief among which is the interaction of a very deformable electron cloud with a polarizing charge, instantiated in cation-π complexes.[6] One of the practical reasons that have hitherto hampered the development of polarizable potential energy functions targeted at numerical simulations is evidently the costly calculation of the induced moments, and the realization that a marginal improvement over conventional, pairwise additive force fields was not necessarily worth the additional computational effort. Yet, the formidable decrease of the computer price/performance ratio over the past decades that benefited the theoretical community by opening new vistas for the numerical simulation of large ensembles of atoms over time scales compatible with the experimentally observed phenomena has also paved the way for the development of more elaborate models for the accurate representation of intermolecular interactions. Convincingly enough, nonpolarizable macromolecular force fields, e.g., Amber,[4] Charmm,[7] Gromos,[8] or Opls-AA,[9] have proven to behave reasonably well, insofar as biologically relevant molecular assemblies, in which induction effects can be safely ignored, are concerned. As the harnessed computational power allows increasingly larger objects of the cell machinery to be tackled in a routine fashion, the pairwise additive approximation remains; however, an intrinsic limitation to the investigation of biophysical processes where the influence of polarization can no longer be neglected without proper justification. For instance, ions permeating membrane channels have been shown to polarize the conduit through which they diffuse,[10,11] thereby altering the charge distribution of the residues pertaining to the conduction pathway, and, hence, the interplay of the permeant with its environment.

More than a renaissance, the research area of polarizable force fields has been recently the theater of an increasing activity, where the relative merits and drawbacks of competing approaches are being explored. One of the earliest routes devised for modeling through-space polarization phenomena in numerical simulations consists of parametrizing the induction forces in terms of atomic quantities, which can be subsequently plugged into molecular mechanics calculations. At the conceptual level, this solution supposes that the electron density response be partitioned into regions of the Cartesian space that correspond to atoms and/or functional groups.[12] It also supposes a truncation of the multipole expansion and a selection of leading terms in the classical expression of the forces exerted between atomic distributions. The popular scheme put forth by Applequist[13] for the construction of models of distributed polarizabilities is based on a self-consistent determination of atomic parameters which are coupled through screened dipole−dipole interactions. Revisited in the couth version of Thole,[14] this heuristic approach bears, however, a marked component of arbitrariness, making the physical interpretation of the derived atomic quantities somewhat arguable. It has been, nonetheless, utilized on several occasions in the statistical simulations of condensed phases, where induction effects were anticipated to play a significant role. In retrospect, it is not clear whether such a partitioning scheme, reduced to an isotropic description of the polarization effects, would increase dramatically the accuracy of the modeled intermolecular interactions, compared to a well parametrized additive force field. In cation-π complexes, for instance, whereas the use of a nonadditive potential energy function appears to improve the accord with the quantum chemical interaction energies bereft of a basis set superposition error[15] (BSSE), inclusion of the latter ironically suggests that the pairwise additive approximation would perform better.[16] Moreover, the overhead imposed by the self-consistent computation of the induced dipole moments in molecular mechanics simulations brings us back to questioning the necessity of turning to polarizable potential energy functions. Much effort, however, has been invested in recent years not only on the front of partitioning the electron density response into distributed polarizabilities[17−22] but also on that of their incorporation in numerical simulations at a lesser computational cost.[23]

In spite of these remarkable advances, the theoretical community appears to be still facing the Gordian knot of increasing the level of sophistication of the current potential energy functions and, hence, the burden of the force evaluation, at the expense of a more extensive sampling of the configurational space. Promising alternative routes to the spatial partitioning of the electron density response are being explored in the context of biomolecular simulations. Among these routes, the Drude shell[24] or dispersion oscillator model relies on a concept devised over a century ago for investigating the charge fluctuation forces in a variety of materials. In a nutshell, it consists of the introduction of massless particles attached to polarizable atoms by means of stiff harmonic springs and bearing a partial charge. It can be shown that the atomic polarizability is a function of both the spring constant and the point charge borne by the so-called Drude particle. In response to an external electric field, the latter is displaced with respect to the atomic core, thereby modifying the molecular charge distribution.[25,26] Models of fluctuating charges constitute yet another promising formalism[27−29] in which the point charges are handled as dynamical variables reflected in the corresponding atomic electronegativities. Conceptually, the electron gas surrounding any nucleus, which is shown to have a chemical potential equal to the negative of the atomic electronegativity, spreads across the entire molecule equalizing the chemical potential at every atomic position. This notion of electronegativity equalization was introduced over 50 years ago by Sanderson[30] and provides a convenient framework for modeling the flow of electrons between atoms as a response to variations of the electric field felt by the participating nuclear sites. Whether these approaches aimed at handling induction effects explicitly in numerical simulations describe the spatial anisotropy of the polarizability with an acceptable accuracy remains, however, unclear.

On account of their overwhelming complexity, compared with the somewhat simpler Drude shell or fluctuating charge models, fully polarizable classical force fields have admittedly not yet come of age to be amenable to numerical simulations of biologically relevant molecular systems over long time-scales. It can be argued, however, that such force fields, if appropriately parametrized, represent the best possible route toward a faithful description of the response electron density upon perturbation by an external electric field.[31−34] In spite of the additional cost implied, which has limited their use so far to mere proofs of concept, fully polarizable classical force fields are expected to become rapidly a relevant competitor to more approximate schemes.

The caveat "appropriately parametrized'' bears some significance in the sense that the prevalent criterion adopted to measure the accuracy of the polarizable models is their propensity to reproduce the induction energy mapped around the molecule. Adopting this philosophy, models truncated to an isotropic point dipole representation, in the spirit of Applequist's prescription,[13] may turn out to be inadequate only because they are incomplete. In the present study, the physically sound, rigorous framework of optimally partitioned electric properties[35,36] (OPEP) is employed for understanding the spatial anisotropy of through-space induction phenomena. In particular, it will be shown that anisotropy can be recovered in models combining isotropic dipole polarizabilities with a zeroth-order charge-flow term between vicinal atoms. To illustrate the critical role played by polarization in intermolecular interactions, two classes of charged complexes will be considered, viz. the cation-π motif resulting from the interaction of benzene with ammonium, and the complexes formed by a calcium ion and a chelating agent, namely water. In the following section, the theoretical formalism is introduced, together with the computational details for the determination of the distributed models of polarizabilities. Next, the performance of the ab initio polarizable force field for reproducing the quantum mechanical interaction energies will be examined in the light of symmetry-adapted perturbation theory[37] (SAPT) calculations, which supplies benchmark values of the polarization contribution. Finally, concluding remarks will be drawn, emphasizing the issue of transferability of the models in classical macromolecular force fields.

Methods

Over 25 years ago, Cox and Williams[38] planted the seed of a method now widely utilized to parametrize the electrostatic term of potential energy functions. In essence, this approach relies on a least-squares fit of atomic charges to the quantum-mechanical electrostatic potential evaluated around the molecule, reminiscing the idea that the former constitutes the fingerprint of the latter.[39−41] Following a similar philosophy, a variety of alternative numerical schemes has been put forth to derive models of distributed polarizabilities based on a least-squares fitting procedure to the polarization potential, i.e., the induction energy associated with the presence of a test charge.[17,20,42] Just like atomic multipole moments can be determined at any given order from the sole knowledge of the reference electrostatic potential, so can atomic polarizabilities, provided that the induction energy has been mapped appropriately around the molecule of interest.[43] Yet, whereas the electrostatic potential at any given point in Cartesian space can be obtained readily from the wave function of a single-point quantum-mechanical calculation, induction energy maps are far more cumbersome to determine. Arguably enough, the most straightforward route is a finite-perturbation approach, whereby the molecule interacts with a nonpolarizable charge, q_k, located at point k. The corresponding induction energy can be expressed as

$$\mathscr{U}_{\mathrm{ind},k} = \varepsilon^{\mathrm{QM}}_{\mathrm{tot},k} - \varepsilon^{\mathrm{QM}}_{0} - q_k V^{\mathrm{QM}}(\mathbf{r}_k) \qquad (1)$$

Here, $\varepsilon^{\mathrm{QM}}_{\mathrm{tot},k}$ stands for the energy of the molecule in the presence of the point charge, which requires one individual quantum-mechanical calculation for each point k of the grid over which the induction energy is mapped. $\varepsilon^{\mathrm{QM}}_{0}$ is the energy of the isolated molecule, and $V^{\mathrm{QM}}(\mathbf{r}_k)$ is the electrostatic potential generated at point k by the isolated molecule.

Quite obviously, the prerequisite of multiple, independent quantum-mechanical calculations imposed by the need for a detailed, accurate picture of the induction energy around the molecule constitutes the main weakness of the finite-perturbation method. Its overwhelming computational cost, rooted in the necessity to include intramolecular electron correlation and employ a sufficiently large basis set to guarantee the faithful reproduction of the molecular polarizabilities, constitutes a stringent limitation of the approach. Whereas the induction energy can be mapped with an appropriate resolution and spatial extension for small, prototypical molecules, the finite-perturbation method becomes rapidly impractical for larger chemical compounds. Given this computational limitation, alternative routes have been explored for faster, yet reliable evaluation of induction energies, chief among which is an elegant method relying upon a single quantum-mechanical calculation carried out at the coupled-perturbed HF (CPHF) or any higher level of approximation.[21] A topological analysis of the response charge density is then performed in the spirit of the "atoms in molecules'' theory[44] to derive the components of the distributed polarizabilities, $\alpha^{ss'}_{l\kappa,l'\kappa'} = \alpha_{l\kappa,l'\kappa'}(\mathbf{r}_s,\mathbf{r}_{s'})$, at a given rank $l,l' \le L$—where L is the highest rank of the components forming what will henceforth be referred to as the model of topologically partitioned electric properties (TPEP).[19,45−47] Such a model, which usually consists of a sizable number of terms, can be utilized to regenerate the induction energy resulting from the polarization of the molecule by the nonpolarizable charge q_k:

$$\mathscr{U}_{\text{ind},k} = -\frac{1}{2} q_k^2 \sum_{s,l,m} \sum_{s',l',m'} T^{ks}_{00,1\kappa}\, \alpha^{ss'}_{l\kappa,l'\kappa'}\, T^{s'k}_{l'\kappa',00} \qquad (2)$$

Here, s and s' denote two polarizable sites of the molecule. $T^{sk}_{l\kappa,00}$ is a matrix element of the electrostatic tensor[48,49] corresponding to multipole component $\{l,\kappa\}$, which gives at point s the electrostatic potential, or its successive derivatives, created by point charge q_k.

In this contribution, models of distributed polarizabilities will be derived with the OPEP suite of programs[35] from maps of induction energies generated using both the finite-perturbation approach and TPEP models.[43] In the case of the cation-π interaction of an ammonium ion with benzene, the induction energy was mapped on grids containing, respectively, 1247 and 3192 points, following the first numerical scheme. For the interaction of a calcium cation, assumed to be nonpolarizable, with water, the induction energy was evaluated on a grid consisting of 3905 points. In addition, models of net atomic charges were derived with the Opep code from the electrostatic potential computed quantum mechanically for the different chemical compounds[39,40] (see Table 1). In all cases, calculations were conducted at the MP2 level of theory with the Sadlej basis set,[50] which supplies polarizability parameters at a favorable quality/cost ratio. Preliminary optimization of the molecular geometries was performed at the MP2/6-311++G(2d,2p) level of approximation, using Gaussian98[51]—MP2/Sadlej computations based on MP2/6-311++G(2d,2p) optimized geometries will be referred to as MP2/Sadlej//MP2/6-311++G(2d,2p). As was demonstrated recently,[36] anisotropy of polarization phenomena can be recovered without the explicit introduction of anisotropic components in the distributed polarizability models, which would not only increase the complexity of the latter but also require a more cumbersome treatment of the corresponding induction forces in numerical simulations. Models combining isotropic dipole- and charge-flow polarizabilities have proven to yield a reasonable reproduction of the target induction energies and molecular quantities and will be utilized in the present investigation (see Table 2). These models will be associated with the Charmm[7] macromolecular force field for the computation of the classical potential energy surfaces delineating the interaction of the ammonium ion with the aromatic ring and that of the calcium ion with water. The quantum-mechanical potential energy surfaces were determined at the MP2/6-311++G(d,p) level of theory, varying the intermolecular distance in 0.1-Å increments and evaluating the BSSE[15] at each step.

For several years, one of the Grail quests for force field developers has been the search for polarizability parameters that could be plugged directly into an existing potential energy function obeying pairwise additivity. It should be remembered, however, that the electrostatic term of the latter exaggerates significantly the polarity of the participating molecules to compensate in an average sense for missing induction effects and, thus, ought to be scaled down accordingly.[52] Given that this implicit polarization scheme relies essentially on the erratic shortcomings of the basis set utilized to derive the point charge models, only a heuristic correction can be applied. A more rational approach consists of determining new sets of atomic charges representative of a true gas phase,[53] which is an easy task for a handful of small organic molecules, as is the case in the present investigation, but admittedly constitutes a tedious endeavor for a complete macromolecular force field. There is an additional complication hitherto only marginally discussed: polarizability parameters are anticipated to depend inherently on the characteristics of the environment and, therefore, ought to be adapted correspondingly.[54−56] In the present work, the description of the electrostatic and the induction contributions of the force field is consistent with a low-pressure gaseous phase, albeit it should be modified to reflect the nature of the surroundings.

Table 1. Models of Net Atomic Charges and Regenerated Molecular Multipole Moments of Benzene, Ammonium, and Water at the MP2/Sadlej//MP2/6-311++G(2d,2p) Level of Approximation[a]

	point charges		molecular multipoles		
				regenerated	MP2/Sadlej
benzene	Q^{C}_{00}	−0.124	Q_{20}	2.804	2.868
	Q^{H}_{00}	0.124	Q_{30}	11.131	11.380
	rmsd	0.222			
	$\Delta\varepsilon$	22.858			
ammonium	Q^{N}_{00}	−0.848	Q_{30}	7.446	7.398
	Q^{H}_{00}	0.462	Q_{40}	6.727	8.071
	rmsd	0.073			
	$\Delta\varepsilon$	0.035			
water	Q^{O}_{00}	−0.672	Q_{10}	−0.747	−0.732
	Q^{H}_{00}	0.336	Q_{20}	−0.189	−0.231
	rmsd	1.308			
	$\Delta\varepsilon$	54.273			

[a] The root-mean-square deviation (rmsd) between the electrostatic potentials determined quantum-mechanically and regenerated from the point charge models is expressed in 10^{-3} au. The corresponding mean error, $\Delta\varepsilon$, is given in percents.[39]

Table 2. Models of Distributed Polarizabilities and Regenerated Molecular Polarizabilities of Benzene, Ammonium, and Water at the MP2/Sadlej//MP2/6-311++G(2d,2p) Level of Approximation[a]

	distributed polarizabilities		molecular polarizabilities		
				regenerated	MP2/Sadlej
benzene	$\alpha^{CC}_{00,00}$	−1.822	$\alpha_{10,10}$	47.537	45.171
	$\alpha^{CH}_{00,00}$	−0.280	$\alpha_{11c,11c}$	89.089	81.401
	$\alpha^{CC}_{1\kappa,1\kappa'}$	7.953	$\alpha_{11s,11s}$	89.089	81.401
	rmsd	0.025			
	$\Delta\varepsilon$	2.737			
ammonium	$\alpha^{NN}_{1\kappa,1\kappa'}$	10.708	$\alpha_{1\kappa,1\kappa'}$	10.708	9.078
	rmsd	0.195			
	$\Delta\varepsilon$	16.684			
water	$\alpha^{OH}_{00,00}$	−0.808	$\alpha_{10,10}$	10.177	9.751
	$\alpha^{OO}_{1\kappa,1\kappa'}$	8.180	$\alpha_{11c,11c}$	11.483	10.063
			$\alpha_{11s,11s}$	8.180	9.542
	rmsd	0.127			
	$\Delta\varepsilon$	7.000			

[a] The root-mean-square deviation (rmsd) between the induction energies determined quantum-mechanically and regenerated from the models of distributed polarizabilities is expressed in 10^{-3} au. The corresponding mean error, $\Delta\varepsilon$, is given in percents.[20]

Potential energy functions are very complex constructs, the building blocks of which are intimately connected. Subtle modifications of the constituent parameters can perturb significantly the delicate balance between the different terms of the force field. Beyond the scaling, or possibly the new derivation of point charges, van der Waals parameters, in principle, ought to be also adjusted to reflect the gas-phase electrostatics and the explicit inclusion of induction effects. Brute force comparison of force field performances upon plugging polarizability parameters, yet without any tuning of the other terms, would evidently render a biased picture and be somewhat unfair to the original pairwise additive potential energy function. Moreover, macromolecular force fields like Amber,[4] Charmm,[7] Gromos, or Opls-AA[9] have not been designed for numerical simulations in the gas phase. The point that the present work intends to make, however, is a demonstration that the proposed models of distributed polarizabilities yield an accurate reproduction of the induction contribution to the total interaction energies. To achieve this objective, the latter was confronted to reference SAPT2 calculations,[37,57] with the 6-311++G(*d*,*p*) basis set, for the selected series of complexes where polarization phenomena are appreciable. In the framework of the SAPT2 approach, the correlated contribution to the interaction energy is nearly equivalent to the supermolecular MP2 correlation energy.

At this stage, for clarity, the relationship between the convoluted contributions of the empirical potential energy function

$$\Delta \mathcal{U}_{\text{tot}}^{\text{MM}} = \Delta \mathcal{U}_{\text{ele}} + \Delta \mathcal{U}_{\text{ind}} + \Delta \mathcal{U}_{\text{damp}} + \Delta \mathcal{U}_{\text{vdW}} \quad (3)$$

and the different terms of an SAPT calculation

$$\Delta \mathcal{U}_{\text{tot}}^{\text{SAPT}} = \Delta \mathcal{U}_{\text{ele}} + \Delta \mathcal{U}_{\text{ind}} + \Delta \mathcal{U}_{\text{exch}} + \Delta \mathcal{U}_{\text{disp}} + \Delta \mathcal{U}_{\text{exch-ind}} + \Delta \mathcal{U}_{\text{exch-disp}} + \delta\text{HF} \quad (4)$$

ought to be established—here, $\Delta \mathcal{U}_{\text{tot}}$ denotes the total interaction energy obtained either from an empirical force field (MM) or from the SAPT scheme. Aside from the pairwise additive approximation adopted by most macromolecular force fields, the choice to describe nonbonded interactions by means of a rudimentary 6-12 Lennard-Jones potential and a Coulomb sum truncated at the monopole level clouds the assignment of a true physical meaning to these contributions. In particular, physical interpretation of individual force field components is difficult when the corresponding parameters are not fitted independently.[58] Under most circumstances, assuming that the molecular charge distribution can be represented accurately by net atomic charges, the electrostatic terms, $\Delta \mathcal{U}_{\text{ele}}$, extracted from force-field and SAPT calculations, are generally comparable, granted that penetration effects are negligible.[59] The same, unfortunately, cannot be said for the so-called van der Waals interactions, $\Delta \mathcal{U}_{\text{vdW}}$. The ad hoc, albeit physically questionable form of the Lennard-Jones potential cannot be interpreted straightforwardly, on a one-to-one basis, in terms of repulsion and dispersion. Lennard-Jones energies, in reality, embrace different terms that can be recovered from an SAPT expansion, namely a dispersion, $\Delta \mathcal{U}_{\text{disp}}$, an exchange, $\Delta \mathcal{U}_{\text{exch}}$, and an exchange-dispersion, $\Delta \mathcal{U}_{\text{exch-disp}}$, contribution. Direct incorporation of induction phenomena in classical potential energy functions raises additional concerns on account of the physically unrealistic forces that thrust the polarizing charge toward the polarizable center. In principle, the classical and the quantum-mechanical induction contributions are comparable, provided that (i) the induction energy is mapped by the model of distributed polarizabilities with an appropriate accuracy and (ii) contamination from the penetration of the electron clouds is avoided. The SAPT expansion involves, however, other terms, which ought to be modeled in the classical, polarizable force field, chief among which is an exchange-induction term, $\Delta \mathcal{U}_{\text{exch-ind}}$. The latter is supplemented by a collection of third and higher order induction and exchange-induction terms. A common route to the description of the exchange-induction contribution consists of damping the interaction of the electric field with the polarizable sites, employing a surrogate empirical function, $\Delta \mathcal{U}_{\text{damp}}$, which does not necessarily compare to the homologue SAPT component. In the present work, use was made of a damping correction that preserves the traceless feature of the interaction tensor.[60,61] The equivalence of the SAPT induction and exchange-induction higher order terms[62]—i.e. δHF, in the classical force field is less obvious, as most of these contributions do not necessarily appear in the parametrization of the polarizability models, viz. hyperpolarizability effects, which are usually neglected by computing the target quantum-mechanical induction energy at grid points lying far enough from the nuclei.

Results and Discussion

Cation-π Interactions. Over the past 20 years, cation-π interactions have progressively emerged as an important component in the subtle balance of noncovalent interactions that determine the three-dimensional structure of proteins.[6,63] They constitute a major driving force in molecular recognition processes, sufficiently strong to compete with the hydration of charged moieties and promote protein−ligand association in hydrophobic cavities formed by aromatic residues—e.g. the binding of acetylcholine to acetylcholinesterase. From an electrostatic perspective, the leading contribution to cation-π interactions is the favorable charge-quadrupole attraction of the charged species toward the π-electron cloud of the aromatic ring. A pure electrostatic description appears, however, to be generally incomplete to supply a faithful, accurate description of cation-π interactions due to the substantial polarizability of aromatic compounds combined with the polarizing character of the positively charged ion.[64] Absence of explicit induction effects in classical representations invariably results in underestimated binding constants, compared to reference quantum-mechanical calculations. Tackling polarization phenomena in cation-π complexes has been endeavored at different levels of sophistication, ranging from rudimentary, ad hoc corrections to the classical pairwise additive force field[63,65] to the explicit incorporation of isotropic polarizability parameters.[16] In retrospect, compared to up-to-date quantum-mechanically determined binding energies, neither route would seem to constitute an optimal solution. Cost-effective short-range corrections to the nonpolarizable potential energy function

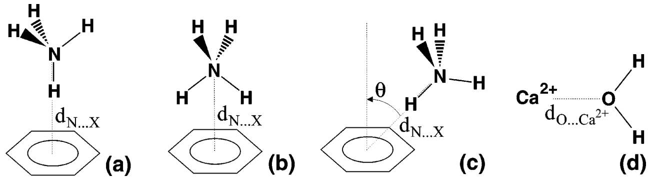


Figure 1. Ammonium-benzene cation-π interaction. (a) Monodentate and (b) bidentate complexes. X refers to the centroid of the aromatic ring. (c) Alternate approach of the cation toward the π-electron cloud of benzene, whereby the N−H chemical bond pointing toward its centroid and the normal to the aromatic plane form a 45° angle. Interaction of a divalent calcium ion with water (d). Approach of the cation toward the donor ligand is considered along the C_2 axis of the latter.

evidently cannot account for multiple cations binding the same π-electron cloud. On the other hand, inasmuch as explicit induction forces are concerned, it is far from clear whether a fully isotropic description of the polarizability is adequate for modeling cation-π interactions correctly. A simple glance at the molecular dipole polarizability of benzene estimated at the MP2/Sadlej level of theory is enough to realize that in-plane deformation of the π-electron cloud by a polarizing charge is considerably larger than it would be in the perpendicular direction (see Table 1).

An analysis of the topologically distributed polarizability model of benzene[45] reveals that a considerable portion of the in-plane polarizability—viz. typically 75%, can be described as interatomic charge-flow, whereas the out-of-plane component is almost entirely due to atomic dipole−dipole polarizabilities. Although the pattern of the topological charge-flow polarizabilities follows the well-known rules of organic chemistry, with a significant contribution between the para carbon atoms and an opposite sign meta-contribution, it can be expected that a simplified OPEP model, consisting of isotropic atomic polarizabilities borne by carbon atoms and retaining only the ortho-type charge-flow between them, is capable of reproducing the essential features of the charge-density response.

The mono- and the bidentate interactions of an ammonium ion with benzene, whereby, respectively, one and two N−H chemical bonds point toward the centroid of the aromatic ring, is depicted in Figure 1. The potential energy surfaces delineating the cation-π interaction determined using the classical Charmm force field, with and without a polarizability correction, are reported in Figures 2 and 3—see also Table 3. Not too surprisingly, the nonpolarizable potential energy function markedly underestimates the strength of the interaction. Macromolecular force fields are, however, targeted at numerical simulations in condensed phases, thus making any brute force comparison with gas-phase quantum-mechanical calculations somewhat arguable. The binding energies determined with the native Charmm force field should, therefore, be seen as a mere indicator. Accord between the values obtained at the MP2/6-311++G(d,p) level of theory and using a polarizable description is somewhat enhanced, but is this comparison necessarily justified?

It is becoming quite clear from eqs 3 and 4 that the ability of the polarizable force field to reproduce the reference

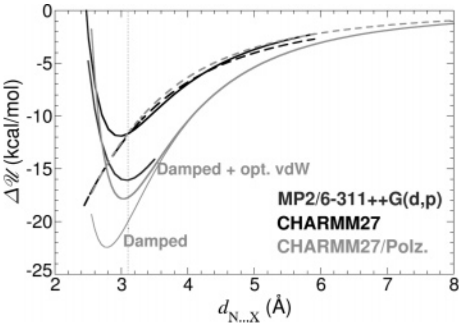


Figure 2. Monodentate motif of the ammonium-benzene cation-π interaction. X refers to the centroid of the aromatic ring. Shown is a comparison of the binding energies determined from BSSE-corrected MP2/6-311++G(d,p) calculations (dark solid line), the classical, nonpolarizable Charmm force field (black lines), and the latter supplemented by a model of distributed polarizabilities (light lines), with and without de novo optimization of the participating Lennard-Jones parameters. The electrostatic contribution to the binding energy is depicted as dashed lines. The vertical dotted line marks the position of the quantum-mechanical energy minimum.

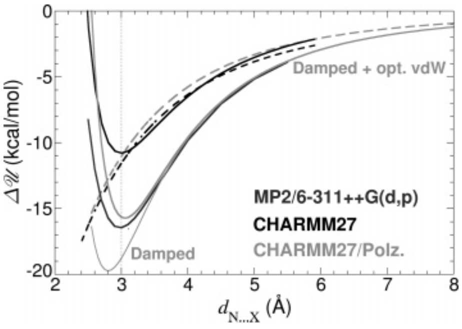


Figure 3. Bidentate motif of the ammonium-benzene cation-π interaction. X refers to the centroid of the aromatic ring. Shown is a comparison of the binding energies determined from BSSE-corrected MP2/6-311++G(d,p) calculations (dark solid line), the classical, nonpolarizable Charmm force field (black lines), and the latter supplemented by a model of distributed polarizabilities (light lines), with and without de novo optimization of the participating Lennard-Jones parameters. The electrostatic contribution to the binding energy is depicted as dashed lines. The vertical dotted line marks the position of the quantum-mechanical energy minimum.

quantum-mechanical binding energies should be appraised on the basis of the individual components, rather than as a whole. In Figures 2 and 3, the electrostatic terms inferred from the potential energy function and from the SAPT expansion are compared at various values of the reaction coordinate, for both the mono- and the bidentate motifs. As can be seen in Table 4, at the minimum of the binding energy, the SAPT electrostatic contribution of the monodentate complex matches exactly the molecular mechanics estimate, viz. −11.6 kcal/mol. This remarkable agreement might be due to the fact that point charges determined using the Sadlej basis set tend to overestimate the multipolar part of the electrostatic potential, compared to MP2 reference calculations with large, triple-ζ basis sets supplemented by diffuse functions.[66] Unfortunately, the accord is less satisfac-

Table 3. Intermolecular Separations[a] in the Mono- and Bidentate Forms of the Ammonium-Benzene Complex and in the Complex Formed by Ca^{2+} with Water, Determined from MP2/6-311++G(*d*,*p*) Potential Energy Surfaces, the Classical Charmm Force Field, and the Ab Initio Polarizable Force Field

	$d_{X\cdots Y}$		
	QM	Charmm	polarizable force field[b]
monodentate	3.1	3.0	3.1
bidentate	3.0	3.0	3.1
$Ca^{2+}\cdots H_2O$	2.3	2.3	2.4

[a] All quantities are expressed in Å. [b] Polarizable force field with a new optimization of the participating Lennard-Jones parameters.

tory for the bidentate motif, the point charge model underestimating the target electrostatic energy by 1.3 kcal/mol. Does it mean that a simple set of net atomic charges is insufficient to capture the subtle electrostatic effects arising from distinct orientational preferences of the cation? This issue can be readily addressed by solving a system of nonlinear equations satisfying the SAPT electrostatic energies for the two complexes, and the unknowns of which are the charges borne by the constituent atoms of the latter. The only way a single set of point charges can discriminate between the mono- and the bidentate motifs is by assigning a charge of −0.135 to the carbon atoms of benzene, comparable to that derived from the MP2/Sadlej wave function, and a physically unrealistic charge of 0.722 to the nitrogen atom of ammonium. Coercing artificially the cation to localize its charge onto the central nitrogen atom deteriorates dramatically the reproduction of the electrostatic potential, thus, calling into question the relevance of such a representation. It would, therefore, appear that a presumptive physically sound model may not be necessarily capable of discriminating between two forms of the same complex. Even though the higher order moments of both benzene and ammonium are described quite accurately by atom-centered point charges, as suggested by Table 1, it is far from clear whether a monopole approximation is legitimate to model the present cation-π interactions, which evidently constitute a challenging test case.

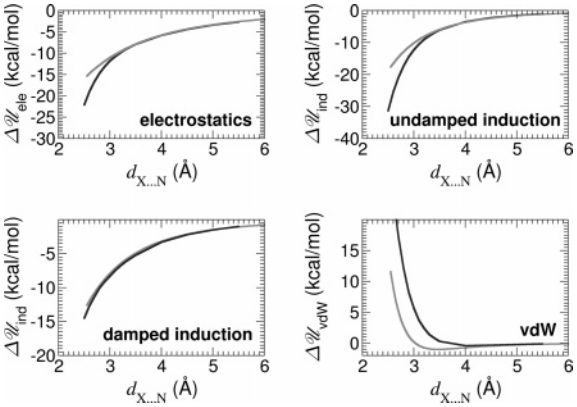


Figure 4. Components of the cation-π interaction energy, determined for the bidentate motif from an MP2/6-311++G-(*d*,*p*) SAPT expansion (dark solid line) and the polarizable potential energy function (light solid line). The undamped induction contribution corresponds to the pure induction term of eqs 3 and 4. At the quantum-mechanical level, the damped induction contribution consists of a sum of induction and exchange-induction terms. At the molecular-mechanical level, it stands for the pure induction component corrected by a damping function. The van der Waals contribution encompasses at the quantum-mechanical level the dispersion, the exchange, and the exchange-dispersion terms of the SAPT expansion. In the classical description, it coincides with the Lennard-Jones potential.

The second term of the SAPT expansion corresponds to the pure induction energy, which, as has been discussed previously, should, in principle, coincide with the contribution arising from the model of distributed polarizabilities. A glimpse at Figure 4, however, indicates otherwise. At short separations of the cation from the π-electron cloud, the polarizable force field underestimates the target SAPT term markedly. At the minimum of the binding energy for the bidentate motif, this discrepancy is still equal to ca. 3.4 kcal/mol. The two profiles delineating the quantum mechanical and the classical distance dependence of $\Delta\mathscr{U}_{ind}$, nevertheless, rapidly merge only 0.3 Å beyond the minimum, suggesting that the long-range behavior of the polarizable model is correct. It is tempting to invoke the incompleteness of the

Table 4. Comparison of the Contributions to the Binding Energies[a] of the Mono- and Bidentate Forms of the Ammonium-Benzene Complex and the Complex Formed by Ca^{2+} with Water, Determined from an SAPT2/6-311++G(*d*,*p*) Expansion and a Polarizable Potential Energy Function

	$\Delta\mathscr{U}_{ele}$		$\Delta\mathscr{U}^{b}_{ind}$		$\Delta\mathscr{U}^{c}_{vdW}$			$\Delta\mathscr{U}_{tot}$		
	SAPT	MM	SAPT	MM	SAPT	MM	δHF[d]	MP2[e]	SAPT	MM[f]
monodentate	−11.6	−11.6	−9.0	−7.9	6.4	1.8 (−0.5)	−2.0	−16.1	−16.2	**−17.7** (−20.0)
bidentate	−13.1	−11.8	−9.1	−8.6	8.3	4.7 (1.1)	−2.5	−16.5	−16.4	**−15.7** (−19.3)
$Ca^{2+}\cdots H_2O$	−51.7	−44.5	−22.1	−23.0	19.8	20.3 (3.6)	−0.8	−53.4	−54.7	**−47.2** (−63.9)

[a] All quantities are expressed in kcal/mol, with respect to the reference quantum-mechanical energy minima of Table 3. [b] The SAPT value includes the pure induction and the exchange-induction contributions. The MM model consists of the pure induction term, supplemented by a damping correction. [c] The SAPT value corresponds to the sum of the exchange, the dispersion, and the exchange-dispersion contributions. The MM value is simply the Lennard-Jones component of the force field. [d] This contribution encompasses the third and higher order induction and exchange-induction terms of the SAPT expansion. It is clearly absent in the MM description. [e] BSSE-corrected MP2/6-311++G(*d*,*p*) interaction energies. [f] The values in bold were determined after de novo optimization of the Lennard-Jones parameters. The values in parentheses correspond to estimates obtained with the standard Lennard-Jones parameters of the force field.

latter to rationalize the observed short-range divergence of the induction energy. In particular, as reported in Table 2, the present models consist of distributed isotropic dipole and charge-flow polarizabilities, and, hence, ignore short-range, higher-order terms. To which extent do they contribute to the faithful description of intermolecular interactions remains unclear. Equally unclear is the influence of charge flows between non-neighboring atoms, which are envisioned to be at play in the well-known Kekule model of benzene.

The disagreement between the *pure* induction energies determined quantum mechanically and by means of the distributed polarizability models, however, casts doubt on the physical interpretation of $\Delta\mathscr{U}_{\text{ind}}$ at short distances. For instance, the classical representation of distributed polarizabilities does not account for any possible overlap of the electron clouds that would damp the electric field felt by the polarizable sites. Yet, the introduction of a damping correction in the interaction tensor raises conceptual difficulties. As an example, the formalism put forth by Thole yields an interaction tensor that is no longer traceless, contrary to the unaltered tensor. More importantly, the damping function proposed by Thole is not continuous, which can be critical at short intermolecular distances. Although, strictly speaking, the ad hoc damping correction, $\Delta\mathscr{U}_{\text{damp}}$, cannot be compared directly with the SAPT exchange-induction term, $\Delta\mathscr{U}_{\text{exch}-\text{ind}}$, it may be contended that the sum of $\Delta\mathscr{U}_{\text{ind}}$ and $\Delta\mathscr{U}_{\text{damp}}$, in the classical description can be related to the sum of $\Delta\mathscr{U}_{\text{ind}}$ and $\Delta\mathscr{U}_{\text{exch}-\text{ind}}$, at the quantum-mechanical level. This is illustrated in the components of Figure 4, which highlight the coincidence of the damped induction profiles over the entire range of cation-π distances explored on the potential energy surface and show that below 3 Å, the undamped induction is considerably stronger than the value derived from the distributed polarizability model. This behavior might be related to the instability of the SAPT induction energy reflecting a divergence of the polarization series. This overestimation is corrected by the exchange-induction term,[67] justifying that the comparison should be done between the damped classical induction energy, on the one hand, and the sum of induction and exchange-induction energies, on the other. Quantitatively, Table 4 reveals that the agreement between the polarizable models and the quantum-mechanical calculations varies from 0.5 to 1.1 kcal/mol for the bidentate and the monodentate complexes, respectively. Interestingly enough, in the light of a Kitaura-Morokuma energy decomposition[68] performed at the HF/6-311+G(*d*,*p*) level of theory for the mono- and the bidentate motifs of the ammonium-benzene complex, a contribution embracing polarization and charge-transfer terms of −9.0 to −9.3 kcal/mol was found.[69] It should be reminded, however, that in a perturbation expansion of the total interaction energy, the charge-transfer term is a short-range part of the induction contribution.[70]

The exact role played by van der Waals interactions in the binding energies is somewhat more difficult to apprehend. As has been emphasized previously, what is generically referred to as the van der Waals contribution can be expressed in the SAPT expansion as the sum of exchange, dispersion, and exchange-dispersion terms. The meaning of van der Waals interactions in a molecular-mechanical description is far more ambiguous, as it embraces in a heuristic, ad hoc function everything from the nonbonded contribution that is neither electrostatic nor induction-related. Assigning a physical meaning to this function and to its individual terms, therefore, constitutes a daunting task. At the beginning of this section, the arbitrariness and the unfair nature of a gross assessment of the Charmm force field to reproduce quantum-mechanical binding energies of cation-π complexes has been underlined. Equally arbitrary is the direct comparison of the binding energies determined quantum mechanically and employing a polarizable and a nonpolarizable force field. The reason is self-explanatory: The van der Waals part of the classical Charmm potential energy function has been optimized for a given set of net atomic charges, assumed to reflect the interaction of the parametrized chemical moieties with an aqueous environment.[7,71] Alteration of the electrostatic contribution to the empirical force field by using point charges appropriate for gas-phase simulations creates an imbalance in the construct, that ought to be corrected by a new optimization of the Lennard-Jones parameters. This rationalizes the noteworthy disagreement between the profiles of Figures 2 and 3 computed at the MP2/6-311++G(*d*,*p*) level of approximation and using the Charmm force field supplemented by atomic polarizabilities. A rapid glance at the van der Waals contributions gathered in Table 4 and Figure 4 suffices to appraise the paramount importance of a proper parametrization of these interactions. At the minimum of the binding energy, the molecular-mechanical interaction energy based on Charmm Lennard-Jones parameters underestimates the repulsion of the nuclei by an amount of 6.9−7.2 kcal/mol for the monodentate and the bidentate complex, respectively. This lack of repulsion between the polarizing cation and the π-electron cloud is illustrated in Figures 2 and 3, where the position of the energy minima is shifted relative to the quantum-mechanical estimates. Interestingly enough, Figure 4 also reveals that the van der Waals profiles obtained from an SAPT expansion and from the classical description have distinct shapes. Whereas the former decays rapidly and exhibits a marginal, shallow minimum around 4 Å, the latter is much smoother, with a pronounced minimum near 3.5 Å. Novel parametrization of the Lennard-Jones interaction potential is, in principle, feasible, based on the wealth of data supplied by the SAPT expansion computed over the entire reaction pathway. Given the geometries of the bidentate cation-π complex and the components of the binding energy at various separations depicted in Figure 4, updated Lennard-Jones parameters can be fitted numerically to the sum of the dispersion, the exchange, and the exchange-dispersion terms of eq 4.

Even though the electrostatic, the induction, and the van der Waals contributions were mimicked optimally by the classical polarizable force field, one could still argue that any attempt to match exactly the quantum-mechanical binding energies is doomed from the onset on account of neglected terms in the potential energy function. These terms, which are referred to as δHF in the SAPT expansion 4, are third and higher order induction and exchange-induction contributions. They are evidently absent from the classical description of the cation-π interaction, as any possible contamination through hyperpolarizability effects of the induction energy has been carefully probed in the fitting procedure of the polarizability parameters.[72] As can be seen in Table 4, the δHF component represents about 10−15% of the total interaction energy and, hence, cannot be ignored in cation-π complexes. This issue, which has been seldom tackled hitherto, further calls into question the promising agreement reached in previous endeavors to model the energetics of such intricate molecular systems.[16,63,65] Granted that hyperpolarizability effects cannot be easily incorporated in the models of distributed polarizabilities without modulating more or less severely the accurate reproduction of the quantum-mechanical induction energy, one possible route to account for the higher order—i.e. nonlinear or many-body, induction and exchange-induction contributions embodied in the δHF term consists of considering the latter in the de novo parametrization of the Lennard-Jones potential. New parameters were optimized following this route and combined to the pure electrostatic and damped induction contributions of the classical force field. The markedly improved accord between the quantum- and the molecular-mechanical estimates of the binding energies is highlighted in Figures 2 and 3. Since the new Lennard-Jones parameters were fitted to the SAPT expansion performed on the bidentate complex, it is not completely surprising that the agreement is somewhat better for the latter than for the monodentate motif. As indicated in Table 4, accuracy in the reproduction of the target quantum-mechanical binding energy for the bidentate complex, within 0.7 kcal/mol, is unprecedented. The discrepancy between the quantum-mechanical and the molecular-mechanical values is, however, somewhat more pronounced for the monodentate complex, viz. 1.6 kcal/mol. Remarkably enough, the newly optimized 6-12 potential causes the position of the molecular-mechanical energy minimum to shift, virtually matching that of the corresponding MP2/6-311++G(*d*,*p*) profiles—see Table 3. Yet, in spite of these improvements, the hierarchy of the associated states of benzene with ammonium still cannot be fully recovered, the monodentate complex emerging 2 kcal/mol below the bidentate complex, when, in principle, their binding energies should be roughly equal. This difference is believed to be rooted in a subtle imbalance between electrostatic and induction contributions that the present polarizable force field cannot capture entirely.

Although the minimum of the binding energy of cation-π complexes generally corresponds to a directional interaction of the polarizing species pointing perpendicularly toward the aromatic ring, approach of the ion may proceed with a different azimuth.[63] It has been seen so far that the polarizable models proposed herein have proven to reproduce the quantum-mechanical binding energies reasonably well. To probe the transferability[73] of the molecular-mechanical potential energy function to other interaction motifs, the 45° approach of ammonium toward benzene was explored (see Figures 1 and 5). In this orientation, on account of steric hindrances, the cation is necessarily coerced to adopt a monodentate-like binding mode. At the MP2/6-311++G(*d*,*p*) level of theory, the association energy corresponding to this approach is about 6.4 kcal/mol higher than the most stable bidentate complex. In glaring contrast with both the bidentate and the monodentate motifs, where the electrostatic component is always stronger than the damped induction term, here, the electrostatic energy is appreciably weaker, which is anticipated to stem from a modulation of the attractive charge-quadrupole interaction by the modified orientation of the cation with respect to the π-electron cloud. As illustrated in Figures 2 and 3, the polarizable force field matches the quantum-mechanical binding energy within 0.9 kcal/mol, which not only is encouraging but also suggests that the classical model is able to capture the anisotropy of induction phenomena.

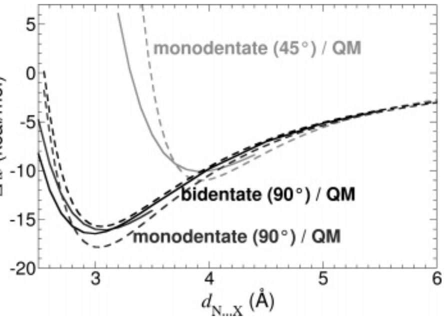


Figure 5. Comparison of the monodentate and the bidentate motifs of the ammonium-benzene complex with an alternate approach of the cation toward the π-electron cloud of benzene, whereby the N−H chemical bond pointing toward its centroid and the normal to the aromatic plane form a 45° angle. Quantum- and molecular-mechanical energies are shown in solid and dashed lines, respectively.

Interaction of a Calcium Ion with Water. Armed with a consistent strategy for modeling induction phenomena in intermolecular interactions, we now delve into a different class of complexes, wherein polarization effects have been shown to be equally sizable. The C_2 association depicted in Figure 1 of a divalent calcium ion with water was examined at the molecular-mechanical level, using the additive pairwise Charmm force field and an ab initio polarizable force field, and quantum mechanically, at the MP2/6-311++G(*d*, *p*) level of approximation. From the onset, it can be observed in Figure 6 that the binding energy of the calcium-water complex is reasonably reproduced with or without explicit polarization. Employing the standard Lennard-Jones parameters of the Charmm potential energy function, attraction is clearly exaggerated, highlighted in Table 4, the van der Waals contribution being underestimated by about 16.2 kcal/mol. It is also noteworthy that, compared with the sum of SAPT dispersion, exchange, and exchange-dispersion terms, the position of the shallow minimum of the classical van der Waals profile is shifted about 0.8 Å toward shorter cation-ligand separations (see Figure 7). It is remarkable that the de novo optimization of the molecular-mechanical 6-12 potential based on the SAPT expansion leads to a virtually flawless reproduction of the target quantum-mechanical van der Waals term. The damped induction contribution is also recovered within chemical accuracy. Yet, the ab initio polarizable potential energy function underestimates by 7.2 kcal/mol the total binding energy. A closer look at the

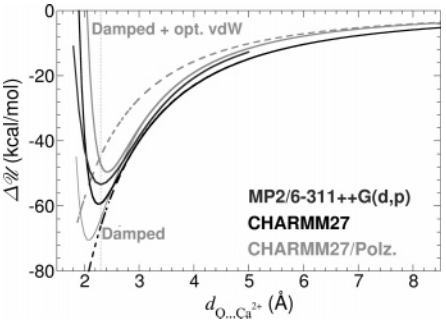


Figure 6. Interaction of a divalent calcium ion with water. Comparison of the binding energies determined from BSSE-corrected MP2/6-311++G(d,p) calculations (dark solid line), the classical, nonpolarizable Charmm force field (black lines), and the latter supplemented by a model of distributed polarizabilities (light lines), with and without de novo optimization of the participating Lennard-Jones parameters. The electrostatic contribution to the binding energy is depicted as dashed lines. The vertical dotted line marks the position of the quantum-mechanical energy minimum.

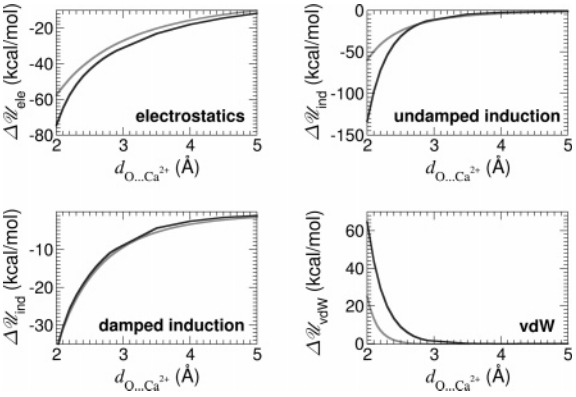


Figure 7. Components of the interaction energy determined for the calcium-water complex from an MP2/6-311++G(*d*,*p*) SAPT expansion (dark solid line) and the polarizable potential energy function (light solid line). The undamped induction contribution corresponds to the pure induction term of eqs 3 and 4. At the quantum-mechanical level, the damped induction contribution consists of a sum of induction and exchange-induction terms. At the molecular-mechanical level, it stands for the pure induction component corrected by a damping function. The van der Waals contribution encompasses at the quantum-mechanical level the dispersion, the exchange, and the exchange-dispersion terms of the SAPT expansion. In the classical description, it coincides with the Lennard-Jones potential.

components of the latter reveals that this discrepancy is likely to be rooted in a flawed molecular-mechanical description of the electrostatic contribution, about 7.2 kcal/mol lower than its SAPT counterpart. Remembering that three-point charge models of water are unable to mimic the large atomic quadrupole borne by the central oxygen atom and generally yield errors in the reproduction of the molecular electrostatic potential on the order of 50% (see Table 1), this result is not completely unexpected. Again, we are faced with a stringent test case that underscores the limitations of atom-centered point charge models and urges us to explore the possibilities of extending them by the inclusion of higher order multipole expansion effects. In an attempt to address this issue, Piquemal et al. have recently delved into the interaction of a calcium ion with the more sophisticated Amoeba[74] water model.[75]

Paradoxically, replacing the gas-phase MP2/Sadlej charges by those of the TIP3P model of water, i.e., by −0.834 on the oxygen atom, increases dramatically the classical electrostatic contribution from −44.5 to −55.3 kcal/mol, hence improving the overall accord on the total binding energy. The adverb "paradoxically" is utilized here on purpose: it is not really surprising that by boosting the polarity of the water molecule, the absence of higher-order moments is compensated in an artificial fashion, concealing the incompleteness of the point charge model. The true paradox lies in the impression that an error in the parametrization of the electrostatic potential, viz. an incomplete model truncated to the monopole term of the multipole expansion, can be somehow corrected by another error, viz. the use of a charge distribution representative of a condensed phase rather than a low-pressure gaseous state.

Conclusion

The theoretical grounds that underlie the development of an ab initio polarizable force field are reported in this contribution. The key ingredient of the proposed potential energy function is a model of implicitly interacting polarizabilities derived numerically from the induction energy mapped around a molecule. Employing a combination of atom-centered isotropic dipole plus charge-flow polarizabilities, the anisotropy of induction phenomena is essentially recovered, thereby obviating the need for the explicit incorporation of cumbersome anisotropic dipole polarizabilities.[36] One might wonder, however, whether the faithful reproduction of the induction energy, the dielectric fingerprint of the molecule, necessarily guarantees an accurate description of intermolecular interactions. To tackle this question, use was made of the pairwise additive Charmm force field, in which models of distributed polarizabilities were plugged. Yet, seamless introduction in a "plug-and-play" fashion of explicit polarization phenomena rapidly proved to be an elusive goal, as newly added contributions to the potential energy perturb the delicate, almost precarious balance of the original, nonpolarizable force field. As a result, direct comparison of the unaltered and the polarizable potential energy functions, in the absence of appropriate correction, is inherently biased. Equally biased is the assessment of these force fields based on gas-phase, high-level quantum-mechanical determinations of potential energy surfaces. What distinguishes, however, the ab initio polarizable force field from its pairwise additive version lies in its physically sound contributions that can be readily compared to the successive terms of an SAPT expansion. In the context of gas-phase quantum-mechanical calculations, the electrostatic component of the polarizable force field is expected to match its SAPT homologue, provided that the electrostatic potential around the molecule is properly described by the distribution of net atomic charges, and penetration of the electron clouds can be safely neglected.[34,59] As has been shown in the present work, the pure induction term of the force field supplemented by a damping correction to account for possible overlaps of the participating electron clouds is also expected to coincide with the sum of the induction and the exchange-induction contributions of the SAPT expansion. Evidently enough, altering only partially the nonbonded section of the pairwise additive Charmm force field is inconsistent. This is reflected by an exaggeratedly weak repulsion of the atoms, which can only be corrected through de novo optimization of the constituent Lennard-Jones parameters. The difficulty to develop a balanced polarizable potential energy function, capable of reproducing gas-phase, quantum-mechanical energetics within chemical accuracy, is further magnified by the necessity to take into account higher order terms of the SAPT expansion, which have no real equivalence in the classical description. Case in point—the third and higher-order induction and exchange-induction contributions to the binding energy of the ammonium-benzene complex can be as large as −2.5 kcal/mol, hence, suggesting that derivation of a polarizable force field sufficiently precise for modeling cation-π interactions may easily turn out to be a fateful venture. Such higher order contributions cannot be ignored and ought to be taken into account in the new parametrization of the 6-12 Lennard-Jones potential. Adopting this strategy, which, in the present case, constitutes a proof of concept for gas-phase complexes, the binding energies determined for the monodentate and the bidentate interaction of ammonium with benzene matched reasonably well the MP2/6-311++G(d,p) estimates, albeit hierarchy of the two states is inverted. These results provide a cogent illustration of the difficulties faced by the theoretician when modeling the subtle balance of electrostatic, induction, and van der Waals contributions that drive cation-π interactions—arguably one of the most challenging cases for assessing the performance of a polarizable force field. In this sense, interaction of a calcium ion with water constitutes a somewhat lesser challenge. Since this interaction is predominantly governed by electrostatic and, to a lesser extent, by induction contributions, suboptimal description of the van der Waals term is less critical than in cation-π complexes. This example raises also the question as to whether simple point charge models will ever be satisfactory or if higher atomic multipoles ought to be included as well. It still remains that de novo parametrization of the Lennard-Jones potential is the key to an improved agreement with the quantum-mechanical binding energies. Either for cation-π interactions or association of a divalent cation with a donor ligand, the very encouraging results reported here underline the strength of ab initio polarizable force fields for handling accurately induction phenomena. The strategy developed represents a significant step forward in the race for modeling explicitly polarization effects in molecular systems, opening exciting new vistas for numerical simulations of condensed phases.

Acknowledgment. Georg Jansen is gratefully acknowledged for stimulating discussions. The authors thank Krzysztof Szalewicz for supplying the SAPT code. They are indebted to the Crvhp, Vandœuvre-les-Nancy, France, and to the Centre de Supercomputació de Catalunya (CESCA) Barcelona, Spain, for provision of generous amounts of computational time. This work was supported by the IngeMol project of the Agence National pour la Recherche, by a joint CNRS/UIUC collaborative grant, and by the Spanish Ministerio de Educación y Ciencia (grant CTQ2005-08797-C02-01/BQU). K.S. is supported by the Humboldt Foundation.

References

(1) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269−6271.

(2) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab initio molecular orbital theory;* Wiley-Interscience: New York, 1986.

(3) Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. *J. Comput. Chem.* **1993**, *14*, 1240−1249.

(4) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. C.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(5) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671−15682.

(6) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303−1324.

(7) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(8) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656−1676.

(9) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(10) Noskov, S. Y.; Berneche, S.; Roux, B. *Nature* **2001**, *431*, 830−834.

(11) Bucher, D.; Raugei, S.; Guidoni, L.; Dal Peraro, M.; Rothlisberger, U.; Carloni, P.; Klein, M. L. *Biophys. Chem.* **2006**, *124*, 292−301.

(12) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065−1082.

(13) Applequist, J. *Acc. Chem. Res.* **1977**, *10*, 79−85.

(14) Thole, B. T. *J. Chem. Phys.* **1981**, *59*, 341−350.

(15) Davidson, E. R.; Chakravorty, S. J. *Chem. Phys. Lett.* **1994**, *217*, 48−54.

(16) Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4177−4178.

(17) Nakagawa, S.; Kosugi, N. *Chem. Phys. Lett.* **1993**, *210*, 180−186.

(18) Le Sueur, C. R.; Stone, A. J. *Mol. Phys.* **1994**, *83*, 293−308.

(19) Ángyán, J. G.; Jansen, G.; Loos, M.; Hättig, C.; Hess, B. A. *Chem. Phys. Lett.* **1994**, *219*, 267−273.

(20) Celebi, N.; Ángyán, J. G.; Dehez, F.; Millot, C.; Chipot, C. *J. Chem. Phys.* **2000**, *112*, 2709−2717.

(21) Dehez, F.; Chipot, C.; Millot, C.; Ángyán, J. G. *Chem. Phys. Lett.* **2001**, *338*, 180−188.

(22) Misquitta, A. J.; Stone, A. J. *J. Chem. Phys.* **2006**, *124*, 024111.

(23) Wang, W.; Skeel, R. D. *J. Chem. Phys.* **2005**, *123*, 164107.

(24) Drude, P. *Lehrbuch der Optik. 1. Ausgabe;* Verlag von S. Hirzel: Leipzig, 1900.

(25) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556−7560.

(26) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185−5197.

(27) Rappé, A. K.; Goddard, W. A., III *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(28) Rick, S. W.; Berne, B. J. *J. Phys. Chem. B* **1997**, *101*, 10488−10493.

(29) Patel, S.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1−15.

(30) Sanderson, R. T. In *Chemical Bonds and Bond Energy;* Academic Press: New York, 1976; p 15.

(31) Cieplak, P.; Caldwell, J. W.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1048−1057.

(32) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497−1506.

(33) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515−1531.

(34) Piquemal, J. P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.

(35) Ángyán, J. G.; Chipot, C.; Dehez, F.; Hättig, C.; Jansen, G.; Millot, C. *J. Comput. Chem.* **2003**, *24*, 997−1008.

(36) Chipot, C.; Ángyán, J. G. *New J. Chem.* **2005**, *29*, 411−420.

(37) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887−1930.

(38) Cox, S. R.; Williams, D. E. *J. Comput. Chem.* **1981**, *2*, 304−323.

(39) Chipot, C.; Maigret, B.; Rivail, J. L.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 10276−10284.

(40) Ángyán, J. G.; Chipot, C. *Int. J. Quantum Chem.* **1994**, *52*, 17−37.

(41) Francl, M. M.; Chirlian, L. E. *Rev. Comput. Chem.* **2000**, *14*, 1−31.

(42) Nakagawa, S. *Chem. Phys. Lett.* **1997**, *278*, 272−277.

(43) Chipot, C.; Dehez, F.; Ángyán, J. G.; Millot, C.; Orozco, M.; Luque, F. J. *J. Phys. Chem. A* **2001**, *105*, 11505−11514.

(44) Bader, R. F. W. *Atoms in Molecules − A Quantum Theory;* Oxford University Press: London, 1990.

(45) Hättig, C.; Jansen, G.; Hess, B. A.; Ángyán, J. G. *Can. J. Chem.* **1996**, *74*, 976−987.

(46) Jansen, G.; Hättig, C.; Hess, B. A.; Ángyán, J. G. *Mol. Phys.* **1996**, *88*, 69−92.

(47) Hättig, C.; Jansen, G.; Hess, B. A.; Ángyán, J. G. *Mol. Phys.* **1997**, *91*, 145−160.

(48) Stone, A. J. Classical electrostatics in molecular interaction. In *Theoretical models of chemical bonding*; Maksić, H., Eds.; Springer-Verlag: Berlin, 1991; Vol. 4, pp 103−131.

(49) Hättig, C.; Hess, B. A. *Mol. Phys.* **1994**, *81*, 813−824.

(50) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995.

(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, C.; Gonzalez, M.; Challacombe, P. M.; Gill, W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98 Revision A.7;* Gaussian Inc.: Pittsburgh, PA, 1999.

(52) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208−6219.

(53) Chipot, C. *J. Comput. Chem.* **2003**, *24*, 409−415.

(54) Morita, A. *J. Comput. Chem.* **2002**, *23*, 1466−1471.

(55) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 9903−9906.

(56) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153−168.

(57) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorska, M.; Jeziorski, B.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszyński, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wheatley, R. J.; Wormer, P. E. S.; Zuchowski, P. S. *SAPT2006: An ab initio program for many-body symmetry-adapted perturbation theory calculations of intermolecular interaction energies. Sequential and parallel versions;* Department of Physics and Astronomy, University of Delaware: Newark, DE 19716 and Department of Chemistry, University of Warsaw: ul. Pasteura 1, 02-093 Warsaw, Poland, 2006.

(58) Claverie, P. Elaboration of approximate formulas for the interaction between large molecules: Application to organic chemistry. In *Intermolecular Interactions: From Diatomics to Biopolymers;* Pullman, B., Eds.; Wiley-Interscience: New York, 1978; Vol. 1, p 69.

(59) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300−7306.

(60) Jensen, L.; Åstrand, P. O.; Osted, A.; Kongsted, J.; Mikkelsen, K. V. *J. Chem. Phys.* **2002**, *116*, 4001−4010.

(61) Use was made of the IM-SQRT damping function proposed by Jensen et al. in ref 60, viz. eq 16. A damping parameter $\phi_p = 0.085$ has been determined for the calcium ion.

(62) Patkowski, K.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **2006**, *125*, 154107.

(63) Minoux, H.; Chipot, C. *J. Am. Chem. Soc.* **1999**, *121*, 10366−10372.

(64) Cubero, E.; Luque, F. J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5976−5980.

(65) Chipot, C.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* 2998−3005.

(66) Piquemal, J. P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem A* **2003**, *107*, 10353−10359.

(67) Patkowski, K.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **2006**, *125*, 154107.

(68) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325−340.

(69) Aschi, M.; Mazza, F.; Di Nola, A. *J. Mol. Struct. (Theochem*) **2002**, *587*, 177−188.

(70) Stone, A. J. *The theory of intermolecular forces;* Clarendon Press: Oxford, 1996.

(71) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(72) Dehez, F.; Soetens, J. C.; Chipot, C.; Ángyán, J. G.; Millot, M. *J. Phys. Chem. A* **2000**, *104*, 1293−1303.

(73) Geerke, D. P.; van Gunsteren, W. F. *J. Phys Chem. B* **2007**, *111*, 6425−6436.

(74) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(75) Piquemal, J. P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511.

CT700156A

JCTC Journal of Chemical Theory and Computation

Polarizable Empirical Force Field for the Primary and Secondary Alcohol Series Based on the Classical Drude Model

Victor M. Anisimov,[†] Igor V. Vorobyov,[†] Benoît Roux,[‡] and
Alexander D. MacKerell, Jr.[*,†]

Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201, and Institute of Molecular Pediatric Sciences, Gordon Center for Integrative Science, University of Chicago, 929 East 57th Street, Chicago, Illinois 60637

Received April 28, 2007

Abstract: A polarizable empirical force field based on the classical Drude oscillator has been developed for the aliphatic alcohol series. The model is optimized with an emphasis on condensed-phase properties and is validated against a variety of experimental data. Transferability of the developed parameters is emphasized by the use of a single electrostatic model for the hydroxyl group throughout the alcohol series. Aliphatic moiety parameters were transferred from the polarizable alkane parameter set, with only the Lennard-Jones parameters on the carbon in methanol optimized. The developed model yields good agreement with pure solvent properties with the exception of the heats of vaporization of 1-propanol and 1-butanol, which are underestimated by approximately 6%; special LJ parameters for the oxygen in these two molecules that correct for this limitation are presented. Accurate treatment of the free energies of aqueous solvation required the use of atom-type specific $O_{alcohol}-O_{water}$ LJ interaction terms, with specific terms used for the primary and secondary alcohols. With respect to gas-phase properties the polarizable model overestimates experimental dipole moments and quantum mechanical interaction energies with water by approximately 10 and 8%, respectively, a significant improvement over 44 and 46% overestimations of the corresponding properties in the CHARMM22 fixed-charge additive model. Comparison of structural properties of the polarizable and additive models for the pure solvents and in aqueous solution shows significant differences indicating atomic details of intermolecular interactions to be sensitive to the applied force field. The polarizable model predicts pure solvent and aqueous phase dipole moment distributions for ethanol centered at 2.4 and 2.7 D, respectively, a significant increase over the gas-phase value of 1.8 D, whereas in a solvent of lower polarity, benzene, a value of 1.9 is obtained. The ability of the polarizable model to yield changes in the dipole moment as well as the reproduction of a range of condensed-phase properties indicates its utility in the study of the properties of alcohols in a variety of condensed-phase environments as well as representing an important step in the development of a comprehensive force field for biological molecules.

Introduction

Alcohol moieties are one of the most ubiquitous classes of functional groups, representing building blocks of proteins, nucleic acids, lipids, and carbohydrates as well as being found in a wide range of industrial chemicals, including pharmaceuticals. For example, hydroxyls are present in the amino acids serine, threonine, and tyrosine, and the presence of the hydroxyl group at the 2′ position of the ribose ring in RNA leads to its unique properties as compared to DNA and hydroxyls which dominate the structure and function of carbohydrates. Notable is the presence of both polar and nonpolar moieties in their structures, allowing alcohols to participate both in hydrophobic and hydrophilic interactions. To better understand the properties of alcohols a number of theoretical chemistry studies have been undertaken,[1−9]

[*] Corresponding author phone: (410)706-7442; fax: (410)706-5017; e-mail: alex@outerbanks.umaryland.edu.

[†] University of Maryland.

[‡] University of Chicago.

10.1021/ct700100a CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/15/2007

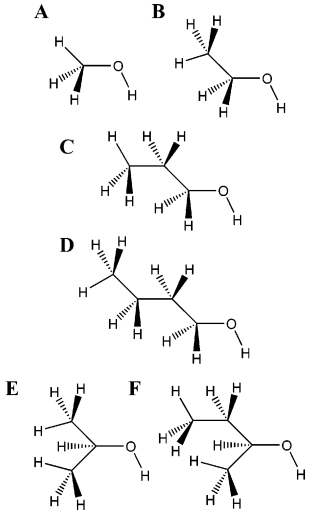


Figure 1. Model compounds used in the parameter optimization. Primary alcohols (A) methanol, (B) ethanol, (C) 1-propanol, and (D) 1-butanol and secondary alcohols (E) 2-propanol (isopropyl alcohol) and (F) 2-butanol.

including empirical force field based studies investigating their condensed-phase properties. To date a majority of these have been based on nonpolarizable (fixed-charge additive) models, such as those available in the popular all-atom biomolecular force fields CHARMM,[10,11] Amber,[12] and OPLS-AA,[4,13] among others.[3,14,15] The additive models use fixed partial atomic charges, and in all cases it is necessary to overestimate the gas-phase dipole moment of alcohols by approximately 40% in order to accurately treat the condensed phase (including alcohols in aqueous solution). Despite their usefulness, the assumption of additivity in the treatment of electrostatic interactions prevents an accurate treatment of the full range response in polar and nonpolar environments where the alcohol function groups exist in biomolecules. To overcome the limitation of additive empirical force fields, models that include explicit treatment of electronic polarizability are being developed, and a number of classical polarizable models for alcohols have been presented.[16−23] These models have been based on induced dipoles,[16−18,24] fluctuating charges,[20−22] and the Drude model[19,23] to treat the electronic polarizability. In some of the models the internal parameters were transferred directly from the corresponding additive models with few remaining parameters being optimized to reproduce condensed-phase properties.[18,19,24] In other models parameters for hydroxyl groups are set to be unique for the particular alcohol being studied and cannot be regarded as transferable across the alcohol series.[21−23] In the present work we extend these efforts via the development of a polarizable model for the alcohol series (Figure 1) with an emphasis on maximizing the transferability of the developed parameters to biological macromolecules. In addition, a systematic, iterative approach is applied to rigorously optimize all aspects of the force field parameters to maximize the overall accuracy of the model.

The present work follows the hierarchical approach toward the optimization of force field parameters that was originally developed for the CHARMM all-atom biomolecular force field.[11] This work builds on the classical Drude polarizable models developed for water,[25,26] the alkane series,[27] aromatics,[28] and ethers.[29] As before the optimization of all necessary parameters including atomic charges, atomic polarizabilities, internal equilibrium parameters, force constants, torsion potentials, and Lennard-Jones terms is undertaken. In the polarizable alcohol series the alkyl group parameters are transferred directly from the alkanes with only the methyl group in methanol being partially optimized as described below.

Methods

The induced polarization framework employed in this work is based on the classical Drude oscillator model as described previously.[30] According to this model each non-hydrogen atom is described by two point charges q_{core} and δ connected by a harmonic spring with a force constant of k_D. The sum of the two point charges yields the partial atomic charge q_A associated with atom A (i.e., $q_A = q_{\text{core}} + \delta$). The host atom in the Drude model is also the center of the Lennard-Jones (LJ) radius, whereas the Drude particle typically does not carry LJ parameters (although an LJ parameter can be assigned in principle). Placement of an atom in an external field E causes a displacement of the charged Drude particle, which gives rise to an induced dipole $\mu = \alpha E$; the displacement (x) of Drude particles from their corresponding atomic centers in response to the external field is opposed by the restoring force of the harmonic spring $F_{\text{harm}} = -k_D x$, which defines the polarizability $\alpha = \delta^2/k_D$ of an atom. When there are many polarizable atoms responding to a field E, the calculation of the total electrostatic interactions can be achieved by relaxing the charged Drude particles iteratively until self-consistency. Alternatively, in the case of MD simulations an extended Lagrangian may be applied allowing the treatment of the electronic degrees of freedom as dynamic variables.[30,31]

Electrostatic interactions of Drude particles with other charged centers involving 1,4 pairs and beyond are treated according to Coulomb's Law as commonly used in classical force fields.[32] 1,2 and 1,3 intramolecular electrostatic interactions between covalently connected atoms, which are typically turned off in additive models, are reintroduced in the polarizable model at the level of dipole−dipole interactions. Here, the Drude particles inherit the position index of their corresponding host atoms. The presence of 1,2 and 1,3 dipole interactions increases the internal polarizability of molecules as well as the anisotropy of that polarizability.[33] Effective inclusion of the 1,2 and 1,3 interactions requires their scaling to avoid polarization catastrophe. Scaling is performed using the approach of Thole[34,35] where the interaction energy of the induced dipoles on the atomic centers i and j is calculated according to the modified Thole scheme

$$\frac{\delta_i\delta_j}{r_{ij}}\left[1-\left(1+\frac{\bar{r}_{ij}}{2}\right)\cdot\exp(-\bar{r}_{ij})\right] \qquad (1)$$

where the normalized distance is defined as

$$\bar{r}_{ij} = a\frac{r_{ij}}{\sqrt[6]{\alpha_i\alpha_j}} \qquad (2)$$

Table 1. Specification of Connolly Surfaces for Perturbation Ion and Grid Point Placement

surface no.	vdW scale factor	density factor	dist to atoms	dist to pert. ions	type
1[a]	1.3	5.0	0.0	0.0	ions(bonds)
	1.3	5.0	1.4	2.0	ions(gaps)
2	2.2	1.1	1.5	2.0	ions(bonds)
	2.2	1.1	2.0	2.0	ions(gaps)
3	4.0	0.1	2.0	2.0	ions(bonds)
4[a]	1.3	6.0	0.0	2.0	grids
5	2.2	1.1	1.0	2.0	grids
6	3.0	1.3	1.0	2.0	grids
7	5.0	0.6	1.0	2.0	grids
8	6.0	0.2	1.0	2.0	grids

[a] Only in the vicinity of polar atoms.

and r_{ij} is the distance between the interacting centers, α_i is the atomic polarizability of center i, and a is a Thole parameter with a default value of 2.6, originally selected to produce the correct polarizability anisotropy ratio of benzene.[34,35]

Optimization of the electrostatic parameters is performed using the FITCHARGE module in the program CHARMM[10,11] and is an extension of our previously published protocol.[36] Modifications include the placement of an additional "near" grid and the inclusion of virtual charged particles at the position of oxygen lone-pairs (LP).[33] The effective charge of the oxygen atom is moved entirely to the corresponding LP sites (two LPs per oxygen atom), while the polarizability is retained on the atomic center. Initial optimization of charges and polarizabilities was done by fitting to quantum mechanical (QM) electrostatic potential (ESP) maps. One "unperturbed" EPS map is calculated for the isolated molecule of interest, and a number of additional "perturbed" maps are calculated in the presence of small point charges (0.5*e*) placed at different locations to probe the polarization response. Grids defining the ESP were placed on concentric surfaces at multiples of the van der Waals (vdW) radii of atoms. The perturbating point charges as required for the production of perturbed ESPs are placed along bonds and at additional locations to approximate an isotropic distribution around the molecule, using the parameters listed in Table 1. In Table 1 the vdW scale factor indicates the distance of the Connolly surfaces from the atomic centers in multiples of the corresponding atomic vdW radius. The total number of points on each surface is a multiple of the "density factor". The "distance to atoms" parameter does not allow placement of a perturbation ion or a grid point at a distance less than the specified value to any atom in the molecule. The "distance to perturbation ions" parameter prevents newly added perturbation ions being closer to other ions than the specified distance. Finally, "type" descriptor specifies the way the points are generated. The type "ions (bonds)" means that the perturbation ions are placed along vectors extended from covalent bonds, while "ions (gaps)" places ions in vacant regions on a surface between the ions originally placed based on the bonds criteria. The type "grids" indicates placement of grid points. The present work extends our previous methodology by adding a layer (#1) of perturbation ions and a corresponding layer of grid points (#4) at a vdW scale factor of 1.3, thereby taking into account the electrostatic response at distances corresponding to direct hydrogen bonds, as described below. Further, an additional layer (#5) of grid points at distances corresponding to the separation of heavy atoms involved in hydrogen bonds was added.

Initial guesses of the charges for ESP fitting were based on the CHARMM22 force field.[11] Initial values for the polarizabilities were obtained from the additive atomic polarizabilities of Miller[37] modified for the present nonhydrogen polarizability model, as described previously.[36] Fitting was performed on monomer geometries optimized at the MP2(fc)/6-31G(d) level of theory[38] using the Gaussian 03 package[39] with ESPs calculated using the B3LYP functional[40−44] with the aug-cc-pVDZ basis set.[45] Gauche and trans conformations of ethanol and isopropyl alcohol were included in the fitting. The positions of the LPs were determined by manually varying the LP geometry to minimize the rms error between the QM and empirical ESPs and to reproduce the relative interaction energies of alcohols with water for different orientations.

QM calculations of alcohol−water complexes were done by constraining the geometry of the alcohols to the corresponding MP2(fc)/6-31G(d) optimized geometry and water geometry to that of the SWM4-NDP model.[25] The position of the water relative to the alcohol molecule was optimized at the MP2(fc)/6-31G(d) level of theory for the interaction distance, C−O...H_{water} angle, and H(O)...C−O...H_{water} torsion. Single point LMP2/cc-pVQZ[46,47] calculations were performed on the minimum energy geometry to obtain the interaction energy using the program Jaguar (Schrodinger Inc.) as previously described.[48] Water−alcohol orientations used in the present study are shown in Figure 2. Empirical alcohol−water interactions were performed on preliminarily relaxed geometry of alcohols with the water orientation obtained from corresponding QM calculations. Only the interaction distance was optimized in the empirical calculations with other geometrical parameters held fixed at their initial values.

QM calculations of the alcohol complexes with rare gas atoms were performed using the MP2(fc)/6-31G(d) optimized geometry of the corresponding alcohols. The relative position of a rare-gas atom was adopted from alcohol−water interactions with additional positions added to probe the alkyl carbon atoms. Minimum interaction energies and distances for rare gas−alcohol complexes were determined using interaction distance scans with 0.01 Å increments. Energy of the complex was evaluated at the MP3(fc)/6-311++G(3d,3p) level of theory for each point on the scan path.[49] The rare gas atom placement is illustrated in Figure 3.

Empirical force field calculations were performed with the program CHARMM. Energy minimizations of model compounds in the gas phase were performed with the adopted basis Newton−Raphson minimizer (ABNR)[10,50] to a final rms gradient of 10^{-5} kcal/(mol * Å). All gas-phase calculations were performed using infinite cutoffs.

Equilibrium parameters and force constants associated with the bonds, valence, and torsion angles were optimized targeting the mean values of geometric parameters from the survey of crystal structures,[51] QM geometries, and QM vibrational spectra of the model compounds. The QM

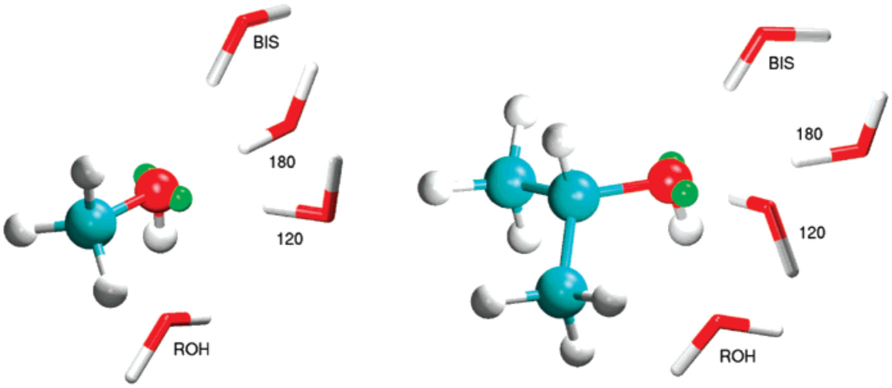


Figure 2. Interaction orientations between water and the alcohols using methanol (left) and 2-propanol (right) as examples. Lone-pair sites are shown in green. BIS-orientation C−O−Hw angle is 115.8, 117.4, and 117.2° for methanol, ethanol, and isopropyl alcohol, respectively; 120-orientation C−O−Hw angle is 104.4, 108.3, and 105.1°, respectively; 120-orientation H(O)−C−O−Hw torsion is 108.0, 131.3, and 118.0°, respectively.

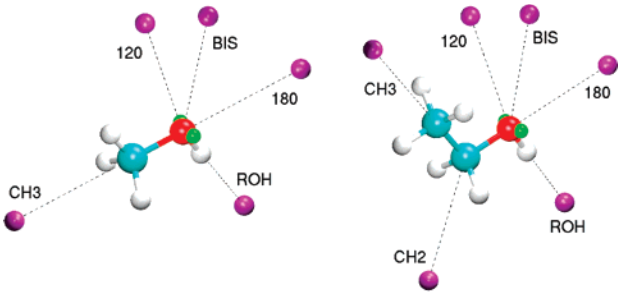


Figure 3. Interaction orientation of the rare gases with the alcohols methanol (left) and ethanol (right). Rare-gas atoms are shown in purple, and lone-pair sites are shown in green.

calculations were performed at the MP2/6-31G(d) level, and a scale factor of 0.9434 was applied to vibrational modes to account for limitations in the level of theory.[52] Second derivatives of energy with respect to atomic coordinates of real atoms were obtained numerically with the position of the Drude particle self-consistently adjusted to every change in coordinates of the real atoms. Potential energy decomposition analysis was performed using the MOLVIB utility[53] in CHARMM. Internal coordinate assignment was done according to Pulay et al.[54]

Target data for optimization of the dihedral parameters were torsion energy profiles obtained from QM calculations. Dihedral angle scans were performed in 10° increments for the torsion angle with subsequent geometry relaxation performed at the MP2(fc)/6-31G(d) level, which was followed by single-point energy evaluation at the MP2/cc-pVTZ[55] level. A similar procedure was repeated in the force field calculations where the torsion angle of interest was scanned, while other geometrical parameters were fully relaxed. Dihedrals were restrained to their target values using harmonic force constants of 10^5 kcal/(mol * Å^2), and energy minimizations were performed with the ABNR method to rms gradients of 10^{-5} kcal/(mol * Å). Empirical torsion parameters were optimized to minimize the difference between the potential energy profile and the MP2/cc-pVTZ data.

Molecular dynamics (MD) simulations of condensed phases were performed at a constant pressure of 1 atm with cubic periodic boundary conditions using the velocity Verlet integrator that includes treatment of Drude particles via an extended Langrangian.[30] The integration time step was 1 fs for both polarizable and additive simulations with temperatures maintained at 298.15 K using the Nose-Hoover thermostat[56] with a relaxation time of 0.1 ps applied to all real atoms. A modified Andersen-Hoover barostat[30,57] with a relaxation time of 0.1 ps was used to maintain the system at constant pressure. The SHAKE algorithm was used to constrain covalent bonds involving hydrogens.[58] Lennard-Jones (LJ) interactions were treated explicitly out to 12 Å with force switch smoothing[59] applied over the range of 10−12 Å. Nonbond pair lists were maintained out to 14 Å, and the long-range correction for LJ interactions[60] was applied in the condensed-phase simulations. Electrostatic interactions were treated using particle mesh Ewald (PME) summation[61] with a coupling parameter 0.34 and sixth-order spline for mesh interpolation. The extended Lagrangian double-thermostat formalism[30] was used in all polarizable MD simulations where a mass of 0.4 amu was transferred from real atoms to the corresponding Drude particles. The amplitude of Drude oscillations was controlled with a separate low-temperature thermostat at 1 K to simulate near-SCF conditions.[30]

Pure solvent MD simulations included 128 alcohol molecules. To obtain convergent results 5 independent MD simulations were run for 250 ps with different initial velocities being assigned to the particles. The first 50 ps of the simulations were treated as equilibration, and the final 200 ps were used for the analysis. Averages were obtained from the 5 independent simulation averages which were also used to calculate the standard errors for the calculated properties. Heats of vaporization, ΔH_{vap}, and molecular volume, V_m, were determined following the standard procedure.[27] Gas-phase simulations were performed using Langevin dynamics in the SCF regimen with infinite cutoffs for nonbonded interactions. The friction coefficient of 5 ps^{-1} was applied to all atoms except for Drude particles. The gas-phase simulations were performed on all 128 individual monomers extracted from the respective equilibrated alcohol box from the condensed-phase MD simulations. The simulations were run for 250 ps for each molecule with the resultant energies obtained from last 200 ps. The gas-phase energy was the average of the averages from the 128 monomer simulations.

Radial distribution functions, isothermal compressibilities, and self-diffusivities were calculated for the alcohols from condensed-phase MD trajectories. Isothermal compressibilities were calculated from

$$\beta_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T = \frac{\langle \partial V^2 \rangle}{V k_B T} \tag{3}$$

according to Klauda et al.,[62] where V is the volume, $\langle V^2 \rangle$ is the volume fluctuation, and k_B is Boltzmann's constant. The slope of the mean squared displacement versus time was used to determine the self-diffusivity for the periodic boundary condition, D_{PBC}. The self-diffusivity was corrected for system-size effects using the hydrodynamic model of Yeh and Hummer[63] of a particle surrounded by a solvent with viscosity, η,

$$D_S = D_{PBC} + \frac{k_B T \xi}{6\pi\eta L} \tag{4}$$

where L is the cubic box length, and $\xi = 2.837297$. The shear viscosities were taken at their experimental values.

Free energies of aqueous solvation (relative to gas phase), ΔG_{sol}, were obtained as a sum of nonpolar, ΔG_{np}, and electrostatic, ΔG_{elec}, contributions via a free energy perturbation (FEP) approach[64,65] according to the step-by-step staged protocol developed by Deng and Roux.[66]

$$\Delta G_{sol} = \Delta G_{np} + \Delta G_{elec} \tag{5}$$

The nonpolar contribution was obtained from the perturbation formula,[64,65] where the free energy change ΔG, corresponding to the change in the potential energy from U_i to U_j, can be calculated as an average over the ensemble of configurations generated with the potential energy U_i:

$$\Delta G = -kT \ln\left\langle \exp\left(-\frac{U_j - U_i}{kT}\right)\right\rangle_{(U_i)} \tag{6}$$

The nonpolar contribution was calculated with all atomic and Drude charges of the solute set to 0. In the protocol the nonpolar term was decomposed into dispersive, U^{dis}, and repulsive contributions, U^{rep}, using the Weeks, Chandler, and Andersen scheme.[67]

$$U^{np}_{uv}(X,Y,\xi) = U^{rep}_{uv} + \xi U^{dis}_{uv}(X,Y) \tag{7}$$

The dispersive contribution was calculated using a linear coupling scheme with the coupling parameter ξ such that the interaction energy $U_{uv}(X,Y)$ between solute u with coordinates X and solvent v with coordinates Y was calculated as ξ was changed from 0 to 1 in increments of 0.1. The repulsive term, due to its sharp r^{12} dependence, cannot be treated accurately via a linear perturbation and instead was transformed into a soft-core potential. It was calculated in multiple stages with a staging parameter s. The staging parameter s was set to 0.0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. The free energy contributions from simulations using different staging parameters were summed. The weighted histogram analysis method (WHAM)[68] was used to obtain the dispersive and repulsive contributions to free energies from the simulations. The electrostatic component of the free energy of hydration was computed by decoupling a molecule of the solute from the solvent by thermodynamic integration (TI)[69−71]

$$\Delta G_{elec} = \int_0^1 d\lambda \left\langle \frac{dU(\lambda)}{d\lambda} \right\rangle \tag{8}$$

where the coupled state (λ=1) corresponds to a simulation where the solute is fully interacting with the solvent, and the uncoupled state (λ=0) corresponds to a simulation where the solute dose not interact with the solvent. In the perturbations λ was changed from 0 to 1 in 0.05 increments with a 0.1 window size and half of the window overlapping with the previous window. Each contributing term to the free energy was obtained as a difference in the free energy of the solute in water and in vacuum.

For the free energy calculations gas-phase simulations were performed using Langevin dynamics with SCF Drudes as described above. Aqueous-phase calculations were performed with the alcohol molecule solvated in a box of 250 SWM4-NDP polarizable water[26] molecules and restrained to the center of mass of the box by a harmonic potential with a force constant of 0.5 kcal/(mol * Å^2) acting on all solute atoms except Drudes. The system was then subjected to a 110 ps NPT simulation at 298.15 K and 1 atm pressure at each value of the coupling/staging parameter. The FEP analysis was performed on the final 100 ps of these dynamics runs. The reported free energy value was averaged over five independent runs each performed with individual seed numbers. Corresponding nonpolarizable simulations were performed using the TIP3 water model.[72] During free energy calculation the molecular dynamics simulations did not include long-range correction (LRC) for dispersion forces. However, the latter were estimated from 50 ps MD simulations of a single alcohol molecule placed in a box of 250 water molecules. The MD calculations were performed using the protocol described for simulations of the condensed phase. The energy due to LRC was calculated for the fixed configuration of the solute−solvent system as a difference of the van der Waals energy contribution calculated using 10 and 30 Å nonbonded cutoffs and averaged for 30 snapshots each written at 1 ps time intervals.

The static dielectric constants ϵ of the neat alcohols were calculated from the total

$$\epsilon = \epsilon_\infty + \frac{4\pi}{3\langle V\rangle k_B T}(\langle M^2\rangle - \langle M\rangle^2) \tag{9}$$

where the dipole moment M denotes the fluctuations of the box,[25,30,73] $<V>$ is the average volume of the box, and ϵ_∞ is the high-frequency dielectric constant. Time series of M were obtained from 5 independent simulations of 5 ns using data from the last 4 ns of each simulation following the previously discussed protocol.[27] The high-frequency contribution ϵ_∞, representing the dielectric constant at the limit of infinitely high frequency of light, is estimated from the Clausius-Mossotti equation

$$\frac{\epsilon_\infty - 1}{\epsilon_\infty + 1} = \frac{4\pi\alpha}{3V_m} \qquad (10)$$

where α is the gas-phase molecular polarizability, and V_m is the molecular volume.

Results and Discussion

Optimization of the alcohol parameters targeted the compounds methanol, ethanol, and 2-propanol (isopropyl alcohol), with greater emphasis placed on ethanol for the primary alcohols. The derived parameters were then validated on 1-propanol, 1-butanol, and 2-butanol to test their transferability. This was followed by additional optimization on the oxygen LJ parameters for 1-propanol and 1-butanol to obtain better agreement with experimental molecular volume and enthalpy of vaporization (see below). The molecules, which include both primary and secondary alcohols, are shown in Figure 1. Hydroxyl groups located on tertiary carbon atoms are not common in biological macromolecules and, therefore, were not considered in this study. For the alcohol series the aliphatic carbon and hydrogen parameters previously determined in our laboratory[27] were applied directly with the exception of the electrostatic parameters on $CH_{(1-3)}$ fragments directly adjacent to the oxygen and of the LJ parameters for the methyl group in methanol. The remaining parameters were optimized as part of the present work.

The parameter optimization was performed using the following protocol. The electrostatic part of the polarizable model was initially derived from fitting to QM unperturbed and perturbed ESP maps with the geometry of the model compounds being fixed at their corresponding QM values. The internal parameters were initially taken from the additive CHARMM22 force field. Next, optimization of the LJ parameters was undertaken based on condensed-phase simulations and rare gas−model compound interactions. After the LJ parameters were initially determined (within 5% of target heat of vaporization and molecular volume) the internal parameters were updated to reproduce target geometries and vibrational spectra. Next the torsion parameters were optimized to reproduce the QM dihedral potential energy surfaces. After the first round of optimization was completed, the atomic charges and polarizabilities were manually adjusted to reproduce QM data on interactions with water as well as dipole moments and condensed-phase properties. The LJ parameters were then reoptimized, and the internal parameters, including the dihedral parameters, were correspondingly updated. These steps were repeated until convergence.

Intramolecular Parameters. Target data for the equilibrium bonded parameters were intramolecular geometries obtained from surveys of the Cambridge Crystallographic Database.[51] Structural data involving hydrogen atoms were obtained from MP2(fc)/6-31G(d) gas-phase optimized geometries. Equilibrium geometries of the model compounds for the final parameters are summarized in Table 2 along with the corresponding target data. The empirical model shows good overall agreement with the target data. The maximum deviation is 0.04 Å for the C−C bond (C−COH) and 0.7° for the C−C−C angle. The C−C bond was not optimized due to adopting the hierarchical approach for parameter development requiring the parameters for the previously defined alkane atom types be preserved. Such constraint is an important precondition for maintaining transferability of the developed parameters, and it also helps reduce the number of parameters to be optimized.

Reproduction of vibrational spectra along with potential energy surfaces for rotation about selected dihedrals was used to optimize the force constants. Presented in Tables S1−S3 of the Supporting Information are the Drude and target QM vibrational spectra for methanol, ethanol, and 2-propanol. Inspection of those results shows the agreement of both the magnitudes of the frequencies and the assignments to be excellent. The largest difference, the CO torsion in the IR-spectrum of methanol, was due to final adjustment of the associated parameter based on the energy surface, as follows. Final optimization of the dihedral parameters was based on the reproduction of QM energy surfaces. Shown in Figure 4 are the surfaces for the three molecules. It is evident that the Drude model satisfactorily reproduces the QM surfaces and is significantly better than CHARMM22, which was originally optimized targeting lower level QM data. The level of agreement of the Drude model for both the vibrational and dihedral surfaces indicates that the alcohols will sample the correct intramolecular conformations during MD simulations.

Electrostatic Model. Atoms in classical molecular mechanics are traditionally treated as point charges. This leads to a certain degree of arbitrariness in the derivation of partial atomic charges from electrostatic fitting that cannot be eliminated due to the inherent ambiguities in partitioning the electron distribution with respect to a set of atomic centers. Therefore the point charge fitting to a QM electrostatic potential is often conducted under restraints enforcing that the derived charges follow chemical intuition.[74] The CHARMM additive force field[11] was developed according to a slightly different methodology. Because of the aforementioned limitations in the fitting of atomic charges to electron distributions, plus additional uncertainty as to how well gas-phase fitted charges would work in the condensed-phase, the additive CHARMM force field treated the atomic charges as adjustable parameters. These terms were then optimized to reproduce energetics and geometries of test molecules interacting with water along with condensed-phase properties. This approach was successfully validated in the development of CHARMM22 and CHARMM27 additive force fields.[11,75] In the classical Drude polarizable model the number of electrostatic parameters has increased due to the inclusion of atomic polarizabilities. To address this additional complexity the Drude electrostatic parameter determination protocol in CHARMM has been extended to include ESP fitting.[36] However, as the final goal of the resultant force field is the reproduction of condensed-phase properties as well as atomic details of gas-phase interactions with water, manual corrections of the ESP fitted values are considered acceptable.

An additional factor that has to be taken into account during the optimization of electrostatic parameters is the placement of virtual charged particles representative of

Table 2. Equilibrium Geometries of Alcohols in the Trans Conformations[a]

	methanol		ethanol		isopropyl alcohol		
	Drude	target	Drude	target	Drude	target	RMSD
C−O[b]	1.43	1.40 (0.05)	1.43	1.42 (0.04)	1.43	1.43 (0.03)	0.02
C−C[c]			1.53	1.48 (0.05)	1.53	1.50 (0.03)	0.04
O−H[d]	0.97	0.97	0.97	0.97	0.97	0.97	0.00
C−H[d]	1.11	1.09	1.11	1.10	1.12	1.09	0.02
C−C−O[b]			111.4	111.5 (3.2)	110.4	109.7 (2.8)	0.5
C−C−C[b]					113.2	112.5 (2.6)	0.7
C−O−H[c]	107.6	107.4	107.5	107.6	106.7	106.6	0.1
H−C−O[c]	109.9	110.3	110.7	110.9	103.6	103.8	0.3
H−C−H[c,d]	109.1	108.6	107.3	107.6			0.4

[a] Bond lengths, Å; valence angles, deg; data are shown for the carbon atom adjacent to oxygen. [b] Cambridge crystallographic survey based on 2037 hits for methanol, 913 hits for ethanol, and 130 hits for isopropyl alcohol; CSD version 5.27 (Nov 2005); standard deviations are shown in parentheses. [c] MP2(fc)/6-31G(d) gas-phase optimized geometry. [d] Not optimized; corresponding parameters were directly transferred from the polarizable alkanes.

oxygen lone-pairs (LP). Inclusion of LPs was motivated by the inability of force field models without them to accurately reproduce the angular dependence of ESP maps in the case of hydrogen bond acceptors.[33,76] The charge on LP sites and their geometry were initially determined from the ESP fitting procedure to reduce the rms error during fitting. The resulting fitted charges, polarizabilities, and LP geometry were then tested on interactions with water, and LP positions were further adjusted to better reproduce the local anisotropy of interactions with water. Because calculation of the interactions with water requires LJ parameters on the hydroxyl the CHARMM22 values were used as an initial guess. In later stages of the optimization the LP position and atomic charges were re-evaluated based on interactions with water each time a new set of oxygen LJ parameters became available from the condensed-phase optimization (see below). This elaborate optimization procedure was performed for ethanol only with the derived LP positions applied without change to the other alcohols. The final oxygen lone-pair positions were as follows: distance between the oxygen atom center and the lone-pair center = 0.35 Å; C−O−LP angle = 110°; virtual torsion angle H(O)−C−O−LP = ±91°.

Placement of water molecules around the hydroxyl group in the primary and secondary alcohols is illustrated in Figure 2 and the interaction energies and distances are presented in Tables 3 and 4, respectively, for both the additive and polarizable models. Additive CHARMM22 force field results are presented as a representative example of those for an additive alcohol force field. Examination of the CHARMM22 energy differences, ΔE, show them to be significantly more favorable than the QM target data and the balance among the interaction orientations to be poor with the differences ranging from −0.6 to −2.2 kcal/mol for ethanol (Table 3). The more favorable interaction energies are expected given the need to overpolarize the effective fixed charges in the additive model to account for the lack of explicit polarizability. However, the limitation in the balance of the interactions for different placements of water molecule around the test molecule is due to limitations in the anisotropy of the electrostatic representation. This problem was also seen in the Drude model without LPs.[33] Therefore the Drude model of alcohols was extended to include LPs on the oxygen atom, yielding a more anisotropic electrostatic model. This anisotropy was further extended by assigning anisotropic polarizability on the oxygen atom. This anisotropic polarizability was introduced by treating the Drude force constant (k_D) of the oxygen as a tensor, as previously described.[33] The X-axis of the tensor is defined along the C−O bond; the Y-axis goes through the oxygen atom perpendicular to the plane created by the C−O−H atoms; the Z-axis is orthogonal to the X- and Y-axes. Increased stiffness of the Drude constant along the C−O bond (X-axis) reduces the overestimation of alcohol−water interaction in the 180-orientation (e.g., 180 in Table 3 for the C22 models). In addition, reducing the force constant along the Y-axis effectively increases oxygen polarizability along the lone-pair directions (e.g., 120 in Table 3 and Figure 2). More details of the impact of the inclusion of LPs and anisotropy on methanol and other molecules with hydrogen bond acceptors is presented in ref 33. Thus, by increasing the force constant along the X-direction and decreasing it along the Y-axis leads to improvements in the balance of interactions of the hydroxyl with water. The final values of the tensor are $k_{Dxx} = 600$, $k_{Dyy} = 400$, and $k_{Dzz} = 500$ kcal/(mol * Å^2) with the same anisotropic Drude constants used for all the alcohols. In addition, the use of anisotropic polarizability leads to a more accurate representation of the polarization response around the hydroxyl group.[33]

As discussed above the additive model systematically overestimates the water−alcohol interactions energies (i.e., too favorable), and there is an imbalance in the treatment of the energies as a function of orientation. These effects are indicated by the large negative ΔE_{C22} values in Table 3 for the former and the greater values of the RRMS$_{C22}$ values for the latter in comparison with the polarizable model. In the Drude model both of these problems are largely alleviated. There is still a tendency for the interaction energies to be slightly too favorable, though the magnitude is generally much less than with the additive model. While such a problem may have contributions from the level of theory used in the ab initio calculations, the need to overestimate the interaction energies was necessary to obtain the correct pure solvent properties (see below). It should be noted that the data in Tables 3 and 4 represent the use of off-diagonal LJ terms for the O$_{hydroxyl}$−O$_{water}$ interactions, as discussed in the following section. Ideally, the need to overestimate the

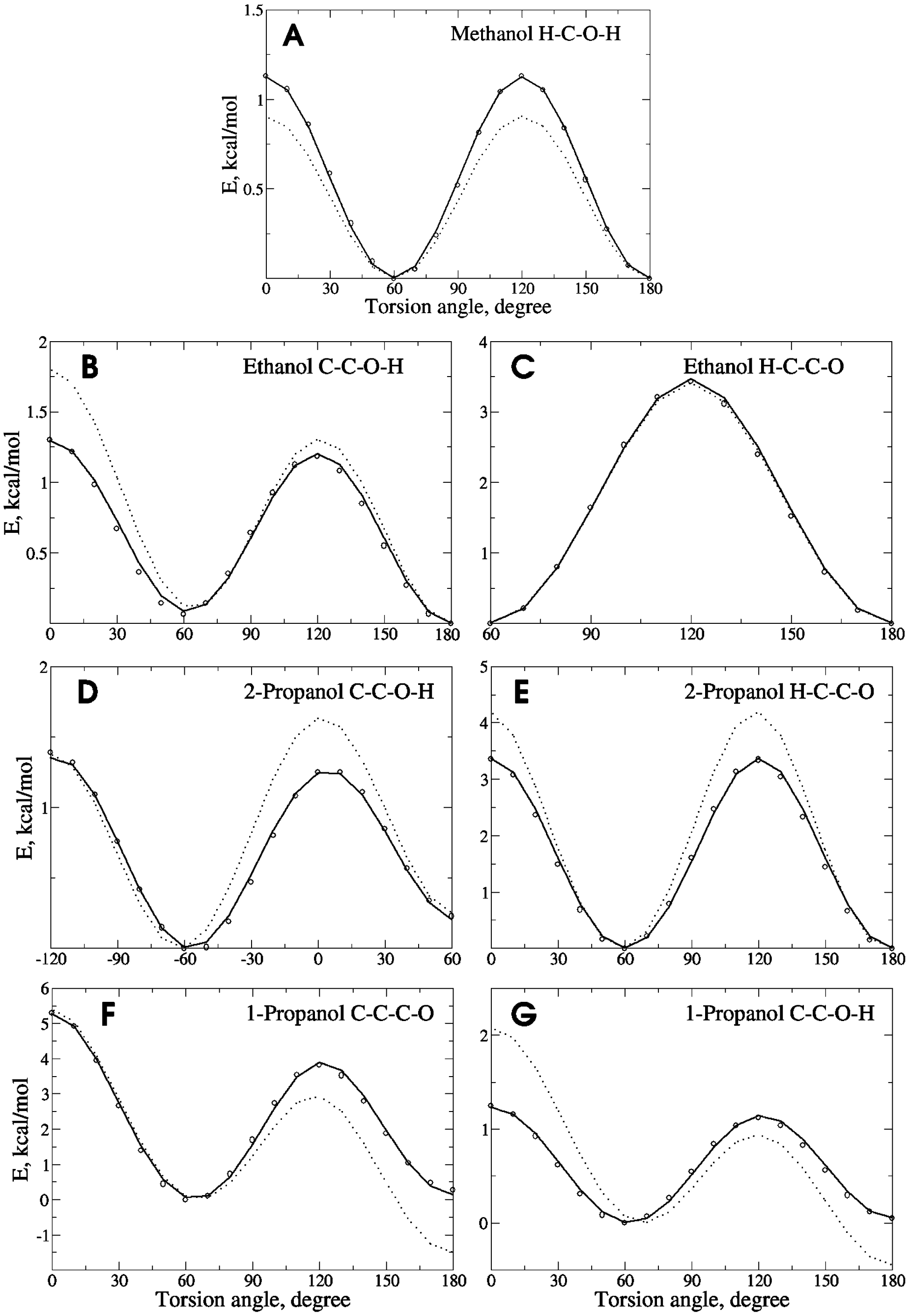


Figure 4. Potential energy surfaces for rotation of selected dihedrals in methanol (A), ethanol (B and C), 1-propanol (D and E), and 2-propanol (F and G). Data are included from the QM (circles), Drude (solid line), and CHARMM22 (dotted line) models.

gas-phase interactions with water is not required for a polarizable model; future investigations will address this result. With respect to the balance of the interactions, the Drude model behavior is satisfactory, with the tendency to overestimate the 180-orientation significantly decreased with respect to the additive model. Thus, the polarizable alcohol model that incorporates oxygen atom charge anisotropy (due to LPs) and oxygen polarization anisotropy more accurately reproduces the change in interaction energy with water as a function of orientation as compared to the additive model.

Due to parameter correlation, the optimization of the electrostatic terms has to be repeated whenever the LJ parameters are changed, because new LJ parameters alter the interactions with water. Therefore, the LJ optimization from condensed-phase simulations and electrostatic model optimization steps are repeated until convergence as judged by the agreement with the target condensed-phase data typically being 2% or less. This iterative optimization procedure leads to maximizing agreement with the condensed-phase properties, while gas-phase interactions are sacrificed to some extent. However, inclusion of the water interaction data assures that the model satisfactorily describes atomic details of hydrogen bonding as discussed above, an attribute that is anticipated to have paramount influence on the utility of the model in biomolecular simulations.

As mentioned above, small corrections to the ESP fitted charges were necessary during the parameter optimization. The manual adjustment of charges addresses two issues. First,

Table 3. Alcohol−Water Minimum Interaction Energy Differences Relative to QM Data[d]

	E_{QM}[a]	ΔE_{C22}[b]	ΔE_{Drude}[b]	ΔE_{C22}, %	ΔE_{Drude}, %	$RRMS_{C22}$[c]	$RRMS_{Drude}$[c]
MeOH						0.44	0.19
BIS	−4.40	−1.83	−0.34	42	8		
180	−2.09	−2.29	−0.71	110	34		
120	−4.90	−1.12	−0.22	23	4		
ROH	−4.12	−2.09	−0.33	51	8		
EtOH						0.63	0.11
BIS	−4.82	−1.72	−0.70	36	15		
180	−2.33	−2.23	−0.50	96	21		
120	−4.80	−0.61	−0.52	13	11		
ROH	−4.24	−2.07	−0.39	49	9		
1-PrOH						0.69	0.12
BIS	−5.01	−1.38	−0.29	28	6		
180	−2.64	−1.86	−0.13	70	5		
120	−5.09	−0.21	−0.01	4	0		
ROH	−4.40	−1.95	−0.31	44	7		
2-PrOH						0.93	0.21
BIS	−4.68	−1.59	0.25	34	−5		
180	−2.46	−2.31	−0.24	94	10		
120	−4.76	0.16	0.13	−3	−3		
ROH	−3.99	−1.83	0.28	46	−7		

[a] QM calculations are performed at the LMP2/cc-pvQZ//MP2/6-31G* level of theory. [b] $\Delta E^{i}_{model} = E^{i}_{int}(model) - E^{i}_{int}(QM)$, where $E^{i}_{int}(model)$ is the interaction energy corresponding to the CHARMM22 or Drude models for the ith orientation. [c] Relative rms error calculated for the difference $\Delta E^{i}_{alcohol} - \Delta E^{av}_{alcohol}$, where $\Delta E^{av}_{alcohol}$ is the average difference between model and QM calculations for a given alcohol molecule. [d] Energies in kcal/mol. See Figure 2 for interaction orientations. Results for the polarizable model include off-diagonal (i.e., NBFIX) $O_{alcohol}$...O_{water} LJ parameters.

Table 4. Alcohol−Water Minimum Interaction Distance Differences Relative to QM Data[d]

	R_{QM}[a]	ΔR_{C22}[b]	ΔR_{Drude}[b]	ΔR_{C22}, %	ΔR_{Drude}, %	$RRMS_{C22}$[c]	$RRMS_{Drude}$[c]
MeOH						0.05	0.05
BIS	1.98	−0.14	−0.08	−7	−4		
180	2.12	−0.23	−0.14	−11	−7		
120	1.95	−0.11	−0.09	−6	−5		
ROH	1.95	−0.13	−0.01	−7	−1		
EtOH						0.06	0.04
BIS	1.98	−0.12	−0.09	−6	−5		
180	2.12	−0.24	−0.13	−11	−6		
120	1.97	−0.08	−0.10	−4	−5		
ROH	1.95	−0.13	−0.02	−7	−1		
1-PrOH						0.06	0.04
BIS	1.98	−0.12	−0.09	−6	−5		
180	2.12	−0.24	−0.13	−11	−6		
120	1.97	−0.08	−0.09	−4	−5		
ROH	1.95	−0.13	−0.02	−7	−1		
2-PrOH						0.07	0.04
BIS	1.97	−0.13	0.00	−7	0		
180	2.10	−0.23	−0.04	−11	−2		
120	1.96	−0.05	−0.01	−3	−1		
ROH	1.97	−0.09	0.06	−5	3		

[a] QM calculations are performed at the LMP2/cc-pvQZ//MP2/6-31G* level of theory. [b] $\Delta R^{i}_{model} = R^{i}_{min}(model) - R^{i}_{min}(QM)$, where $R^{i}_{min}(model)$ is the minimum energy distance corresponding to the CHARMM22 or Drude models for the ith orientation. [c] Relative RMS error calculated for the difference $\Delta R^{i}_{alcohol} - \Delta R^{av}_{alcohol}$, where $\Delta R^{av}_{alcohol}$ is the average difference between model and QM calculations for a given alcohol molecule. [d] Distance in Å. See Figure 2 for interaction orientations. Results for the polarizable model include off-diagonal (i.e., NBFIX) $O_{alcohol}$...O_{water} LJ parameters.

it circumvents the present limitation that the ESP fitting procedure be applied to only one test molecule at a time (although simultaneous fitting of multiple conformations is performed in this work). Ideally, for a series of compounds for which transferable parameters are desired, they all should be fit simultaneously, including the application of equality restraints on those groups of atoms that should be transferable (e.g., the hydroxyl O and H atoms for the primary alcohols). The second issue is related to the requirement that the derived transferable charges and polarizabilities should ultimately reproduce as correctly as possible the interaction energies of the model compounds with water. Conceivably, one could implement an extended global optimization procedure simultaneously including the ESP for a set of molecules in

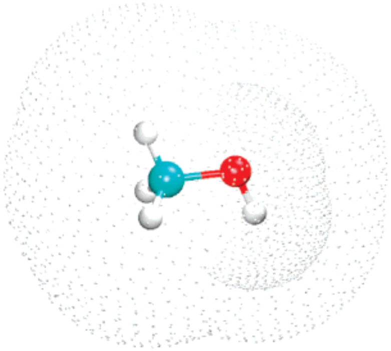


Figure 5. Image of the additional Connolly surfaces around methanol at vdW scale factors of 1.3 and 2.2. See Table 1 for details of all grid surfaces and perturbation ion positions used in the electrostatic parameter fitting.

multiple conformations as well as their interactions with water molecules in specific hydrogen binding configurations to achieve such goal. However, it is not trivial to perform such global optimization with all the different target data weighted appropriately. For the sake of simplicity, in the present work manual adjustment of the electrostatic parameters was performed, addressing the two issues in two steps. To evaluate the robustness of this approach as well as the transferability of the parameters additional calculations were performed on alcohols not included in the training set.

Another modification to the ESP fitting procedure[36] is the addition of several new Connolly surface layers for placement of perturbation ions and grid points. This was motivated by the inability of the published procedure to yield an electrostatic model that reproduced condensed-phase properties of alcohols. Application of the original approach produced electrostatic models that significantly (about 2 kcal/mol) underestimated the enthalpy of vaporization of ethanol (experimental value 10.11 kcal/mol), such that LJ parameter optimization could not correct the deficiency of the derived charge model. Comparison of the hydroxyl charges for the additive CHARMM22 (q_O=−0.66; q_H=0.43) with those derived using the original[36] grid (q_O=−0.436; q_H=0.312) indicates substantial underestimation of polarity of the hydroxyl group, leading to underestimation of the electrostatic contribution to the enthalpy of vaporization. To correct for this deficiency the fitting protocol was extended to include an additional "near" grid and perturbation ions in the vicinity of polar atoms thereby increasing the contribution of polar atoms in ESP fitting. The location of this grid is shown in Figure 5. The ethanol charges derived from the extended grid fitting procedure (q_O=−0.462; q_H=0.355) show an increased local dipole of the hydroxyl group and indeed allowed identification of LJ parameters that yielded satisfactory condensed-phase properties. Despite the improvement brought by the near grid the fitting procedure gave different charges for methanol, ethanol, and 2-propanol (Table 5), whereas to enforce the parameter transferability a single set of charges was required. Therefore, the fitted charges and polarizabilities were subjected to manual adjustment, leading to the final optimized charges for alcohols of $q_O = -0.46$ (q_{LP}=−0.23); $q_H = 0.36$, which are quite close to the ESP fitted values. Basically, the ESP fitting provided a good initial guess for the electrostatic model although empirical adjustment was required to derive the final fully balanced electrostatic model.

Table 5. ESP Fitted Atomic Charges (q) and Polarizabilities (α) of the Model Compounds

	methanol		ethanol[a]		2-propanol[a]	
atom	q	α	q	α	q	α
O	0.000	0.90	0.000	0.44	0.000	0.99
lone-pair	−0.242	0.00	−0.231	0.00	−0.223	0.0
H (O)	0.352	0.00	0.355	0.00	0.365	0.0
C (O)	−0.006	1.92	−0.021	1.45	−0.003	1.08
H (CO)	0.046	0.00	0.064	0.00	0.084	0.00
C_{alk} (CH_2)	n/a[b]	n/a[b]	n/a[b]	n/a[b]	n/a[b]	n/a[b]
H_{alk} (CH_2)	n/a[b]	n/a[b]	n/a[b]	n/a[b]	n/a[b]	n/a[b]
C_{alk} (CH_3)	n/a[b]	n/a[b]	−0.18	2.05	−0.18	2.05
H_{alk} (CH_3)	n/a[b]	n/a[b]	0.06	0.00	0.06	0.00

[a] Gauche and trans conformations were fitted simultaneously. [b] n/a indicates not applicable.

The final electrostatic parameters are presented in Table 6. Two additional adjustments were made to the electrostatic model prior to finalization. As discussed previously,[25,26] it appears to be necessary to empirically scale gas-phase polarizabilities by a factor smaller than 1.0 to yield accurate properties of polar molecules for the condensed phase. This was necessary for the SWM4-NDP water model and is applied to account for increased Pauli exclusion that occurs in the condensed phase due to surrounding molecules in the environment over that in the gas phase. Further support for the reduced polarizabilities are studies on macromolecules in the condensed phase, where unscaled values can lead to polarization catastrophe.[25,77−79] Accordingly, the polarizabilities of aliphatic moieties in the alcohols were scaled by 0.7 consistent with the scaling applied to the SWM4-NDP water model. This yielded a polarizability of 1.4 for the CH_3 group (alkane value 2.05) and 1.2 for the CH_2 group (alkane value 1.66). A sp3 carbon atom is considered an alkane type if it is not covalently bound to a heteroatom. The electronic properties of carbon atoms directly connected to heteroatoms are influenced by the electronegative character of such atoms and, therefore, are included in ESP fitting. Therefore, the polarizability of the carbon atom connected to oxygen was taken as the average from the ESP fitted values for methanol, ethanol, and 2-propanol and scaled by 0.7 giving the final polarizability value of 1.0. The oxygen atom was treated differently. Here the polarizability was set to the Miller value of 1.0[36] and not scaled as the unscaled value was required to reproduce the dipole moments, the interactions with water, and the condensed-phase properties all in the context of enforcing transferability across the alcohols studied. Importantly, tests indicated that the oxygen polarizability was not causing electrostatic collapses in either the pure solvent or aqueous environment MD simulations.

Dipole moments for the final electrostatic models are shown in Table 7 along with CHARMM22 and target experimental and QM values. The fixed-charge additive CHARMM22 values are approximately 44% larger than the target values, which is necessary for the additive model to reproduce condensed-phase properties. The agreement with

Table 6. Final Atomic Charges (*q*) and Polarizabilities (α) of the Model Compounds

	methanol		ethanol		2-propanol		1-propanol		2-butanol	
atom	*q*	α	*q*	α	*q*	α	*q*	α	*q*	α
O	0.00	1.0	0.00	1.0	0.00	1.0	0.00	1.0	0.00	1.0
lone-pair	−0.23	0.0	−0.23	0.0	−0.23	0.0	−0.23	0.0	−0.23	0.0
H (O)	0.36	0.0	0.36	0.0	0.36	0.0	0.36	0.0	0.36	0.0
C (O)	−0.14	1.0	−0.06	1.0	0.00	1.0	−0.06	1.0	0.00	1.0
H (CO)	0.08	0.0	0.08	0.0	0.10	0.0	0.08	0.0	0.10	0.0
C_{alk} (CH_2)	n/a[a]	n/a[a]	n/a[a]	n/a[a]	n/a[a]	n/a[a]	−0.12	1.2	−0.12	1.2
H_{alk} (CH_2)	n/a[a]	n/a[a]	n/a[a]	n/a[a]	n/a[a]	n/a[a]	0.06	0.0	0.06	0.0
C_{alk} (CH_3)	n/a[a]	n/a[a]	−0.18	1.4	−0.18	1.4	−0.18	1.4	−0.18	1.4
H_{alk} (CH_3)	n/a[a]	n/a[a]	0.06	0.0	0.06	0.0	0.06	0.0	0.06	0.0

[a] n/a indicates not applicable.

Table 7. Dipole Moments for Alcohols in the Trans Conformation[a]

alcohol	μ (C22)	μ (Drude)	μ (exp)	μ (QM)	$\Delta\mu$ (C22), %	$\Delta\mu$ (Drude), %
MeOH	2.38	1.83	1.70	1.72	40	8
EtOH	2.36	1.81	1.71	1.63	38	6
2-PrOH	2.43	1.87	1.58	1.73	54	18
2-BuOH	2.42	1.79	...	1.76	38	2
1-PrOH	2.35	1.82	1.55	1.54	52	17
1-BuOH	2.36	1.80	1.66	1.60	42	8
average	n/a[b]	n/a[b]	n/a[b]	n/a[b]	44	10

[a] Units in Debye, QM dipole moments at the MP2(fc)/aug-cc-pVQZ// MP2(fc)/6-31G(d) level, and percent differences with respect to the experimental data. Experimental data are from ref 87. [b] n/a indicates not applicable.

experimental gas-phase dipole moments is improved in the polarizable model; however, the values are still overestimated by 10%. This is consistent with the polarizable model yielding more favorable interactions energies with water (Table 3) as required to reproduce the condensed-phase properties throughout the alcohol series. A similar approach was utilized by Gao[18] in development of the PIPF polarizable model for alcohols by treating the dipole moment as an adjustable parameter to improve agreement with condensed-phase properties and is necessary to obtain the targeted condensed-phase properties as discussed below.

Final atomic charges and polarizabilities, presented in Table 6, represent a balance required to reproduce target data of all alcohols throughout the series with primary importance in reproducing condensed-phase properties. Grid scans (Tables S8−S10, Supporting Information) of point charges and atomic polarizability parameters performed in the vicinity of the optimized parameter values confirm their optimal choice. Attempts to improve the agreement with the target data through variations of the charges do show improvement for individual properties, but the overall agreement with the multiple target properties considered becomes poorer. For example, decreasing the value of point charges followed by a corresponding increase in atomic polarizabilities would improve agreement with gas-phase dipole moment, but it would also negatively impact the already too favorable dielectric susceptibility of 1-butanol. Therefore, the current electrostatic parameters represent a balance for the entire alcohol series within the limitations of a transferable parameter set for the hydroxyl group. Such a constraint is a necessary prerequisite for subsequent transfer of the developed parameters to corresponding fragments of biological macromolecules.

Lennard-Jones Parameters. Optimization of LJ parameters represents the most difficult aspect of empirical force field development as this term impacts the strength of ionic and hydrogen bond interactions as well as dispersion types of interactions. Adjustment of the LJ parameters was performed to reproduce experimental molecular volumes and enthalpies of vaporization of the neat alcohols, with the relative values of the LJ parameters checked via interactions with rare gases, as previously performed.[49] All alkyl groups were constrained to the previously determined alkane LJ values,[27] and the polar hydrogen LJ parameters were constrained to a well depth of 0.01 kcal/mol and radius, $R_{min}/2$, of 0.4 Å. Such LJ parameters on the polar H introduce a repulsive potential on the hydrogen atom to diminish the possibility of overpolarization during hydrogen or ionic bonding interactions. Based on these assumptions only the oxygen LJ parameters were optimized subject to the constraint that the same oxygen LJ parameters were to be utilized throughout the alcohol series, with the only exception being the LJ parameters of the methyl group in methanol. This strategy was selected to facilitate parameter transferability to a biomacromolecular force field and to limit the overall number of atom types in the force field. However, such a limitation will diminish the quality of the fit for the individual alcohols, though the compromise to be made is moderate (see below). Final optimized values of the oxygen well depth and $R_{min}/2$ were 0.15 kcal/mol and 1.765 Å, respectively. The derived oxygen LJ parameters are relatively close to the polarizable SWM4-NDP water model LJ parameters where the well depth is 0.21 kcal/mol and $R_{min}/2$ is 1.79 Å. Additionally well depth parameters on the methyl carbon and hydrogen in methanol were optimized to obtain better agreement with the experimental molecular volume, which was otherwise too large. The LJ radii on the methyl atoms were constrained to the alkane values,[27] as condensed-phase properties of methanol were considerably less sensitive to changes in the LJ radii than in the well depths, ϵ. Final values of well depth for methanol were $\epsilon(C) = 0.11$ and $\epsilon(H) = 0.035$ kcal/mol. The Lennard-Jones parameters on other aliphatic C,H atoms were preserved at their alkane values.[27]

One of the common assumptions in empirical force fields is the use of combining or mixing rules to convert LJ

Table 8. Pure Solvent Properties of Neat Alcohols[d]

property	MeOH[a]	EtOH	2-PrOH	2-BuOH	1-PrOH	1-BuOH
V_m(C22)	69.18	99.09	128.41	157.25	128.98	157.37
	(0.34)	(0.31)	(0.39)	(0.56)	(0.40)	(0.42)
	2.90%	2.20%	0.50%	3.00%	3.40%	3.50%
V_m(Drude) final[b]	67.21	97.11	125.79	153.36	125.78	153.09
	(0.15)	(0.19)	(0.30)	(0.32)	(0.20)	(0.49)
	0.20%	−1.60%	0.50%	0.80%	0.70%	0.00%
V_m(Drude) alternative[c]	66.02	95.42	124.11	152.13	124.06	151.87
	(0.17)	(0.28)	(0.46)	(0.41)	(0.66)	(0.59)
	−1.5%	−2.9%	−0.3%	−0.6%	−0.1%	−1.8%
V_m(exp)	67.23	96.92	127.79	152.65	124.78	152.05
ΔH_{vap}(C22)	9.11	10.20	10.81	10.63	11.22	12.42
	(0.04)	(0.05)	(0.06)	(0.33)	(0.29)	(0.27)
	1.80%	0.90%	−0.30%	−10.50%	−1.10%	−0.70%
ΔH_{vap}(Drude) final[b]	8.94	10.07	10.99	11.76	10.55	11.68
	(0.04)	(0.06)	(0.07)	(0.18)	(0.13)	(0.19)
	−0.10%	−0.40%	1.20%	−1.00%	−7.00%	−6.60%
ΔH_{vap}(Drude) alternative[c]	9.58	10.86	11.83	12.51	11.36	12.49
	(0.04)	(0.05)	(0.08)	(0.19)	(0.13)	(0.19)
	+7.0%	+7.4%	+9.0%	+5.3%	+0.2%	−0.1%
ΔH_{vap}(exp)	8.95	10.11	10.85	11.88	11.34	12.51

[a] Methanol LJ well-depth: carbon $\epsilon = 0.11$, hydrogen $\epsilon = 0.035$. [b] Final polarizable model oxygen LJ $\epsilon = 0.15$, $R_{min}/2 = 1.765$. [c] Alternative polarizable model oxygen LJ $\epsilon = 0.15$, $R_{min}/2 = 1.74$. [d] Heats of vaporization in kcal/mol and molecular volumes in Å^3; values in parentheses are the standard deviations; percent differences are with respect to experiment. Experimental data are from ref 87.

parameters for individual atom types to those for atom pairs.[80−82] This assumption is largely for convenience avoiding the need to individually determine LJ terms for each atom type pair. However, in the present work it was observed that the LJ parameters of the hydroxyl oxygen that yielded good pure solvent condensed-phase properties lead to both the interaction energies with individual water molecules as well as the free energies of aqueous solvation being too favorable. This motivated the use of a specific, or off-diagonal (ie. NBFIX), LJ term for the $O_{alcohol}$...O_{water} atom pairs, with individual terms for the primary and secondary hydroxyl oxygens. As shown above in Tables 3 and 4 and presented below, this leads to good agreement for the aforementioned properties and will be part of the present alcohol force field.

Computed and experimental properties for the neat liquids are summarized in Table 8. Data were obtained for the training set molecules, methanol, ethanol, and 2-propanol as well as for the test molecule 2-butanol, 1-propanol, and 1-butanol. Overall, the level of agreement of the polarizable model is quite good, especially with methanol, ethanol, 2-propanol, and 2-butanol. For all the compounds the molecular volumes are generally improved over CHARMM22, the exception being 2-propanol. Concerning the heats of vaporization both the polarizable and additive models are good for methanol, ethanol, and 2-propanol, with CHARMM22 significantly underestimating the value for 2-butanol, while the polarizable model significantly underestimates the heats for 1-propanol and 1-butanol. Thus, it appears that neither of the models is capable of accurately reproducing pure solvent properties for the full series of primary and secondary alcohols. While this may be associated with constraints on the number of parameters optimized to ensure transferability, similar problems have been observed in polarizable alcohol force fields based on both induced dipole and fluctuating charge models.[18,21,22] In the induced dipole model, the heats of vaporization were typically in good agreement with experiment, while molecular volumes were overestimated for the smaller alcohols and underestimated for the larger compounds.[18] With the fluctuating charge force field distinct parameters were used for methanol[21] and ethanol[22] to obtain agreement with experiment. Thus, despite the inclusion of polarizability, it appears that the current form of the energy function, combined with limitations associated with the need to develop transferable parameters, is not capable of accurately treating the full series of primary and secondary alcohols.

A natural extension of the present parametrization will be to model longer aliphatic-chain alcohols. To this end, the development of LJ parameters for the oxygen targeting the 1-propanol and 1-butanol pure solvent properties was undertaken. This model only differed in those LJ parameters; the remainder of the force field was maintained. Results in Table 8 (the alternative LJ model) show the second LJ model to yield good agreement for 1-propanol and 1-butanol. This second model, with a well depth = 0.15 kcal/mol and $R_{min}/2 = 1.74$ Å on the alcohol oxygen, is recommended for use on long-chain primary alcohols. Also included in Table 8 are results using that alternative LJ model for ethanol, 2-propanol, and 2-butanol, for comparison with the original model. Basically, the utility of the alternative LJ set for hydroxyl oxygen is limited to long-chain primary alcohols only.

To minimize the impact of parameter correlation during the optimization of LJ parameters, interactions with rare gases were monitored to facilitate optimization of the relative values of the LJ parameters. The target data from the rare gas interactions were the rms fluctuations about the average difference (or ratios) for the minimum interaction energies and distances. Use of this target data allows for the LJ parameters to produce interactions that are systematically

Table 9. Differences between Empirical and QM Values and RMS Fluctuations about the Average Differences and Ratios for the Rare Gas Interactions with Methanol and Ethanol[a]

	orientation							
	MeOH+He		MeOH+Ne		EtOH+He		EtOH+Ne	
CHARMM22	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}
BIS	−0.12	−0.01	0.03	0.15	−0.18	−0.03	0.02	0.26
180	0.03	0.02	0.14	0.23	0.00	0.02	0.12	0.25
120	−0.32	−0.03	−0.08	0.16	−0.36	−0.06	−0.10	0.17
ROH	−0.33	0.01	−0.16	0.37	−0.36	0.01	−0.15	0.38
CH3	0.19	0.06	0.20	0.34	0.00	0.02	0.10	0.37
CH2	n/a[b]	n/a[b]	n/a[b]	n/a[b]	−0.15	−0.01	0.03	0.28
RMS	0.20	0.03	0.15	0.10	0.15	0.03	0.10	0.07

	orientation							
	MeOH+He		MeOH+Ne		EtOH+He		EtOH+Ne	
Drude	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}	ΔR_{min}	ΔE_{int}
BIS	−0.14	−0.02	−0.03	0.11	−0.20	−0.02	−0.04	0.23
180	0.02	0.02	0.10	0.22	0.00	0.02	0.10	0.24
120	−0.33	−0.04	−0.13	0.12	−0.39	−0.05	−0.16	0.15
ROH	−0.35	0.01	−0.28	0.28	−0.37	0.01	−0.28	0.31
CH3	0.16	0.03	0.17	0.29	−0.02	0.01	0.08	0.37
CH2	n/a[b]	n/a[b]	n/a[b]	n/a[b]	−0.26	−0.03	−0.08	0.23
RMS	0.20	0.03	0.16	0.08	0.15	0.03	0.13	0.07

[a] Interaction energy differences (kcal/mol) and distance differences (Å) are calculated as $X^{model}-X^{QM}$. [b] n/a indicates not applicable.

offset from the QM data to indicate that the relative LJ parameters for different atom types are properly balanced while accounting for limitations in QM methods to treat dispersion interactions.[49] Shown in Table 9 are the rms fluctuations for CHARMM22 and the final polarizable model for methanol and ethanol, with additional details supplied in Table S4 of the Supporting Information. The rms values for the two models are similar, indicating that the polarizable model did not improve the balance of the LJ parameters using interactions with rare gases as a metric. Analysis of the individual minimum interaction distance differences shows those associated with direct interactions with the lone-pairs (120 interaction) and hydroxyl hydrogen (ROH interaction, Figure 3) to be significantly shorter for the empirical models as compared to the target QM data. Such difference suggests that the empirical models may benefit from assignment of LJ parameters on lone-pair sites and the hydroxyl hydrogens. However, in the present model LJ parameters are not applied to the lone-pairs for simplicity, and a standard LJ parameter is applied to the polar hydrogen to facilitate transferability of the model.

Additional Condensed-Phase Properties of the Alcohols. Additional properties of the pure solvents investigated include the dielectric constants, diffusion constants, isothermal compressibilities, and structural features based on radial distribution functions. In addition, free energies of solvation of the final models were evaluated.

An important feature of a model is proper treatment of the dielectric constant as this term is important in dictating the energetics of dissolution of solutes in the alcohols. Dielectric constants for both the Drude and additive models as well as the experimental data are presented in Table 10. The additive model systematically underestimates the dielectric constant on an average percent difference of −35.9%, an inherent limitation of the model due to the lack of explicit polarizability, as previously described for alkanes.[27] With the Drude model, larger dielectric constants are obtained, leading to systematically better agreement with experiment with an average percent difference of −2.3%. As the internal parameters in the additive and polarizable models are similar the improvement in description of dielectric constant is clearly due to the explicit description of electronic polarization. Small underestimations are observed for ethanol and 2-propanol, the results for methanol, 1-propanol, and 2-butanol are quite good, whereas 1-butanol overestimates the experimental data. Computations of dielectric constant show that this is a very slow converging property. In the current study five independent condensed-phase runs of 5 ns were performed for each alcohol molecule yield a total of 25 ns of simulation time. Longer simulations may be beneficial to obtain more reliable estimates; however, the present results already illustrate the advantage of the polarizable model over the additive one.

Table 10. Dielectric Constant

alcohol	C22[2]	Drude, ϵ_∞[a]	Drude, ϵ[b]	exp, ϵ[b]
MeOH	17.2 (0.1)	1.5	30.1 (0.1)	32.61
EtOH	18.8 (0.3)	1.6	21.4 (0.2)	24.85
2-PrOH	13.7 (0.1)	1.7	17.6 (0.5)	19.26
2-BuOH	7.8 (0.1)	1.7	15.8 (0.4)	15.94
1-PrOH	15.2 (0.2)	1.6	19.5 (1.1)	20.52
1-BuOH	10.8 (0.1)	1.7	21.2 (0.7)	17.33
av % diff	−35.9	n/a[c]	−2.3	n/a[c]

[a] ϵ_∞ estimation from the Clausius-Mossotti equation using experimental polarizabilities of alcohols. [b] $T = 298.15$ K; Experimental data from ref 87. Av % diff is the average of the percent difference with respect to the experiment over the six alcohols studied. [c] n/a indicates not applicable.

Table 11. Self-Diffusion Coefficients of Alcohols, (D_{tot}), 10^{-5} cm^2/s

	D_{PBC}		D_{corr}	D_{tot}		exp
alcohol[a]	C22	Drude		C22	Drude	
MeOH	2.16 (0.16)	2.03 (0.44)	0.56	2.72	2.59	2.4
EtOH	1.08 (0.15)	0.93 (0.13)	0.25	1.33	1.18	1.0
2-PrOH	0.60 (0.13)	0.52 (0.07)	0.12	0.72	0.64	0.6
1-PrOH	0.66 (0.13)	0.66 (0.16)	0.13	0.79	0.79	0.6

[a] Data calculated as in reference: methanol,[21,88−90] ethanol,[22,88−91] isopropyl alcohol,[89,90] and n-propanol.[89−91] D_{pbc} is the direct result obtained from MD simulation involving periodic boundary condition. D_{corr} is the correction for the system-size effect according to eq 4.

Table 12. Isothermal Compressibility of Alcohols, MPa^{-4}

alcohol	temp, K	C22	Drude	exp[a]
MeOH	313.15	11.98 (0.96)	8.68 (0.97)	13.83
EtOH	293.15	10.12 (1.08)	10.03 (0.69)	11.19
EtOH	343.15	15.24 (1.44)	16.58 (1.53)	15.93
2-PrOH	313.15	13.62 (2.13)	11.71 (1.42)	13.32
1-PrOH	273.15	8.73 (0.99)	8.17 (1.01)	8.43

[a] Experimental data from ref 87.

Self-diffusion coefficients and isothermal compressibilities for four of the alcohols are presented in Tables 11 and 12. Diffusion constant data include the values obtained directly from the periodic boundary simulations, a correction for PBC effects on the diffusion,[63] and the total values along with the experimental values. The results demonstrate that the polarizable model to be an improvement over the additive one, although there is a tendency to slightly overestimate the experimental values. Concerning the isothermal compressibilities the calculated results show better agreement with the additive model for methanol and 2-propanol, while the polarizable model is equivalent or better for ethanol and 1-propanol. Overall, there is a tendency for the polarizable model to underestimate the compressibilities, with the exception of ethanol at 343 K.

Radial distribution functions (RDF) for the methanol, ethanol, and 2-propanol were obtained to analyze structural features of those pure solvents. O−H and O−O radial distribution functions for these solvents along with coordination numbers as a function of distance are shown in Figure 6. Comparison of the polarizable and additive $g(r)$s show minor but systematic differences. In all cases the first peak is higher in the polarizable model, indicating a higher degree of structural organization. The peak is also shifted to shorter distances in the polarizable model. For example, the first peak in the O−H RDF shifts in from 1.87 to 1.81 Å upon going from the additive to the polarizable model for all studied alcohols. A similar shift from 2.81 to 2.75 Å occurs in the O−O RDFs. The computed RDFs of the O−O and O−H distances in ethanol are in satisfactory agreement with experimental data. Neutron scattering[83] and X-ray diffraction[84] show the first peak in the O−O RDF to occur in the range 2.7−2.8 Å. The calculated coordination numbers are 2.00 and 1.99 to the first minimum occurring at 3.54 and 3.61 Å in the O−O RDFs for the polarizable and additive model, respectively, which compare well with the experimental value of 2.0 reported at 3.0 Å, the location of the minimum in the experimental work. Empirical O−H coordination numbers out to the first minima integrate to 0.97 at 2.68 Å and 0.98 at 2.64 Å for additive and polarizable models, respectively. These are in good agreement with neutron scattering data for liquid ethanol where oxygen is surrounded by 0.95 hydroxyl atoms up to the first minima at 2.1 Å. Beyond the first peak the $g(r)$s are similar for the two solvents. Thus, upon going from the additive to the polarizable model there is a significant change in the O−O and O−H pair correlation function for ethanol and 2-propanol, although the second peaks and the coordination numbers are similar for the two models.

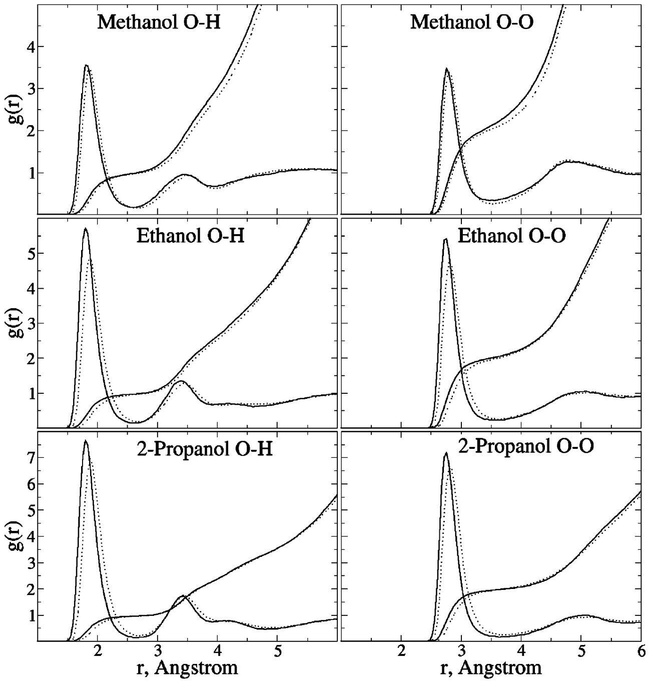


Figure 6. Radial distributions functions of the pure solvents of ethanol and 2-propanol for both the CHARMM22 additive and Drude polarizable force fields. Results for the O−H (left panels) and O−O (right panels) intermolecular interactions are shown. Solid line: polarizable model; dotted line: CHARMM22 model.

While the optimization process was dominated by the pure solvent properties, free energies of aqueous solvation values were also considered during the optimization, though their weight in the selection of the final parameter set was less than that assigned to the pure solvents. Table 13 shows the additive, Drude, and experimental free energies of solvation for the alcohols studied; the free energies include a long-range correction (LRC) for the truncation of the LJ atom−atom interactions. The CHARMM22 results are in good agreement with experiment, though they tend to be too favorable than the experimental data by 4−19%. In the polarizable model this tendency was enhanced when LJ parameters based on the pure solvent simulations were used (see Table 13, ΔG_{uncorr}), with the largest percent difference being 34% for 2-butanol (Table S5 of the Supporting Information). To overcome this trend atom type $O_{alcohol}$−O_{water} LJ terms were developed. These terms lead to improved agreement for the interactions of the alcohols with water (Tables 3 and 4) and for the free energies of solvation (Table 13). During this process it was observed that the $O_{alcohol}$−

Table 13. Free Energies of Solvation, kcal/mol

	CHARMM22			Drude				
alcohol	LRC[a]	ΔG_{solv}	%diff	LRC[a]	ΔG_{uncorr}	ΔG_{solv}	%diff	exp ΔG_{solv}
MeOH	−0.20	−4.98 (0.08)	−3	−0.26	−5.20 (0.19)	−4.64 (0.18)	−9	−5.11[b]
EtOH	−0.31	−5.34 (0.12)	7	−0.35	−5.66 (0.31)	−4.97 (0.13)	−1	−5.01[b]
2-PrOH	−0.39	−5.07 (0.09)	7	−0.45	−6.06 (0.23)	−4.82 (0.16)	1	−4.76[b]
2-BuOH	−0.45	−4.93 (0.27)	8	−0.57	−6.11 (0.18)	−4.75 (0.43)	4	−4.57[c]
1-PrOH	−0.42	−5.33 (0.24)	10	−0.46	−5.38 (0.16)	−4.85 (0.15)	0	−4.83[b]
1-BuOH	−0.53	−5.60 (0.21)	19	−0.57	−5.72 (0.16)	−4.67 (0.23)	−1	−4.72[b]
av			9				−1	

[a] The long-range correction (LRC) is estimated for dispersion forces. In the polarizable model, ΔG_{uncorr} represents free energy obtained using the standard combining rule for intermolecular LJ interactions, and ΔG_{solv} is the free energy that included an off-diagonal (i.e., NBFIX) for the $O_{alcohol}$...O_{water} LJ parameters (R_{min}=3.60, ϵ=0.18 for primary alcohols, and R_{min}=3.60, ϵ=0.21 for secondary alcohols). [b] Experimental results as reported in ref 92. [c] Reference 93.

O_{water} LJ term developed based on ethanol led to the ΔG_{solv} values still too favorable for the secondary alcohols by −0.5 to −0.6 kcal/mol (not shown). Therefore, a secondary alcohol specific $O_{alcohol}$...O_{water} LJ term was optimized, yielding the results shown in Tables 3, 4, and 13. The resulting pair specific $O_{alcohol}$...O_{water} LJ terms were (ϵ_{ij}=0.18 kcal/mol; R_{min_ij}=3.58 Å for the primary alcohols and ϵ_{ij}=0.21 kcal/mol; R_{min_ij}=3.60 Å for the secondary alcohols) replacing the corresponding terms (ϵ_{ij}=0.17788 kcal/mol; R_{min_ij}=3.5519 Å for both primary and secondary alcohols) generated by the combining rule. While the use of a specific $O_{alcohol}$−O_{water} LJ terms for the primary and secondary alcohols deviates from the goal of a fully transferable alcohol model, the significant improvement in the aqueous solvation warrants this decision.

With respect to previous studies ΔG_{solv} values of −4.88 and −4.08 kcal/mol for methanol and ethanol, respectively, were obtained by Deng and Roux for the CHARMM22 force field using a spherical solvent boundary potential (SSBP).[66] Pande and co-workers[85] used periodic boundary conditions (PBC) simulations with thermodynamic integration along with extensive sampling (with 5 ns per window versus 100 ps in the present work) to obtain free energies of solvation of −4.59 and −4.22 kcal/mol for the CHARMM22 methanol and ethanol models, respectively. Comparison with the CHARMM22 additive results shows the present results to be more favorable than the published values by 0.4−1.1 kcal/mol. Further analyses indicate that the variations are associated with methodological differences. The present computations were done using PBC with 250 water molecules, atom-based truncation, LJ switch truncation from 10 to 12 Å, and treatment of long-range electrostatic interactions via PME. The computations of Deng and Roux were done using 100 water molecules, LJ switch truncation from 10 to 12 Å, treatment of long-range electrostatics with extended electrostatics, and the SSBP continuum model. The computations of Pande and co-workers used PBC with 900 water molecules, group-based truncation, and treatment of both LJ and electrostatic interactions with switch truncation over 10−12 Å (i.e., no long-range electrostatic correction). These results emphasize the sensitivity of free energy perturbation calculations to differences in computational methodology.

Beyond methodological difference, the impact of the conformation of the alcohol on the obtained ΔG_{solv} was considered. This was performed by calculating the free energy of solvation of ethanol with the hydroxyl in either the gauche (60°) or trans (180°) conformation via inclusion of a harmonic restraint on the C−C−O−H dihedral of 1000 kcal/mol/rad.[2] The resulting ΔG_{solv} values for the gauche and trans states, assuming the same LRC corrections reported in Table 13, were −5.01 and −5.61 kcal/mol, respectively, for CHARMM22 and −4.48 and −5.52 kcal/mol, respectively, for the Drude model. Thus, the relative g versus t conformations may lead to significant differences in the obtained free energy of solvation. While ethanol stayed in the trans conformation in the present study, with the same behavior presumably occurring in the studies discussed in the preceding paragraph, the potential impact of the conformation of the hydroxyl should be noted.

It is interesting to note that the observed trends in hydration free energies of ethanol as a function of conformation is opposite to the gas-phase dipole moments of polarizable ethanol (μ_{gauche}=1.96 D, μ_{trans}=1.81 D). Such a difference indicates that the relative free energy of solvation of ethanol is dictated by the higher degree of availability of the hydroxyl group in the trans conformation to intermolecular hydrogen bonds rather than the intrinsic dipole moment. This speaks to the importance of the proper treatment of intermolecular interactions in the gas versus condensed phases, including proper polarization contribution, and their impact on condensed-phase properties.

RDFs of water with ethanol and 2-propanol were analyzed to check the impact of the polarizable model on structural properties in solution with respect to CHARMM22. Analysis of Figure 7 shows there to be a significant difference between the polarizable and additive force fields. In the $O_{alcohol}$−H_{water} RDFs the polarizable model has the first peak shifted to longer values versus the additive model, while the opposite is true for the $H_{alcohol}$−O_{water} RDF. In addition, with the $O_{alcohol}$−H_{water} RDF there are differences in the first minimum as well as the second peak. With the secondary alcohol, 2-propanol, even larger differences between the additive and polarizable models are observed, with the largest change in the $H_{alcohol}$−O_{water} RDF followed by the $O_{alcohol}$−O_{water} RDF. Thus, significant differences in the atomic details of the additive versus polarizable models are observed for an aqueous solution.

Explicit inclusion of electronic polarizability is anticipated to allow for more accurate modeling as a function of the polarity of the environment. To see if such effects occur in

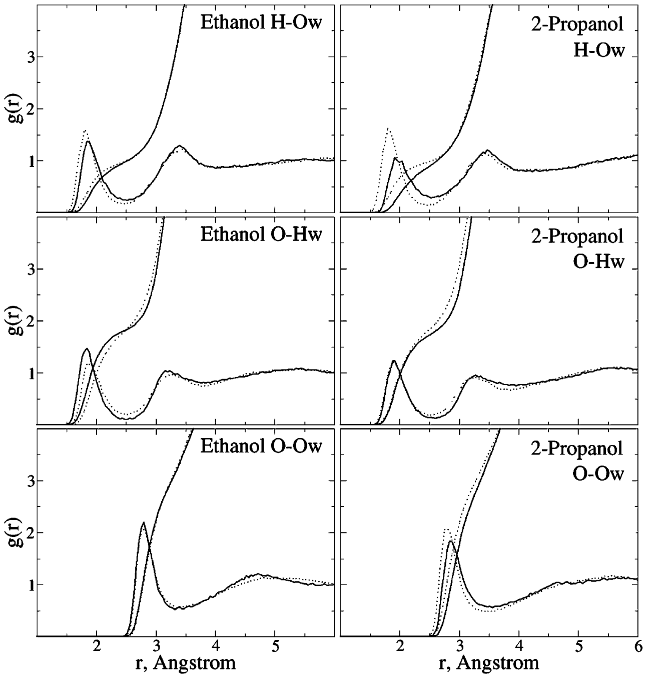


Figure 7. Radial distributions functions of ethanol and 2-propanol in aqueous solution for both the CHARMM22 additive and Drude polarizable force fields. Solid line: polarizable model; dotted line: CHARMM22 model.

the Drude polarizable force field dipole distributions were obtained from MD simulations in the gas phase, the pure solvent, and in aqueous solution for both the additive and polarizable models (Figure 8). Analysis of the figures shows extreme differences. The additive model has the three distributions clustered together with the maximum close to 2.4 D for both ethanol and 2-propanol. Such similar distributions are expected due to the lack of explicit polarizability, with the value of 2.4 being significantly larger than the gas-phase experimental value for both molecules, as required to implicitly overpolarize the molecule to obtain reasonable condensed-phase properties. With the polarizable model, significant differences are seen as a function of environment. From the MD simulations in the gas phase, which may be considered a hydrophobic environment, the distribution is centered around the gas-phase experimental values (Table 7). Upon going to the pure solvent an upshift occurrs in the distribution which is centered around 2.4 and 2.5 D for ethanol and 2-propanol, respectively. Upon moving to the more polar aqueous environments additional upshifting occurs, where the distributions are now centered around 2.7 and 2.9 D for ethanol and 2-propanol, respectively. The value for ethanol is in good agreement with a Carr−Parrinello MD prediction of the dipole moment of ethanol of 3.1 D for ethanol solvated in water.[86] Interestingly, the implicitly overpolarized additive model has a distribution similar to that of the polarizable model in the pure solvent, consistent with the satisfactory agreement for ΔH_{vap} for both models.

Further comparison of the dipole moment distributions was obtained from 500 ps MD simulation of ethanol in a box of 128 benzene molecules. The simulation using the polarizable model yields an average ethanol dipole moment of 1.86 D, close to the gas-phase value, while the additive model yields a value of 2.34 D. This result suggests that a model where polarizability changes as a function of environment has a distinct advantage over the fixed charge additive model.

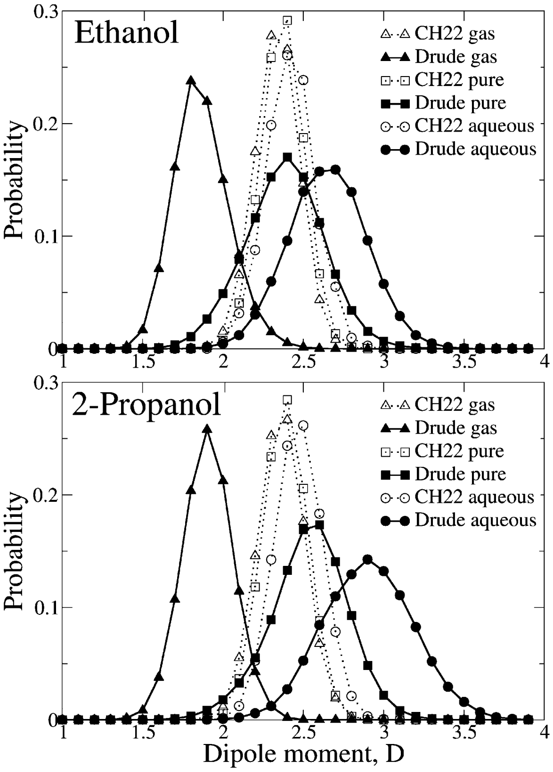


Figure 8. Dipole moment distributions of ethanol and 2-propanol in the gas phase, in pure solvents, and in aqueous solution for both the CHARMM22 additive and Drude polarizable force fields. Solid line: polarizable model; dotted line: CHARMM22 model.

Also of interest is the impact of the presence of the alcohol on the electrostatic properties of the surrounding water molecules. This was analyzed by calculating the average dipole moment of water molecules as a function of $O_{alcohol}-O_{water}$ distance for ethanol in water (Figure 9). While the overall change in the dipole moment is 0.2 D, there is a clear trend for the water dipole moment to decrease from the bulk value in the vicinity of the first minimum in the $O_{alcohol}-O_{water}$ RDF followed by an increase upon moving in to shorter distance. The overall trends in the average dipole moment of water molecules interacting with ethanol as H-bond donors or acceptors are similar, though minor differences are present. It is observed that the variations in the total dipole of water molecules relax back to the average bulk value only after the second hydration shell (4−5 Å away from the solute). Incorporating such subtle effects clearly requires going beyond the mean field picture provided by effective additive force fields, in which all the water molecules are modeled with the same electrostatic charge distribution. These results further indicate the power of a polarizable model for the investigation of condensed-phase properties.

Conclusions

An empirical polarizable force field for the alcohol series has been developed. Explicit incorporation of electronic polarizability via the classical Drude oscillator facilitates a

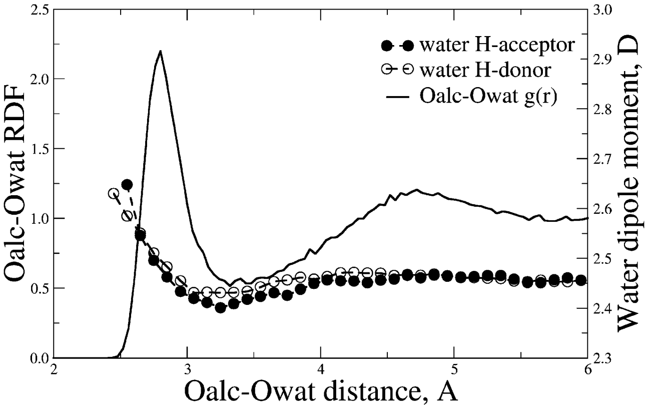


Figure 9. Average water dipole moment around ethanol as a function of the $O_{alcohol}-O_{water}$ distance overlaid on the $O_{alcohol}-O_{water}$ radial distribution function (solid line) of ethanol in water. Open circles indicate water molecules acting as H-bond donors, and filled circles indicate water molecules acting as H-bond acceptors.

more realistic response of the model compounds to the degree of polarity of the environment, with the molecular dipole changing significantly upon transition from hydrophobic to hydrophilic environments. This represents a considerable improvement over the additive model of alcohols, indicating the polarizable model to yield a better balance of the types of interactions dictating structural and thermodynamic properties of condensed phases. Significant improvement is obtained in prediction of dielectric susceptibility of liquid alcohols due to the explicit incorporation of electronic polarizability. Notable improvement over the additive model in the prediction of self-diffusivity of alcohols is also obtained. The potential energy profiles for rotation about selected bonds show the Drude model to accurately reproduce the high-level QM correlated energy maps. The Drude model is in good agreement with experiment for both pure solvent and aqueous solvation properties, though the agreement with the free energies of solvation required the use of atom type specific terms for the $O_{alcohol}-O_{water}$ LJ interactions, including individual terms for the primary and secondary alcohols. This represents a departure of the goal of transferability where the LJ parameters for the hydroxyls are identical in all the molecules and the LJ parameters for the aliphatic moieties were transferred directly from the alkanes.[27] In contrast, transferable LJ parameters were used in the additive model yielding a level of agreement with experiment typically better than that for the Drude model before the use of the atom type specific LJ terms. This poorer agreement is suggested to be due to additional sensitivity of the polarizable model to changes in the polarity of its environment leading to the constraint of transferability having a more adverse impact on that model versus than on the additive model, where the fixed-charge model diminishes the sensitivity of the model to the environment (including changes in the neighboring atoms forming the intramolecular "environment" of a given atom type). Supporting this are results from previously published polarizable models of alcohols where it was observed that the same LJ parameters could not be used for methanol and ethanol to yield agreement with experiment[18,21,22] and in a second study on a polarizable alcohol series where constraining the LJ parameters on the hydroxyl to be identical lead to a systematic variation of the pure solvent molecular volumes with respect to experiment.[18] Addressing this issue will be beneficial in gaining an understanding about effective ways to improve the predictive potential of empirical force fields.

Acknowledgment. Financial support for the work was received from the NIH (GM51501 and GM72558), and computational support was obtained from the DOD ACS Major Shared Resource Computing and PSC Pittsburgh Supercomputing Center.

Supporting Information Available: Comparison of empirical and target QM IR-spectra, data on interaction with rare gases, optimized parameter values, parameter scans in vicinity of the optimized values, and free energy computation details (Tables S1−S11). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wensink, E. J. W.; Hoffmann, A. C.; Maaren, P. J. v.; Spoel, D. v. d. Dynamic properties of water/alcohol mixtures studied by computer simulation. *J. Chem. Phys.* **2003**, *119*, 7308−7317.

(2) Jorgensen, W. L. Optimized intermolecular potential functions for liquid alcohols. *J. Phys. Chem.* **1986**, *90*, 1276−1284.

(3) Allinger, N. L.; Chen, K.-H.; Lii, J.-H.; Durkin, K. A. Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. *J. Comput. Chem.* **2003**, *24*, 1447−1472.

(4) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(5) González, M. A.; Bermejo, F. J.; Enciso, E.; Cabrillo, C. Hydrogen bonding in condensed-phase alcohols: some keys to understanding their structure and dynamics. *Philos. Mag.* **2004**, *84*, 1599−1607.

(6) Taylor, R. S.; Shields, R. L. Molecular-dynamics simulations of the ethanol liquid−vapor interface. *J. Chem. Phys.* **2003**, *119*, 12569−12576.

(7) Saiz, L.; Padro, J. A.; Guardia, E. Structure and Dynamics of Liquid Ethanol. *J. Phys. Chem. B* **1997**, *101*, 78−86.

(8) Noskov, S. Y.; Kiselev, M. G.; Kolker, A. M.; Rode, B. M. Structure of methanol-methanol associates in dilute methanol-water mixtures from molecular dynamics simulation. *J. Mol. Liq.* **2001**, *91*, 157−165.

(9) Kiselev, M.; Ivlev, D. The study of hydrophobicity in water−methanol and water−tert-butanol mixtures. *J. Mol. Liq.* **2004**, *110*, 193−199.

(10) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(11) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen,

D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(12) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(13) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(14) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.* **1999**, *20*, 730−748.

(15) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(16) Caldwell, J. W.; Kollman, P. A. Structure and Properties of Neat Liquids Using Nonadditive Molecular Dynamics: Water, Methanol, and N-Methylacetamide. *J. Phys. Chem.* **1995**, *99*, 6208−6219.

(17) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(18) Gao, J.; Habibollazadeh, D.; Shao, L. A Polarizable Intermolecular Potential Function for Simulation of Liquid Alcohols. *J. Phys. Chem.* **1995**, *99*, 16460−16467.

(19) Noskov, S. Y.; Lamoureux, G.; Roux, B. Molecular Dynamics Study of Hydration in Ethanol - Water Mixtures Using a Polarizable Force Field. *J. Phys. Chem. B* **2005**, *109*, 6705−6713.

(20) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from ab Initio Quantum Chemistry. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(21) Patel, S.; Brooks, C. L. III A nonadditive methanol force field: Bulk liquid and liquid-vapor interfacial properties via molecular dynamics simulations using a fluctuating charge model. *J. Chem. Phys.* **2003**, *122*, 024508, 1−10.

(22) Patel, S.; Brooks, C. L. III Structure, thermodynamics, and liquid-vapor equilibrium of ethanol from molecular-dynamics simulations using nonadditive interactions. *J. Chem. Phys.* **2005**, *123*, 164502, 1−12.

(23) Yu, H.; Geerke, D. P.; Liu, H.; van Gunsteren, W. F. Molecular dynamics simulations of liquid methanol and methanol-water mixtures with polarizable models. *J. Comput. Chem.* **2006**, *27*, 1494−1504.

(24) Dang, L. X.; Chang, T.-M. Many-body interactions in liquid methanol and its liquid/vapor interface: A molecular dynamics study. *J. Chem. Phys.* **2003**, *119*, 9851−9857.

(25) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. A simple polarizable model of water based on classical Drude oscillators. *J. Chem. Phys.* **2003**, *119*, 5185−5197.

(26) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418*, 245−249.

(27) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. Polarizable Empirical Force Field for Alkanes Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B* **2005**, *109*, 18988−18999.

(28) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Jr. Polarizable Empirical Force Field for Aromatic Compounds Based on the Classical Drude Oscillator. *J. Phys. Chem. B* **2007**, *111*, 2873−2885.

(29) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. Additive and Classical Drude Polarizable Force Field for Linear and Cyclic Ethers. *J. Chem. Theory Comput.* **2007**, *3*, 1120−1133.

(30) Lamoureux, G.; Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.* **2003**, *119*, 3025−3039.

(31) Iftimie, R.; Minary, P.; Tuckerman, M. E. Ab initio molecular dynamics: Concepts, recent developments, and future trends. *PNAS* **2005**, *102*, 6654−6659.

(32) MacKerell, A. D., Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.

(33) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2006**, *2*, 1587−1597.

(34) Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.* **1981**, *59*, 341−350.

(35) van Duijnen, P. T.; Swart, M. Molecular and Atomic Polarizabilities: Thole's Model Revisited. *J. Phys. Chem. A* **1998**, *102*, 2399−2407.

(36) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2005**, *1*, 153−168.

(37) Miller, K. J. Additivity methods in molecular polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542.

(38) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153*, 503−506.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford,

S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(40) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(41) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785−789.

(42) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(43) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate,spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.

(44) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(45) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(46) Saebo, S.; Pulay, P. Local Treatment of Electron Correlation. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213−236.

(47) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. Pseudospectral localized Møller−Plesset methods: Theory and calculation of conformational energies. *J. Chem. Phys.* **1995**, *103*, 1481−1490.

(48) Huang, N.; MacKerell, A. D., Jr. An ab Initio Quantum Mechanical Study of Hydrogen-Bonded Complexes of Biological Interest. *J. Phys. Chem. A* **2002**, *106*, 7820 -7827.

(49) Yin, D.; MacKerell, A. D., Jr. Combined ab initio/empirical approach for optimization of Lennard-Jones parameters. *J. Comput. Chem.* **1998**, *19*, 334−348.

(50) Chu, J.-W.; Trout, B. L.; Brooks, B. R. A super-linear minimization scheme for the nudged elastic band method. *J. Chem. Phys.* **2003**, *119*, 12708−12717.

(51) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci*. **2002**, *58*, 370−379.

(52) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Mller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502−16513.

(53) Kuczera, K.; Wiorkiewicz-Kuczera, J. *MOLVIB program*; 1991.

(54) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives. *J. Am. Chem. Soc.* **1979**, *101*, 2550−2560.

(55) Dunning, T. H. J. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(56) Evans, D. J.; Holian, B. L. The Nose−Hoover thermostat. *J. Chem. Phys.* **1985**, *83*, 4069−4074.

(57) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384−2393.

(58) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(59) Steinbach, P. J.; Brooks, B. R. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* **1994**, *15*, 667−683.

(60) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U.K., 1987.

(61) Darden, T. A.; York, D. M.; Pedersen, L. G. Particle mesh Ewald: An N•log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(62) Klauda, J. B.; Brooks, B. R.; MacKerell, A. D., Jr.; Venable, R. M.; Pastor, R. W. An ab Initio Study on the Torsional Surface of Alkanes and Its Effect on Molecular Simulations of Alkanes and a DPPC Bilayer. *J. Phys. Chem. B* **2005**, *109*, 5300−5311.

(63) Yeh, I. C.; Hummer, G. System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions. *J. Phys. Chem. B* **2004**, *108*, 15873−15879.

(64) Simonson, T. Free Energy Calculations. In *Computational Biochemistry and Biophysics*; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker: New York, 2001; p 169.

(65) Kollman, P. A. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.

(66) Deng, Y.; Roux, B. Hydration of Amino Acid Side Chains: Nonpolar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with Explicit Water Molecules. *J. Phys. Chem. B* **2004**, *108*, 16567−16576.

(67) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces,in Determining the Equilibrium Structure of Simple Liquids. *J. Chem. Phys.* **1971**, *54*, 5237−5247.

(68) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(69) Straatsma, T. P.; Berendsen, H. J. C.; Postma, J. P. M. Free energy of hydrophobic hydration: A molecular dynamics study of noble gases in water. *J. Chem. Phys.* **1986**, *85*, 6720−6727.

(70) Pearlman, D. A. A Comparison of Alternative Approaches to Free Energy Calculations. *J. Phys. Chem.* **1994**, *98*, 1487−1493.

(71) Pearlman, D. A. Free energy derivatives: A new method for probing the convergence problem in free energy calculations. *J. Comput. Chem.* **1994**, *15*, 105−123.

(72) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(73) Neumann, M.; Steinhauser, O. Computer simulation and the dielectric constant of polarizable polar systems. *Chem. Phys. Lett.* **1984**, *106*, 563−569.

(74) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(75) Foloppe, N.; MacKerell, A. D., Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **2000**, *21*, 86−104.

(76) Dixon, R. W.; Kollman, P. A. Advancing beyond the atom-centered model in additive and nonadditive molecular mechanics. *J. Comput. Chem.* **1997**, *18*, 1632−1646.

(77) Yu, H.; Gunsteren, W. F. v. Accounting for polarization in molecular simulation. *Comput. Phys. Commun.* **2005**, *172*, 69−85.

(78) Harder, E.; Kim, B.; Friesner, R. A.; Berne, B. J. Efficient Simulation Method for Polarizable Protein Force Fields: Application to the Simulation of BPTI in Liquid Water. *J. Chem. Theory Comput.* **2005**, *1*, 169−180.

(79) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **2004**, *25*, 1504−1514.

(80) Halgren, T. A. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827−7843.

(81) Waldman, M.; Hagler, A. T. New combining rules for rare gas van der waals parameters. *J. Comput. Chem.* **1993**, *14*, 1077−1084.

(82) Al-Matar, A. K.; Rockstraw, D. A. A generating equation for mixing rules and two new mixing rules for interatomic potential energy parameters. *J. Comput. Chem.* **2004**, *25*, 660−668.

(83) Benmore, C. J.; Loh, Y. L. The structure of liquid ethanol: A neutron diffraction and molecular dynamics study. *J. Chem. Phys.* **2000**, *112*, 5877−5883.

(84) Narten, A. H.; Habenschuss, A. Hydrogen bonding in liquid methanol and ethanol determined by x-ray diffraction. *J. Chem. Phys.* **1984**, *80*, 3387−3391.

(85) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* **2003**, *119*, 5740−5761.

(86) van Erp, T. S.; Meijer, E. J. Ab initio molecular dynamics study of aqueous solvation of ethanol and ethylene. *J. Chem. Phys.* **2003**, *118*, 8831−8840.

(87) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 84th ed.; CRC Press: 2003; p 2616.

(88) Karger, N.; Vardag, T.; Lüdemann, H.-D. Temperature dependence of self-diffusion in compressed monohydric alcohols. *J. Chem. Phys.* **1990**, *93*, 3437−3444.

(89) Yu, Y.-X.; Gao, G.-H. Study on self-diffusion in water, alcohols and hydrogen fluoride by the statistical associating fluid theory. *Fluid Phase Equilib.* **2001**, *179*, 165−179.

(90) Partington, J. R.; Hudson, R. F.; Bagnall, K. W. Self-diffusion of Aliphatic Alcohols. *Nature* **1952**, *169*, 583−584.

(91) Meckl, S.; Zeidler, M. D. Self-diffusion measurements of ethanol and propanol. *Mol. Phys.* **1988**, *63*, 85−95.

(92) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute-Water Clusters. *J. Chem. Theory Comput.* **2005**, *1*, 1133−1152.

(93) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Accessible Surface Areas as a Measure of the Thermodynamic Parameters of Hydration of Peptides. *PNAS* **1987**, *84*, 3086−3090.

CT700100A

JCTC Journal of Chemical Theory and Computation

Recipe of Polarized One-Electron Potential Optimization for Development of Polarizable Force Fields

Setsuko Nakagawa,*,[†,‡] Pekka Mark,[‡] and Hans Ågren[‡]

Department of Human Life and Environment, Kinjo Gakuin University, Omori, Moriyama-ku, Nagoya 463-8521, Japan, and Department of Theoretical Chemistry, Royal Institute of Technology, S-106 91 Stockholm, Sweden

Received June 2, 2007

Abstract: Polarized one-electron potential (POP) optimization is a powerful and practical method to determine multicenter dipole polarizabilities that can be used for constructing polarizable force fields. The POP optimization is similar to the widely used electrostatic potential (ESP) optimization to determine the partial charges of molecules. However, while the ESP optimization targets the electrostatic potentials on a molecular surface, the POP optimization targets the change of electrostatic potentials on molecular surfaces which are induced by the field of a test charge on the molecular surface. Since only additional one-electron integrals for the test charge are required for the estimation of the surface potentials, the change of electrostatic potentials has been named "polarized one-electron potentials". We show that in the POP optimization, both an explicitly interacting polarizability model and an implicitly interacting polarizability model can be used for the determination of the multicenter polarizabilities. In the explicitly interacting model, intramolecular induced dipole−induced dipole interaction is mutually included in the process of the POP optimization, but the interaction is not included in the implicitly interacting model. In the implicitly interacting polarizability model, a combined model of isotropic atom polarization and anisotropic bond polarization is shown to provide the best fitting results for nucleic acid bases which show large polarization anisotropy. A simple scaling model to the chemical bond has been newly proposed for the explicitly interacting polarizability model. We show that the simple model can be applied to molecular simulations without any damping of exponential type in the intramolecular induced dipole interaction. A detailed procedure for determination of the multicenter dipole polarizability by the POP optimization is also presented.

Introduction

Classical molecular simulations constitute an indispensable tool in theoretical studies of biomolecular systems such as proteins, nucleic acids, and membranes as well as physicochemical systems.[1−7] Molecular mechanics (MM) force fields are commonly used for molecular dynamics simulations to investigate dynamical structures and thermodynamical properties of the biomolecular systems.[8−11] Nowadays MM and quantum mechanics (QM) methods are often combined, QM/MM,[12−14] as a significant tool to study reaction mechanisms of enzymes. So far additive two-body potential functions have been mainly used in the MM force fields for biomolecular systems. The gap of potential quality between the QM and MM is, however, still quite large. Although the average polarization taking place in the condensed phase is included, the two-body potential does not respond to the electron density redistribution due to the ambient electric field. Potential functions that respond to a nonhomogeneous environment such as biomolecular systems can be made by the explicit inclusion of polarization. Computationally, the

* Corresponding author fax: +81 52 798 0370; e-mail: naka@kinjo-u.ac.jp.

[†] Kinjo Gakuin University.

[‡] Royal Institute of Technology.

10.1021/ct700132w CCC: $37.00
Published on Web 10/06/2007

implementation of a polarizable force field that adds a polarization term to the existing two-body potential is becoming a realistic proposition as the operation speed and memory capacity of the computers are much improved. This fact has indeed advanced the development of the polarizable force fields for biomolecular systems.[15−24]

The response of a molecule to an external polarizing field is essentially nonlocal, and the multicenter polarizabilities are not observable physical quantities. Until now, several polarization models have been reported. They can be classified in polarization models of two principal types,[25] namely as an explicitly interacting polarizability model and an implicitly interacting polarizability model. In the explicitly interacting model, intramolecular induced dipole−induced dipole interaction is included directly, but the intramolecular interaction is not included in the implicitly interacting model. The former model has been proposed, for instance, from an empirical method that is a simple approximation for predicting and rationalizing average molecular polarizabilities.[26−29] The latter model has been studied using quantum mechanics as distributed polarizabilities.[30]

An inducible point dipole (PD) model where the induced dipole−induced dipole interaction in the molecule is explicitly included has often been used for the polarizable force fields. The interacting atomic dipole model by Applequist et al. was employed in the AMBER/ff02 force field.[16,17] The intramolecular fields will be severely damped if they are included. In their approach the intramolecular electric fields from atoms separated by one and two chemical bonds were excluded. Thole improved the interacting atomic dipole model by introducing the modification of a dipole field tensor.[27] The predicted anisotropy of molecules was significantly improved by the modification. One of the damping models was employed for the AMOEBA polarizable force field.[18,19] Furthermore, a classical Drude oscillator model has been proposed, that is a variation of the PD model. Each polarizable atom is then represented by a pair of point charges of opposite sign bound by a stiff spring. This model with the short-range damping was employed in the CHARMM program.[23] Because very large force constants are used for the bonds between each atom and its Drude particle, the Drude particle stays close to the atom. The Drude model gives almost the same calculation result as the inducible PD model in the numerical value.

The microscopic electronic response of polarization has been studied using quantum mechanical calculations. It is noted that the intramolecular polarization is included implicitly at the QM level. The induced dipoles represent directly the electron cloud that changes by an external electric field. Accordingly, the induced dipoles can be simply added in the implicitly interacting polarizability models. The distribution of molecular polarizability which is estimated from the QM calculation has been studied based on a general theory developed by Stone.[30] Jansen et al. presented a robust scheme to calculate the distributed polarizabilities and distributed multipoles based on the partitioning of the physical space into atomic regions.[31] However, this scheme encountered difficulties for use in MM calculations, because of the manifold of parameters that have to be included. The dipole molecular polarization was distributed approximately on the centroids of localized orbitals by Gamer and Stevens.[32] Such distributed polarizabilities have been adopted in the SIBFA polarizable mechanics procedure.[24]

The fluctuating charge (FQ) model, where the principle of electronegativity equalization is used, is also categorized as the implicitly interacting model. This model has been adopted for the polarizable force field for proteins.[22] The FQ model may be computationally the most efficient model, but as this model has limitations in the spatial distribution of induced charges it has been combined with the PD model.[20]

A quite powerful and practical method to determine the multicenter polalizabilities was proposed in 1993.[33] In this method, first, the changes of electrostatic potentials mapped on a molecular surface induced by an external electric field point charge (test charge) are estimated using a QM method. A series of potential maps changed by a test charge put on an appropriate molecular surface is then required. Next, the multicenter polarizabilities are optimized in order to reproduce the surface potentials derived from the QM calculations. The change of electrostatic potentials is named polarized one-electron potentials (POP), because additional one-electron integrals are required by a test charge. In this methodology, the POP optimization is similar to electrostatic potential (ESP) optimization that is widely used to determine the partial charges of molecules.[34] While the ESP optimization targets the electrostatic potentials on the molecular surface, the POP optimization targets the change of electrostatic potentials on this surface. The POP optimization is applicable to both of the explicitly interacting model and the implicitly interacting model.

The POP optimization method was first applied for a water molecule.[33] Subsequently the anisotropic contribution of polarization and the transferability of multicenter polarizabilities were studied.[35,36] Two models of the isotropic atom polarization and the anisotropic polarization along the chemical bond (anisotropic bond polarization) were employed, and it was shown that the combination of the isotropic and anisotropic polarization brings good results in the POP optimization. This resembles the combination of FQ and PD.[20] The polarized model potential (PMP) function composed of Coulomb, van der Waals, and polarization terms was constructed for methanol and nucleic acid bases based on the parameters derived from the ESP and POP optimization methods.[37,38] The potential energy surfaces estimated using high-level ab initio molecular orbital (MO) calculations were reproduced quite well by the PMP function.

Ángyán et al. proceeded detailed studies and presented a formulation for distributed multipoles.[39] In their method, the energies induced by a test charge are mapped for the target of optimization instead of POP. The similar optimization method using surface POP was independently developed by Kaminski et al.[20] They used a dipolar probe which mimics liquid water instead of a single test charge. Recently, the polarization parameters of Drude particles and atomic charges were derived by the surface potential optimization method using a test charge.[23] Thus, the optimization that uses the test charge(s) has become a standard method to obtain a variety of polarization parameters from the QM calculations. However, since the high-level QM calculations are still time-consuming for larger molecules, the polarization model and the polarizability parameters have not yet been fully developed.

In this study, the explicit and implicit interacting polarizability models are investigated by using the POP optimization. Four nucleic acid bases which were used in the previous work[38] are studied further. As the nucleic acid bases show large polarization anisotropy, they are good model molecules for this research. The aim of this work is to assess the polarization models and to offer a protocol of the POP optimization for development of polarizable force fields.

Method

Electrostatic Potential Optimization and Polarized One-Electron Potential Optimization. Electrostatic potential optimization has been used as a practical method to determine partial charges of molecules.[34] The electrostatic potentials at several points ($\mathbf{R}_l$) on an appropriate molecular surface are evaluated from the wave function of an isolated molecule and are used as the reference of the charge optimization. The electrostatic potential is rigorously defined by the quantum mechanical expression

$$V^{\mathrm{QM}}(\mathbf{R}_l) = \sum_i \frac{Z_i}{|\mathbf{R}_l - \mathbf{R}_i|} - \int \frac{\rho(\mathbf{r})}{|\mathbf{R}_l - \mathbf{r}|}\,\mathrm{d}\mathbf{r} \qquad (1)$$

Here the first term represents the electrostatic contribution from the nuclear charges Z_i located at positions $\mathbf{R}_i$, and the second term represents the electrostatic potential originating from the electron density $\rho(\mathbf{r})$ throughout the whole space. In the evaluation of the electrostatic potential the wave function of the unperturbed Hamiltonian (H_0) is kept frozen (ϕ_0) under perturbation.

In the classical picture, the electron density is approximated by discrete point charges (q_i). The classical electrostatic potential given as

$$V^{\mathrm{CM}}(\mathbf{R}_l) = \sum_i \frac{q_i}{|\mathbf{R}_l - \mathbf{R}_i|} \qquad (2)$$

is estimated on the same molecular surface. The Levenberg−Marquardt nonlinear optimization procedure[40] is used to minimize the following target function in order to determine the fractional point charges:

$$\chi^2 = \sum_l^L [\mathrm{V}^{\mathrm{CM}}(\mathbf{R}_l) - \mathrm{V}^{\mathrm{QM}}(\mathbf{R}_l)]^2 \qquad (3)$$

The polarized one-electron potential optimization method is used to determine the multicenter polarizabilities.[33,35,36] In this methodology, the POP optimization is similar to the ESP optimization. To evaluate the polarization effect, it is necessary to relax the wave function (ϕ_k) under the perturbed Hamiltonian ($H_0 + H_\mathrm{k}$). Here, a molecule is perturbed by an external electric field test charge (q_k). When the molecule is polarized, the one-electron potential is modified. In quantum mechanics, we can evaluate the change in the one-electron potential ($\Delta \mathrm{V}_k^{\mathrm{QM}}(\mathbf{R}_l)$) by the polarization as follows:

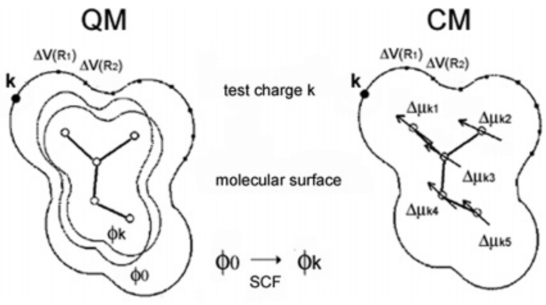


Figure 1. Schematic representation of POP optimization.

$$\Delta \mathrm{V}_k^{\mathrm{QM}}(\mathbf{R}_l) = \int \frac{\Delta\rho_k(\mathbf{r})}{|\mathbf{R}_l - \mathbf{r}|}\,\mathrm{d}\mathbf{r} \qquad (4)$$

Here $\Delta\rho_k(\mathbf{r})$ is the difference between the electron densities obtained from the frozen and relaxed wave functions (ϕ_0 and ϕ_k). Since only one-electron integrals for the test charge are required for the estimation of the surface potentials of ϕ_k, the change of electrostatic potentials was named polarized one-electron potentials. The QM calculation with the test charge is the most time-consuming step, because it requires large sets of points to sample the space around the molecule appropriately. However, the two electron integrals can be reused repeatedly. This was a useful method in the age when the operation speed of the computer was not too fast.

On the other hand, in the classical picture the difference is approximated as several discrete fractional charges (Δq_{ki}) as follows:

$$\Delta \mathrm{V}_k^{\mathrm{CM}}(\mathbf{R}_l) = \sum_i \frac{\Delta q_{ki}}{|\mathbf{R}_l - \mathbf{r}_i|} \qquad (5)$$

Since the sum of the discrete charges should be zero, an induced dipole model is introduced. The inducible point dipoles, the fluctuating charges, and the Drude oscillator models can be treated for the discrete fractional charges. Here, isotropic atom induced dipoles (induced charges of $-\Delta q^{\mathrm{a}}$ and $+\Delta q^{\mathrm{a}}$) and/or anisotropic bond induced dipoles (induced charges of $-\Delta q^{\mathrm{b}}$ and $+\Delta q^{\mathrm{b}}$) are used for the implicitly interacting polarization model. In Figure 1, the schematic representation of the POP optimization for the isotropic atom induced dipole model is presented. For the explicitly interacting model inducible point dipoles ($\Delta\boldsymbol{\mu}^I$) are used.

The nonlinear optimization procedure[40] is used to minimize the following quantity in order to determine the induced dipoles:

$$\chi^2 = \sum_j^J \sum_k^K \sum_l^L [\Delta \mathrm{V}_{jk}^{\mathrm{CM}}(\mathbf{R}_l) - \Delta \mathrm{V}_{jk}^{\mathrm{QM}}(\mathbf{R}_l)]^2 \qquad (6)$$

Here, j and k are the strength and position of the test charge, respectively. A test charge is placed on the molecular surface defined by an envelope of 1.8 times the van der Waals radius of the atoms when not specified, and test charges of −0.5 e and +0.5 e are used ($J = 2$) unless specified.

The multicenter polarizabilities of the nucleic acid bases are optimized to reproduce the polarized one-electron potentials obtained from the MP2/6-31+G* wave function.[41−43] The geometries of the nucleic acid bases are optimized at the MP2/6-31G** level. The numbers of test charge places (K) of adenine, cytosine, guanine, and thymine are 222, 194, 229, and 226, respectively. These numbers are the same as in the previous work.[38] The one-electron potential change by the polarization is estimated on the same positions of test charges. The number of the estimated points (L) was equally taken with K though neither K nor L had necessarily to be the same. In the example of adenine, the POP data of $2 \times 222 \times 222$ points were used for the fitting.

Implicitly and Explicitly Interacting Polarizability Models. The locally induced dipoles are used for the discrete charges. Since the intramolecular polarization is included at the quantum mechanical level, a simple polarization model to respond to the test charge can be used. This model is an implicitly interacting induced charge model. Two types of the induced dipoles are considered: an isotropic atom induced dipole ($\Delta\boldsymbol{\mu}^{a}_{km}$) and an anisotropic bond induced dipole ($\Delta\boldsymbol{\mu}^{b}_{km}$). The former shows a spatial movement of the charges around the atom, and the latter shows the movement of the charges along the chemical bond. The polarization anisotropy of the molecule can be introduced in the bond polarizabilities more clearly though the molecular anisotropy can be shown to some degree by the isotropic atom polarization. The induced dipoles are expanded by power series of the electric fields ($\mathbf{F}_{km}$) at the centers (m) of the dipoles, which are produced by the test charge. Here, the higher order terms are truncated, because the energetic contribution of the hyperpolarizability is expected to be small in the molecular interactions treated here. The induced dipoles are defined as

$$\Delta\boldsymbol{\mu}^{a}_{km} = \Delta q^{a}_{km}\mathbf{r}_{m_1} - \Delta q^{a}_{km}\mathbf{r}_{m_2} \approx \alpha^{a}_{m}\mathbf{F}_{km} \tag{7}$$

$$\Delta\boldsymbol{\mu}^{b}_{km} = \Delta q^{b}_{km}\mathbf{r}_{m_a} - \Delta q^{b}_{km}\mathbf{r}_{m_b} \approx \alpha^{b}_{m}(\mathbf{F}_{km}\cos\theta_{km}) \tag{8}$$

where Δq^{a}_{km} and Δq^{b}_{km} are the induced charges representing the isotropic atom and anisotropic bond induced dipoles, respectively, and $\mathbf{r}_{m1}$, $\mathbf{r}_{m2}$, $\mathbf{r}_{ma}$, and $\mathbf{r}_{mb}$ are the positional vectors of the locally induced dipole. α denotes the multicenter polarizability, and θ_{km} is an angle between the electric field vector and the chemical bond direction. α is the optimization parameters of eq 6. Because the treatment of the isotropic atom dipoles and the anisotropic bond dipoles does not need the setting of local coordinates, the handling is easy in the MM calculation. For the isotropic atom induced dipole moments, $|\mathbf{r}_{m1}-\mathbf{r}_{m2}|$ is set to 1.0 bohr. Because the induced dipole of 1.0 bohr is sufficiently buried in the van der Waals surface of the molecule, the spatial movement of the charges can be expressed well. For the anisotropic bond induced dipoles the induced charges are placed on the atoms of the bond. Because the treatment of adding the induced charge to the atomic charge is possible, the calculational efficiency can be improved in the MM calculations. The models using the isotropic atom and anisotropic bond induced dipoles are called model a and model b, respectively. The combined model is called model ab here.

An explicitly interacting polarizability model is also used for the POP optimization, namely the atom dipole interaction model proposed empirically by Applequist et al. in order to determine atomic polarizabilities from a set of experimental molecular polarizabilities.[26] The atoms are here regarded as isotropically polarizable points located at their nuclei, interacting via the fields of their induced dipoles. The induced dipole is given by using test charge field ($\mathbf{F}_{km}$) as

$$\Delta\boldsymbol{\mu}^{I}_{km} = \alpha^{I}_{m}[\mathbf{F}_{km} - \sum_{n\neq m}^{N}\tilde{\mathbf{T}}_{mn}\Delta\boldsymbol{\mu}^{I}_{n}] \tag{9}$$

where $\tilde{\mathbf{T}}_{mn}$ is the dipole field tensor and $\Delta\boldsymbol{\mu}^{I}_{n}$ is the induced dipole moment in the molecule. This model is called model A here.

We also employed Thole's modification of the intramolecular dipole interaction for repairing the deficiency of infinite polarization by the cooperative interaction between two induced dipoles in the direction of the line connecting the two.[27] The dipole field tensor is modified using the damping coefficients (λ_3 and λ_5) as follows

$$\tilde{\mathbf{T}}_{mn} = \lambda_3\frac{\tilde{\mathbf{1}}}{{r_{mn}}^{3}} - 3\lambda_5\frac{\mathbf{r}_{mn}\otimes\mathbf{r}_{mn}}{{r_{mn}}^{5}} \tag{10}$$

where $\tilde{\mathbf{1}}$ is the unit tensor, r_{mn} is the distance between atoms m and n, and $\mathbf{r}_{mn}$ is the vector connecting atoms m and n. Various forms of the modification which is related to a charge distribution were investigated by Thole.[27] Van Duijnen and Swart investigated further the linear and exponential type dampings.[28] The forms of the exponential type damping are

$$\lambda_3 = 1 - \left(\frac{a^2u^2}{2} + au + 1\right)\exp(-au) \tag{11}$$

$$\lambda_5 = 1 - \left(\frac{a^3u^3}{6} + \frac{a^2u^2}{2} + au + 1\right)\exp(-au) \tag{12}$$

where u is $r_{mn}/(\alpha_m\alpha_n)^{1/6}$, and a is the damping factor (1.9088). This is called model T with the damping type Exp 1 (model T1).

Ren and Ponder adopted different damping forms in their AMOEBA polarizable force field as follows:[18,19]

$$\lambda_3 = 1 - \exp(-au^3) \tag{13}$$

$$\lambda_5 = 1 - (au^3 + 1)\exp(-au^3) \tag{14}$$

They assigned 0.39 as the damping factor. This model is called model T with damping type Exp 2 (model T2).

A new damping model is tested here. In the molecular mechanics calculations, electrostatic interactions of atoms separated by one bond (1−2) and by two bonds (1−3) are always neglected. The electrostatic interactions of atoms separated by three bonds (1−4) are usually scaled by 0.5. In the same way the dipole field tensor $\tilde{\mathbf{T}}_{mn}$ is simply scaled. This model is called model S. The best scaling factors are investigated here.

We also consider the Drude oscillator model used in the CHARMM program in order to incorporate the polarizable force field.[23] The dipole field tensor is then scaled using the following factor

$$S_{mn} = 1 - \left(\frac{au}{2} + 1\right)\exp(-au) \tag{15}$$

where a is 2.6. The scaling is applied for the 1−2 and 1−3 induced dipole−induced dipole interaction. This model is called model D with the damping type Exp 3 (model D3). The force constant of the atom-Drude bonds is 1000 kcal/mol/Å^2.

The effects of permanent dipoles (or permanent charges) in the molecule can be omitted as pointed out by Appleuist et al. since these do not affect the net moment induced by an external field.[26] However, in the development of polarizable force fields for relatively large molecules such as proteins and DNA, a consistent treatment for inter- and intramolecular polarizations might be required for induced dipole-permanent charge interactions. The intramolecular induced dipole-permanent charge interaction is studied for nucleic acid bases using the simple scaling such as model S.

Evaluation of POP, Induction Energy, and Induced Dipole Moment. In order to evaluate the polarization model and the intramolecular damping, the root-mean-square deviations (rmsd) of POP is estimated as

$$\mathrm{rmsdPOP} = \sqrt{\frac{\sum_{j}^{J}\sum_{k}^{K}\sum_{l}^{L}[\Delta \mathrm{V}_{jk}^{\mathrm{CM}}(\mathbf{R}_l) - \Delta \mathrm{V}_{jk}^{\mathrm{QM}}(\mathbf{R}_l)]^2}{JKL}} \tag{16}$$

The relative rms deviation (rrmsd) of POP from the QM values is also estimated as the percentage.

The induction (polarization) energies are used for the evaluation of the models. In quantum mechanics the induction energy (IE) is given as

$$\Delta U_k^{\mathrm{QM}} = \langle \phi_k | H_0 + H_k | \phi_k \rangle - \langle \phi_0 | H_0 + H_k | \phi_0 \rangle = E_k - E_0 - q_k V_k^{\mathrm{QM}} \tag{17}$$

where E_k and E_0 are the total energies of the molecule in the presence and in the absence of the test charge, respectively. V_k^{QM} is the electrostatic potential at point k of the molecule without the test charge. In classical mechanics the induction energy of the isotropic induced dipoles is given as

$$\Delta \mathrm{U}_{jk}^{\mathrm{CM}} = -\frac{1}{2}\sum\nolimits_m \Delta\boldsymbol{\mu}_m \mathbf{F}_m^{jk} \tag{18}$$

The root-mean-square deviation of the induction energies is computed as

$$\mathrm{rmsdIE} = \sqrt{\frac{\sum_{j}^{J}\sum_{k}^{K}(\Delta \mathrm{U}_{jk}^{\mathrm{CM}} - \Delta \mathrm{U}_{jk}^{\mathrm{QM}})^2}{JK}} \tag{19}$$

The relative rms deviation of the induction energy from the QM values is also estimated. In the study of optimally partitioned electric properties (OPEP) by Ángyán et al. these induction energies are mapped for the target of optimization instead of POP.[25,39]

The induced dipole moments of the molecule are also used for the evaluation of the models. The QM induced dipole moment of molecule is given as follows

$$\Delta\boldsymbol{\mu}_{jk}^{\mathrm{QM}} = \boldsymbol{\mu}^{jk} - \boldsymbol{\mu}^0 \tag{20}$$

where $\boldsymbol{\mu}^{jk}$ and $\boldsymbol{\mu}^0$ are the dipole moments of the molecule in the presence and in the absence of the test charge, respectively, while in classical mechanics the induced dipole moment of molecule is given as

$$\Delta\boldsymbol{\mu}_{jk}^{\mathrm{CM}} = \sum_m \Delta\boldsymbol{\mu}_m^{jk} \tag{21}$$

The root-mean-square deviation of the induced dipole moments is computed as

$$\mathrm{rmsdIDM} = \sqrt{\frac{\sum_{j}^{J}\sum_{k}^{K}(\Delta\boldsymbol{\mu}_{jk}^{\mathrm{CM}} - \Delta\boldsymbol{\mu}_{jk}^{\mathrm{QM}})^2}{JK}} \tag{22}$$

Intermolecular Interaction of the Polarizable Model Potential Function. The optimized multicenter polarizabilities are applied for the estimation of the interaction energies of the nucleic acid base pairs. The polarizable model potential (PMP) function which consists of a electrostatic term (E_{es}), a van der Waals term (E_{vdw}), and a polarization term (E_{plz}) is used for the estimation of the total energy (E_{PMP}) of an interacting molecular system.[37]

$$E_{\mathrm{PMP}} = E_{\mathrm{es}} + E_{\mathrm{vdw}} + E_{\mathrm{plz}} \tag{23}$$

The electrostatic energy is represented by the Coulomb form using the permanent partial charges of the molecules. The charges were optimized by the ESP optimization. The van der Waals (vdw) interactions are represented by the Lennard-Jones potential. The intramolecular charge−charge and the intramolecular vdw interactions are not taken into account in the energy because they are canceled in the rigid structure. The charge and vdw parameters of the nucleic acid bases are taken from the previous work.[38]

The polarization energy is expressed as

$$E_{\mathrm{plz}} = -\frac{1}{2}\sum_{i}^{N} \Delta\boldsymbol{\mu}_i \mathbf{F}_i^{\,0} \tag{24}$$

Here, $\Delta\boldsymbol{\mu}_i$ is the induced dipole moment of site i, and $\mathbf{F}_i^0$ is the electrostatic field at site i due to the permanent charges of all other sites belonging to different molecules. The induced dipole moments are calculated self-consistently as follows:

$$\Delta\boldsymbol{\mu}_i = \alpha_i(\mathbf{F}_i^{\,0} - \sum_j \tilde{\mathbf{T}}_{ij}\Delta\boldsymbol{\mu}_j) \tag{25}$$

Here, α_i is the polarizability of site i. In the implicitly interacting polarizability model, site j is not in the molecule containing site i. The intramolecular polarizations of induced dipole−induced dipole and the intramolecular polarization of induced dipole-permanent charge are not taken into account. In the explicitly interacting polarizability model the intramolecular induced dipole−induced dipole polarization is taken into account according to the damping type. In the present work the intramolecular polarization of induced dipole-permanent charge is not taken into account unless specified. The dipole field tensor is given as follows:

$$\tilde{\mathbf{T}}_{ij} = \frac{\tilde{\mathbf{1}}}{r_{ij}^{\;3}} - \frac{3\mathbf{r}_{ij}\otimes\mathbf{r}_{ij}}{r_{ij}^{\;5}} \qquad (26)$$

For the induced dipole−induced dipole damping the following consistent formula for intra- and intermolecular interaction can be used.

$$\tilde{\mathbf{T}}_{ij} = \lambda_3\frac{\tilde{\mathbf{1}}}{r_{ij}^{\;3}} - 3\lambda_5\frac{\mathbf{r}_{ij}\otimes\mathbf{r}_{ij}}{r_{ij}^{\;5}} \qquad (27)$$

An iterative procedure is used to solve eq 25. Convergence is achieved when the deviation of the induced dipole moments from two sequential iterations falls to within 0.00025 Debye/site.

Evaluation of Surface Electrostatic Potentials of Complex Molecules. The surface electrostatic potential of a complex molecule is calculated from the nuclear charge Z_i and the electron density $\rho(\mathbf{r})$ as follows:

$$\mathrm{V}^{\mathrm{QM}}(\mathbf{R}_l) = \sum_i \frac{Z_i}{|\mathbf{R}_l - \mathbf{R}_i|} - \int \frac{\rho(\mathbf{r})}{|\mathbf{R}_l - \mathbf{r}|}\,\mathrm{d}\mathbf{r} \qquad (28)$$

Here, the complex molecule AB is treated as a supermolecule. The surface potential for complex molecule using the PMP function is calculated from atomic charges and induced dipoles as follows

$$\mathrm{V}^{\mathrm{PMP}}(\mathbf{R}_l) = \sum_i \frac{q_i}{|\mathbf{R}_l - \mathbf{r}_i|} - \sum_j \frac{\Delta\boldsymbol{\mu}_j \cos\theta_j}{|\mathbf{R}_l - \mathbf{r}_i|} \qquad (29)$$

where θ_j is an angle between the vector $\mathbf{R}_l - \mathbf{r}_j$ and the induced dipole vector $\Delta\boldsymbol{\mu}_j$. The rms deviation of the electrostatic potentials is given by

$$\mathrm{rmsdESP} = \sqrt{\frac{\sum_l^L [\mathrm{V}^{\mathrm{PMP}}(\mathbf{R}_l) - \mathrm{V}^{\mathrm{QM}}(\mathbf{R}_l)]^2}{L}} \qquad (30)$$

The relative rms deviation of ESP from the QM values is estimated as the percentage.

All of the ab initio MO calculations are done with the Gaussian03 computer program.[43]

Results

Polarization Models. Using the implicitly interacting polarizability model and the explicitly interacting polarizability model, the multicenter polarizabilities of four nucleic acid bases were determined by POP optimization. MP2/6-31+G* was used for the calculations of polarized one-electron potentials on the molecular surfaces, because this approach gave molecular polarizabilities relatively close to the experimental values in solution as shown in the previous study.[38] The polarization reduction by the exchange repulsion in the condensed phase might be taken into account by this basis set.

Table 1. Root-Mean-Square Deviation of Polarized One-Electron Potentials in kcal/mol, Induction Energies in kcal/mol, and Induced Dipole Moments in Debye of Various Polarization Models for Four Nucleic Acid Bases

model	damping type	molecule	rmsd (rrmsd %)		
			POP	IE	IDM
a		A	1.4 (31)	0.4 (15)	0.6 (19)
		T	1.2 (29)	0.3 (14)	0.5 (11)
		C	1.3 (32)	0.4 (14)	0.6 (8)
		G	1.4 (32)	0.4 (16)	0.7 (9)
b		A	1.6 (36)	1.0 (34)	0.6 (20)
		T	1.0 (29)	0.7 (27)	0.5 (10)
		C	1.4 (34)	0.8 (33)	0.6 (8)
		G	1.4 (33)	0.8 (31)	0.6 (9)
ab		A	0.6 (13)	0.3 (11)	0.2 (5)
		T	0.4 (9)	0.2 (8)	0.1 (2)
		C	0.5 (12)	0.3 (10)	0.2 (3)
		G	0.5 (12)	0.3 (11)	0.2 (2)
A		A	1.2 (27)	0.9 (32)	0.4 (12)
		T	0.8 (20)	0.6 (23)	0.2 (5)
		C	1.0 (24)	0.8 (29)	0.3 (4)
		G	1.1 (25)	0.8 (29)	0.4 (5)
T1	Exp 1	A	0.6 (13)	0.3 (10)	0.2 (7)
		T	0.5 (12)	0.2 (8)	0.2 (4)
		C	0.5 (12)	0.2 (9)	0.2 (2)
		G	0.5 (13)	0.3 (10)	0.2 (3)
T2	Exp 2	A	0.7 (15)	0.4 (13)	0.3 (8)
		T	0.7 (17)	0.3 (14)	0.3 (6)
		C	0.9 (20)	0.5 (18)	0.3 (4)
		G	0.7 (17)	0.4 (14)	0.3 (4)
D3	1-2,1-3 Exp 3	A	1.1 (25)	0.4 (14)	0.5 (15)
		T	1.0 (26)	0.3 (12)	0.5 (9)
		C	1.2 (28)	0.3 (13)	0.5 (7)
		G	0.6 (15)	0.3 (10)	0.3 (4)
S	1-2 0.1	A	0.5 (11)	0.3 (12)	0.1 (4)
		T	0.5 (13)	0.2 (9)	0.2 (4)
		C	0.5 (12)	0.3 (11)	0.2 (2)
		G	0.5 (10)	0.3 (11)	0.1 (2)

The root-mean-square deviations of polarized one-electron potentials, induction energies, and induced dipole moments of various polarization models for the four nucleic acid bases are shown in Table 1. In the implicitly interacting induced charge model, model ab shows the best results in the POP fitting. The relative rms deviations of model ab are 9−13%. The rms deviation of induction energies and induced dipole moments were only 8−11% and 2−5%, respectively. These IE and IDM results estimated from the optimized induced charges showed better results compared with the results of POP that inspected the electronic density change in detail. The model ab has shown quite good results for acetylene, ethylene, and benzene.[35] The combination of isotropic induced dipoles at atom centers with anisotropic induced dipoles along bonds is significant for describing the molecular polarization with large anisotropy.

Table 2. Multicenter Polarizability in au of Implicitly Interacting Induced Dipole Models

		isotropic atom polarizability (au)				anisotropic bond polarizability (au)				
model	molecule	C	N	H	O	C−C	C−H	C−N	N−H	C−O
a	A	4.555	10.122	1.983						
	T	6.614	5.502	2.121	6.983					
	C	7.688	6.004	2.354	7.576					
	G	6.532	7.600	1.891	7.689					
b	A					23.61	10.37	18.87	5.96	
	T					19.03	8.93	14.32	7.02	23.40
	C					19.01	5.35	21.43	4.80	22.08
	G					10.01	6.31	21.05	4.48	23.82
ab	A	0.464	9.047	0.762		15.88	5.73	9.50	−0.46	
	T	9.853	0.866	3.025	13.279	9.85	3.02	10.18	0.87	13.28
	C	4.272	3.850	1.718	3.850	10.07	0.58	12.67	−0.16	13.92
	G	2.268	5.693	1.177	5.781	8.63	3.51	11.34	−0.41	11.63
additive (empirical)[a]		6.93	8.34	2.75	5.68					

[a] The atomic polarizabilities of C (alkane), N (nitrile), H (alkane), and O (carbonyl) are shown.[26]

Model A, which is the explicitly interacting polarizability model without damping, shows better results in comparison with the additive model (model a) in the POP fitting. However, the induction energies were overestimated. In the explicitly interacting model, model T1 and model S show significantly similar results compared to model ab in the POP fitting. The results of model S are the best ones when scaling 0.1 is given to the 1−2 bonds. The results of model T2 and model D3 are somewhat worse.

The optimized polarizabilities of the implicitly interacting polarization model are shown in Table 2. The parameters of each atomic species have been optimized in model a. H, C, N, and O polarizability parameters of four nucleic acid bases were almost converged, though the values C and N of adenine depart somewhat. Similar convergence was found for alkanes and alcohols in the POP optimization.[36] The parameters of model a can be compared with the empirical values of the additive model listed by Applequist.[26]

The optimized polarizabilities of the explicitly interacting polarization model are shown in Table 3. The parameters of each atomic species have been optimized. The atomic polarizabilities of model A are smaller than those of model a, since the mutual induction enhances the molecular polarizability in model A. The convergence of parameters is relatively good in models A and S. The atomic polarizabilities changed according to the type of the damping.

Table 3. Atomic Polarizability in au of Explicitly Interacting Induced Dipole Models

			atomic polarizability (au)			
model	damping type	molecule	C	N	H	O
A		A	4.759	3.417	0.736	
		T	4.325	3.235	1.107	3.021
		C	4.511	3.566	0.764	2.932
		G	4.905	3.133	0.827	2.158
T1	Exp 1	A	7.937	11.967	0.756	
		T	10.379	7.331	1.278	7.711
		C	11.986	9.674	0.718	7.703
		G	8.265	11.027	0.531	8.671
T2	Exp 2	A	2.022	15.488	1.959	
		T	4.388	10.690	3.375	9.572
		C	3.531	14.401	2.625	9.345
		G	1.815	14.632	1.632	10.626
D3	1-2,1-3	A	2.136	13.367	1.650	
	Exp 3	T	5.883	5.424	2.614	7.848
		C	7.295	5.611	2.917	8.163
		G	7.714	7.494	1.222	7.275
S	1-2 0.1	A	6.654	8.302	0.992	
		T	7.308	5.202	1.440	7.221
		C	8.316	6.945	1.029	6.834
		G	6.854	7.493	0.901	7.588
empirical						
A[a]			6.93	8.34	2.73	5.68
T[b]	Exp 1		8.794	6.670	3.059	5.648

[a] The atomic polarizabilities of C (alkane), N (nitrile), H (alkane), and O (carbonyl) are shown.[27] [b] Atomic polarizabilities fitted to the original 16 molecules.[28]

Condition of POP Optimization. Since the results of model S were good in the explicitly interacting polarizability model though the damping form was quite simple, further investigations were performed using this model. In the first paper on the POP optimization we investigated the strength of field point charges using water molecule.[33] It was shown that the polarizabilities are changed exponentially and that the deviations are greatly increased in the highly positive field. In Figure 2 the effects of a field test charge for guanine are shown. The optimized polarizabilities are almost constant from −1.5 e to +0.5 e. However, they change rapidly when +1.0 e is exceeded. In the region between −1.0 e and +0.5 e the rms deviation of POP is less than 11% but increases to 26% for +1.0 e. In this study simultaneous fittings were performed using −0.5 e and +0.5 e leading to good results (10%) as shown in Table 1. Very similar results were obtained using the test charge of +0.1 e as shown in Figure 2. To reduce computational time the single test charge of +0.1 e might be a good choice.

The effect of the van der Waals surface in which the test charge was located was investigated. Here, the same surface points were used to evaluate the polarized one-electron potentials. The single test charge of +0.1 e was used. In Figure 3 the changes in atomic polarizabilities are plotted for the vdw radius of the atoms. The almost constant atomic polarizabilities were obtained in the region from 1.6 to 3.0 times the vdw radius of atoms. The rms deviations are less

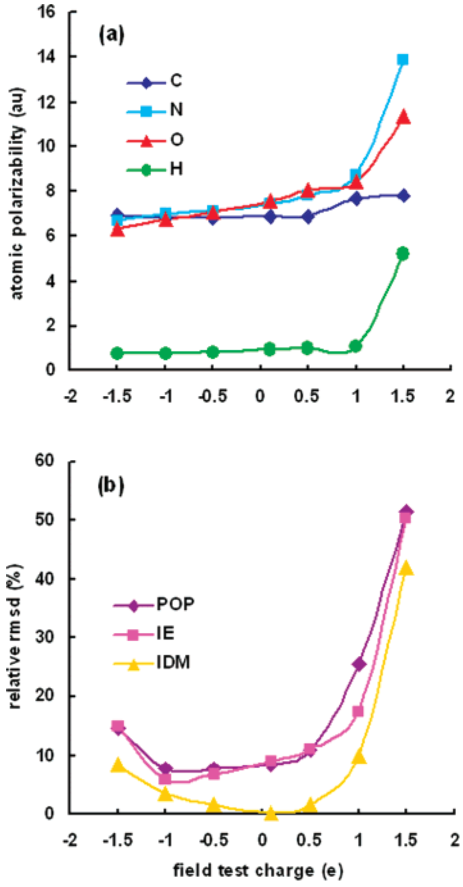


Figure 2. (a) Variations of atomic polarizabilities of guanine for field test charges. (b) Relative root-mean-square deviations of polarized one-electron potential fitting, induction energies, and induced dipole moments.

than 10%. The POP optimization did not succeed at the surface of the 1.4 times of the vdw radius, because of the penetration of the test charge to the electron cloud. So far we have used the 1.8 times the vdw radius. This single surface choice is reasonable.

Simple Scaling Model. The scaling effect for the induced dipole−induced dipole interactions was studied in detail using model S of guanine. The effect of atoms separated by one bond (1−2) is shown in Figure 4. The atomic polarizabilities changed greatly by the scaling. The best result of the POP optimization was obtained with the scaling of 0.1−0.2. The best result of induction energy was obtained in the scaling of 0.0. Thus, it is necessary to scale down the induced dipole−induced dipole interactions between atoms making the chemical bonds.

The effect of atoms separated by two bonds (1−3) is shown in Figure 5. Here, the 1−2 interaction has been scaled by zero. The changes of atomic polarizabilities are found to be quite minor. The best result of the POP optimization was obtained for the scaling of 1.0. A complete inclusion of the 1−3 interaction is necessary contrary to the 1−2 interaction. Moreover, a complete inclusion of the 1−4 interaction or more improves somewhat the rms deviation. Thus, model S in which the 1−2 interactions are scaled by 0.1 works well. In the linear model of Thole atoms separated by two bonds were located in the nondamping region. The complete inclusion of 1−3 or more is consistent with Thole's linear model.

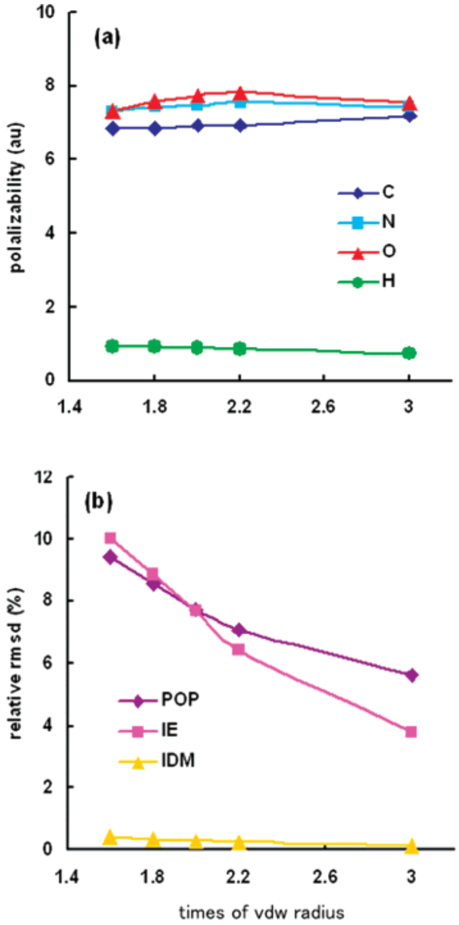


Figure 3. (a) Variation of atomic polarizabilities of guanine for times of van der Waals radius of atoms. (b) Relative root-mean-square deviations of the polarized one-electron potential fitting, induction energies, and induced dipole moments.

Permanent Charge-Induced Dipole Interactions within the Molecule. The intramolecular interactions for permanent charge-induced dipole were studied using model S of guanine. The scaling effect for 1−3, 1−4, and 1−5 or more are shown in Figure 6. The scaling combination of 0.0, 0.5, and 1.0 for 1−3, 1−4, and 1−5 or more were tested. Here, the 1−2 interaction was neglected (zero scaling). The atomic polarizabilities changed greatly by the scaling as shown in Figure 6(a). Even the inclusion of 1−5 or more interaction causes the 30% deterioration of the rms deviation of POP. Because the effect of permanent charges is large, the adjustment of the induced dipoles seems not to be possible. An exclusion of the intramolecular permanent charge-induced dipole interactions is necessary contrary to the case of induced dipole−induced dipole interactions.

Apart from the above effect, an inclusion of the intramolecular permanent charge-induced dipole interaction influences the permanent charges obtained from the ESP optimization. In the ESP optimization process, this can be corrected for by the inclusion of induced dipole moments in addition to the permanent charges.[16,43] However, the intrinsic deterioration by the intramolecular permanent charge-induced

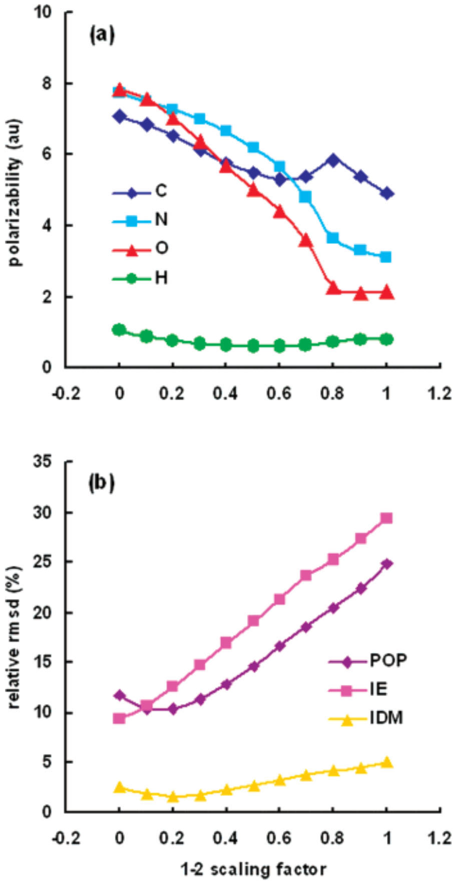


Figure 4. (a) Variation of atomic polarizabilities of guanine by 1-2 bond scaling of simple model. (b) Relative root-mean-square deviations of the polarized one-electron potential fitting, induction energies, and induced dipole moments.

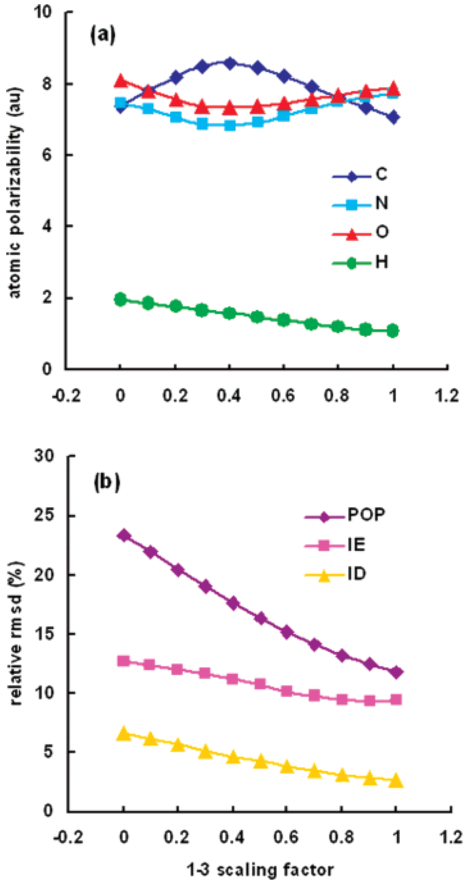


Figure 5. (a) Variations of atomic polarizabilities of guanine by 1-3 bond scaling of simple model. (b) Relative root-mean-square deviations of polarized one-electron potential fitting, induction energies, and induced dipole moments.

dipole interaction cannot be corrected. Thus, the correction of the permanent charges by the effect of the intramolecular induced dipoles was not executed in our ESP optimization.

Application of Polarization Models for Nucleic Acid Base Interactions. Five types of nucleic acid base complexes were studied: the Watson−Crick adenine−thymine pair (AT-wc), the Hoogsteen adenine−thymine pair (AT-h), the Watson−Crick cytosine−guanine pair (CG-wc), the stacked adenine−thymine pair (AT-s), and the stacked cytosine−guanine pair (AT-s). The optimized geometries and the quantum mechanical interaction energies of the base pairs are taken from the previous work.[38] The target interaction energies were calculated by using MP2/6-311++G(3df,2pd) for the hydrogen bond base pairs and by using MP2/6-311++G(2d,2p) for the stacked base complexes. The basis set super position error (BSSE) corrected interaction energies[45] are shown in Table 4.

The classical energies were estimated by using the polarized model potential function. The atomic charges and van der Waals parameters are taken from our previous work.[38] The polarization energies were estimated by using the implicitly interacting model and the explicitly interaction model. The interaction energies and polarization energies of the five nucleic acid complexes are shown in Table 4. For the empirical T1 model in Table 4, the empirical atomic polarizabilities reported by van Duijnen and Swart were used: H 3.0588 au; C 8.7979 au; N 6.6704 au; and O 5.6480 au.[28] Model ab, model T1, and model S show quite good results for AT-cw, AT-h, and CG-wc. Model A and the empirical model T1 show slightly too high values for AT-wc and AT-h. Model T2 and the empirical model T1 show similar high polarization energies for CG-wc. The empirical parameters were fitted to the experimental molecular polarizabilities in gas phase. Since the polarizabilities are reduced in the condensed phase, the use of empirical values might have a tendency to overestimate the polarization energy.[38] Model b and model D3 show a little too low values. The contribution of the polarization effect is small in the stacked base complexes. All polarization models studied here show similar polarization energies for the stacked complexes.

The dipole moments of the complexes are shown in Table 4. The dipole moment estimated by the electrostatic term of PMP (Ees) shows the state that has not polarized. The dipole moments were relatively well reproduced except for model A applied to the hydrogen bond type base pairs. The relative rms deviations of surface ESP are also shown in Table 4. The relative rms deviations were then less than 12% except for models a, T2, and D3 of CG-wc. The electron distribution change by the formation of the complex as well as the interaction energies is well represented by the implicitly interacting polarizability model and the explicitly interacting polarizability model.

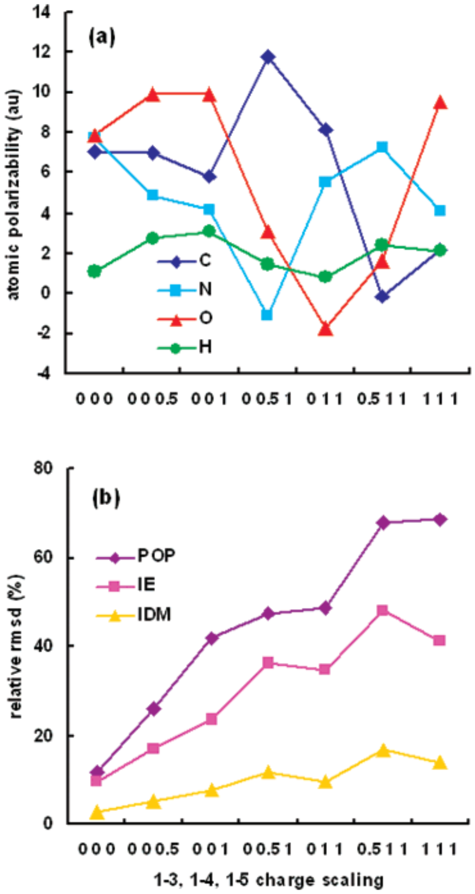


Figure 6. (a) Variations of atomic polarizabilities of guanine by 1-3, 1-4, and 1-5 or more bond scaling for charge-induced dipole interactions. (b) Relative root-mean-square deviations of the polarized one-electron potential fitting, induction energies, and induced dipole moments.

Discussion

The POP optimization is a powerful and practical method to determine multicenter dipole polarizabilities that can be used for determining polarizable force fields. This method can be applied to any molecule that can be calculated by quantum mechanics. A test charge used in POP optimization mimics a nonuniform external field induced by a nearby ion. Since the change in the electron density is a target of the optimization, the higher order multipole contributions in the single-center expansion are all included. It is readily possible to implement this method in both the implicitly interacting and the explicitly interacting polarizability models.

In the implicitly interacting model, the ab model including isotropic atom polarizabilities and anisotropic bond polarizabilities showed good results. The anisotropic contribution is directly calculated by the anisotropic polarizabilities along the chemical bond. On the other hand, the anisotropy is included in the convergence process of the induced dipole−induced dipole interactions in the explicitly interacting model. The explicitly interacting model needs substantially more computer time compared with the implicitly interacting model. The T1 and S models have shown great results at the same level as model ab. Although the treatment of the induced dipoles is quite different from the implicitly interacting model, the polarization anisotropy of molecule is similarly included in the explicitly interacting model.

When the multicenter polarizability parameters of each atom and/or each bond were individually relaxed, we obtained parameters that differed from what can be expected by chemical intuition. In this study, the parameters of each atomic species and/or each bond have been optimized. The rms deviation hardly deteriorated at all though the number of parameters decreased by the restraint to the atomic species and/or the bond species. In the empirical approach by Thole, one polarizability is adopted for each atom irrespective of chemical environment. The empirical atomic polarizabilities of Thole were somewhat different from those determined by the POP optimization, but comparatively good results were obtained in both of the parameter sets for the studied nucleic acid base complexes. In the ESP optimization, the charge restraints for deriving atomic charges have often been introduced in order to obtain transferable parameters for the conformation.[34] Although the multicenter polarizabilities are determined for each molecule by using POP optimization and applied in the molecular mechanics calculations, it seems that a small number of transferable polarization parameters can be obtained by the restraint for atomic species especially in the explicitly interacting model as inferred from the empirical approach of Thole.[27]

When the multicenter polarizabilities of a small molecule are used as segments of a large molecule, it is necessary to exclude the induced dipole−induced dipole interactions within the segment in the implicitly interacting model. This manipulation is somewhat troublesome for the treatment of intrasegment interaction. On the other hand, it is not necessary to exclude this contribution in the explicitly interacting model. Thus, the consistent treatment of inter- and intramolecular polarization is possible in the explicitly interacting model. Especially model S is suitable for polarizable force fields because the complicated damping calculations of exponential type can be avoided.

The electron density change induced by a test charge is reflected purely in the polarized one electron potentials. The polarization by the intramolecular atomic charges is not originally included in the density change itself. Those are included in the atomic charges that are determined by the ESP optimization. When the multicenter polarizabilities of small molecule are used as the segments of a large molecule, it is necessary to exclude the atomic charge-induced dipole interactions within the segment. Thus, the consistent treatment between the intermolecular polarization and intramolecular polarization is difficult for the atomic charge-induced dipole interactions. Ren and Ponder suggested a treatment of the polarization group in which the net charges are small,[18] but such grouping is difficult for the conjugate system such as nucleic acid bases. In the construction of the AMBER polarizable force field (ff02), atomic charges were optimized using the electrostatic potential around a molecule, in which the electrostatic potential, created by the induced dipoles and atomic charges, is subtracted.[16,43] Such a correction for the double counting of the permanent charge-induced dipole is valid for the estimation of electrostatic energy, but the induction energy is not corrected. In the POP optimization

Table 4. Nucleic Acid Base Interaction Energies in kcal/mol, Dipole Moments in Debye, and rms Deviations of Electrostatic Potentials in kcal/mol

		AT-wc				CG-wc				AT-h			
model	damping type	E	E_{plz}	dipole moment	rmsd ESP	E	E_{plz}	dipole moment	rmsd ESP	E	E_{plz}	dipole moment	rmsd ESP
a		−15.5	−4.4	1.4	1.2 (10)[b]	−28.6	−9.3	6.3	2.5 (16)	−17.5	−4.2	6.3	1.1 (10)
b		−14.7	−3.6	1.4	1.2 (10)	−27.5	−8.2	6.5	1.6 (10)	−16.8	−3.5	6.4	1.3 (12)
ab		−15.5	−4.4	1.2	1.3 (11)	−28.4	−9.1	6.4	1.7 (11)	−17.6	−4.3	6.1	1.3 (12)
A		−17.8	−6.7	2.2	1.3 (11)	−29.1	−9.9	5.4	1.3 (9)	−18.9	−5.6	6.7	1.3 (11)
T1	Exp 1	−16.5	−5.4	1.3	1.1 (10)	−29.6	−10.4	6.3	1.6 (11)	−18.5	−5.2	6.1	1.1 (10)
T1 empirical	Exp 1	−17.4	−6.2	1.4	1.0 (8)	−31.8	−12.5	6.4	1.9 (12)	−19.3	−6.0	6.3	0.9 (8)
T2	Exp 2	−15.8	−4.7	1.8	1.1 (9)	−34.3	−15.1	6.5	2.4 (15)	−17.8	−4.4	6.6	1.0 (9)
D3	1-2,1-3 Exp 3	−14.3	−3.1	1.8	1.2 (10)	−24.8	−5.5	5.7	2.6 (17)	−16.2	−2.9	6.6	1.2 (11)
S	1-2 0.1	−15.2	−4.1	1.3	1.2 (10)	−28.9	−9.6	6.5	1.6 (10)	−17.3	−4	6.2	1.1 (10)
Ees+Evdw		−11.1		2.2	1.3 (11)	−19.3		4.6	2.4 (16)	−13.3		6.8	1.2 (11)
MP2/6-31+G* [a]		−13.8		1.6	0.6 (5)	−27.5		6.1	0.7 (4)	−14.3		6.4	0.6 (5)
MP2/6-311++G(3df,2pd)[a]		−15.7		1.5		−30.0		5.8		−17.5		6.1	

		AT-s				CG-s			
model	damping type	E	E_{plz}	dipole moment	rmsd ESP	E	E_{plz}	dipole moment	rmsd ESP
a		−9.8	−0.6	6.9	1.2 (8)	−10.7	−1.0	4.5	1.5 (9)
b		−9.6	−0.4	6.6	1.1 (8)	−10.6	−0.9	4.3	1.1 (7)
ab		−9.8	−0.6	6.7	0.8 (6)	−10.8	−1.1	4.3	1.0 (7)
A		−9.8	−0.6	6.9	1.2 (8)	−10.5	−0.9	4.2	0.9 (6)
T1	Exp 1	−9.8	−0.5	6.7	0.9 (6)	−10.8	−1.1	4.3	1.1 (7)
T1 empirical	Exp 1	−9.7	−0.5	6.7	0.9 (6)	−10.7	−1.0	4.3	1.1 (7)
T2	Exp 2	−9.7	−0.5	6.8	0.8 (6)	−10.9	−1.2	4.3	1.1 (7)
D3	1-2,1-3 Exp 3	−9.9	−0.7	6.6	0.9 (7)	−10.6	−0.9	4.3	1.2 (8)
S	1-2 0.1	−9.7	−0.5	6.7	0.8 (6)	−10.8	−1.1	4.2	1.0 (6)
Ees+Evdw		−9.2		7.0	1.4 (10)	−9.7		4.7	1.7 (11)
MP2/6-31+G* [a]		−5.4		6.4	0.5 (3)	−8.4		4.1	0.6 (4)
MP2/6-311++G(2d,2p)[a]		−8.4		6.2		−11.4		3.8	

[a] The BSSE corrected interaction energies are taken from ref 38. [b] Rrmsd %.

the effect from the atomic charges should be excluded, and the atomic charge-induced dipole interactions within the segment should be excluded as much as possible in the evaluation of the induction energies. Therefore, in the explicitly interacting model the treatment for the intrasegmental atomic charge-induced dipole interaction has a completely different necessity from the treatment for the intrasegmental induced dipole−induced dipole interaction. On the other hand, in the implicitly interacting polarization model both of the intrasegmental interactions must be excluded.

The intermolecular interaction energy can be divided into electrostatic (ES), exchange repulsion (EX), polarization (PL), and charge-transfer (CT) terms in the calculations of HF level.[47] The dispersion force (DIS) is estimated by the difference between the HF energy and the energy including electron correlation effect. In the polarizable model potential (PMP) function used here a polarization (plz) term is added to an existing pair potential function which consists of an electrostatic (es) term and a van del Waals (vdw) term. The es and plz terms correspond to the ES and PL terms, respectively. The vdw term represents the EX and the DIS. An explicit CT term is not included in the PMP function. However, it is difficult to obtain a one-to-one correspondence between energies obtained by the quantum and the classical calculations, because the atomic charges and the multicenter polarizabilities were derived in this study from the calculations of the MP2 level that include the effect of electron correlation. In the previous study on methanol using the PMP function, a quite good agreement of the interaction energies had been shown for ion−methanol and methanol−ion−methanol systems.[37] The ions of Cl^-, Na^+, and Mg^{2+} were used in that study. In the ion−methanol complex the energy decomposition results showed that the ES, PL+CT+R, and EX+def+DIS roughly correspond with the es, plz, and vdw of the PMP function, respectively. Here, the PL+CT+R shows the sum of the energies of PL, CT, and the residual. The EX+def+DIS shows the sum of the energies of EX, deformation, and DIS. Recently, Donchev et al. showed that the MP2 correction of ES and EX for DIS works well in the development of their quantum mechanical polarizable force field though the MP2 correction of induction energy for DIS was neglected.[48] The vdw term (L-J potential) of PMP function shows the EX and DIS. The inclusion of the plz term to the present pair potential may partly double-count the dispersion because the multicenter polarizabilities were derived from the MP2 calculations. In the development of the PMP function, the parameters of the vdw term are treated as adjustable and lose their original physical meaning.[37,38]

In the previous study on the nucleic acid bases it was pointed that the multicenter polarizabilities estimated by POP optimization partly include the ability of CT.[38] Concerning the CT contribution, Chelli et al. suggested that classical polarizable force fields underpolarize in a hydrogen-bond model system of water.[49] On the other hand, it has been suggested that the overpolarization occurs in the condensed phase by the neglect of coupling between many-body exchange and polarization.[50] In the POP optimization the reference quantum mechanical calculations of nucleic acid bases were computed at the MP2/6-31+G* level since the calculated molecular polarizabilities were relatively close to the experimental values in the condensed phase.[38] For example, the theoretical and the experimental molecular polarizabilities of guanine are 96.8 au and 91.8 au, respectively. The largest theoretical value reported is 106.7 au.[46] Since the polarization of the molecule is reduced by the electron repulsion and is increased somewhat by the CT, the choice of MP2/6-31+G* might be adequate for the estimation of polarization energy in the condensed phase.

The present polarizable force fields still have problems and should be investigated further.[48,51] However, the results from the QM study reproduced the interaction energies and the charge density changes though the same parameters were applied to quite different systems (hydrogen bond base pair, stacked base pair, and ion base complex). Especially, the interactions of the complexes including ions are excellently improved in comparison with the pair potential as shown in the previous work.[38] It seems that these results give more encouragement to the development of polarizable force fields.

A systematic development of high quality polarizable force fields is possible by the POP optimization. The development of polarizable force fields for proteins and nucleic acids has already been started by using POP optimization. Our intention is to further elaborate this recipe for the development of polarizable force fields.

Acknowledgment. The authors are grateful to Prof. N. Kosugi of Institute for Molecular Science for his helpful discussions and Prof. M. Yamakawa of Kinjo Gakuin University for his technical support. This work was supported by the Grant-in-Aid for Scientific Research (C) of Japan Society for the Promotion of Science.

References

(1) Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 321.

(2) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III *Acc. Chem. Res.* **2000**, *33*, 889.

(3) Giudice, E.; Lavery, R. *Acc. Chem. Res.* **2002**, *35*, 350.

(4) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430.

(5) Saiz, L.; Klein, M. L. *Acc. Chem. Res.* **2002**, *35*, 482.

(6) Nakagawa, S.; Yu, H.-A.; Karplus, M.; Umeyama, H. *Proteins: Struct., Funct., Genet.* **1993**, *16*, 172.

(7) Maekawa, K.; Ishikawa, S.; Ishida, H.; Nakagawa, S.; Ohkubo, K.; Yamabe, T. *Mol. Eng.* **1998**, *8*, 9.

(8) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(9) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Chio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(10) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem Soc.* **1988**, *110*, 1657.

(11) MacKerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946.

(12) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.

(13) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

(14) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.

(15) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.

(16) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048.

(17) Wang, Z. -X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781.

(18) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.

(19) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

(20) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515.

(21) Harder, E.; Kim. B.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2005**, *1*, 169.

(22) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1504.

(23) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153.

(24) Gresh, N.; Šponer, J. E.; Špačková, N.; Leszczynski, J.; Šponer, J. *J. Phys. Chem. B* **2003**, *107*, 8669.

(25) Chipot, C.; Ángyán, J. G. *New J. Chem.* **2005**, *29*, 411.

(26) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952.

(27) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.

(28) van Duijnen, P. Th.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399.

(29) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8543.

(30) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065.

(31) Jansen, G.; Hättig, C.; Hess, B. A.; Ángyán, J. G. *Mol. Phys.* **1996**, *88*, 69.

(32) Garmer, D. R.; Stevens, W. J. *J. Phys. Chem.* **1989**, *93*, 8263.

(33) Nakagawa, S.; Kosugi, N. *Chem. Phys. Lett.* **1993**, *210*, 180.

(34) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(35) Nakagawa, S. *Chem. Phys. Lett.* **1995**, *246*, 256.

(36) Nakagawa, S. *Chem. Phys. Lett.* **1997**, *278*, 272.

(37) Nakagawa, S. *J. Phys. Chem. A* **2000**, *104*, 5281.

(38) Nakagawa, S. *J. Comput. Chem.* **2007**, *28*, 1538.

(39) Ángyán, J. G.; Chipot, C.; Dehez, F.; Hättig, C.; Jansen, G.; Millot, C. *J. Comput. Chem.* **2003**, *24*, 997.

(40) Fletcher, R. *Practical methods of optimization*; Wiley: New York, 1980; Vol. 1, p 10276.

(41) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(42) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O; Rabuck, A. D.; Raghavachari, K.; Foresman, J.B.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03 (Revision A.1)*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(44) Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. *J. Comput. Chem.* **1999**, *20*, 704.

(45) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(46) Jasien, P. G.; Fitzgerald, G. *J. Chem. Phys.* **1990**, *93*, 2554.

(47) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325.

(48) Donchev, A. G.; Galkin, N. G.; Pereyaslavets, L. B.; Tarasov, V. I. *J. Chem. Phys*. **2006**, *125*, 244107.

(49) Chelli, R.; Schettino, V.; Procacci, P. *J. Chem. Phys.* **2005**, *122*, 234107.

(50) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 9903.

(51) Engkvist, O.; Åstrand, P.-O.; Karlström, G. *J. Phys. Chem.* **1996**, *100*, 6950.

CT700132W

JCTC Journal of Chemical Theory and Computation

Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand−Macromolecule Complexes. A Bottom-Up Strategy

Nohad Gresh,*,† G. Andrés Cisneros,‡ Thomas A. Darden,‡ and Jean-Philip Piquemal*,§

Laboratoire de Pharmacochimie Moléculaire et Cellulaire, U648 INSERM, UFR Biomédicale, Université René-Descartes, 45, rue des Saints-Pères, 75006 Paris, France, Laboratory of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, and Laboratoire de Chimie Théorique, Université Pierre-et-Marie-Curie, UMR 7616 CNRS, case courrier 137, 4, place Jussieu, 75252 Paris, France

Received June 4, 2007

Abstract: We present an overview of the SIBFA polarizable molecular mechanics procedure, which is formulated and calibrated on the basis of quantum chemistry (QC). It embodies nonclassical effects such as electrostatic penetration, exchange-polarization, and charge transfer. We address the issues of anisotropy, nonadditivity, and transferability by performing parallel QC computations on multimolecular complexes. These encompass multiply H-bonded complexes and polycoordinated complexes of divalent cations. Recent applications to the docking of inhibitors to Zn-metalloproteins are presented next, namely metallo-β-lactamase, phosphomannoisomerase, and the nucleocapsid of the HIV-1 retrovirus. Finally, toward third-generation intermolecular potentials based on density fitting, we present the development of a novel methodology, the Gaussian electrostatic model (GEM), which relies on ab initio-derived fragment electron densities to compute the components of the total interaction energy. As GEM offers the possibility of a continuous electrostatic model going from distributed multipoles to densities, it allows an inclusion of short-range quantum effects in the molecular mechanics energies. The perspectives of an integrated SIBFA/GEM/QM procedure are discussed.

Introduction

The realm of applications of computational chemistry is considerably expanding owing to steady advances in computer power. This benefits high-level ab initio and DFT quantum chemistry (QC) as well as molecular mechanics (MM) and dynamics (MD). It is anticipated that complexes of many hundreds of thousands of atoms will soon lend themselves to MM/MD simulations. This is a compelling incentive for refining the interaction energy potential. One example is provided by the docking of competing drugs or inhibitors in the recognition site of a protein or nucleic acid target. The correct ranking of the drugs in terms of their relative affinities depends upon binding energy differences that can be smaller than the relative errors in the interaction energies ΔE_{int}: it is therefore critical to reduce the margins of uncertainty by refining ΔE_{int}. The most sought-after refinement is by explicit addition of a polarization energy contribution, E_{pol}, to integrate the principal determinant of nonadditivity. The development of 'polarizable' molecular mechanics (PMM) is presently the object of intense efforts worldwide, as attested by the publication of review papers

* Corresponding author e-mail: nohad.gresh@univ-paris5.fr (N.G.), jpp@lct.jussieu.fr (J.-P.P.).

† Université René-Descartes.

‡ National Institute of Environmental Health Sciences.

§ Université Pierre-et-Marie-Curie.

10.1021/ct700134r CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/27/2007

on a nearly yearly basis since 2001,[1−5] and the present dedicated volume.

Inclusion of an explicit E_{pol} contribution to compute interaction energies between small molecules of biological interest in the gas phase was pioneered in the mid-1960s by Claverie, Rein, and co-workers, giving rise to the so-called 'monopole-bond polarizability approximation', which used CNDO/2 derived atomic charges to compute the electrostatic contribution E_{el} and the polarizing field.[6] An MM formulation and integration of E_{pol} to compute protein−ligand-solvent complexes was due to Warshel and Levitt in the very first implementation of QM/MM methodology.[7a] This work introduced the evaluation of E_{pol} in condensed phases taking into account iteratively the interaction between the induced dipoles of all the molecules in the system. The impact of E_{pol} on ΔE_{int} can be essential not only in complexes with one or more charged species but also in multiply H-bonded complexes, as exemplified by simulations of water clusters or liquid water (ref 8 and references therein).

E_{pol} is generally determined by computing induced dipoles with distributed polarizabilities. Although this list is not exhaustive, and apart from SIBFA,[9] this is done by the MOLARIS,[7b] EFP,[10] ORIENT,[11] ASP-W,[12] SDFFIII,[13] NEMO,[14,15] OPEP,[16] AMOEBA,[8a,b] AMBER,[17] TCPE,[18] Langlet et al.,[19] and Dang-Chang[20] potentials. The polarizabilities are either scalar or tensor quantities. E_{pol} can also be computed in the context of fluctuating charge models[21] or, more recently, using the Drude model.[22] The electrostatic field is screened in several MM potentials. This was done for the first time in ref 7a. SIBFA resorts to a screening by means of a Gaussian damping function. Other potentials resort to a formalism due to Thole[23a,b] or to an alternative Gaussian framework.[23c−e]

At this point it is important to recall that MM refinements have also borne on the other ΔE_{int} MM contributions. The most important ones bore on the electrostatic contribution E_{el} upon implementing higher-order distributed multipoles (see refs 9−16 and 19 and ref 5 for discussion). We will denote below by the acronym APMM (anisotropic polarizable molecular mechanics) MM procedures which resort to distributed multipoles to compute E_{el}, on account of the strong anisotropy features that they confer to it.

As stressed in our previous review papers[5,24] a molecular mechanics methodology aiming to reproduce QC results should have the following features:

(1) Separability. The intermolecular interaction energy ΔE_{int} should be expressed under the form of distinct separate contributions. Each contribution should be formulated and calibrated in order to closely reproduce its QC counterpart obtained from energy-decomposition analyses.[25]

(2) Anisotropy. ΔE_{int} and its individual contributions should be able to reproduce the fine angular features of their QC counterparts, upon performing in- and out-of-plane variations in the approach of one molecule to another.

(3) Nonadditivity. ΔE_{int} and its individual contributions must be able to mirror the extent of nonadditivities of their QC counterparts upon passing from bi- to multimolecular complexes. In the latter, the total interaction energies can differ substantially from the corresponding summed pairwise interactions, being either larger or smaller in magnitude, namely in cooperative as opposed to anticooperative complexes respectively.

(4) Transferability. The MM potential having been calibrated on a limited training set to reproduce QC results should then be validated on a diversity of bimolecular complexes and then on multimolecular complexes without having to alter the initial calibration. Upon passing to flexible molecules, it should be able to address the issue of multipole transferability that was raised by Faerman and Price.[26]

Separability of ΔE_{int} into five distinct contributions is an essential feature of the SIBFA procedure. In the present review, following the Methods section, we will investigate the extent to which requisites 2−4 above are met. This will be followed by presentations of recent SIBFA applications to molecular recognition problems.

The last section will summarize the recent advances in the development of the Gaussian electrostatic model (GEM), a force field based on density fitting.[27] This method resorts to Hermite Gaussian densities derived from ab initio calculations on molecules or molecular fragments. These densities constitute a continuous electrostatic model connecting distributed multipoles and electron densities.[27] They are used instead of the distributed multipoles at all levels allowing a direct inclusion of short-range quantum effects by means of the computation of electrostatic and repulsion integrals. Thus GEM takes into account nonclassical contributions such as the penetration energy and enables the computation of the main overlap-dependent contribution, namely short-range exchange-repulsion. As the polarization and charge-transfer contributions have been coded in the spirit of SIBFA, the use of such fitted Hermite Gaussian densities[27] should lead to further integration and merging of SIBFA and GEM toward third-generation molecular mechanics potentials.

Formulation of the SIBFA Procedure. The SIBFA intermolecular interaction energy is formulated as a sum of five contributions

$$\Delta E_{int} = E_{MTP} + E_{rep} + E_{pol} + E_{ct} + E_{disp} \qquad (1)$$

denoting respectively the electrostatic multipolar (E_{MTP}), short-range repulsion (E_{rep}), polarization (E_{pol}), charge-transfer (E_{ct}), and dispersion (E_{disp}) contributions. The analytical forms of these contributions are given in the original papers,[9,28] and we only review here their essential features.

Electrostatic from Distributed Multipoles. *Inclusion of Penetration Effects.* E_{MTP} is computed with multipoles (up to quadrupoles) that are distributed on the atoms and bond barycenters. They are extracted from the molecular orbitals (MOs) of a given molecule or molecular fragment by a procedure developed by Vigné-Maeder and Claverie.[29] The derivation of distributed multipoles was pioneered in the early 1970s by Dreyfus and Claverie concerning ab initio MOs[30] and by Rein concerning MOs resulting from Iterative Extended Huckel Theory computations.[31] It is useful to recall in the present context that the first applications of ab initio distributed multipoles to compute gas-phase ΔE_{int} in biologically relevant complexes[32] had been published in 1979−1982, where the Dreyfus-Claverie procedure was used. The methodology was employed on the following molecular recognition problems: the preferential Ca(II) versus Mg(II) binding in 1:2 complexes with the polar head of an anionic phospholipid, phosphatidyl serine;[33] the preferential binding of tetramethylammonium versus monomethylammonium in the binding site of a phosphorylcholine antibody;[34] the binding of nucleic acid bases by amino acid side chains;[35] and cation-selective binding by valinomycin,[36] nonactin,[37] and calcimycin[38] ionophores. In its latest refinements, E_{MTP} has been augmented with an explicit penetration term, E_{pen}.[39] This was shown to afford for a closer match to the Coulomb contribution, E_C, which is obtained from energy-decompositions analyses of the ab initio intermolecular interaction energies. Together with the developments by Vigné-Maeder and Claverie,[29] important advances to derive ab initio multipoles from ab initio QC MOs were pioneered in the early 1980s due to contributions of the groups of Stone et al.,[40] Pullman et al.,[41] Sokalski et al.,[42] and Karlstrom et al.[43] An interesting development is the availability on the Web of the OPEP suite of Fortran programs, interfaced to a user-friendly package to derive both distributed multipoles and polarizabilities.[16a] It can be also noted that promising results have been obtained using Bader's Atom in Molecules[16c] approach by Popelier et al.[16b] and as we will discuss latter using density fitting techniques.[27b] However, apart from ref 39, the sole other explicit introduction of E_{pen} into a multipoles treatment was within the context of the effective fragment potential (EFP) methodology[44a,b] implemented in GAMESS.[44c]

Short-Range Exchange-Repulsion. E_{rep} is formulated as a sum of bond−bond, bond−lone pair, and lone pair−lone pair interactions. An S^2/R representation has been used since 1994[28b−e] following earlier proposals by Murrell and Teixeira-Dias.[45] Here S denotes an approximation of the overlap between localized MOs (LMOs) of the interacting partners. Hybridization is on chemical bonds as well as on the lone pairs. R is the distance between the LMO centroids. Following the E_{MTP} refinements with inclusion of the E_{pen} term, E_{rep} is augmented with an S^2/R^2 term.[28e,39b]

Consistent Treatment of Induction: Polarization, Exchange-Polarization, and Charge-Transfer Energies. In SIBFA, the induction is equivalent to the HF or DFT E_{deloc} contribution (see ref 25e and references therein).

In E_{pol} the polarizing field is computed with the same permanent multipoles as E_{MTP}. The field is screened by a Gaussian function that depends on the distance between the two interacting centers. Such a screening embodies part of short-range effects including exchange-polarization.[23d] The contribution of the induced dipoles to the field is computed by a self-consistent iterative procedure. Since 1991, the polarizabilities are tensors that are distributed on the bond barycenters and on the heteroatom lone pairs and are derived from the LMOs of the considered molecule or molecular fragment by a procedure due to Garmer and Stevens.[46] As such, both distributed multipoles and polarizabilities can be obtained from one ab initio computation performed on a molecule or constitutive molecular fragment. Each molecular entity is stored in the SIBFA library of fragments and used for subsequent assembly of molecules or molecular complexes. Usually extracted from GAMESS[44c] computations at the HF level, they can also be calculated at the DFT level.[25d]

E_{ct} is derived from the development of a formula due to Murrell et al.[47] This contribution was explicitly integrated into ΔE_{int} in 1982−1986.[48,28a] A coupling with electrostatics was subsequently introduced.[28b] That is, the ionization potential, I_A, of the electron donor, on the one hand, and the electron affinity, A_M, and 'self-potential', V_M, of the electron acceptor, on the other hand, are modified by the electrostatic potential that each undergoes in the complex. These include the effect of the induced dipoles along with those of the permanent multipoles, thereby introducing a coupling with polarization. Such modifications of I_A, A_M, and V_M were essential to account for the very strong anticooperative character of E_{ct} in polycoordinated complexes of divalent cations.

To ensure for a correct inclusion of second-order polarization effects, both E_{pol} and E_{ct} components are fitted upon their RVS[25b] or CSOV[25c−e] counterparts as the two approaches do not violate the Pauli principle conserving antisymmetrized wave functions.[23d,25d] Concerning the polarization, one can compare its first iteration directly to the RVS results. Furthermore, the fully relaxed SIBFA energy can be related to the fully relaxed Morokuma polarization[25a] even though the latter approach does not embody exchange-polarization and can be seen has an upper bound to the polarization energy.[23d]

$$E_{pol}(\mathrm{SIBFA}) + E_{ct}(\mathrm{SIBFA}) \sim E_{deloc}(\mathrm{HF/DFT}) = \Delta E(\mathrm{HF/DFT}) - E_c - E_{exch-rep}$$

$$E_{pol}(\mathrm{SIBFA, prior\ to\ iterating}) \sim E_{pol}(\mathrm{RVS/CSOV})$$

Dispersion and Exchange-Dispersion Components. Finally, E_{disp} is computed as a sum of $1/R^6$, $1/R^8$, and $1/R^{10}$ terms.[49] Directionality effects are accounted for by the introduction of lone-pairs under the form of fictitious atoms. An exchange-dispersion term was also introduced. For H-bonded complexes, E_{disp} was initially calibrated on the basis of symmetry-adapted perturbation theory (SAPT)[25f] energy-decomposition analyses.

Treatment of Flexible Molecules. A flexible molecule is assembled from its constitutive rigid fragments. Following the procedure published in ref 9, the intramolecular (conformational) energy is computed as the sum of all intermolecular, interfragment interactions, using a formulation related to eq 1. Two successive fragments are connected along X−H and H−Y bonds, where X and Y denote heavy atoms. Conformational changes take place by rotations around junction bond X−Y. The multipoles of the H atoms and of the barycenters of the X−H and H−Y bonds that belonged to the upstream and the downstream fragments, respectively, disappear and are redistributed on three centers: atoms X and Y and the midpoint of the newly formed X−Y bond. SIBFA was originally validated by comparisons with QC in a series of conformational studies of small organic molecules.[9,50] The 1985 paper[50] reported gas-phase conformational studies of the Gly and Ala dipeptides and comparisons with QC computations done in parallel on representative conformers. Ensuring consistency with gas-phase QC results is a requisite prior to simulations on larger systems and accounting for solvation effects. Reference 50 constituted to our knowledge the very first such study on peptides that used distributed multipoles and polarizabilities. This is worth recalling at this point, in view of the anticipated surge of such studies that should now resort to this kind of approach.

Calculation of Solvation Energies ΔG_{solv}. ΔG_{solv} is computed using the Langlet-Claverie (LC)[51a] procedure interfaced in SIBFA.[51b] It is formulated as a sum of electrostatic, polarization, repulsion, dispersion, and cavitation contributions. The electrostatic term is the energy due to the interaction between the electrostatic potential V created by the distributed multipoles of the solute and a fictitious charge density distributed on the cavity surface S. The charge density at a given point of S is a function of the solvent dielectric constant and of the scalar product of the electric field due the solute multipoles and of the unitary vector normal to the surface at that point. The polarization energy of each solute polarizable center is a function of its polarizability and the square of the reaction field created on that center by the charge density. Following the derivation by Huron and Claverie,[51c,d] the repulsion and dispersion terms are computed as sums of repulsion and dispersion energy volume integrals. The sums run on the solute atoms i, on the one hand, and on the solvent types of atoms j, on the other hand. The cavitation energy is computed as a sum of contributions from intersecting spheres, centered on the solute atoms. Following a formulation due to Pierotti,[51e] it is a function of a quantity d, which is the sum of the diameters of the considered atom-centered sphere and of the solvent sphere.

The possibility of constructing large, flexible molecules upon resorting to the multipoles and polarizabilities of their constitutive fragments enabled the addressing of a diversity of molecular recognition problems in 1985−1990. These bore on complexes of DNA with nonintercalating ligands[52] as well as intercalating drugs,[53] complexes of calmodulin central helix with phenothiazine drugs,[54] and selective binding of metal cations and biogenic amines by ionophores.[55] Subsequently, the availability of the restricted variational space analysis (RVS) procedure[25b] was instrumental to enable refinements of the E_{rep}, E_{pol}, and E_{ct} contributions. Together with the integration of the Langlet-Claverie continuum reaction field procedure to compute ΔG_{solv} using distributed ab initio multipoles, these have in turn enabled performing energy balances for the complexes of inhibitors with Zn-metalloenzymes. This was earlier exemplified in studies of the complexes of thermolysin with mercaptocarboxylate and phosphoramidate inhibitors.[56] The need for a balanced treatment of solvation and interaction energies was emphasized as early as 1976,[7a] and treatments in the context of classical electrostatics encompassing solvent effects were developed by Warshel and co-workers.[7c] The last section of this review paper will summarize some of the most recent applications in this domain.

Further Refinements. *(a) Quadrupolar Polarizability and Back-Donation Charge Transfer.* Significant improvements in the representation of the monovalent Cu(I) cation were as follows:[57] the inclusion of its quadrupolar polarizability (QP) in addition to the dipolar one, to express the additional dependency of Cu(I) polarization energy upon the gradient of the electrostatic field; and the inclusion of charge transfer from Cu(I) to its ligands, in addition to the one taking place from the ligands to the cation.

(b) Handling of Open-Shell Metal Cations. Significant progress to represent open-shell metal cations took place in 2003,[58] upon integrating ligand field (LF) effects in SIBFA using an effective Hamiltonian in the framework of the angular overlap model (AOM).[59] The SIBFA-LF procedure was applied to polyligated Cu(II) complexes and was shown to enable close reproductions of QC calculations. An essential result was the preferential stabilization of square-planar arrangements in tetraligated Cu(II) complexes, in marked contrast to the tetrahedral arrangements preferentially stabilized in tetraligated Zn(II) complexes.[28e]

Results and Discussion

I. Are the Essential Features of the QC Contributions Reproduced? Most validation computations reported in this paper have resorted to the CEP 4-31G(2d) basis set due to Stevens et al.[60] This ensures for consistency, since the distributed multipoles and polarizabilities were derived from QC computations on the fragments that used this very basis set. Furthermore, we have observed extremely close correlations between results obtained with this basis set and those obtained with more extended basis sets, such as the 6-311G** or LACV3P** ones. This is illustrated in the present paper in the case of complexes of two Zn-metalloenzymes, β-lactamase and phosphomannoisomerase, with their inhibitors. Thus as commented later in this paper we could observe persistent parallelisms in the evolutions of ΔE(QC) as a function of the structure of the competing inhibitor-metalloenzyme model complexes as well as closely similar magnitudes in the CEP 4-31G(2d) versus LACV3P** ΔE(QC) values.

(1) Anisotropy. The anisotropy features are illustrated below upon monitoring the angular dependencies of QC versus SIBFA energy contributions in two representative examples. The first is the complex of methanethiolate with the Zn(II) cation, and the second is that of carboxylate with water. Methanethiolate is the side chain of deprotonated Cys residues, which constitute an essential Zn-ligating entity in proteins. It is also encountered in the structure of several Zn-metalloenzyme inhibitors. The carboxylate anion is the most ubiquitous anion in biological systems and interacts with a diversity of polar, cationic entities as well as the majority of biologically relevant metal cations. It is therefore essential to evaluate how well the orientation sensitivity of the QC energy and its contributions can be translated by their APMM counterparts. In both cases, the Zn−S$^-$ or the H(w)−O distances of approach are held fixed, and stepwise variations are done on the angle of approach θ = C−S−Zn or C−O−H(w).

(a) Zn-Methanethiolate. This complex was previously investigated in the course of the refinements of the SIBFA E_{rep}, E_{pol}, and E_{ct} contributions.[28b] We report in Supp. Info 1

Table 1. Interaction Energies (kcal/mol) in Four Cyclic Water Tetramers[a]

	a		*b*		*c*		*d*	
	ab initio	SIBFA	ab initio	SIBFA	ab initio	SIBFA	ab initio	SIBFA
E_1	+2.2	+2.2	−5.1	−5.1	−6.1	−6.6	−6.4	−6.2
E_{pol}(RVS)/E_{pol}*	−10.3 (−3.8)	−10.9 (−4.3)	−4.3 (−0.6)	−4.0 (−0.7)	−2.4 (+0.8)	−2.1 (+0.9)	−6.4 (−1.9)	−4.6 (−1.3)
E_{pol}(KM)/E_{pol}	−14.0 (−7.3)	−15.8 (−7.7)	−5.1 (−0.9)	−5.1 (−1.2)	−2.8 (+0.7)	−2.3 (+0.9)	−5.7 (−1.8)	−6.4 (−1.9)
E_{ct}	−8.0 (−0.6)	−6.6 (−1.3)	−3.5 (0.0)	−3.2 (−0.1)	−2.5 (+0.4)	−2.6 (+0.1)	−3.7 (−0.3)	−3.7 (+0.2)
δE(MP2)/E_{disp}	−11.9 (+0.1)	−11.5	−8.7 (0.0)	−7.3	−8.4 (+0.1)	−6.3	−10.2 (+0.4)	−8.5
ΔE(MP2)/ΔE_{tot}	**−30.0**	**−31.9**	**−21.9**	**−20.7**	**−19.6**	**−17.8**	**−25.9**	**−23.9**

[a] See text for definitions. Nonadditivities are given in parentheses. Negative values indicate cooperativity.

in the Supporting Information the corresponding evolutions in light of the latest refinements.[28e]

(b) Formate-Water. The angularity features of this complex have been analyzed and reported in a former study[28c] and, regarding the newest SIBFA refinements, for E_{rep} in a recent paper.[39b] For completeness, Supp. Info 2 in the Supporting Information displays the corresponding evolutions of the second-order RVS contributions and of their SIBFA counterparts. Both E_{pol} and E_{ct} now have shallower behaviors than in the methanethiolate-Zn(II) complex.

(c) Stacked Formamide Dimer. This complex is commented on in Supp. Info 3 in the Supporting Information.

The anisotropy features of E_{pol} stem from the Garmer-Stevens polarizabilities, which are tensors rather than scalars. Furthermore, heteroatoms are endowed with off-centered lone-pair polarizabilities. The corresponding E_{pol}(lp) is maximized when a polarizing center approaches closer to the location of the lone pair centroid. The necessity of off-centered as opposed to atom-centered polarizabilities was recently shown in studies of water-chain complexes designed to maximize the cooperativity response.[23d]

The energy minimizations of the multimolecular complexes reported below used the ‘Merlin’ software.[61]

(2) Nonadditivity. In multimolecular complexes, the total interaction energy is not equal to the summed pairwise intermolecular interactions between individual molecules. Thus, the magnitude of ΔE_{int} can be larger than such a sum: *cooperativity* is a feature of the majority of multiply H-bonded complexes or chains. It can, alternatively, be smaller in magnitude than it. *Anticooperative* complexes are mostly encountered in the polycoordinated complexes of a charged species, particularly in the complexes of divalent metal cations. It is critical for polarizable potentials to account equally well for both features. While this has been recognized for a long time, there have been surprisingly few QC analyses of the energy origins of nonadditivity, δE_{nadd}: i.e., to what an extent could δE_{nadd} be traced back essentially to the second-order contributions, what are the separate contributions stemming from E_{pol} and from E_{ct}, and how well could the APMM contributions reproduce the nonadditive behaviors of their QC counterparts. RVS energy-decompositions on multimolecular complexes are an invaluable asset for such a quantification.

(a) Cooperativity. QC and SIBFA studies were performed on multiply hydrogen-bonded water oligomers[62] and models of peptide H-bonded networks.[63] The amounts of QC-computed cooperativities were closely reproduced by SIBFA. RVS analyses showed δE_{nadd} to originate predominantly from E_{pol}, while E_{ct} contributed little to it, and the SIBFA analyses were fully consistent with the RVS ones.

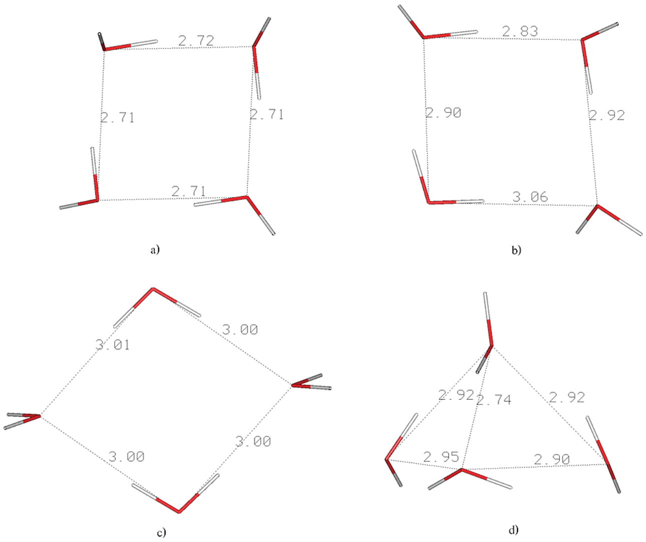


Figure 1. Representation of the four cyclic water tetramers *a*−*d*.

As an illustration, Table 1 reports a comparison between QC and SIBFA results on four cyclic water tetramers initially designed by Hodges et al.[64] and further considered by Masella et al.[62b] to probe nonadditivity from QC computations and how well these could be translated by polarizable molecular mechanics. These tetramers are represented in Figure 1. In the first, *a*, each water acts in an alternating pattern as an H-bond acceptor to one neighbor and as an H-bond donor to the other. In the second, *b*, one of these waters acts as an H-bond acceptor from both its neighbors, with one of the neighbors acting as an H-bond donor to its own two neighbors. In *c*, two opposite waters act as double H-bond donors, while the two other opposite waters act as double H-bond acceptors. *d* is an alternating three-dimensional arrangement. Table 1 shows a close numerical agreement of QC and SIBFA values in terms of total energies as well as individual contributions, the ΔE(MP2) and ΔE(SIBFA) ordering being the following: $a > d > b > c$. It is instructive to compare the amounts of anticooperativity of E_{pol} and E_{ct}, as given in parentheses in Table 1. E_{pol} is the essential determinant of nonadditivity, consistent with ref 64. Complex *c* is the sole anticooperative complex, with similar QC and SIBFA δE_{nadd} values. E_{pol}(KM) and E_{pol}(RVS) denote the values of E_{pol} that result from the Kitaura-Morokuma[25a] and the RVS[25b] energy decomposition analyses, respectively. E_{pol}* and E_{pol} denote the values of

Table 2. RVS and SIBFA Interaction Energies (kcal/mol) in Four 12−20 Water Clusters

	number of waters							
	12		16		16 (MC)		20	
	SIBFA*	RVS	SIBFA*	RVS	SIBFA*	RVS	SIBFA*	RVS
E_{MTP}*/E_c	−167.6	−168.5	−230.9	−231.4	−179.5	−179.8	−293.2	−294.3
E_{rep}*/E_{exch}	151.9	151.4	207.9	207.5	149.8	149.9	263.6	263.2
E_1	−15.8	−17.1	−23.1	−23.9	−29.7	−29.9	−30.6	−31.1
E_{pol}*/E_{pol} RVS	−30.6	−34.7	−42.0	−47.8	−32.7	−35.5		
E_{pol}/E_{pol}	−41.3	−44.7	−56.5	−61.7	−44.1	−45.1	−71.3	−78.6
E_{ct}	−22.1	−23.1	−30.2	−31.3	−22.6	−23.1	−37.3	−39.4
Δ*E*(SIBFA)/Δ*E*(RVS)	**−79.2**	**−80.1**	**−109.8**	**−110.4**	**−96.4**	**−94.8**	**−139.2**	**−139.1**

SIBFA polarization in which the polarizing field is computed with the sole permanent multipoles and with the permanent + induced dipoles, respectively. As discussed in ref 62, E_{pol}*(SIBFA) has close numerical values to E_{pol}(RVS), and E_{pol}(SIBFA) has values close to E_{pol}(KM). Such agreement also carries over to the corresponding δE_{nadd} values. E_{ct} is weakly nonadditive, its δE_{nadd} values being the largest in absolute magnitude for the most strongly bound tetramer *a*.

Table 2 reports the results of parallel RVS and SIBFA computations on four 12−20 water clusters.[39b] It is instructive to re-emphasize the impact of second-order terms in such complexes. Complexes *a, b,* and *d* are three-dimensional aggregates in three-dimensional cubic arrangements having 12, 16, and 20 water molecules, respectively, and complex *c* is a small aggregate extracted from an ongoing Monte Carlo (MC) simulation on a water box of $n = 64$ molecules. The numerical values of E_{pol}(SIBFA) outweigh those of the summed first-order contributions E_1, for which the large stabilizing values of E_{MTP} are strongly opposed by those of E_{rep}, on account of the shortening of the O−O H-bonding distances (in the 2.7−2.9 Å range) due to cooperativity. In fact, for all three cubic arrangements, *a, b,* and *d,* even E_{ct}(SIBFA) has larger absolute values than E_1. All these trends are found in the RVS computations. For all four complexes, Δ*E*(SIBFA) reproduces Δ*E*(RVS) with a relative error <2%. As in Table 1 above, a close correspondence is seen between E_{pol}(RVS) and E_{pol}*(SIBFA), on the one hand, and E_{pol}(KM) and E_{pol}(SIBFA), on the other hand. E_{pol}(KM)/E_{pol}(SIBFA) have larger magnitudes than E_{pol}(RVS)/E_{pol}*(SIBFA), a signature for cooperativity.

(b) Anticooperativity. The first concurrent RVS and SIBFA computations on polycoordinated cation complexes were performed in the course of SIBFA refinements and bore on polyhydrated complexes of Zn(II), Mg(II), Ca(II), and Cd(II).[65] These were followed by studies on polycoordinated Zn(II) complexes in ‘hard’ and ‘soft’ binding protein binding sites[56b,66a] as well as in Zn(II)-metalloenzyme sites including different inhibitor anionic moieties.[56b] The presence of two anions in these sites resulted in very large increases of the magnitudes of Δ*E* and its contributions. The SIBFA computations were nevertheless able to closely reproduce the QC Δ*E* values, in terms of both the total energies and their individual contributions. Subsequent analyses of anticooperativity were done on complexes of formate with penta- and hexahydrated Zn(II) complexes[66b] and on the above-mentioned polycoordinated Zn(II) complexes.[65,66a] In these studies the values of QC and SIBFA E_1, E_{pol}, and E_{ct} were compared to their summed values in the separate pairwise complexes that make up the multimolecular complexes. While E_1 showed very little nonadditivity, E_{pol} and mostly so E_{ct} were strongly anticooperative. It was observed that E_{pol}(SIBFA) reproduced well the anticooperativity of E_{pol}(RVS), while E_{ct}(SIBFA) somewhat overestimated that of E_{ct}(RVS), particularly upon accumulation of negatively charged ligands (up to four) in the first Zn(II) coordination shell. The anticooperativity of E_{ct}(SIBFA) could be reduced by a very simple concerted change of Zn-parameters to allow for the best match to E_{ct}(RVS) upon passing from the monoligated $[Zn-H_2O]^{2+}$ complex to the hexaligated $[Zn(H_2O)_6]^{2+}$ one (see ref 28e for details). As compared to ref 66a, this then resulted in a notably closer agreement of E_{ct}(SIBFA) values to the E_{ct}(RVS) ones in the representative complexes of Zn(II) with three and four methanethiolate ligands.[28e] This leaves open the issue of the nonadditivity of the contribution of correlation to Δ*E*, $\delta\Delta E_{corr}$(MP2), in polycoordinated Zn(II) complexes, while in contrast E_{disp}(SIBFA) is purely additive. Inclusion of triple-dipole interactions[67] could be considered in future studies to endow E_{disp}(SIBFA) with nonadditivity.

The correspondence between QC and SIBFA computations is illustrated below in two examples. The first is that of Zn(II) complexes with six water molecules, and the second is a binuclear Zn(II) complex with a metallo-*β*-lactamase binding site.

In Supp. Info 4 in the Supporting Information are reported the results of parallel QC and SIBFA computations that bore on three competing complexes of Zn(II) with six water molecules.

Binuclear Zn(II) Binding Sites. These sites constitute stringent tests for APMM procedures because dramatic enhancements of nonadditivity can be expected. This is due to the proximity of the two cations (in the 3−4.5 Å range) and to the buildup of charged and highly polarizable ligands. Previously investigated complexes[66a,28e] bore on models of Gal4, a binuclear Zn-finger with six cysteinate residues, and on Zn(II)-metallo-*β*-lactamase, an enzyme responsible for the acquired resistance of bacteria to antibiotics. High-resolution X-ray diffraction studies on the *B. fragilis* strain[68] showed the first Zn(II) to be ligated by three His side chains and a hydroxy anion, while the second was ligated by three anionic residues: the hydroxy, an aspartate, and a cysteinate as well as by one His side chain and a water molecule.

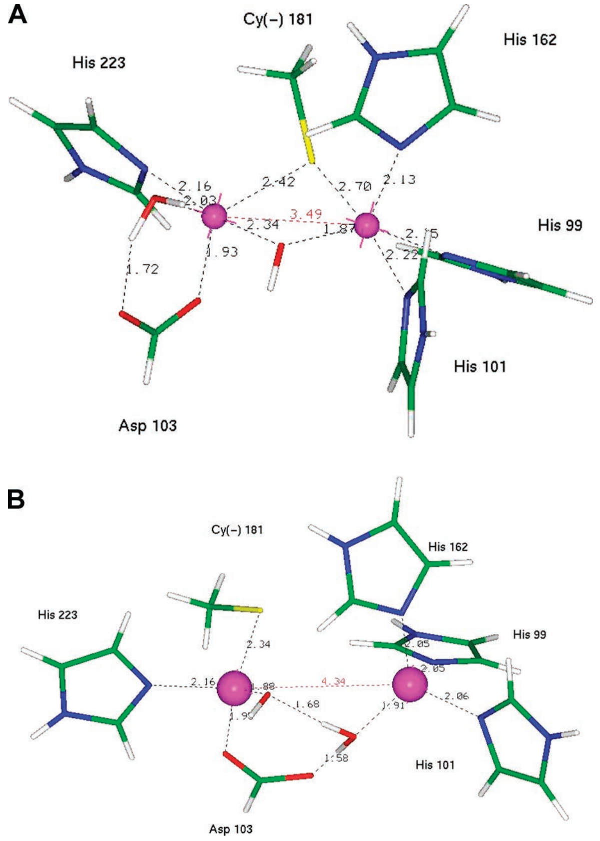


Figure 2. Representation of the complexes with two Zn(II) cations in the binding site of metallo-β-lactamase at the Zn−Zn distances of (a) 3.5 Å and (b) 4.3 Å Reprinted with permission from Gresh et al. *Journal of Computational Chemistry* **2005**, *26*, 1113. Copyright 2005 John Wiley.

Starting from the X-ray structure, SIBFA energy minimizations were performed, after constraining the Zn−Zn distances at 3.0, 3.5, and 3.8 Å (structures *a*−*c*). QC energy minimizations were subsequently performed starting from the SIBFA minima. While these confirmed the shallow dependence of ΔE upon the Zn−Zn distance that was found by SIBFA, they also derived an alternative minimum, denoted as *d*, with the two Zn cations now at >4 Å; the His-bound Zn(II) is now bound to water instead of hydroxy, as a consequence of proton transfer that took place during QC energy minimization. The other cation is now bound to all three anionic ligands and to one His side chain.[69] Complex *d* was reprocessed and energy-minimized using SIBFA and standard internal SIBFA fragment coordinates.[28e] Complexes *b* and *d* are represented in parts a and b, respectively, of Figure 2. The results of concurrent parallel SIBFA computations and RVS analyses at the SIBFA minima are reported in Table 3. In keeping with the results from the previous HF energy minimizations, the RVS analysis shows complexes *b* and *d* to have very close ΔE values, differing by 6 out of 1200 kcal/mol, namely less than 1%. Such a small difference is due to compensations of large energy differences between individual contributions. Thus E_1 favors *b* over *d* by a large amount (57 kcal/mol), while both $E_{\rm pol}$ and $E_{\rm ct}$ favor *d* over *b* by a total of 64 kcal/mol. The SIBFA computations have very close agreements with the RVS ones. These concern the numerical values of the total energies as well as of their individual contributions, the opposed trends of first- versus second-order contributions, and the *d* > *b* energy ordering. Such trends remain the same if the LACV3P** basis set[70] is used instead of the CEP 4-31G(2d) one as well as upon going to correlated levels, namely, DFT, LMP2,[71] or MP2.

(3) Transferability. Interactions Involving Flexible Molecules. There are several aspects to transferability. The first is the need for a molecular mechanics potential to be applied on a diversity of complexes other than the ‘training set’ on which it was initially calibrated. The separability feature of an APMM potential, whose individual contributions are each formulated on the basis of quantum chemistry, should, if their formulations are correct, ensure such transferability. Thus, e.g., if water is properly calibrated on the basis of a limited training set of water dimer complexes, it should be possible to subsequently investigate not only all possible water dimer complexes but also water oligomers of virtually any size as well. Extension of the calibration to any other chemical entity should enable the investigation of all possible complexes that involve this entity in combination with all other ones present in the library. In SIBFA, such ‘entities’ are the constitutive molecular fragments with their internal geometries and distributed multipoles and polarizabilities, which are stored in a library of fragments. Another aspect of transferability relates to the recurrence of well-defined atomic ‘species’ within the molecular fragments. Each atom is identified according to its hybridization state, the number and nature of its neighbors, and the net charge and type of fragment to which it belongs. As an example, O atoms can be assigned as belonging to a hydroxyl or ether-like group, to a carbonyl, a carboxylate, a phosphate, or to a methoxy group, etc. According to its class, a given O is given effective radii for $E_{\rm rep}$, $E_{\rm pol}$, $E_{\rm ct}$, and $E_{\rm disp}$. These radii are calibrated once and for all to reproduce the radial behavior of the corresponding RVS contribution on a model bimolecular complex. There is a third aspect to transferability that is critical to handling flexible molecules of arbitrarily large size, ranging from pharmacologically relevant ligands up to macromolecules. Such molecules are assembled from their constitutive fragments given the knowledge of the sequence, the length of the junction bond, and the torsion angle along that bond. The multipoles are redistributed along the junction bond following a procedure published in ref 9. This gives rise to the following issue: what is the loss of accuracy due to assembling. That is, is it possible to account in terms of interaction energies for the fact that the multipoles on the fragments undergo changes in their intensities upon integration in a large molecule? With the increase of computer power, it becomes now possible to perform an ab initio computation on large molecular entities of 200 atoms and more and derive their distributed multipoles and polarizabilities. Denoting by A-B a saturated chemical bond between heavy atoms A and B, a large molecule can be subsequently split into smaller fragments by breaking bond A-B and replacing it by two junction bonds A-H* and H*-B, with two fictitious hydrogen atoms H* having null multipoles along the A-B direction, the A-B distance being the same as in bond A-B. This enables for rotations around A-B of the two newly created fragments. How then to energetically

Table 3. Interaction Energies (kcal/mol) in the β-Lactamase Binding Sites[a]

	a		*b*		*c*		*d*	
	ab initio	SIBFA	ab initio	SIBFA	ab initio	SIBFA	ab initio	SIBFA
E_c/E_{MTP}	−1351.8	−1373.4	−1346.3	−1367.1	−1330.7	−1364.7	−1321.0	−1345.4
E_{exch}/E_{rep}	362.3	393.9	344.3	370.0	350.4	390.5	375.9	398.8
$\boldsymbol{E_1}$	**−989.5**	**−979.5**	**−1002.0**	**−996.2**	**−980.4**	**−974.2**	**−945.1**	**−946.6**
E_{pol}(RVS)/E_{pol}*	−223.9	−224.9	−203.3	−202.5	−209.9	−216.6	−252.9	−250.2
E_{pol}(HF)/E_{pol}	−184.9	−165.7	−173.6	−152.4	−185.6	−172.9	−216.9	−199.2
E_{pol}(Zn(II))	−6.1	−3.7	−6.0	−3.6	−7.8	−5.4	−8.0	−3.4
E_{ct}	−56.8	−65.5	−57.2	−66.0	−60.9	−61.7	−75.2	−70.6
E_{ct^*}	−35.7		−36.5		−40.1		−56.3	
BSSE	−21.1		−20.7		−20.8		−19.0	
$\boldsymbol{\Delta E}$	**−1210.2**	**−1207.0**	**−1212.1**	**−1211.9**	**−1206.0**	**−1203.4**	**−1218.8**	**−1213.0**
$\boldsymbol{\Delta E}$**(MP2)/**$\boldsymbol{\Delta E_{tot}}$	**−1327.6**	**−1324.0**	**−1324.3**	**−1323.5**	**−1313.5**	**−1311.2**	**−1325.7**	**−1325.1**
δE(MP2)/E_{disp}	−117.4	−116.1	−112.2	−110.8	−107.5	−107.1	−106.9	−111.6
ΔE(HF/LACV3P**)	−1241.0		−1242.6		−1237.4		−1248.3	
ΔE(LMP2)	−1270.5		−1270.6		−1270.2		−1272.5	
δE(LMP2)	−29.5		−28.0		−32.8		−24.2	
ΔE(B3LYP/LACV3P**)	−1292.1		−1292.7		−1284.9		−1296.6	

[a] *a*−*c*: standard complexes from the *B. fragilis* binding site; *d*: complex derived from HF energy minimization. In *a*−*c*, the Zn−Zn distances are 3.0, 3.5, and 3.8 Å, respectively. In *d*, the Zn−Zn distance is 4.3 Å. The electrostatic potential used in the computation of E_{ct} is computed with a full multipolar expansion and with the induced dipoles. In ref 28e, it was mistakenly limited to the sole monopoles.

account for the fact that the multipoles on the fragments undergo changes in their intensities upon conformational changes so as not to bias any particular set of conformers. Such an issue was raised for the first time by Faerman and Price[26] upon constructing oligopeptides from the multipoles of their constitutive fragments.

In perturbation or variation theories, the impact of changes of multipole intensities due to complex formation is translated by the second-order contributions E_{pol} and E_{ct}, while the first-order electrostatic contribution is computed with the multipolar distributions that retain the intensities they have in the isolated molecule or molecular fragment. The electrostatic field giving rise to the polarization contribution is itself computed with the permanent multipolar distribution augmented with induced dipoles derived by a self-consistent iterative procedure. We have extended this representation to the case of intramolecular interactions. Since the inception of the SIBFA procedure,[9] these are computed as the sum of intermolecular interactions between the constitutive fragments of the molecule. In the procedure that is presently used, E_{MTP} is computed with junction multipoles that are redistributed along the junction bond, namely its origin, its extremity, and its barycenter. These junction multipoles do not interact with the two connected fragments, since such interactions are large and constant. To compute E_{pol}, on the other hand, an alternative set of multipoles is used, for which no redistribution along the junctions is done. In this fashion, each individual fragment retains the net charge it has prior to the assembling procedure, namely 0 if neutral, −1 if anionic, and 1 if cationic, whereas it is not retained following redistribution. This prevents an imbalance of E_{pol} between two successive fragments that have lost their net charges, and that could be amplified in the complete molecule due to the nonadditivity of E_{pol}. It was also necessary to prevent overlaps involving the H atoms belonging to the X−H junction bonds. Such bonds were shrunk by carrying back the end H atoms on the X atom whence the bond originates. Finally, upon computing the intermolecular interactions between flexible molecules, *inter*- and *intra*molecular *inter*-fragment interactions have to be computed simultaneously and consistently as a single integrated energy. This need is a consequence of the nonadditivity of E_{pol} and E_{ct}. It illustrates the connections between nonadditivity and transferability. ΔE_{int} between two or more interacting molecules can be subsequently derived by subtracting from such a total energy all sums of interfragment interactions within each individual molecule.

An illustration of the manner flexible molecules are constructed from their fragments is given in Figure 3a,b. Parts a and b relate respectively to the assembly of the five first amino acids of protein Fak (focal adhesion kinase), a target for the design of antitumor drugs, and of an inhibitor belonging to the pyrrolopyrimidine series (de Courcy et al., to be published). Part a represents the first ten fragments making up the backbone. The side chains are assembled after completion of the 140 amino acid backbone. Thus Asp414 is built out from its methane and formate fragments, Tyr415 from methane, benzene, and phenol, etc. Part a also gives the numbering of the atoms that takes into account the presence of the additional centers along the chemical bonds. All individual peptide and nucleic acid fragments being stored in a library with the relevant information concerning the internal geometry, the types of atoms, the distributed multipoles and polarizabilities on proteins and nucleic acids can be constructed using software that uses in addition the information regarding the sequence and torsional angles. Part b shows the inhibitor as constructed from its constitutive pyrimidine, sp2 amine, benzene, water, methane, and formate fragments. To account for conjugation effects, a prior QC computation was performed on an aminopyrimidine molecule, which was then broken up into pyrimidine and HNH_2, these two entities retaining the same multipolar expansion as in the original molecule, the fictitious H atoms on their junctions having null multipoles, and while the junction

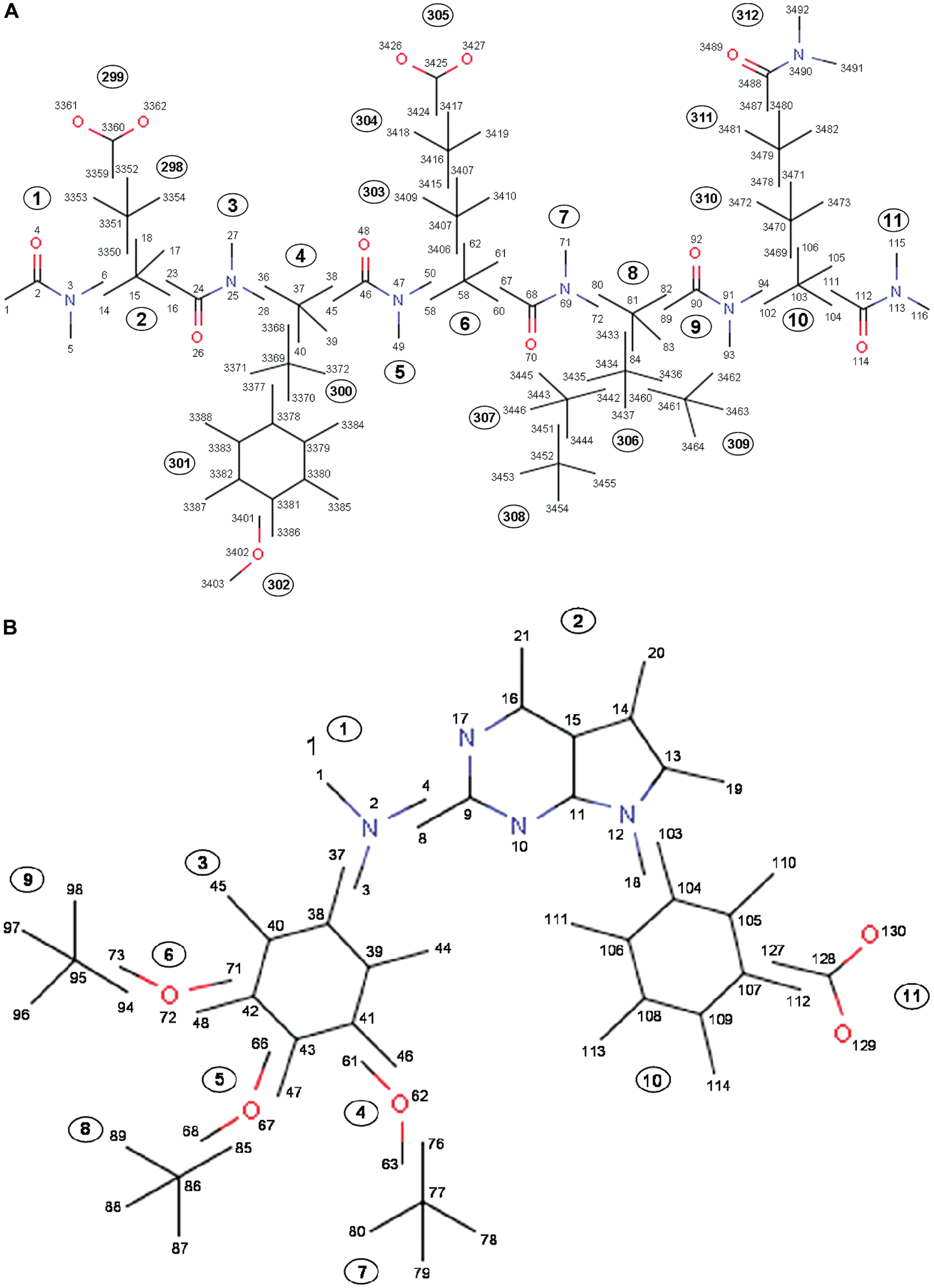


Figure 3. (a) Fragments making up the backbone of the five first amino acids of protein Fak (focal adhesion kinase). (b) Fak protein inhibitor as constructed from its constitutive pyrimidine, sp2 amine, benzene, water, methane, and formate fragments.

bonds CH and HN have each half of the multipoles of the broken C−N bond. The other fragments already belong to the library of SIBFA fragments. Thus a new molecule can be constructed from fragments that are already present in the library. If this is not the case, a QC computation is done on it enabling to derive its distributed multipoles and polarizabilities, and the fragment can be stored for future uses. Most QC computations are done with the GAMESS package.[44c] In the general case, SIBFA energy minimizations are done in internal coordinates. Conformational changes thus take place by torsions around the junction bonds. The approach of a given molecule toward another is governed by six intermolecular variables. Molecular dynamics are done in Cartesian coordinates, while standard bond lengths and valence angles are enforced by stretching and bending harmonic restraints.

Our first studies on the intermolecular interactions of flexible molecules bore on the complexes of Zn(II) with glycine and the glycine zwitterion,[72] on the one hand, and with α- and β-mercaptocarboxamides, on the other hand.[73] The latter constitute the Zn-binding moieties of several potent Zn-metalloenzyme inhibitors.[74] Following the procedures outlined above, it was possible to closely reproduce the QC values of Zn(II) binding in different configurations of approach or as a function of the zwitterionic state[72] and its conformational dependencies.[73] These studies were extended to complexes of Cu(I) with flexible molecules involved in the formation of supramolecular assemblies[57] and to those of Cu(II) with a new class of HIV-1 inhibitors that can fit the protease dimer binding site.[75] SIBFA was also used to study of the conformation-dependent intermolecular interactions of the triphosphate anion, the tetra-anionic end of ATP, with Zn(II) used as a probe.[76] The results are commented on in Supp. Info 5 in the Supporting Information. We next considered the high-resolution X-ray structure of the complex of HPPK with a nonhydrolyzable ATP analog, that has one central ester O replaced by methylene.[77] The results are commented on in Supp. Info 6 in the Supporting Information.

Conformational Studies of Oligopeptides. Test on the Alanine Tetrapeptide. Most previous analyses of transferability had borne on charged flexible ligands and their interactions with divalent cations. The predominant effects of divalent cation binding on the ligands involved the polar/charged heteroatoms and their connecting bonds, since these were the most exposed to the incoming cation and involved simultaneously the mutual interactions between these sites. The junction bonds, being less accessible, were expected to play a lesser role. It was then important to evaluate the impact of the approximations done for the handling of the interfragment junctions in the case of neutral molecules and in the absence of external charge. This is exemplified by the case of oligopeptides of alanine, which has the simplest side chain, namely a methyl group. The oligopeptide backbones are assembled in SIBFA as a succession of formamides and methyl groups, and the Ala side chain is represented by a methyl group.[50] For pure intramolecular interactions, the interactions involving the junction bonds are expected to have weights comparable to those involving the nonjunction bonds or the atoms. For the evaluation of the SIBFA conformational energies in such molecules, we have in ref 78 computed the energies of ten alanine tetrapeptide conformers, that were used by Beachy et al.[79a] to benchmark standard molecular mechanics potentials against ab initio computations. The structures of these ten conformers are recalled in Supp. Info 7 in the Supporting Information. Starting from these, SIBFA energy minimizations were performed as a function of the ϕ, ψ, and χ dihedral angles with fixed standard internal coordinates. At the converged minima, single-point QC computations were performed with three different basis sets: CEP 4-31G(2d), 6-311G**, and cc-pvtz(-f). The results are reported in Table 4a,b. Table 4a reports the QC results at the HF level and the SIBFA ones in the absence of the E_{disp} contribution. Table 4b reports the results in the presence of correlation, namely at the DFT level with different functionals for the exchange-correlation terms, namely Becke88/Perdew 86,[80] PLAP3,[81] K2-BVWN,[82] and B3LYP;[83] at the LMP2 level;[71] and at the MP2 level. The SIBFA results are given with two different scalings of E_{disp} by 1.0 and by 0.8. The latter value was previously found[28d] to enable the reproduction by ΔE_{tot}(SIBFA) of the −5.1 kcal/mol water−water dimerization energy that resulted from a large basis set MP2 study of this dimer by Feyereisen et al.,[79b] with ΔE(SIBFA) in the absence of E_{disp} providing a very close agreement to the corresponding HF value by these authors (−3.9 versus −3.6 kcal/mol, respectively). Table 4a shows δE(SIBFA) to give the same ordering of conformer stability as the CEP 4-31G(2d) and 6-311G** basis sets. The δE(SIBFA) values are close to those found with the CEP 4-31G(2d) basis set. Such agreements also carry out to the cc-pVTZ(-f) basis set, with a maximal error of 1.4 kcal/mol for high-lying conformer 7, and an rms of 0.7 kcal/mol. These results indicate that the introduction of E_{pol} in pure intramolecular interaction energies, with the same calibration as for intermolecular interactions, could be realized in a balanced fashion, without overestimating the stabilities of the most folded conformers. The values of δE^*(SIBFA), namely without E_{pol}, have a downgraded agreement with the δE(QC) ones. The values of δE_{mono}(SIBFA), computed by limiting E_{MTP} to the sole monopole−monopole term, even though in the presence of E_{pol}, have an even worse agreement. Thus, in the framework of SIBFA, explicit introduction of a polarization contribution is clearly insufficient to restore the agreement with QC computations if the electrostatic contribution were to be limited to the sole monopole−monopole term. Table 4b shows that correlation brings a reduction of the δE values, the folded conformations having their relative stabilities improved with respect to the extended ones. However, the extent of δE reduction depends upon the procedure, the basis sets, and, for the DFT computations, upon the exchange-correlation functional as well. The LMP2 computations bring δE reductions that are intermediate between the DFT and MP2 ones. The results of Table 4b were commented on in more detail in ref 78. It is observed that the δE_{tot}(SIBFA) values with a scaling of 0.8 for E_{disp} (conform to the value adopted in ref 28d concerning the water dimer) agree best with the 6-311G** LMP2 calculations, with which they give a 1.3 kcal/mol rms. It is presently difficult to trace back to a specific contribution the origin of the 0.7−1.3 kcal/mol rms increase upon passing from the uncorrelated to correlated levels, since there are no QC energy-decomposition analyses for intramolecular interactions. It could be instructive in future calculations to resort to correlated rather than uncorrelated multipoles and polarizabilities to construct the fragments as recently initiated for intermolecular interaction energies.[39b] The results of Table 4 should not be compared to those published by Beachy et

Table 4. Ala Tetrapeptide: (a) Values of the HF and SIBFA (without the Dispersion Component) Conformational Energy Differences δE and (b) Values of the DFT, LMP2, MP2 Quantum-Chemical, and SIBFA Conformational Energy Differences δE^a

(a)

	ab initio HF			SIBFA		
conformer	4-31G(2d)	6-311G**	cc	δE	δE^b	δE_{mono}
1	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0	1.1	1.3	0.8	1.0	−0.7
3	10.2	9.3	10.5	11.3	14.9	6.1
4	3.2	3.1	3.4	3.3	3.2	1.9
5	7.3	7.3	7.6	7.7	6.5	6.3
6	7.5	6.1	7.7	8.1	9.8	9.2
7	13.4	12.2	13.7	12.3	13.4	12.4
8	17.6	16.2	17.8	18.9	21.8	22.0
9	30.0	28.4	29.8	29.6	35.7	27.4
10	28.2	26.6	28.5	28.9	34.6	38.1

(b)

conformer	DFT B88/PD86	PLAP3	K2	B3LYP/6-311G**	B3LYP/ cc	LMP2 6-311G**	cc	MP2 6-311G**	SIBFA *c*	SIBFA *d*
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.7	1.0	0.7	0.7	1.0	0.6	3.4	−0.1	0.3	0.4
3	6.9	11.4	7.7	6.6	7.7	5.8	10.8	1.5	3.4	4.8
4	3.5	3.7	2.8	2.7	3.0	1.9	2.1	1.1	2.3	2.5
5	7.7	8.1	7.0	7.2	7.3	5.8	5.8	4.4	6.5	6.7
6	7.2	9.5	6.8	5.4	6.6	3.9	4.2	0.8	3.6	4.5
7	11.5	14.7	11.2	9.8	11.1	7.5	9.7	3.1	5.2	6.6
8	15.3	19.9	15.0	13.6	15.0	11.7	16.3	6.6	10.7	12.1
9	22.3	33.0	23.9	22.5	23.7	21.5	28.3	15.4	17.5	19.8
10	24.0	32.5	24.0	21.4	23.6	17.6	20.2	10.7	17.5	19.7

[a] Single-point ab initio computations are performed on the SIBFA minima. The δE values (kcal/mol) are computed with respect to the energy of the most stable conformer taken as energy zero. [b] δE: SIBFA energy value in the absence of E_{pol}. [c] A multiplicative factor of 1 is used for the E_{disp} component. [d] A multiplicative factor of 0.8 is used for the E_{disp} component.

al. since as mentioned above, energy minimization was only along the torsion angles, while valence angles and bond lengths were not relaxed. Toward this aim, angle bending and bond stretching force constants have to be recalibrated in the framework of SIBFA. Because the formulation of the energy is different than in standard molecular mechanics procedures, such constants can significantly differ from the ‘classical’ ones. This was actually undertaken regarding the peptide sp^3 C$_\alpha$-centered angle, and the results were commented on.[78] While the ten Ala tetrapeptide conformers had been reinvestigated for the first time in the context of polarizable potentials,[84] the results reported in ref 78 were the first such investigation that used distributed ab initio multipoles and polarizabilities. The very first conformational studies of dipeptides that resorted to distributed multipoles date back to 1985, during the inception of SIBFA.[50] The polarizabilities then used were scalar polarizabilities, and the contribution of E_{pol} was smaller than in the present studies; this was due to the use of much smaller basis sets. Further studies on the Ala dipeptide as well as on β-turn forming peptides were published in 1998 using the CEP 4-31G(2d) basis set as a follow-up to the 1995 SIBFA refinements.[85] Addressing the issue of multipole transferability leading to that of an appropriate representation of interfragment E_{pol} and E_{ct} was done subsequently[78] which then led to the study reported here.

At this stage the existence of dependencies between anisotropy, nonadditivity, and transferability is worth mentioning. Such dependencies thus exist *between anisotropy and nonadditivity*. A recent example was provided by water chains of up to 12 molecules.[23d] Thus off-center lone pair polarizabilities not only are a determinant of anisotropy but also enhance cooperativity due to their closer distances to the polarizing partners. By contrast, atom-centered polarizabilities give rise to underestimated E_{pol} values with respect to QC computations. There are also dependencies *between nonadditivity and transferability* as occurs upon handling flexible molecules, namely regarding the issue of multipole transferability. Thus it was shown that both nonadditive E_{pol} and E_{ct} contributions, which resort to permanent multipoles and induced dipoles, enabled for the accounting of the impact of changes in multipole intensities upon building a large molecule from fragments and upon conformational changes. On the other hand, the existence of connections *between separability and transferability* is not clear. While such connections are obvious in the case of intermolecular interactions between rigid fragments, they could be questioned for intramolecular interactions. In this case, separability of the contributions could only be considered regarding the interfragment interaction energies. Thus as was shown above for the Zn(II) complexes of triphosphate,[76] while ΔE(QC) can be correctly reproduced by ΔE(SIBFA), this is not the case for the individual contributions.

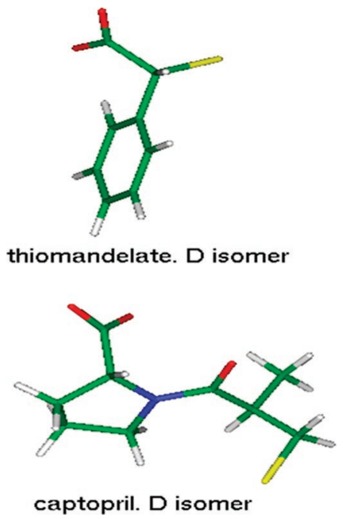
II. Extension to Molecular Recognition Problems. In addition to the above-mentioned applications to Cu(I) and Cu(II) complexes in the context of supramolecular chemistry, SIBFA was applied to the following systems:

Toward APMM Applications to DNA and RNA. The binding of hydrated Zn(II) and Mg(II) cations to guanine, adenine, and the G-C and A-T base pairs was investigated in parallel by SIBFA and QC, showing close numerical agreements in ΔE_{int} values.[86] Direct as well as through-water binding of the cations to the bases was investigated. SIBFA was able to account for the significant cooperativity (−15 kcal/mol) of Zn(II) binding to the G-C base pair. These studies were extended to 5′-guanosine monophosphate, a basic building block of DNA/RNA helices.[87] With the ribose in either a C2′endo or a C3′endo conformation, three competing binding modes were investigated. They involved the following: (a) simultaneous cation binding to both phosphate O_1 and guanine N_7; (b) direct binding to O_1 and through-water binding to N_7; (c) and, conversely, through-water binding to O_1 and direct binding to N_7. At both HF and DFT levels, close agreements were observed between the SIBFA and the QC energy values, both regarding the magnitudes of the binding energies and the ranking of the different binding modes. These studies will be extended to oligonucleotides of increasingly larger sizes and to their complexes with metal cations and ligands.

Toward de Novo Predictions of the Conformations of Short Zn-Metallo-Oligopeptides. We have resorted to a hierarchical procedure which, starting from random conformations, selects candidate conformers by a Monte Carlo approach with a potential of mean-force[88] and then postprocesses them using SIBFA.[89] This procedure was applied to the 18-residue Zn-finger of the HIV-1 nucleocapsid protein having a CCHC core (three Cy$^-$ residues and a His one) and its CCHH mutant. rms deviations of the C_α backbones of 3.5 Å and in the 2.2−3 Å range were found for these two Zn-fingers, respectively. Extensions of the procedure to include algorithms for global minimum searches[90] will be considered for future applications.

Complexes of Zn-Metalloproteins with Inhibitors. The targeted proteins are two bacterial enzymes, the Zn-metallo-β-lactamase from *B. fragilis* and phosphomannoisomerase (PMI) from *C. albicans*, and the C-terminal Zn-finger from HIV-1 nucleocapsid.

(a) Complexes of Metallo-β-Lactamase (MBL) with Captopril and Thiomandelate Mercaptocarboxylate Inhibitors. There are presently no inhibitors with sufficient affinity to MBL so as to be clinically useful, which raises a serious health concern. Mercaptocarboxylate derivatives endowed with micromolar affinity to MBL could be used as possible leads for the design of more efficient inhibitors. These are D- and L-captopril and D- and L-thiomandelate (Figure 4). While binding to *B. fragilis* MBL is known to occur upon removal of the Zn-chelating hydroxy anion and its replacement by one or by both anionic moieties of the inhibitor, there was no high-resolution structural information regarding the actual structures of their complexes with MBL. We have in refs 91 and 92 modeled a 108-residues model of MBL on the basis of the high-resolution X-ray structure by Concha et al. of uninhibited MBL.[68] Thiomandelate was built from methanethiolate, methane, benzene, and formate fragments. Captopril was built from methanethiolate, methane, proline, and formate fragments. Energy minimization (EM) was performed on the side chains of the residues making up the binding site, on all inhibitor torsion angles, and on the six inhibitor intermolecular variables as well as on the positions of the two Zn(II) cations. Different starting positions for EM were chosen, that were obtained from an exploratory docking that used constrained MD with the Accelrys software and the Cff91 force field,[93a] the constraints corresponding to enforcements of mono- or bidentate binding.


thiomandelate. D isomer

captopril. D isomer

Figure 4. Molecular structures of D-captopril and D-thiomandelate. Reprinted with permission from Antony et al. *Journal of Computational Chemistry* **2005**, *26*, 1131. Copyright 2005 John Wiley.

Thiomandelate Complexes. Seven and four distinct complexes were characterized for D- and L-thiomandelate, respectively.[92] Figure 5a represents the d-I D-thiomandelate complexes. In d-I, thiomandelate binds monodentately to the two Zn(II) cations through its S$^-$ atom, and the carboxylate binds to the Asn193 side chain. d-IIb is a bidentate binding mode in which the carboxylate binds to one Zn(II) cation. d-III is an alternative binding mode in which the carboxylate has replaced S$^-$ in the Zn(II)-chelating position and binds simultaneously through its second O atom to the Asn193 side chain. At the converged unconstrained SIBFA minima, the energy balances were completed upon computing the solvation energy ΔG_{solv} using the Langlet-Claverie Continuum reaction field procedure.[51a] Energy balances including ΔG_{solv} were more favorable for D-thiomandelate than for L-thiomandelate binding, consistent with experimental results, and for both isomers, more favorable for mono- than bidentate binding.

D- and L-Captopril Complexes. The competing modes can be either monodentate with binding of the sole S$^-$ to the two Zn(II) cations, or bidentate, involving additional Zn(II)-

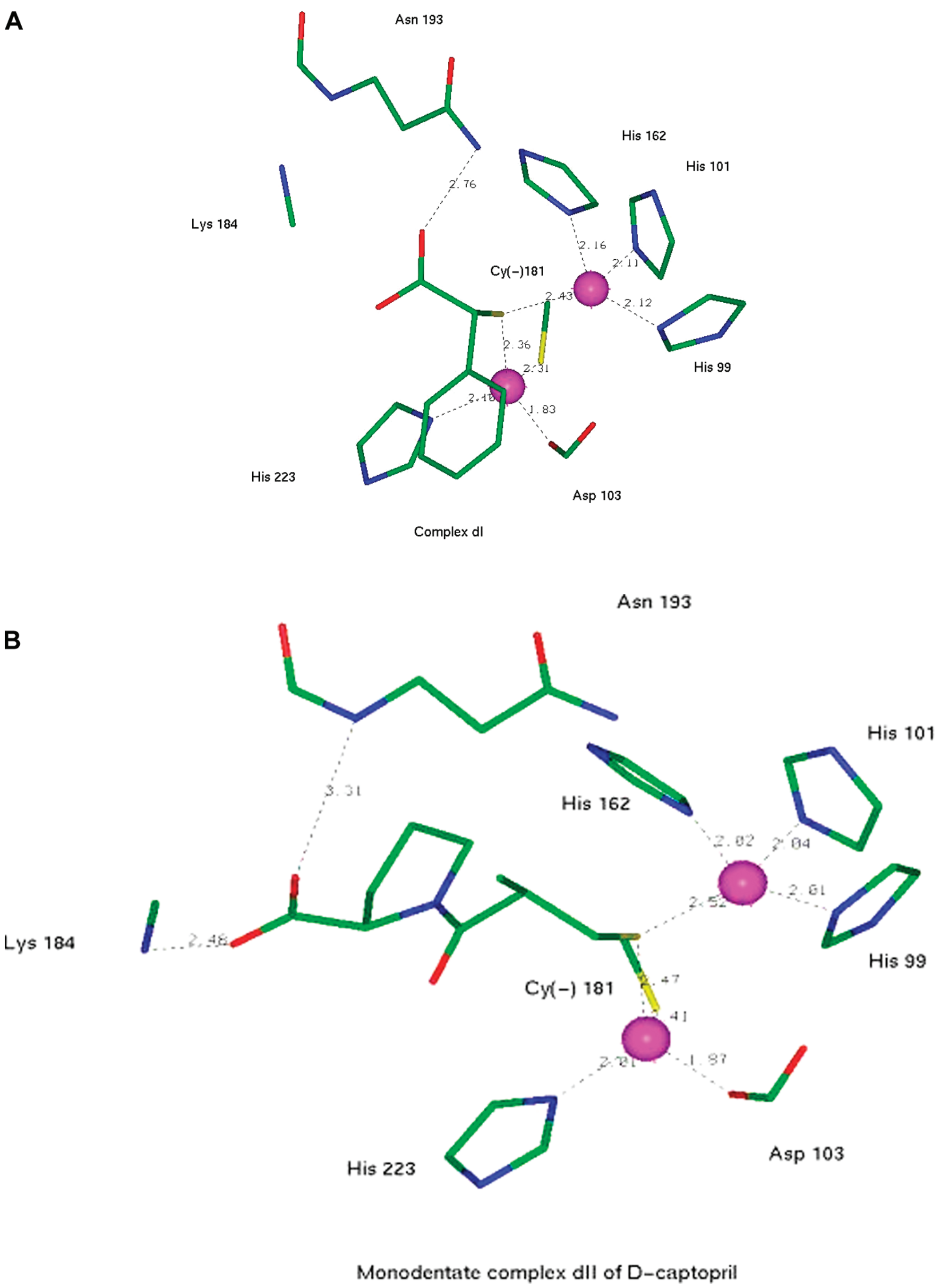


Figure 5. (a) Representative complexes of D-thiomandelate with metallo-β-lactamase. Reprinted with permission from Antony et al. *Journal of Computational Chemistry* **2005**, *26*, 1131. Copyright 2005 John Wiley. (b) Representative complexes of D-captopril with metallo-β-lactamase. Reprinted with permission from Gresh *Current Pharmaceutical Design* **2006**, *12*, 2121. Copyright 2006 Bentham Science Publisher, Ltd.

binding by either the formate or the carbonyl group. Up to nine distinct complexes could be characterized, as discussed in more detail in the preceding papers.[5,91] Thus monodentate complex d-II is stabilized, in addition to Zn(II) chelation by S^-, by interactions of the carboxylate with both the Lys184 side chain and the Asn193 main chain (see Figure 5b). In complex d-III, it is the carbonyl that now interacts with the An193 main chain. In complex d-IV, the carbonyl binds to one Zn(II) cation, and the formate is bound to the Lys184 side chain. In complex d-VI, the formate binds simultaneously to the Zn(II) cation and the Lys184 side chain, while the carbonyl binds the Asn193 side chain. The energy balances showed D-captopril to be more favorably bound by MBL than L-captopril, consistent with experimental results, and that the best binding mode was monodentate mode d-II. Although as mentioned above, there are no X-ray structures of *B. fragilis* MBL complexes with captopril, it is worth mentioning that a high-resolution X-ray structure on the complex of a MBL from a *P. aeruginosa* strain with a mercaptocarboxamide inhibitor analogous to D-captopril had shown very similar binding modes: monodentate binding of S^- to the two Zn(II) cations, and the terminal carboxylate simultaneously bound to the side chain of Lys161 and the main chain of Asn167, two residues that occupy positions

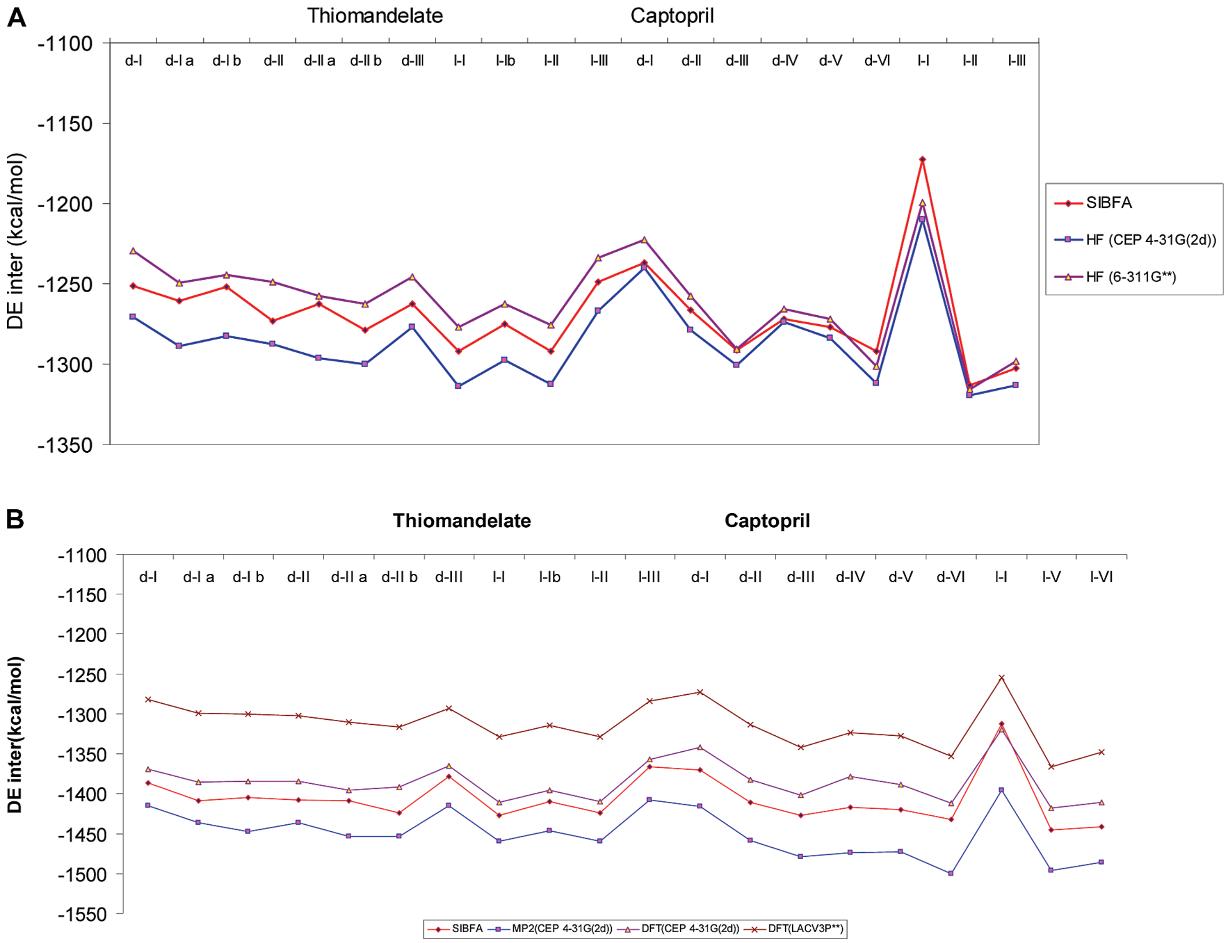


Figure 6. (a) Compared evolutions of Δ*E*(SIBFA) and Δ*E*(HF) in the 19 complexes of captopril and thiomandelate with the two Zn(II) cations and the eight residues modeling the metallo-*β*-lactamase binding site. SIBFA vs HF interation energies (kcal/mol). (b) Compared evolutions of ΔE_{tot}(SIBFA), Δ*E*(MP2), and Δ*E*(DFT) in the 19 complexes of captopril and thiomandelate with the two Zn(II) cations and the eight residues modeling the metallo-*β*-lactamase binding site. Values (kcal/mol) of Δ*E*(SIBFA) with E_{disp} and correlated quantum-chemical interatcion energies. Reprinted with permission from Antony et al. *Journal of Computational Chemistry* **2005**, *26*, 1131. Copyright 2005 John Wiley.

similar to the respective Lys184 and Asn193 ones of *B. fragilis* MBL.[94] While this should lend credence to the APMM calculations, an equally demanding test relates to comparing the ΔE_{int} values to parallel Δ*E*(QC) ones in model binding sites extracted from the cavity. Such models total 98 atoms, a size rendering them amenable to QC computations. The binding cavity has a very high local concentration of ionic charges. In addition to the two Zn(II) dications at 3.5 Å from one another, these include the anionic charges of Asp104 and Cy^{-}181, those of the inhibitor methanethiolate and formate groups, and the cationic charge of Lys184. Thus, similar to the kinase binding site, very important nonadditivity effects can be anticipated, underlining again the need to correctly account for the simultaneous interplay of inter- and intramolecular polarization and charge transfer. The SIBFA/QC comparisons were done at both uncorrelated and correlated levels. At the HF level, Δ*E*(SIBFA) was compared to Δ*E*(HF) using either CEP 4-31G(2d) or LACV3P** basis sets. At the correlated level, ΔE_{tot}(SIBFA) was compared to Δ*E*(DFT) with both basis sets and to Δ*E*(MP2) with the CEP 4-31G(2d) basis set. Such comparisons are discussed below together with those done for the thiomandelate complexes.

We have regrouped in Figure 6a,b the captopril and thiomandelate results under the form of graphs representing the evolutions of Δ*E*(SIBFA) and Δ*E*(QC) values for all 20 complexes, namely d-I up to l-III for captopril and d-I up to l-III for thiomandelate. Figure 6a shows Δ*E*(SIBFA) to have values consistently intermediate between the Δ*E*(HF) ones with the CEP 4-31G(2d) and LACV3P** basis sets, with the sole exception of the highest-lying complex l-I. The SIBFA curve shows very good agreement with the QC one, except at the level of complex d-II for thiomandelate. This is because d-II is computed in SIBFA to have a more favorable Δ*E* than d-IIa, while the reverse occurs with the HF calculations. Such an inversion involves differences of 10 kcal/mol out of 1260, namely less than 1%. At the correlated level, ΔE_{tot}(SIBFA) has values intermediate between the MP2 and the DFT ones with the CEP-431G(2d) basis set. The SIBFA curve displays very good correlation with the QC ones (Figure 6b).

Such results are highly encouraging, notwistanding further SIBFA refinements. They could be used to benchmark other polarizable molecular mechanics procedures. The structures of the 20 complexes are available as Supporting Information to ref 92 as well as on the Web at http://www.lct.jussieu.fr/pagesperso/jpp/SIBFA.html.

As concerns the energy balances done in the 108-residue model, we wish to note that while the D isomers of both captopril and thiomandelate are predicted to be the better-bound isomers, the energy differences between competing complexes are likely to be overestimated, since the interaction energy values (without ΔG_{solv}) represent enthalpies, not free energies, as they do not presently include entropy effects due to the reduction of translational and rotational motions of the ligand upon complex formation (for a recent discussion, see ref 93b). The captopril versus thiomandelate energy balances should not be presently compared, essentially because of a different calibration of ΔG_{solv} that was adopted in the thiomandelate study, conforming to the one used in ref 87.

(b) Complexes of Phosphomannoisomerase (PMI) to 5-Phospho-D-arabinohydroxamate and 5-Phospho-D-arabinonate Inhibitors. PMI is a Zn(II)-dependent isomerase that catalyzes the reversible isomerization of D-mannose 6-phosphate and D-fructose-6-phosphate. It plays an essential role in the metabolism of bacteria and microorganisms. It is involved in several pathologies, such as leishmaniasis, cystic fibrosis, and opportunistic infections in immuno-depressed individuals.[95] There are no PMI inhibitors presently in use clinically. 5-Phospho-D-arabinohydroxamate **(**5-PAH, Figure 7a) was recently reported as the most potent PMI inhibitor, displaying nanomolar affinity.[96] Replacing hydroxamate by carboxylate yielding 5-phospho-D-arabinonate (5-PAA, Figure 7a**)** resulted in loss of inhibitory potency. Zn(II) binding was experimentally shown to occur through hydroxamate rather than phosphate, despite the latter's dianionic character. We have performed SIBFA energy minimizations on the complexes of 5-PAH and 5-PAA with a 164-residue model of PMI.[97] We used the X-ray crystal structure of uninhibited PMI[98] as a starting point. As in the MBL studies, the PMI backbone was held rigid, and the side chains of the residues making up the binding site were relaxed. Two different starting points were considered, with either hydroxamate/carboxylate or phosphate bound to the Zn(II)-binding site. The non-Zn(II)-bound anionic moiety interacted with two basic residues, Arg304 and Lys310, at the entrance of the receptor cavity. In addition, the 5-PAH minima were used as new starting points for energy minimization of the 5-PAA complexes and conversely. This yielded a total of eight complexes. One more 5-PAH complex was investigated in which bidentate Zn(II)-binding of hydroxamate through both O atoms was enforced and subsequently relaxed. These energy minimizations were performed first in vacuo and then refined resumed by including ΔG_{solv}(LC) in the total energies. In modes *A* and *A*', hydroxamate is bound bidentately and monodentately, respectively. Bidentate binding occurs at the expense of Zn(II) binding to His285. In mode *B*, phosphate binding displaces both His residues from Zn(II). In mode *B*′, on the other hand, phosphate is bound to Zn(II) only indirectly, namely through Lys136 that is itself H-bonded to Zn-bound Glu138. Figure 7b gives a representation of the most stably bound complex of 5-PAH complex with PMI, namely *A*′, limited to the binding site. It was similarly found that in the 5-PAA complexes the carboxylate could bind either directly to Zn(II) or indirectly through the Lys136-Glu138 salt bridge. Figure 7c represents the most stably bound complex of 5-PAA, which corresponds to mode *C*′.

In agreement with experiment, the final energy balances indicated 5-PAH to have a significantly larger affinity than 5-PAA and that Zn(II) binding should occur through hydroxamate/carboxylate rather than phosphate. However, as in the MBL case, it was necessary to validate the values of ΔE_{int} by comparisons with parallel QC computations on the model binding site, now encompassing up to 140 atoms. The results reported in Table 5 indicate at both uncorrelated and correlated levels close agreements with the QC results. As for the model MBL complexes, the nine structures could serve to benchmark other PMM approaches. They are provided as Supporting Information for ref 97 as well as at the above-mentioned Web site. Extensions of the present work are in the design of novel 5-PAH analogs, in order to further improve their binding affinities.

(c) Binding of a Mercaptobenzamide Thioester to the C-Terminal Zn-Finger of HIV-1 Nucleocapsid. The HIV-1 nucleocapsid (NCp7) plays a pivotal role in HIV-1 metabolism. It has two highly structured Zn-binding domains with the C(X2)C(X4)H(X4)C motif (where X is any amino acid). It is a potential target for the development of novel antiviral drugs, because, in contrast to the HIV-1 protease and reverse transcriptase, mutations can impair its structure and function. This has led to the design of 'Zn-ejector' molecules that can disrupt Zn(II) binding.[99] Recently, mercaptobenzamide thioesters have been designed.[99d,e] One compound, denoted as C-247 (Figure 8a) has an S-connected carbonyl group that could make a covalent bond with the S^- atom of a Zn-coordinating NCp7 residue. Thus if the proximity between the carbonyl C and one Cys S were sufficient (in the 3.0−3.6 Å range), and if the S−C−O angle were adequate, a covalent bond could be formed, entailing loss of Zn-binding. We have performed SIBFA energy minimization on the binding of compound C-247 with residues Arg32-Gln53 of the C-terminal Zn-finger.[100] Both main-chain and side-chain torsion angles were relaxed. One of the most stable structures, represented in Figure 8b, complies with such requirements. It is stabilized by a double H-bond of the carboxamide chain with the side chain of Gln45 and by partial stacking of the benzene ring over the Trp37 ring. Two additional H-bonds are between the Lys34 main-chain carbonyl and the end carboxamide side chain and between the Gln45 side-chain N and the thioester carbonyl O. The energy balances including ΔG_{solv} are reported in Table 6. They are computed as the difference between the minimized energies of the C-247−NCp7 complex, on the one hand, and those of isolated C-247 and NCp7, separately minimized prior to complexation, on the other hand. They indicate the predominant role of the second-order terms and, in particular, of E_{pol} and E_{disp}, in complex stabilization. By contrast, E_1 is destabilizing. As a continuation of this work, we will seek

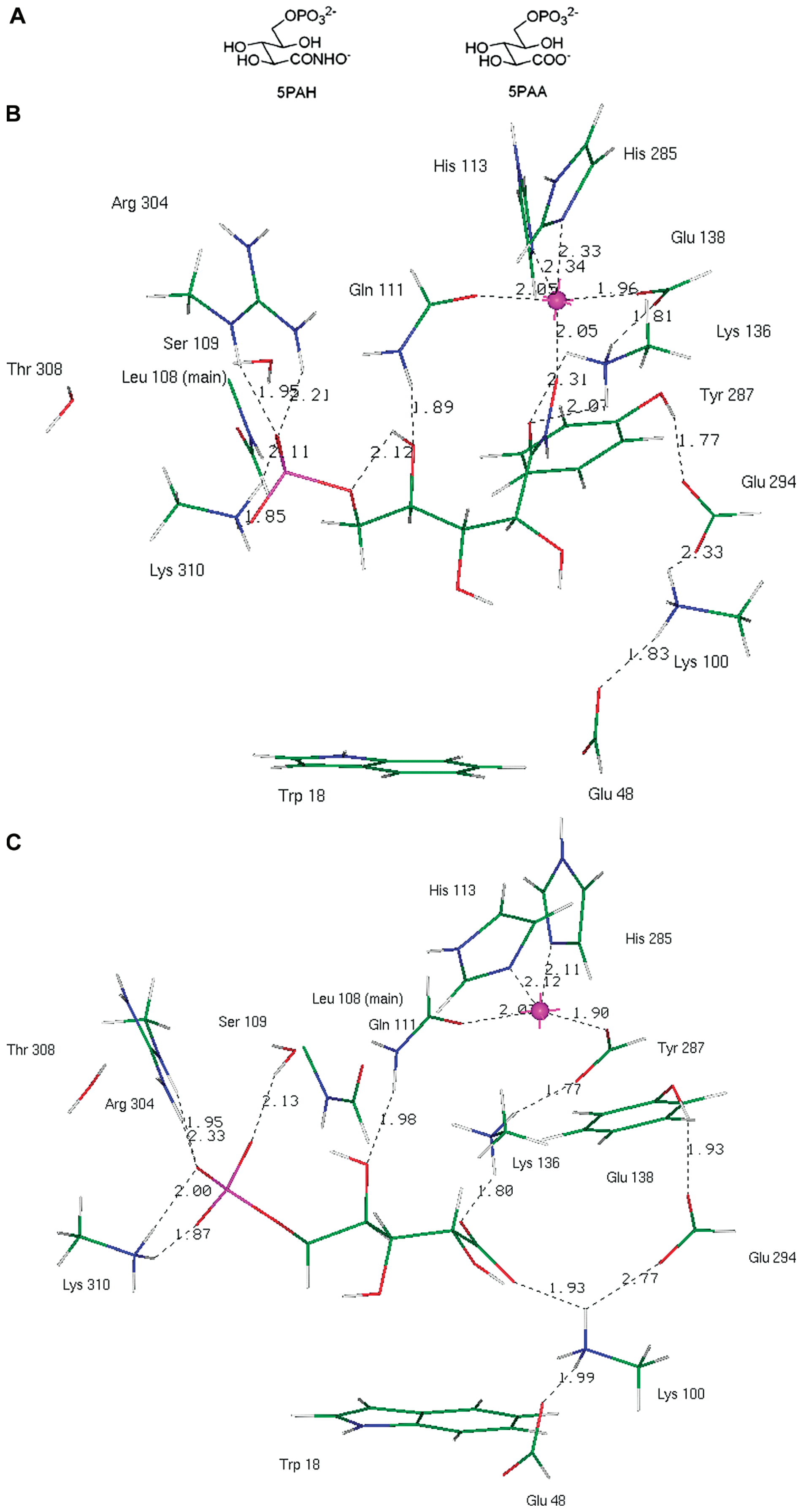


Figure 7. Representation of 5-PAH and 5-PAA PMI inhibitors as well as representative complexes of 5-PAH with the model binding site of PMI. Reprinted with permission from Roux et al. *Journal of Computational Chemistry* **2007**, *28*, 938. Copyright 2007 John Wiley.

Table 5. Interaction Energies (kcal/mol) of the Bifunctional Inhibitors in the Model Binding Site (MBS) Consisting of 14 Residues (See Text) Extracted from Their PMI Complexes in the Two Competing Arrangements[a]

	PMI-5PAH					PMI-5PAA				
	A	A′	A″	B	B′	C	C′	D	D′	PMI
E_{MTP}	−1396.2	−1417.3	-1383.8	−1377.1	−1341.9	−1359.9	-1300.7	−1353.3	−1367.4	−625.6
E_{rep}	270.3	270.5	261.4	269.9	264.5	254.6	266.7	284.8	277.7	170.6
E_1	−1125.9	−1146.7	−1122.4	−1107.2	−1077.4	−1105.3	−1033.0	−1068.5	−1089.7	−454.9
E_{pol}	−122.0	−123.4	−122.8	−113.7	−89.2	−122.8	−147.5	−113.5	-102.5	−110.4
E_{ct}	−40.0	−25.1	−40.2	−33.5	−46.5	−39.8	−42.5	−40.0	−39.2	−30.9
Δ*E*	**−1287.9**	**−1310.8**	**−1285.4**	**−1242.9**	**−1224.6**	**−1267.9**	**−1224.0**	**−1222.0**	**−1231.4**	**−596.3**
Δ*E*[b]	**−1283.6**	**−1310.8**	**−1278.2**	**−1243.2**	**−1240.4**	**−1250.7**	**−1227.5**	**−1221.0**	**−1233.7**	**−601.6**
Δ*E*[c]	**−1315.0**	**−1344.9**	**−1308.6**	**−1264.9**	**−1266.1**	**−1278.7**	**−1252.0**	**−1242.4**	**−1256.6**	**−618.8**
E_{disp}	−86.1	−87.0	−85.4	−79.0	−76.5	−79.6	−78.4	−78.5	−76.6	−57.3
ΔE_{tot}	**−1374.0**	**−1397.8**	**−1370.8**	**−1321.9**	**−1301.1**	**−1347.5**	**−1302.4**	**−1300.5**	**−1308.0**	**−653.6**
Δ*E*(DFT)[c]	**−1358.5**	**−1386.9**	**−1349.8**	**−1295.1**	**−1300.9**	**−1324.2**	**−1295.4**	**−1288.0**	**−1299.9**	**−653.0**

[a] (a) 5PAH with hydroxamate in the Zn-binding site; (b) 5PAH with phosphate in the Zn-binding site; (c) 5PAA with carboxylate in the Zn-binding site; (d) 5PAA with phosphate in the Zn-binding site; (e) unligated PMI with one water molecule replacing the inhibitor in the Zn(II) coordination sphere. [b] CEP 4-31G(2d) basis set. [c] LACV3P** basis set.

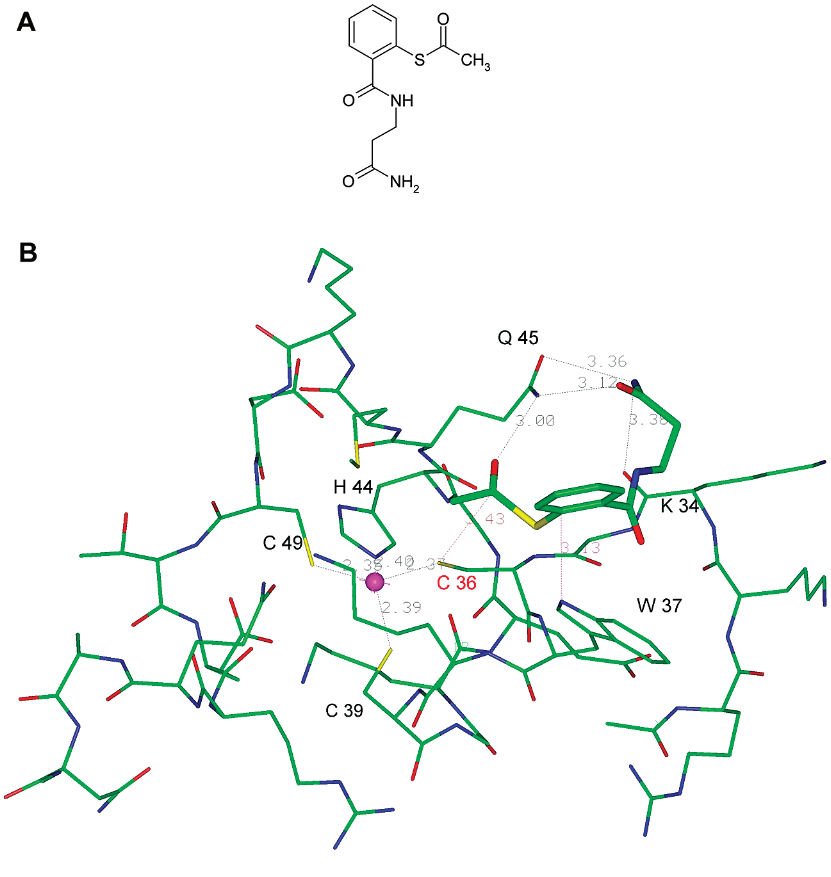


Figure 8. (a) Molecular structure of a 2-mercaptobenzamide thioester inhibitor (compound C-247) of the HIV-1 nucleocapsid protein. (b) Representation of the complex of inhibitor C-247 with the second Zn-finger of HIV-1 NCp7.

to improve the binding energies of C-247 by local modifications. We are also simultaneously performing QM/MM studies to elucidate the mechanism of S−C bond formation using the structure of Figure 8b as a starting point.

III. Toward Condensed Phase and Higher Accuracy: The Gaussian Electrostatic Model. As quantum calculations are able to give quantitative results and have shown the importance of short-range effects on intermolecular interaction energies, the development of the SIBFA equations constitutes a notable step toward a quantitative description of intermolecular interactions in molecular mechanics enabling a separate reproduction of the individual physical components of the total interaction energy. However, as SIBFA attempts to mimic the anisotropy of the density, a second more natural option can be by means of interacting frozen electron densities. As demonstrated several years ago by Kim and Gordon[101] for atom−atom potentials based on Density Functional Theory, these could improve the description of short-range quantum effects.

Indeed, some of us recently introduced[27] a methodology termed Gaussian Electrostatic Model (GEM) which is able to compute molecular interaction energies in the spirit of

Table 6. Interaction Energies (kcal/mol) of C-247 with the Arg32-Asn55 Zn-Finger of HIV-1 NCp7

	complex	finger	C-247	summed
E_{MTP}	−3647.5	−3411.8	−183.5	−3595.3
$E_{MTP}{}^{a}$	−52.2			
E_{rep}	2773.0	2613.0	88.4	2701.4
$E_{rep}{}^{a}$	71.6			
E_1	−874.5	−798.8	−95.1	−893.9
$\boldsymbol{E_1{}^{a}}$	**19.4**			
E_{pol}	−543.6	−507.2	−20.7	−527.9
$\boldsymbol{E_{pol}{}^{a}}$	**−15.7**			
E_{ct}	−79.4	−69.3	−0.3	−69.6
$\boldsymbol{E_{ct}{}^{a}}$	**−9.8**			
E_{disp}	−951.3	−855.7	−53.9	−909.6
$\boldsymbol{E_{disp}{}^{a}}$	**−41.7**			
E_{tor}	59.1	53.3	+4.2	57.5
$E_{tor}{}^{a}$	1.6			
E_{tot}	−2389.7	−2177.7	−165.8	−2343.5
$\boldsymbol{\delta E_{tot}{}^{a}}$	**−46.2**			
ΔG_{solv}	−572.7	−560.0	−39.7	−599.7
$\boldsymbol{\delta\Delta G_{solv}{}^{a}}$	**+27.0**			
$\boldsymbol{\delta E_{tot} + \delta\Delta G_{solv}}$	**−19.2**			

[a] After subtraction of the energies of the Zn-finger and of C-247 separately minimized.

the SIBFA approach but using the formalism of density fitting[102] (DF) methods usually devoted to the fast evaluation of Coulomb integrals for ab initio codes. We present here an overview of recent achievements concerning GEM. We will first summarize the initial steps of the development by addressing the important issue of the calculations of the required integrals to derive intermolecular Coulomb energies from fitted densities.[103] Results of a first GEM version that calculates intermolecular interaction energies from isolated monomer electron densities will then be detailed.[27a] To conclude, a generalized GEM density fitting scheme[27b] will be presented as well as its extension to periodic boundary conditions (PBC)[27b] and to QM/MM.[104]

(I) Methods. (A) From a Density Fitting Procedure to Intermolecular Coulomb Energies. We have used the formalism of the variational density fitting method,[102] an approach which is usually devoted to a fast approximation of the Coulomb interaction.

This method relies on the use of an auxiliary Gaussian basis set (ABS) to fit the molecular electron density obtained from a relaxed one-electron density matrix using a linear combination of atomic orbitals (LCAO).[105]

$$\tilde{\rho} = \sum_{k=1}^{N} x_k k(r) \approx \rho = \sum_{\mu\nu} P_{\mu\nu}\phi_\mu(r)\phi_\nu^*(r) \qquad (2)$$

The determination of the coefficients requires the use of a modified singular value decomposition (SVD) procedure in which the inverse of an eigenvalue is set to zero if it is below a certain cutoff.[27,102]

Using the fitted electronic densities, it has been shown[103] that it is possible to accurately compute the intermolecular Coulomb interaction energy (see eq 3) from frozen monomer densities in the direct spirit of ab initio energy decomposition schemes (see for example refs 25b,e).

$$E_{\text{Coulomb}} = \frac{Z_A Z_B}{r_{AB}} - \int \frac{Z_{A\tilde{\rho}^B}(\mathrm{r}_B)}{r_{AB}}\,\mathrm{d}r - \int \frac{Z_{B\tilde{\rho}^A}(\mathrm{r}_A)}{r_{AB}}\,\mathrm{d}r + \int \frac{\tilde{\rho}^A(\mathrm{r}_A)\tilde{\rho}^B(\mathrm{r}_B)}{r_{AB}}\,\mathrm{d}r \qquad (3)$$

By using density fitting, both long-range multipolar and short-range penetration electrostatic energies (missing in a distributed multipole treatment) are included, the errors being relatively small compared to reference ab initio data using the same density matrices.[103]

All the required integrals (electron−electron and electron−nuclear) were computed based on the McMurchie-Davidson recursions[106]enabling the use of higher angular moment Gaussian functions if required. It is important to point out that the formalism also enables an accurate representation of both electrostatic potentials and fields (Figure 9)

(B) From a Density Fitting Procedure to Intermolecular Interaction Energies. The reproduction of total interaction energy from fitted densities was studied based on the capability of the DF approach to compute accurate intermolecular Coulomb energies, thereby offering the possibility of a direct application of the methodology to molecular mechanics.[27a]

The total interaction was computed as the sum of four *separate* contributions: electrostatic (Coulomb), exchange-repulsion, polarization, and charge transfer. The central idea is that each contribution should match its Density Functional Theory (DFT) counterpart obtained using the Constrained Space Orbital Variation (CSOV) approach[25d,e] at the DFT level.

$$\Delta E_{\text{tot}} = E_{\text{Coulomb}} + E_{\text{exch-repulsion}} + E_{\text{pol}} + E_{\text{ct}} = E_{\text{Frozen Core}} + E_{\text{pol}} + E_{\text{ct}} \qquad (4)$$

At this point, no long-range dispersion contribution was added since we focused on reproducing DFT/ B3LYP[83] interaction energies, but the SIBFA E_{disp} contribution could also be included.[39]

As mentioned above E_{Coulomb} is directly computed from the integrals computed using the fitted densities. Extending the approach, we followed an idea put forth by Wheatley and Price[107a] and computed a two-body exchange repulsion based on the overlap model. This model relies on the observed proportionality between the exchange-repulsion energy and[107b] the overlap of the charge density, the calculation of the latter quantity being straightforward in the framework of our density fitting approach.

$$E_{\text{exch-repulsion}} \approx KS_\rho \qquad (5)$$

where

$$S_\rho = \int \rho_a(r)\rho_b(r)\mathrm{d}r \approx \int \tilde{\rho}_a(r)\tilde{\rho}_b(r)\mathrm{d}r \qquad (6)$$

The value of the parameter K can be easily determined and corresponds to the slope of a linear regression of the overlap of charge density versus the corresponding ab initio exchange-repulsion energy values. Finally, the charge transfer and polarization energies were computed following the SIBFA

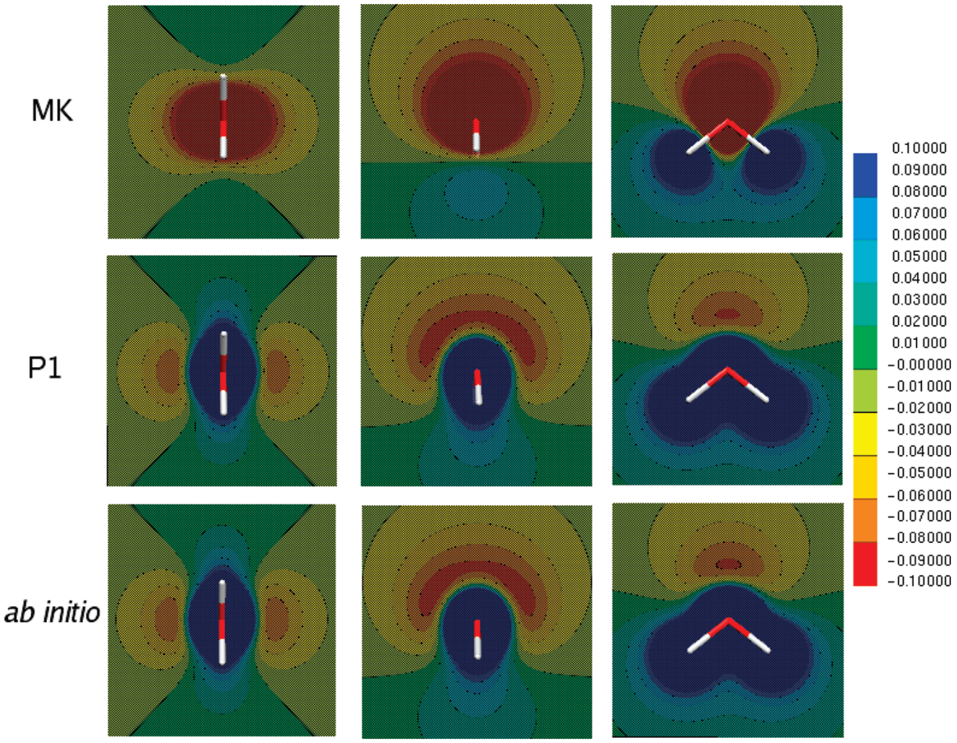


Figure 9. Electrostatic potential maps for the water molecule calculated from Merz−Kollman-generated charges MK, GEM fitted density, and ab initio calculation. All errors are in kcal/mol (see ref 101 for details).

Table 7. Intermolecular Coulomb Energies (in kcal/mol) for Ten Water Dimer Geometries for the GEM-0 Approach Fitted on B3LYP (or CCSD)/aug-cc-pVTZ Densities[a]

level of theory	water dimer geometry									
for E_{Coulomb}	1	2	3	4	5	6	7	8	9	10
CSOV (DFT)	−8.11 (−6.15)	−6.85 (−5.08)	−6.64 (−4.91)	−6.73 (−4.86)	−5.77 (−4.17)	−5.44 (−3.97)	−4.87 (−3.47)	−1.64 (−1.09)	−4.95 (−3.42)	−2.87 (−2.04)
GEM-0 (DFT)	−8.14	−6.89	−6.55	−6.77	−5.77	−5.48	−5.05	−1.77	−4.76	−2.74
CCSD (DCBS)	−7.96	−6.69	−6.48	−6.69	−5.71	−5.33	−4.89	−1.55	−4.77	−2.72
GEM-0 (CCSD)	−8.07	−6.75	−6.55	−6.58	−5.79	−5.56	−5.01	−1.68	−4.66	−2.70
SAPT (CCSD)	−8.02	−6.73	−6.49	−6.70	−5.69	−5.33	−4.96	−1.55	−4.81	−2.70

[a] Results in parentheses are interaction energies from a distributed multipole approach. CCSD reference calculation using the aug-cc-pVTZ basis set are provided and compared to the SAPT results (for details see ref 27a).

equations but using *the electrostatic potentials and fields computed from the fitted densities*.

Results from a First Force Field Implementation: GEM-0. We present here GEM results[27a] using fitted densities with an auxiliary basis set restricted to s-type ($l = 0$) Gaussians on water dimers and water clusters of up to 64 molecules. As the use of s-type Gaussian functions enables the rotation of the frozen fitted monomer densities, we term this method the Gaussian Electrostatic model (GEM-0).[27a] Using a nine-center spherical Gaussian density model for water, we demonstrated that accurate calculations could be performed on electrostatic energies. Table 7 gives results of our model on the ten minima of the total energy surface of the water dimer determined in previous studies[108] to investigate the accuracy of the intermolecular electrostatic energy at the B3LYP/aug-cc-PVTZ level. With respect to QC, a first striking result is that the values of the Coulomb interaction energy are notably improved compared to those from distributed multipoles (without E_{pen})[29] obtained at the same level of theory. If we compare the results to the CSOV references values, we can see in Table 7 that the *penetration energy is recovered* by the molecular mechanics as in our previous study.[103] The average absolute error of the ten configurations is 0.089 kcal/mol. The transferability of the auxiliary coefficients is demonstrated, and each of the dimers is correctly described. Regarding the reproduction of reference exchange-repulsion energies, the results were encouraging. They showed the robustness of the overlap model and were strongly correlated to reference B3LYP ab initio calculations of E_{exch} with a correlation factor of 0.9986 as displayed in Figure 10. The model has an average absolute error of 0.12 kcal/mol as shown in Table 8.

For all ten water dimers, close agreements were similarly found concerning the polarization and charge-transfer contributions (parts a and b, respectively, of Table 9), for which average absolute errors of 0.096 and 0.097 kcal/mol were found with respect to CSOV.

The final step consisted of comparisons of the sums of the GEM-0 energy components to the corresponding DFT interaction energies. For the ten water dimers, and with

Table 8. Intermolecular Exchange-Repulsion Energies (in kcal/mol) for 10 Water Dimer Geometries for GEM-0 Fitted on B3LYP (or CCSD−BD)/aug-cc-pVTZ Densities vs CSOV at the B3LYP/aug-cc-pVTZ Level[a]

level of theory	water dimer geometry									
for $E_{\mathrm{exch-rep}}$	1	2	3	4	5	6	7	8	9	10
CSOV (DFT)	6.84	5.63	5.37	5.08	4.22	3.85	3.59	1.18	3.59	1.89
GEM-0 (DFT)	6.84	5.71	5.36	4.86	3.95	3.95	3.94	1.27	3.64	1.97

level of theory	water dimer geometry									
for $E_{\mathrm{exch-rep}}$	1	2	3	4	5	6	7	8	9	10
SAPT(CCSD)	8.01	6.62	6.31	6.12	5.06	4.64	4.26	1.28	4.35	2.22
GEM-0 (CCSD)	8.05	6.76	6.32	5.77	5.01	4.75	4.54	1.24	4.14	2.19

[a] GEM-0 results in parentheses are exchange-repulsion energies obtained with GEM-0 using auxiliary coefficients obtained by averaging fits of the density using a 10^{-10} cutoff (ref 27a).

Table 9. (a) Polarization Energies (kcal/mol) for Ten Water Dimer Geometries for the GEM-0 Approach Fitted on B3LYP/augcc-pVTZ Densities Compared to CSOV B3LYP/augcc-pVTZ Results (Ref 27a) and (b) Charge-Transfer Energies (kcal/mol) for Ten Water Dimer Geometries for the GEM-0 Approach Fitted on B3LYP/augcc-pVTZ Densities Compared to CSOV B3LYP/augcc-pVTZ Results (Ref 27a)

(a)

	water dimer geometry									
level of theory for E_{pol}	1	2	3	4	5	6	7	8	9	10
CSOV/DFT	−1.33	−1.14	−1.12	−0.69	−0.64	−0.62	−0.37	−0.12	−0.44	−0.28
GEM-0/DFT	−1.22	−1.03	−0.92	−0.55	−0.53	−0.50	−0.27	−0.08	−0.42	−0.29

(b)

	water dimer geometry									
level of theory for E_{ct}	1	2	3	4	5	6	7	8	9	10
CSOV	−1.77	−1.48	−1.42	−0.96	−0.80	−0.68	−0.53	−0.20	−0.54	−0.26
GEM-0/SIBFA	−1.86	−1.42	−1.31	−0.94	−0.73	−0.63	−0.44	−0.11	−0.56	−0.29

Table 10. Total Interaction Energies (kcal/mol) for Ten Water Dimer Geometries for the GEM-0 Approach Fitted on B3LYP/aug-cc-pVTZ Densities Compared to CSOV B3LYP/aug-cc-pVTZ Results Corrected from the Basis Set Superposition Error (Ref 27a)

	water dimer geometry									
level of theory for ΔE_{int}	1	2	3	4	5	6	7	8	9	10
CSOV	−4.39	−3.82	−3.80	−3.38	−3.00	−2.91	−2.36	−0.78	−2.30	−1.56
GEM-0/SIBFA	−4.30	−3.71	−3.28	−3.32	−3.13	−2.88	−1.98	−0.45	−2.29	−1.59

respect to the BSSE-corrected CSOV total interaction energies, a 0.16 kcal/mol average absolute error was obtained, limited to 0.038 kcal/mol in terms of the relative average error (Table 10). Such a result thus confirms this methodology to reproduce realistic interactions.

We have also applied the model to 16−64 water clusters, as extracted from Monte Carlo simulations in ice or in bulk water that resorted to SIBFA. In all cases, the accuracy of the method appears very good. Thus for the 16, 20, and 64 water clusters, the values of the Coulomb interaction energy amounting to −186.84, −309.38, and −449.52 kcal/mol compare closely to the corresponding CSOV values of −186.38, −307.20, and −446.12 kcal/mol, respectively (Table 11). For the exchange-repulsion energies, the model also performs very well with errors below 1% (Table 11). The polarization and charge-transfer terms have also close agreements with available QC results (Table 12). Therefore, in order to evaluate the overall accuracy of our model, we computed the total BSSE-corrected interaction energies at the same level of theory for the 16 and 20 molecule clusters. Relative errors of +3.16 out of −114.02 kcal/mol and of −3 out of −168.1 kcal/mol were found for these two respective clusters, confirming the good transferability of the

Table 11. Coulomb and Exchange-Repulsion Intermolecular Interaction Energies (kcal/mol) for Water Clusters (n = 16, 20, 64) for the GEM Approach Fitted on B3LYP (or CCSD)/aug-cc-pVTZ vs ab Initio CSOV/B3LYP/aug-cc-pVTZ Values[a]

n	E_{coulomb} GEM-0 (DFT)	E_{coulomb} CSOV (DFT)	E_{coulomb} GEM-0 (CCSD)	$E_{\mathrm{exch-rep}}$ GEM-0 (DFT)	$E_{\mathrm{exch-rep}}$ CSOV (DFT)
16	−186.84	−186.38	−184.80	164.95	166.54
20	−309.38	−307.20	−305.84	292.25	292.16
64	−449.52	−446.12	−443.54	336.48	NC

[a] NC = not computed (ref 27a).

Table 12. Polarization Energies (kcal/mol) for Water Clusters ($n = 16, 20, 64$) for the GEM-0 Approach Fitted on B3LYP /aug-cc-pVTZ vs ab Initio CSOV/B3LYP/ aug-cc-pVTZ Values[a]

n	E_{pol} two-body GEM-0	E_{pol} two-body CSOV	E_{pol} GEM-0 (E_{pol} GEM-0 initial guess)	E_{pol} KM/HF (E_{pol} RVS/HF)
16	−30.75	−31.03	−48.53 (−36.82)	−45.11 (−35.50)
20	−47.53	−48.01	−82.79 (−62.60)	−78.6 (NC)
64	−57.97	NC	−77.89 (−64.78)	NC (NC)

[a] For the GEM-0 column, results in parentheses correspond to the polarization energy of the first set of induced dipoles. RVS polarization results are given in parentheses in the KM column. Both are computed at the CCP 4-31G(2d)level. NC = not computed (ref 27a).

different approximations. GEM-0 has been also tested for metals for electrostatic and exchange-repulsion contributions. Accurate results are obtained even at a very short range.[27a] It is important to point out that this density fitting procedure is not limited to Hartree−Fock or DFT energies. Thus Tables 7 and 8 had also shown close agreements[27a] between GEM fitted on relaxed CCSD-Bruckner-Double densities and reference CCSD SAPT calculations.

(C) Extension to Higher Angular Momenta, Computational Speedup, and Periodic Boundary Conditions. At this point, GEM-0 showed a very good accuracy but requires several nonatomic centers as it uses s-type ($l = 0$) Gaussian functions only. In order to reduce the number of sites, an extension of the formalism to higher angular momenta ($l > 0$) was required.

(a) Extension to Higher Angular Momenta: Accuracy of Forces and Energies. One advantage of using fitted densities expressed in a linear combination of Gaussian functions is that the choice of Gaussian functions for the ABS needs not be restricted to Cartesian Gaussians. In order to extend GEM to higher order angular momenta,[27b] we have chosen to use normalized *Hermite Gaussian functions* for the calculation of the intermolecular interactions. Thus, the use of Hermite Gaussians in the calculation of the intermolecular interactions results in improved efficiency by the use of the McMurchie-Davidson (McD) recursion[106] since the expensive Cartesian-Hermite transformation is avoided. Obtaining the Hermite expansion coefficients from the fitted Cartesian coefficients is straightforward since Hermite polynomials form a basis for the linear space of polynomials.

We have also implemented noise reduction techniques[27b] for the fitting procedure in addition to the already discussed cutoff in the eigenvalues during the SVD procedure. Indeed, this method produces undesirable numerical instabilities (noise) when the number of basis functions starts to grow with Gaussian functions as commented above. In addition, we have observed that these instabilities are also present when using only *s*-type spherical functions[27a] albeit to a lower extent. In the present implementation we have opted to use the Tikhonov regularization formalism. Additionally, Jung et al.[109] have recently shown that the use of a damped Coulomb operator $\hat{O} = \mathrm{erfc}(\beta r)/r$ can be used for the fitting procedure. These authors have employed this kernel to localize the integrals in order to increase the calculation speed of three-center Coulomb integrals in a quantum mechanical program. For our purposes, the implemented damped Coulomb operator could be employed to attenuate the near-singular behavior due to long-range interactions.[27b]

Table 13. Relative rms Force Deviation with Respect to CSOV for the Ten Water Dimers (Ref 27b)

	6-31G*			aug-cc-pVTZ		
level of theory	A1	P1	G03	A1	P1	G03
Coulomb	0.06	0.03	0.01	0.15	0.04	0.05
exchange	0.22	0.08	0.07	0.12	0.04	0.07

With such procedures, Coulomb and exchange-repulsion have been calculated with higher angular momenta that allow for a reduction of the number of sites compared to GEM-0 for the ten water dimers as well as representative benzene dimers. Excellent agreement was obtained in all cases for the intermolecular interactions with errors below 0.1 kcal/mol for electrostatic and around 0.15−0.2 kcal/mol for exchange repulsion.[27b] In practice in the MD community, the measure of the accuracy has been the forces since this is the quantity that determines the trajectories. Upon using GEM with three ABSs, such an accuracy could be evaluated by comparing the calculated GEM forces with those obtained with CSOV using the finite difference method.[27b] For both Coulomb and overlap interactions, Table 13 shows that for the ten water dimers small rms deviations are observed between forces calculated with A1, P1,[110] and g03 ABSs compared to the CSOV forces computed at the B3LYP level with the same basis sets, namely 6-31G* and aug-cc-pVTZ. The errors in the exchange-repulsion forces are also very satisfactory considering the simplicity of the overlap model compared to ab initio.

(b) From Densities to Site Multipoles: A Continuous Electrostatic Model. Challacombe et al.[111] have shown that Hermite Gaussians have a simple relation to elements of the Cartesian multipole tensor. Expanding on that work, once the Hermite coefficients have been determined, they may be employed to calculate multipoles centered at the expansion sites.[27b] Thus we have been able to obtain distributed multipoles centered at the ABS's sites that connect naturally with an accurate evaluation of the exact Coulomb interaction energy. This connection will be useful for the direct use of such multipoles into SIBFA as well as in the generation of damping functions[39a] that accounts for the penetration error when using these multipoles. Unlike conventional multipole expansions, the spherical multipole expansion obtained from Hermite Gaussians has an intrinsic finite order, namely, the highest angular momentum in the ABS. This is thus similar to the multipolar expansions derived by Volkov and Coppens.[112]

This connection between multipoles and Hermite densities is important. Indeed, unlike s-type functions ($l = 0$), fitting coefficients with l > 0 (sp, spd ...) are not invariant by rotation. These coefficients must be transformed for each molecular fragment orientation in order to compute interaction energies. Such a transformation can be achieved by defining both a *global* orthogonal coordinate system frame and a *local* orthogonal coordinate frame for each fragment fitting site. Hermite Gaussians in the two coordinate systems can be related using the chain rule.[27b] Such a method has been previously developed for point dipoles and generalized to higher order multipoles.[113] These frame definitions are similar to those in the OPEP code[16a] and could be applied to SIBFA as well. It is important to point out that the same chain rule approach works also for the transformation to scaled fractional coordinates which will be important toward the extension to PBC where the use of Particle Mesh Ewald (PME)[114] requires that the coefficients be transformed to scaled fractional coordinates.

(c) Increasing Computational Efficiency Using Reciprocal Space Methods. Additionally, a significant computational speedup can be achieved using reciprocal space methods.[27b] Indeed, it is possible to split the integrals required for the frozen-core contribution into direct and reciprocal space contributions.

The direct sum corresponds to full computation of integrals between two centers at a distance below a chosen cutoff. Such integrals are computed using a generalized McD recursion applicable to Gaussian derivatives of any smooth function of r and so thus to all the direct space integrals used in this study, i.e., overlap, Coulomb, and damped Coulomb.[27b]

The rest of the integrals are treated using reciprocal space. Three methods were implemented: regular Ewald, Particle Mesh Ewald (PME),[114] and Fast Fourier Poisson (FFP)[115] Denoting by N the number of molecules, since the regular Ewald approach scales as N^2, the use of fast Fourier transformations (FFT) is necessary to improve the scaling and reach $N \log(N)$.

PBC GEM implementation with reciprocal space methods has been tested by calculating the intermolecular Coulomb energies and forces for a series of water boxes $[H_2O]_N$, $N =$ 64, 128, 256, 512, and 1024. The reciprocal space methods are quite efficient. The calculation of the energies and forces for the largest system was tested at the highest accuracy, i.e., the 1024 water box with a very extended g03 ABS which corresponds to 654 336 Hermite coefficients (located on atoms, bonds, and lone pairs) and takes only 34 s with FFP and 42 s with PME (see Figure 11) using rms accuracies of 10^{-4} on a dual Xeon 3.3 Ghz processor.[27b] It is further noted that both reciprocal space methods are highly parallelizable, which would increase computational efficiency.

Moreover, we have recently shown that thanks to a numerical approach to the Hermite fitting[116] using molecular properties calculated on grids as well as an improved splitting procedure for the compact and diffuse functions it was possible to improve the accuracy and so to diminish the number of auxiliary functions. For example, it has been possible to use the small A2 basis set restricted to atoms and to perform a calculation on a 4096 water box GEM calculation fitted on a B3LYP/6-31G* reference level. Such calculation took 2.6 s on *a single processor*, which is about an order of magnitude slower than the corresponding point charges Amber calculations which took 0.2 s on the same computer.

(D) QM/MM and Future Developments. A QM/MM implementation has been recently performed[104] using GEM as the MM force field. This method has been used, parallel to conventional QM/MM using point charges, to evaluate the polarization on the QM subsystem by the MM environment for the ten water dimers. GEM was found to give the correct polarization response compared to reference CSOV polarization energies. By contrast, point charges produced significant underpolarization of the QM subsystem, in several cases actually presenting an opposite sign of the polarization contribution (see Figure 12). This approach prefigures a prospective multilevel implementation of a SIBFA/GEM/QM strategy. Indeed, it is important to mention that results obtained with both PME and FFP can be mixed. This opens up novel possibilities for QM/MM implementation: thus the GEM section proximal to the QM could be calculated with PME or FFP, the remaining MM subsystem could be represented via GEM multipoles, and these could be used for SIBFA and calculated in reciprocal space using PME. As most of the gradients of the SIBFA energy function are available, this opens up the possibility of long condensed-phase SIBFA MD/PBC. Such an implementation[8b] in

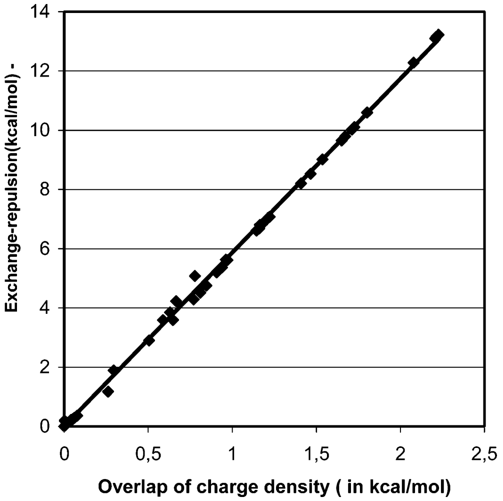


Figure 10. Correlation of the overlap of charge density (kcal/mol) computed with GEM-0 vs exchange-repulsion energy (kcal/mol) obtained at a CSOV/B3LYP/aug-cc-pVTZ level for 200 orientations of the water dimer.

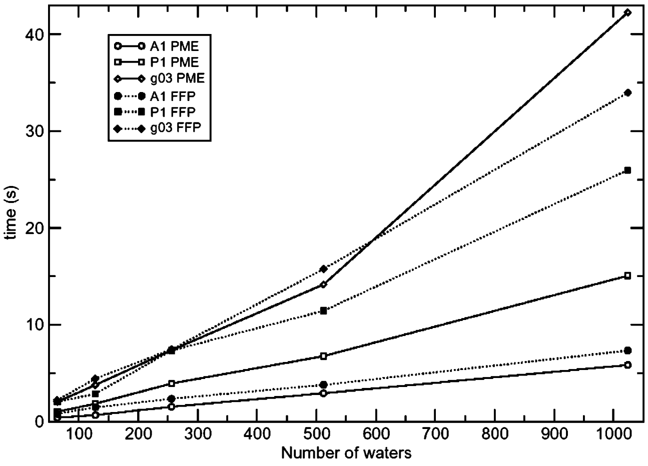


Figure 11. Timings for water boxes with rms force tolerance of 10-4. Closed circles: A1 PME; closed squares: P1 PME; closed diamonds: G03 PME; open circles: A1 FFP; open squares: P1 FFP; open diamonds: G03 FFP.

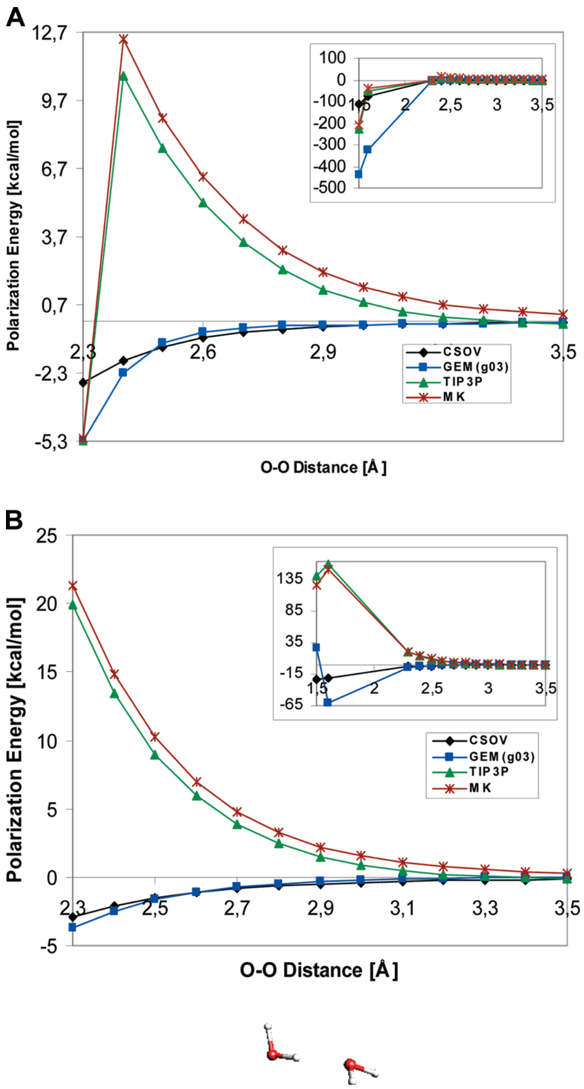


Figure 12. Polarization of the QM water molecule in the geometry of the linear water dimer at various distances for a QM/MM calculation using GEM (molecule A = QM, top; molecule B = QM, bottom). Inset shows a range from 1.5 to 3.5 Å.

AMBER 9.0 was recently achieved by some of us for AMOEBA. We plan to perform it in the context of SIBFA as well.

Present Status of the Software. At this point it is necessary to mention some present possibilities and limitations of the SIBFA software

(a) Timings. An in vacuo single-point computation on the complex of the 5-PAH inhibitor with a PMI model encompassing 164 residues (about 2700 atoms totaling about 8000 centers) requires about 3 min CPU time on a single-processor IBM sp4 computer (there are no cutoffs for the energy computations). Merlin can resort to nongradient minimizers such as ROLL or SIMPLEX or to numerical evaluations of the gradients using the BFGS, the Davidon-Fletcher-Powell, or the Conjugate Gradients Algorithms (see ref 61 for details). For most applications, we found the ROLL algorithm as the most effective, although it entails a significantly larger number of energy evaluations. Energy minimizations on about 200−300 internal variables requiring about 5000 energy computations thus take about 10 days on a single processor. They are postprocessed for one or two additional rounds to ensure for convergence of the energy. Subsequent energy minimizations encompassing ΔG_{solv} are about 6-fold more time-consuming but can now be done on a version of the code that parallelizes on four to eight processors.

(b) Availability of the Gradients. Presently, most analytical gradients have been coded and checked. The principal gradients presently not available are those of E_{ct} and of $\Delta G_{solv.}$ The coding of E_{ct} is underway and will be reported shortly. In the present context, a simplified version of E_{pol} and its gradients has been coded for which, similar to the ΔG_{solv} computations, scalar instead of tensor polarizabilities are used. The availability of the analytical gradients should enable for more efficient searches of the potential energy surface although with a simpler energy function, since the minima could be reprocessed for a last round with the complete function. This availability has also enabled us to start preliminary MD simulations with the simplified SIBFA potential. These will be used to locate alternative docking modes in ligand−macromolecule complexes. On the other hand, however, condensed-phase simulations will depend upon the merging of SIBFA with PME and/or GEM methodologies as discussed below.

Present Scope of Applications. Several ongoing applications of SIBFA bear on the docking of inhibitors with protein targets and are carried out in close collaboration with experimentalists. While Zn-metalloproteins constitute a privileged target, the extension to other targets, such as signaling proteins and kinases, is underway. The size of the target proteins can encompass up to 200 amino acid residues. Optimization of the code to handle larger systems is underway, including its porting to Fortran 90 and parallelization.

Conclusions and Perspectives

The availability of energy-decomposition analyses of QC intermolecular interactions is essential for the development of APMM procedures. We have shown in this review that the separable SIBFA potential can reproduce the anisotropy and nonadditivity features of ΔE_{int}(QC) and of its contributions. A particularly challenging test was provided by binuclear Zn(II) complexes, as in the binding site of bacterial metallo-β-lactamase (MBL).[28e] ΔE_{int}(SIBFA) could closely reproduce ΔE(QC) and the contrasting behaviors of E_1, on the one hand, and of E_{pol} and E_{ct}, on the other hand, in two structurally very distinct and competing arrangements. Multipole transferability is a critical issue in order to be able to handle flexible molecules. We have shown that the separable character of the potential, encompassing both polarization and charge transfer, were necessary to compute intra- and intermolecular interactions of a flexible molecule assembled from rigid fragments. This was illustrated in two extreme cases, divalent cation binding by triphosphate and mercapcarboxamides, on the one hand, and the conformational energies of ten Ala and Glu tetramers, on the other hand. SIBFA has been applied to investigate inhibitor binding to Zn-dependent metalloenzymes. Two recent examples are MBL and phosphomannoisomerase (PMI).[91,92,97] Energy balances were performed including the contribution of continuum ΔG_{solv}(LC) for different inhibitors in several competing arrangements. Validations by parallel QC computations were done on model binding sites of MBL and PMI totaling up to 140 atoms. Twenty and nine complexes were thus evaluated in these respective sites. The evolutions of the SIBFA interaction energies paralleled the QC ones, with relative errors $<3\%$. The last application bore on a nonenzymatic Zn-metalloprotein, the HIV-1 nucleocapsid (NCp7), a novel target for the design of new-generation anti-HIV inhibitors. One of the low-energy minima had the nucleophilic S-connected carbonyl group at an appropriate distance (3.4 Å) and orientation from Cys36 S^- to initiate covalent bond formation followed by Zn-ejection. E_{pol} and E_{disp} were the main contributors to the final energy balances, while E_1 was destabilizing. The SIBFA-derived complexes are being reprocessed by QM/MM procedures, indicating the connectedness between classical MM, APMM, and QM. SIBFA is being extended to a diversity of metal cations. Such extensions benefit from the integration[58] of Ligand Field (LF) effects, on the one hand, and the availability of energy-decomposition procedures[25] and the possibility of quantifying correlation as well as relativistic effects,[117,118] on the other hand. The coupling with Particle Mesh Ewald (PME) methodologies[27b,113,8b] should significantly widen its scope toward large macromolecular complexes and condensed phase. The interface with GEM,[27] which can itself be coupled to QM[104] should give rise in the near future to a multilevel QM/GEM/SIBFA methodology since GEM offers a direct connection between multipoles and densities. This approach could be applied to biomolecular systems such as 4-oxalocrotonate tautomerase.[119,120] GEM offers increased accuracy and full separability of its components as well as improved cooperative effects by the inclusion of native short-range quantum effects. Finally GEM Hermite Gaussian densities can be derived for any element of the periodic classification where ab initio relaxed densities at Hartree−Fock, post Hartree−Fock, or DFT levels are available. In two forthcoming papers, we describe new fitting improvements for the hermites as well as the generalized energy function for small molecules and flexible peptides.

Abbreviations. 5 -PAH, 5-phospho-D-arabinohydroxamate; 5-PAA, 5-phospho-D-arabinonate; ABS, auxiliary basis set; AMBER, assisted model building with energy refinements, AMOEBA, atomic multipole optimized energetics for biomolecular applications; AOM, angular overlap model; APMM, anisotropic polarizable molecular mechanics; ASP-W, anisotropic site potential for water; ATP, adenosine triphosphate; A1,A2, DGauss DFT Coulomb fitting auxiliary basis set; B3LYP, Becke-Lee-Yang Parr functional; BVWN, Becke-Vosko-Wilk-Nusair functional; CEP 4-31(2d), coreless effective potential double-zeta and two 3d polarization functions on heavy atoms; CNDO, complete neglect of differential overlap; CSOV, constrained space orbital variations; DF, density fitting; DFT, density functional theory; EFP, effective fragment potential; EM, energy minimization; FFP, fast Fourier Poisson; GEM, Gaussian electrostatic model; G03, automatically generated Gaussian 03 Coulomb fitting auxiliary basis set; HF, Hartree-Fock; HPPK, dihydropterin pyrophosphokinase; KM, Kitaura-Morokuma energy-decomposition procedure; LACV3P**, Los Alamos compact valence potentials; LCAO, linear combination of atomic orbitals; LF, ligand field effects; LMO, localized molecular orbitals; LMP2, localized Moller−Plesset 2; MBL, metallo-β-lactamase; MC, Monte-Carlo; McD, McMurchie Davidson; MD, molecular dynamics; MM, molecular mechanics; MO, molecular orbitals; MP2, Moller−Plesset 2; NCp7, nucleocapsid protein from HIV-1 retrovirus; NEMO, nonempirical molecular orbital; OPEP, optimally partitioned electric properties; PBC, periodic boundary conditions; PME, particle mesh Ewald; PMI, phosphomannoisomerase; PMM, polarizable molecular mechanics; Pvtz, polarized valence triple zeta; QC, quantum chemistry; QM, quantum mechanics; QP, quadrupolar polarizability; RMS, root mean square; RVS, restricted variational space; SAPT, symmetry-adapted perturbation theory; SDFF, spectroscopically determined force field; SIBFA, sum of interactions between fragments ab initio computed; TCPE, topological and classical many-body polarization effects.

Inception and developments of SIBFA were done in close collaboration with Drs. Pierre Claverie, Alberte Pullman, Jacqueline Langlet, and Claude Giessner-Prettre. We also wish to thank Drs. Walter J. Stevens, Morris Krauss, David R. Garmer, Lee G. Pedersen, and Dennis R. Salahub, for very stimulating collaborations. One of us (N.G.) thanks Benoit de Courcy for his help during preparation of the manuscript. We thank the computer centers of CINES (Montpellier, France), CRIHAN (Rouen, France), CCRE (Paris, France), and NIEHS (RTP, NC) and the Advanced Biomedical Computer Center, NCI-FCRDC, U.S.A. This research was supported in part by CNRS (France), the Ligue Nationale contre le Cancer (Comité Ile-de-France), and by the Intramural Research Program of the NIH and NIEHS (U.S.A.).

Supporting Information Available: Zn-methanethiolate, water-formate, and stacked formamide complexes, interaction energies for Zn(II) complexes with water ligands, interaction of triphosphate with Zn(II) probe cation, and interaction of triphosphate anion with the binding site of HPPK kinase. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Halgren, T. A; Damm W. *Curr. Opin. Struct. Biol*. **2001**, *11*, 236.

(2) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89.

(3) Ponder, J. W.; Case, D. A. *Adv. Protein Chem*. **2003**, *66*, 27.

(4) McKerell, J. J. *Comput. Chem.* **2004**, *25*, 1584.

(5) Gresh, N. *Curr. Pharm. Des.* **2006**, *12*, 2121.

(6) (a) Pullman, B.; Claverie, P.; Caillet, J. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *57*, 1663. (b) Rein, R.; Claverie, P.; Pollack, M. *Int. J. Quantum Chem*. **1968**, *2*, 1129. (c) Claverie, P.; Rein, R. *Int. J. Quantum Chem.* **1969**, *3*, 537.

(7) (a) Warshel, A.; Levitt, M. *J. Mol. Biol*. **1976**, *103*, 227. (b) Lee F. S.; Chu Z. Y.; Warshel A. *J. Comput. Chem*. **1993**, *14*, 161, and references therein. (c) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283.

(8) (a) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933, and references therein. (b) Piquemal, J. P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys*. **2006**, *125*, 054511.

(9) Gresh, N.; Claverie, P.; Pullman, A. *Theor. Chim Acta* **1984**, *66*, 1.

(10) (a) Day, P.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D. R.; Basch, H.; Cohen, D. *J. Chem. Phys*. **1996**, *105*, 1968. (b) Slipchenko, L. V.; Gordon, M. S. *J. Comput. Chem.* **2007**, *28*, 276.

(11) (a) Price, S. L.; Stone, A, J. *Mol. Phys.* **1984**, *51*, 569. (b) Price, S. L.; Stone, A, J. *J. Chem. Soc., Faraday Soc.* **1992**, *88*, 1755.

(12) Millot, C. J.; Soetens, J. C.; Martins, Costa, N. T. C.; Hodges, M. P.; Stone, A. J. *J. Phys. Chem. A* **1998**, *102*, 754.

(13) Mannfors, B.; Mirkin, N. G.; Palmo, K.; Krimm, S. *J. Comput. Chem.* **2001**, *22*, 1933.

(14) Hermida-Ramón, J. M.; Brdarski, S.; Karlström, G.; Berg, U. *J. Comput. Chem.* **2003**, *24*, 161.

(15) (a) Gagliardi, L.; Lindh, R.; Karlstrom, G. *J. Chem. Phys.* **2004**, *121*, 4494. (b) Soderhjelm, P.; Krogh, J. W.; Karlstrom, G.; Ryde, U.; Lindh, R. *J. Comput. Chem*. **2007**, *28*, 000.

(16) (a) Angyan, J. G.; Chipot, C.; Dehez, F.; Hattig, C.; Jansen, G.; Millot, C. *J. Comput. Chem*. **2003**, *24*, 997. (b) Popelier, P. L. A.; Joubert, L.; Kosov D. S. *J. Phys. Chem. A* **2001**, 105, 8254. (c) Bader, R. F. W. *Atoms in Molecules: a Quantum Theory*; Oxford University Press: 1990.

(17) Case, D. A.; Cheatham, T. E., II; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem*. **2005**, *26*, 1668.

(18) (a) Masella, M.; Flament, J.-P. *J. Chem. Phys*. **1997**, *107*, 9105. (b) Masella, M.; Flament, J.-P. *Mol. Phys.* **1998**, *95*, 97. (c) Cuniasse, P.; Masella, M. *J. Chem. Phys*. **2003**, *119*, 1874.

(19) (a) Langlet, J.; Claverie, P.; Caron, F.; Boeuve, J.-C. *Int. J. Quantum Chem*. **1981**, *19*, 299. (b) Derepas, A. L.; Soudan, J. M.; Brenner, V.; Dognon, J. P.; Millié, P. *J. Comput. Chem* **2002**, *23*, 1013.

(20) Dang, L. X.; Chang, T. M. J. *Chem Phys.* **1997**, *106*, 8149.

(21) (a) Rappé, A. K.; Goddard, W. A., III *J. Phys. Chem.* **1991**, *95*, 3358. (b) Liu, Y. P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739. (c) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys*. **1999**, *110*, 741. (d) Chelli, R.; Procacci, P. *J. Chem. Phys.* **2002**, *117*, 9175.

(22) (a) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys*. **2003**, *119*, 5185. (b) Yu, H.; Hansson, T.; van Gunsteren, W. L. *J. Chem. Phys*. **2003**, *118*, 221. (c) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput*. **2006**, *2*, 1587.

(23) (a) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341. (b) van Duijnen, P. T.; Swart, M. J. *Phys. Chem. A* **1998**, *102*, 2399. (c) Chelli, R.; Procacci P. *J. Chem. Phys.* **2002**, *117*, 9175. (d) Piquemal, J. P.; Chelli, R.; Proccaci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, 111, 8170. (e) Elking, D.; Darden, T. A.; Woods R. J. *J. Comput. Chem.* **2007**, *28*, 1261.

(24) Gresh, N. *J. Chim.-Phys. Chim. Biol*. **1997**, *94*, 1365.

(25) (a) Kitaura, K.; Morokuma, K. *Int. J. Quantum Chem.* **1976**, *10*, 325. (b) Stevens, W. J.; Fink W. *Chem. Phys. Lett.* **1987**, *139*, 15. (c) Bagus, P. S.; Hermann, K.; Bauschlicher, C. W., Jr. *J. Chem. Phys*. **1984**, *80*, 4378. (d) Bagus, P. S.; Illas F. *J. Chem. Phys.* **1992**, *96*, 896. (e) Piquemal, J.-P.; Márquez, A.; Parisel, O; Giessner-Prettre, C. *J. Comput. Chem.* **2005**, *26*, 1052.. (f) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887. (g) Langlet, J.; Caillet, J.; Bergès, J. Reinhardt, P. *J. Chem. Phys.* **2003**, *118*, 6157.

(26) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 4915.

(27) (a) Piquemal, J. P.; Cisneros, A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys*. **2006**, *125*, 104101. (b) Cisneros, A.; Piquemal, J. P.; Darden, T. A. *J. Chem. Phys.* **2006,** *125*, 184101.

(28) (a) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem*. **1986**, *29*, 101. (b) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856. (c) Gresh, N.; Leboeuf, M.; Salahub, D. R. In *Modelling the Hydrogen Bond*; ACS Symposium Series 569; Smith, D. A., Ed.; 1994; p 82. (d) Gresh, N.; Guo, H.; Kafafi, S. A.; Salahub, D. R.; Roques, B. P. *J. Am. Chem. Soc.* **1999**, *121*, 7885. (e) Gresh, N.; Piquemal, J.-P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113.

(29) Vigné-Maeder, F.; Claverie, P. *J. Chem Phys.* **1988**, *88*, 4934.

(30) (a) Dreyfus, M. Ph.D. Thesis, University of Paris, 1970. (b) Claverie, P. Ph.D. Thesis, Paris, 1973, CNRS library number A.O. 8214. (c) Claverie, P. In *Localization and Delocalization in Quantum Chemistry*; Chalvet, O., Daudel, R., Diner, S., Malrieu, J. P., Eds.; Reidel: Dordrecht, Vol. II, p 127.

(31) (a) Rein, R.; Rabinowitz, J. R.; Swissler, T. J. *J. Theor. Biol.* **1972**, *34*, 215. (b) Rein, R. *Adv. Quantum Chem*. **1973**, *7*, 335.

(32) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1979**, *Symp. 11*, 253.

(33) Gresh, N. *Biochim. Biophys. Acta* **1980**, *597*, 345.

(34) Gresh, N.; Pullman, B. *Biochim. Biophys. Acta* **1980**, *625*, 356.

(35) Gresh, N.; Pullman, B. *Biochim. Biophys. Acta* **1980**, *608*, 47.

(36) Gresh, N.; Etchebest, C.; de la Luz, Rojas, O.; Pullman, A. *Int. J. Quantum Chem*., *Quantum Chem. Symp.* **1981**, *8*, 109.

(37) Gresh, N.; Pullman, A. *Int. J. Quantum Chem*. **1982**, *22*, 709.

(38) Gresh, N.; Pullman, A. *Int. J. Quantum Chem*., *Quant. Biol. Symp*. **1983**, *10*, 215.

(39) (a) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353. (b) Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824.

(40) (a) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233. (b) Stone A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047.

(41) Etchebest, C.; Lavery, R.; Pullman, A. *Theor. Chim. Acta* **1982**, *62*, 17.

(42) (a) Sokalski, W. A.; Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86. (b) Sokalski, W. A.; Sawaryn, A. *J. Chem. Phys.* **1987**, *87*, 526.

(43) (a) Karlstrom, G.; Linse, P.; Wallqvist, A.; Jonsson, B. *J. Am. Chem. Soc.* **1983**, *105*, 3777. (b) Andersson, M.; Karlstrom, G. *J. Phys. Chem.* **1985**, *89*, 4957.

(44) (a) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, A.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293. (b) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys*. **2003**, *112*, 7300. (c) Schmidt, M. W.; Baldridge, K. K.; Boatz, J.; A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem*. **1993**, *14*, 1347.

(45) Murrell, J. N.; Teixeira-Dias, J. J. C. *Mol. Phys*. **1970**, *19*, 521.

(46) Garmer, D. R.; Stevens, W. J. *J. Phys. Chem. A* **1989**, *93*, 8263.

(47) Murrell, J. N.; Randic, M.; Williams, D. R. *Proc. R. Soc. London, Ser. A* **1966**, *284*, 566.

(48) Gresh, N.; Claverie, P.; Pullman, A. *Int. J. Quantum Chem.* **1982**, *22*, 199.

(49) Creuzet, S.; Langlet, J.; Gresh, N. *J. Chim.-Phys. Phys. Chim. Biol*. **1991**, *88*, 2399.

(50) Gresh, N.; Pullman, A.; Claverie, P. *Theor. Chim. Acta* **1985**, *67*, 11.

(51) (a) Langlet, J.; Claverie, P.; Caillet, J.; Pullman, A. *J. Phys. Chem*. **1988**, *92*, 1617. (b) Langlet, J.; Gresh, N.; Giessner-Prettre, C. *Biopolymers* **1995**, *36*, 765. (c) Huron, M.-J.; Claverie, P.; *J. Phys. Chem.* **1972**, *76*, 2123. (d) Huron, M.-J.; Claverie, P. *J. Phys. Chem.* **1974**, *78*, 1853. (e) Pierotti, R. A. *J. Phys. Chem.* **1965**, *69*, 2813.

(52) Gresh, N. *Biopolymers* **1985**, *24*, 1527.

(53) (a) Chen, K.-X.; Gresh, N.; Pullman, B. *Nucleic Acids. Res.* **1986**, *14*, 3799. (b) Chen, K.-X.; Gresh, N.; Pullman, B. *Mol. Pharmacol*. **1986**, *30*, 279. (c) Chen, K.-X.; Gresh, N.; Pullman, B. *Nucleic Acids. Res*. **1988**, *16*, 3061. (d) Gresh, N.; Pullman, B.; Arcamone, F.; Menozzi, M.; Tonani, R. *Mol. Pharmacol.* **1989**, *35*, 251.

(54) (a) Gresh, N.; Pullman, B. *Mol. Pharmacol.* **1986**, *29*, 355. (b) Gresh, N. *Mol. Pharmacol.* **1987**, *31*, 617.

(55) (a) Gresh, N.; Pullman, A. *New. J. Chem.* **1986**, *10*, 405. (b) Gresh, N. *New. J. Chem.* **1986**, *11*, 61.

(56) (a) Gresh, N.; Roques, B.-P. *Biopolymers* **1997**, *41*, 145. (b) Garmer, D. R.; Gresh, N.; Roques, B.-P. *Proteins* **1998**, *31*, 42.

(57) Gresh, N.; Policar, C.; Giesner-Prettre, C. *J. Phys. Chem. A* **2002**, *106*, 5660.

(58) Piquemal, J. P.; Williams-Hubbard, B.; Fey, N.; Deeth, R. J.; Gresh, N.; Giessner-Prettre, C. *J. Comput. Chem.* **2003**, *24*, 1963.

(59) (a) Schäffer, C. E.; Jørgensen, C. K. *Mol. Phys.* **1964**, *9*, 401. (b) Gerloch, M.; Harding, J.-H.; Wooley, R. G. *Struct. Bonding* **1981**, *46*, 1. (c) Bridgeman, A. J.; Gerloch, M. *Prog. Inorg. Chem.* **1997**, *45*, 179. (e) Burton, V.J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. *J. Am. Chem. Soc.* **1995**, *117*, 8407. (f) Deeth, R. J. *Coord. Chem. Rev.* **2001**, *212*, 11.

(60) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys*. **1984**, *81*, 6026.

(61) Evangelakis, G. A.; Rizos, J. P.; Lagaris, I. E.; Demetropoulos, I. N. *Comput. Phys. Commun*. **1987**, *46*, 401.

(62) (a) Gresh, N. *J. Phys. Chem. A* **1997**, *101*, 8690. (b) Masella, M.; Gresh, N.; Flament, J. P. *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 2745.

(63) Guo, H.; Gresh, N.; Roques, B. P.; Salahub, D. R. *J. Phys. Chem. B* **2000**, *104*, 9746.

(64) Hodges, M. P.; Stone, A. J.; Xantheas, S. J. *Phys. Chem. A* **1997**, *101*, 9163.

(65) Gresh, N.; Garmer, D. R. *J. Comput. Chem.* **1996**, *17*, 1481.

(66) (a) Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W. *J. Comput. Chem.* **2000**, *21*, 1011. (b) Tiraboschi, G.; Roques, B. P.; Gresh, N. *J. Comput. Chem.* **1999**, *20*, 1379.

(67) Axilrod, B. M.; Teller, E. *J. Chem. Phys.* **1943**, *11*, 299.

(68) Concha, N. O.; Rasmussen, B. A.; Bush, K.; Herzberg, O. *Structure (London)* **1996**, *4*, 623.

(69) Krauss, M.; Gilson, H. S. R.; Gresh, N. *J. Phys. Chem. B* **2001**, *105*, 8040.

(70) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.

(71) (a) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914. (b) Murphy, R. B.; Beachy, M. D.; Friesner, R. A. *J. Chem. Phys.* **1995**, *103*, 1481.

(72) Rogalewicz, F.; Gresh, N.; Ohanessian, G. *J. Comput. Chem.* **2000**, *21*, 963.

(73) Tiraboschi, G.; Fournie-Zaluski, M. C.; Roques, B. P.; Gresh, N. *J. Comput. Chem.* **2001**, *22*, 1038.

(74) Roques, B. P.; Noble, F.; Fournié-Zaluski, M. C.; Beaumont, A. *Pharmacol. Rev.* **1993**, *45*, 88.

(75) Ledecq, L.; Lebon, F.; Durant, F.; Giessner-Prettre, C.; Marquez, A.; Gresh, N. *J. Phys. Chem. B* **2003**, *107*, 10640.

(76) Gresh, N.; Shi, G.-B. *J. Comput. Chem.* **2004**, *25*, 160.

(77) Biaszcyk, J.; Shi, G. B.; Yan, H.; Ji, X. *Structure (London)* **2000**, *8*, 1049.

(78) Gresh, N.; Kafafi, S. A.; Truchon, J.-F.; Salahub, D. R. J. *Comput. Chem.* **2004**, *25*, 823.

(79) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908. (b) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 2993.

(80) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053.

(81) Proynov, E. I.; Sirois, S.; Salahub, D. R. *Int. J. Quantum Chem.* **1997**, *64*, 427.

(82) Kafafi, S. A.; El-Gharkawy, E. R. H. *J. Phys. Chem. A* **1998**, *102*, 3202.

(83) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785. (b) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648.

(84) (a) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741. (b) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515.

(85) Gresh, N.; Tiraboschi, G.; Salahub, D. R. *Biopolymers* **1998**, *45*, 405.

(86) Gresh, N.; Sponer, J. *J. Phys. Chem. B* **1999**, *103*, 11415.

(87) Gresh, N.; Sponer, J. E.; Spacková, N.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2003**, *107*, 8669.

(88) (a) Derreumaux, P. J. *Chem. Phys.* **1999**, *111*, 2301. (b) Derreumaux, P. *J. Chem. Phys.* **2002**, *117*, 3499

(89) Gresh, N.; Derreumaux, P. *J. Phys. Chem. B* **2003**, *107*, 4862.

(90) Yun, M. R.; Lavery, R.; Mousseau, N.; Zakrzewska, K.; Derreumaux, P. *Proteins: Struct., Genet., Bioinformatics* **2006**, *63*, 967**.**

(91) Antony, J.; Gresh, N.; Olsen, L.; Hemmingsen, L.; Schofield, C.; Bauer, R. *J. Comput. Chem.* **2002**, *23*, 1281.

(92) Antony, J.; Piquemal, J.-P.; Gresh, N. *J. Comput. Chem.* **2005**, *26*, 1131.

(93) (a) *Accelrys software*; 9645 Scranton Road, San Diego, CA, U.S.A. (b) Wu, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825.

(94) Concha, N. O.; Janson, C. A.; Rowling, P.; Pearson, S.; Cheever, C. A.; Clarke, B. P.; Lewis, C.; Galleni, M.; Frère, J. M.; Payne, D. J.; Bateson, J. H.; Abdel-Meguid, S. S. *Biochemistry* **2000**, *39*, 4288.

(95) (a) Payton, M. A.; Rheinnecker, M.; Klig, L. S.; DeTiani, M.; Bowden, E. *J. Bacteriol.* **1991**, *173*, 2006. (b) Smith, D. J.; Proudfoot, A. E. I.; De Tiani, M.; Wells, T. N. C.; Payton, M. A. *Yeast* **1995**, *11*, 301. (c) Patterson, J. H.; Waller, R. F.; Jeevarajah, D.; Billman-Jacobe, H.; McConville, M. J. *Biochem. J.* **2003**, *372*, 77. (d) Garami, A.; Ilg, T. *J. Biol. Chem.* **2001**, *276*, 6566. (e) Shinabarger, D.; Berry, A.; May, T. B.; Rothmel, R.; Fialho, A.; Chakrabarty, A. M. *J. Biol. Chem.* **1991**, *266*, 2080.

(96) Roux, C.; Lee, J. H.; Jeffery, C. J.; Salmon, L. *Biochemistry* **2004**, *43*, 2926.

(97) Roux, C.; Gresh, N.; Perera, L. E.; Piquemal, J. P.; Salmon, L. *J. Comput. Chem***. 2007**, *28*, 938.

(98) Cleasby, A.; Wonacott, A.; Skarzynski, T.; Hubbard, R. E.; Davies, G. J.; Proudfoot, A. E. I.; Bernard, A. R.; Payton, M. A.; Wells, T. N. C. *Nature Struct. Biol.* **1996**, *3*, 470.

(99) (a) Rice, W. G.; Turpin, J. A.; Huang, M.; Clanton, D.; Buckheit, R. W., Jr.; Covell, D. G.; Wallqvist, A.; McDonnell, N. B.; DeGuzman, R. N.; Summers, M. F.; Zalkow, L.; Bader, J. P.; Haugwitz, R. D.; Sausville, E. A. *Nat. Med.* **1997**, 3, 341. (b) Huang, M.; Maynard, A.; Turpin, J. A.; Graham, L.; Janini, G. M.; Covell, D. G.; Rice, W. G. *J. Med. Chem.* **1998**, *41*, 1371. (c) Goel, A.; Mazur, S. J.; Fattah, R. J.; Hartman, T. L.; Turpin, J. A.; Huang, M.; Rice, W. G.; Appella, E.; Inman, J. K. *Bioorg. Med. Chem. Lett.* **2002**, 12, 767. (d) Turpin, J. A.; Song, Y.; Inman, J. K.; Huang, M.; Wallqvist, A.; Maynard, A.; Covell, D. G.; Rice, W. G.; Appella, E. *J. Med. Chem.* **1999**, 42, 67. (e) Srivastava, P.; Schito, M.; Fattah, R. J.; Hartman, T.; Buckheit, R. W; Turpin, J. A.; Appella, E. *Bioorg. Med. Chem.* **2004**, *12*, 6437.

(100) Miller, Jenkins, L. M.; Hara, T.; Durell, S. R.; Hayashi, R.; Inman, J. K.; Piquemal, J.-P.; Gresh, N.; Appella, E. *J. Am. Chem. Soc.* **2007**, *129*, 11067.

(101) Gordon, R. G.; Kim, Y. S. *J. Chem. Phys*. **1972**, *56*, 3122.

(102) Boys, S. F.; Shavitt, I. A *Fundamental Calculation of the Energy Surface for the System of Three Hydrogens Atoms;* NTIS: Springfield, VA, 1959; AD212985. (b) Dunlap, B. I; Connoly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 4993.

(103) Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A. *J. Chem. Phys.* **2005**, *123*, 044109.

(104) Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A. *J. Phys. Chem. B (Letter)* **2006**, *110*, 13682.

(105) Roothaan, C. C. J. *Rev. Mod. Phys*. **1960**, *23*, 69.

(106) McMurchie, L. E.; Davidson, E. R. *J. Comput. Phys*. **1978**, *26*, 218.

(107) (a) Wheatley, R. J.; Price, S. *Mol. Phys*. **1990**, *69*, 50718. (b) Kita S.; Noda K.; Inouye, H. *J. Chem. Phys*. **1976**, *64*, 3446.

(108) (a) van Duijneveldt-van deRijdt, J. G. C. M.; Mooij, W. T. M.; van Duijneveldt, F. B. *Phys. Chem. Chem. Phys*. **2003**, *5*, 1169. (b) Tschumper G. S.; Leininger M. L.; Hoffman B. C.; Valeev, E. F.; Quack M.; Schaffer, H. F., III *J. Chem. Phys*. **2002**, *116*, 690.

(109) Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6692.

(110) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

(111) Challacombe, M.; Schwegler, E.; Almlöf, J. *Computational Chemistry: Review of Current Trends*; World Scientific: Singapore, 1996.

(112) Volkov, A.; Coppens, P. *J. Comput. Chem*. **2004**, *25*, 921.

(113) (a) Sagui, C.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2004**, *120*, 73. (b) Toukmaji, a.; Sagui, C.; Board, J.; Darden, T. A. *J. Chem. Phys*. **2000**, *113*, 10913.

(114) (a) Essmann, M.; Perera, L.; Berkowitz, M.; Darden, T. A.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.***1995**, *103*, 8577. (b) Sagui, C.; Pedersen, L. E.; Darden, T. A. *J. Chem. Phys.* **2004**, *120*, 73.

(115) York, D.; Yang, W. *J. Chem. Phys*. **1994**, *101*, 3298

(116) Cisneros, G. A.; Elking, D.; Piquemal, J.-P.; Darden, T. A. *J. Phys. Chem. A* **2007**, DOI: 10.1021/jp074817r.

(117) Gourlaouen, C.; Piquemal, J.-P.; Parisel, O. *J. Chem. Phys.* **2006**, *124*, 174311.

(118) Gourlaouen, C.; Piquemal, J. P.; Saue, T.; Parisel, O. *J. Comput. Chem.* **2006**, *27*, 142.

(119) Cisneros, G. A.; Liu, H.; Zhang, Y.; Yang, W. J. *Am. Chem. Soc.* **2003**, *125*, 10384.

(120) Cisneros, G. A.; Wang, M.; Silinski, P.; Yang, W.; Fitzgerald, M. C. *Biochemistry* **2004,** *43*, 6885.

CT700134R

JCTC Journal of Chemical Theory and Computation

Polarization Effects for Hydrogen-Bonded Complexes of Substituted Phenols with Water and Chloride Ion

William L. Jorgensen,* Kasper P. Jensen, and Anastassia N. Alexandrova

Department of Chemistry, Yale University, 225 Prospect Street,
New Haven, Connecticut 06520-8107

Received July 13, 2007

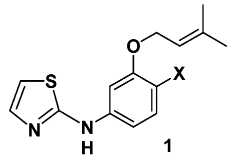
Abstract: Variations in hydrogen-bond strengths are investigated for complexes of nine *para*-substituted phenols (XPhOH) with a water molecule and chloride ion. Results from ab initio HF/6-311+G(d, p) and MP2/6-311+G(d, p)//HF/6-311+G(d, p) calculations are compared with those from the OPLS-AA and OPLS/CM1A force fields. In the OPLS-AA model, the partial charges on the hydroxyl group of phenol are not affected by the choice of *para* substituent, while the use of CM1A charges in the OPLS/CM1A approach does provide charge redistribution. The ab initio calculations reveal a 2.0-kcal/mol range in hydrogen-bond strengths for the XPhOH...OH_2 complexes in the order X = NO_2 > CN > CF_3 > Cl > F > H > OH > CH_3 > NH_2. The pattern is not well-reproduced with OPLS-AA, which also compresses the variation to 0.7 kcal/mol. However, the OPLS/CM1A results are in good accord with the ab initio findings for both the ordering and range, 2.3 kcal/mol. The hydrogen bonding is, of course, weaker with XPhOH as acceptor, the order for X is largely inverted, and the range is reduced to ca. 1.0 kcal/mol. The substituent effects are found to be much greater for the chloride ion complexes with a range of 11 kcal/mol. For quantitative treatment of such strong ion−molecule interactions the need for fully polarizable force fields is demonstrated.

Introduction

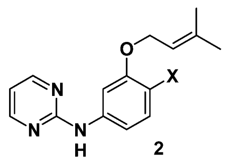
In our development of non-nucleoside inhibitors of HIV-1 reverse transcriptase (NNRTIs), high sensitivity to substitution at the 4-position in the phenyl ring has been found for the thiazole series **1** and the pyrimidines **2**, as summarized in Table 1.[1,2] Specifically, for the thiazoles, there is a 50-fold enhancement in activity as the substituent X is made more electronegative in going from X = H to CN, while a 1500-fold enhancement is obtained in the pyrimidine series. The structures of the complexes of such NNRTIs with HIV-RT have been well established through X-ray crystallography and computation.[3−5] A key feature is a short hydrogen bond between the amino group of the NNRTI and the carbonyl oxygen of Lys101 of HIV-RT (Figure 1). The β-nitrogen in the heterocycle is also in a longer hydrogen bond with the backbone NH of Lys101.

Questions that then arise are (a) how sensitive are such hydrogen-bond strengths to substitution in the phenyl rings and (b) are such effects adequately reflected in the force-field calculations that are often used to examine the energetics of protein−ligand binding.[6−8] For example, successful guidance of lead-optimization by performing free-energy perturbation calculations to predict the effects of changes in

Table 1. Anti-HIV Activity (EC_{50} in μM for Protection of MT-2 Cells) for Compounds **1** and **2**[a]



1



2

X	1	2
H	10.0	30.0
CH_3	3.0	2.8
Cl	0.30	0.20
CN	0.21	0.02

[a] References 1 and 2.

* Corresponding author e-mail: william.jorgensen@yale.edu.

10.1021/ct7001754 CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/04/2007

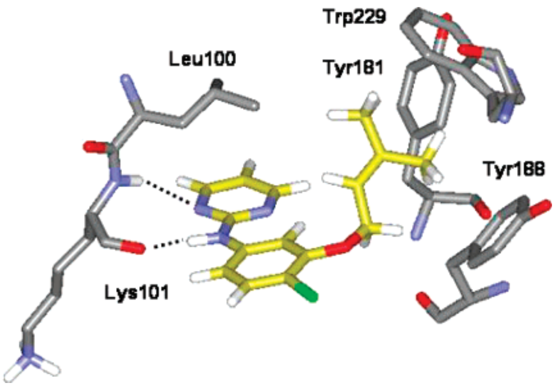
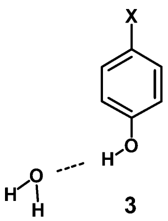


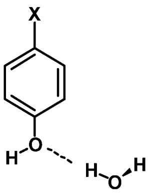
Figure 1. Partial computed structure for **2** (X = Cl) bound to HIV-1 reverse transcriptase highlighting the hydrogen bonds with the backbone of Lysine101. Carbon atoms of the ligand are colored gold for clarity.

substituents on rings and of choices of heterocycles on binding affinities is expected to require proper representation of such effects.[1,9,10] Electronic polarization is a central issue here since change in X alters the charge distribution including for the key hydrogen-bond donating hydrogen of the ligand.[7,11] Related effects on acidities of substituted benzoic acids and phenols led to the development of the Hammett equation and the σ and σ^- substituent constants.[12]

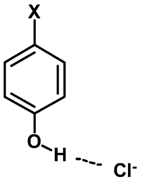
As summarized below, these ideas were pursued by performing ab initio calculations on prototypical hydrogen-bonded systems and comparing the results to those obtained from the nonpolarizable OPLS-AA force field[13] and its OPLS/CM1A variant,[8] which incorporates polarized partial atomic charges that are obtained from the quantum mechanical CM1A procedure.[14] The CM1A method, which is based on AM1 wave functions, was derived to reproduce dipole moments for organic molecules in the gas phase.[14] The CM1A charges when enhanced by 14% (1.14*CM1A) were also found to perform well for computing free energies of hydration of 25 diverse organic molecules[15] in explicit TIP4P water[16] with all other force-field parameters taken from OPLS-AA. The systems chosen for initial study of substituent effects on hydrogen bonding are the phenol−water and phenol−chloride ion complexes, **3**−**5**.



3



4



5

Computational Details

The principal interest here is comparison of ab initio and force-field predictions for the effects of the substituents X on the hydrogen-bond strengths. Ab initio and density functional theory calculations were carried out with Gaussian 03, and all geometrical degrees of freedom were optimized for the complexes and separated components.[17] In Table 2, results for the PhOH...OH$_2$ (**3**) complex and water dimer are compared at the HF/6-31G(d), B3LYP/6-31G(d), HF/6-311+G(d, p), and MP2/6-311+G(d, p)// HF/6-311+G(d, p) levels. The interaction energies at the latter two levels bracket what is accepted as the true value for the water dimer, −5.4 ± 0.7 kcal/mol from experiment[18] and −5.1 ± 0.2 kcal/mol from theory.[19] The ΔE results for the phenol−water complex are also similar to those from the highest-level calculations in a prior study, i.e., −7 to −8 kcal/mol at the MP2/aug-cc-pVDZ level.[20] Thus, the substituent effects were explored with the HF/6-311+G(d, p) and MP2/6-311+G(d, p) calculations. Counterpoise corrections have not been made since they are expected to show little variation with the choice of substituent X.

Table 2. Computed Interaction Energies (kcal/mol) and OO Distances (Å)[a]

	PhOH---OH$_2$		(H$_2$O)$_2$	
method	$-\Delta E$	r(OO)	$-\Delta \boldsymbol{E}$	r(OO)
HF/6-31G(d)	7.35	2.901	5.62	2.971
B3LYP/6-31G(d)	9.70	2.808	7.68	2.861
HF/6-311+G(d, p)	6.24	2.940	4.83	3.000
MP2/6-311+G(d, p)[b]	8.13	(2.940)	5.91	(3.000)
MP2/6-311++G(2d, 2p)[c]			5.44	2.911

[a] For A + H$_2$O → A−H$_2$O, $\Delta E = E(\text{A}-\text{H}_2\text{O}) - E(\text{A}) - E(\text{H}_2\text{O})$. [b] Using HF/6-311+G(d, p) optimized structures. [c] Reference 19a.

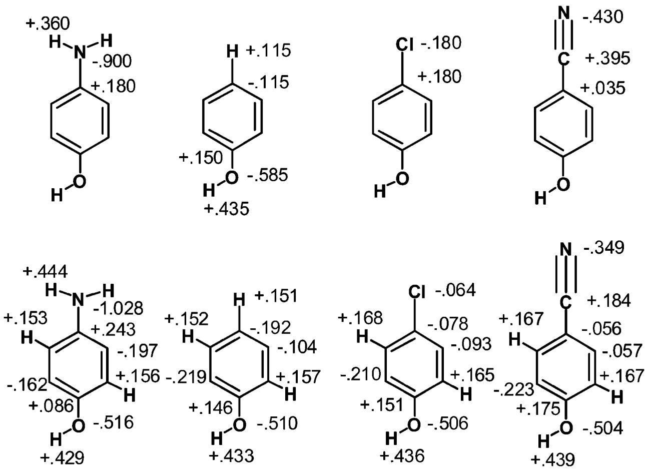


Figure 2. Examples of OPLS-AA (top) and OPLS/CM1A (bottom) atomic charges for substituted phenols. All unsubstituted phenyl C and H atoms have charges of −0.115 e and +0.115 e in the OPLS-AA model. The CM1A charges for neutral molecules are scaled by a factor of 1.14 for the OPLS/CM1A force field.

The corresponding force-field calculations were carried out first with the substituted benzenes described with the OPLS-AA force field[21] and with the water molecule represented by the TIP4P model.[16] Complete energy minimizations were carried out with the BOSS program[22] except that the internal geometry of the water molecule is fixed in the TIP4P model, $r(\text{OH}) = 0.9572$ Å and $\angle\text{HOH} = 104.52°$. Notably, in the reported OPLS-AA model for *para*-substituted benzenes, the net charge on the substituent plus attached benzene carbon atom is zero.[21] This permits transferability that simplifies the modeling of arbitrary substituted benzenes, but it ignores associated polarization effects. Thus, the partial atomic charges on the COH group of all *para*-substituted phenols are the same (Figure 2). Though testing for numerous mono- and disubstituted benzenes has revealed modest

Table 3. Computed Interaction Energies (−Δ*E*, kcal/mol) for Complexes **3**

X	σ^{a}	HF[b]	MP2[c]	OPLS-AA	OPLS/CM1A
NH_2	−0.57	5.90	7.79	7.19	7.03
CH_3	−0.14	6.05	7.98	7.00	7.23
OH	−0.38	6.18	8.10	7.22	7.40
H	0.0	6.24	8.13	7.09	7.28
F	0.15	6.65	8.57	7.51	7.88
Cl	0.24	6.87	8.74	7.50	7.95
CF_3	0.53	7.24	9.20	7.67	8.34
CN	0.70	7.64	9.54	7.59	8.64
NO_2	0.81	7.94	9.72	7.66	9.33
mue		1.90	(0)	1.26	0.74

[a] Hammett σ constant (ref 12). [b] HF/6-311+G(d, p). [c] MP2/6-311+G(d, p)//HF/6-311+G(d, p).

average errors for computed free energies of hydration (0.5 kcal/mol) and pure liquid heats of vaporization (1.0 kcal/mol) and densities (0.02 g/mL),[21,23] differential polarization effects are expected to be more apparent upon examination of specific hydrogen-bond strengths as in protein−ligand binding.

For the OPLS/CM1A approach,[8] OPLS-AA parameters are used except for the partial atomic charges, which are obtained from the CM1A method.[14] A sequence of geometry optimizations and CM1A calculations is performed with a BOSS script until the charges are converged. For neutral molecules, it is noted again that the CM1A charges are scaled by a factor of 1.14 for use in the OPLS/CM1A force field.[15] For ions, the CM1A charges are not scaled to avoid nonphysical net charges.[24] The program also symmetrizes the charges for equivalent atoms, e.g., the charges are averaged for equivalent methyl hydrogens or the *ortho* carbons and hydrogens in Figure 2. Without the symmetrization, artifacts arise for general molecular modeling such as introduction of spurious minima in conformational searching. Optimization of the complexes is then performed with the converged charges and with the internally rigid TIP4P water molecule. Further polarization of the charge distribution for the substituted benzenes upon complex formation with water is not carried out. For the much stronger phenol−chloride ion interactions, the importance of a full treatment of polarization effects is considered below.

Results and Discussion

XPhOH−Water Complexes 3. The computed interaction energies Δ*E* for the complexes **3** with the phenol as the hydrogen-bond donor are summarized in Table 3, and the optimized OO distances are in Table 4. The trend in the ab initio results is as expected with electron-withdrawing substituents acidifying the hydroxyl group, increasing the hydrogen-bond strengths (Table 3), and decreasing the hydrogen-bond lengths (Table 4). However, there are fine points. For example, the π-donating character of the amino substituent outweighs its σ-withdrawing character to yield the weakest hydrogen bond. For fluorine and chlorine, the opposite pattern seems to be operative as the hydrogen bonds are stronger than for phenol (X = H) in those cases. As shown in Figure 3, the ab initio hydrogen-bond strengths roughly follow the trend of Hammett σ constants, though this is not fully expected in view of the differences in the processes, i.e., substituent effects on phenol−water hydrogen-bond strengths and on acidities of substituted benzoic acids in aqueous solution.

Table 4. Computed Oxygen−Oxygen Distances (Å) for Complexes **3** and **4**[a]

X	HF[b]	OPLS-AA	OPLS/CM1A
NH_2	2.955, 3.029	2.734, 2.767	2.759, 2.844
CH_3	2.945, 3.044	2.734, 2.766	2.757, 2.851
OH	2.947, 3.028	2.733, 2.769	2.760, 2.837
H	2.940, 3.050	2.728, 2.767	2.754, 2.850
F	2.936, 3.036	2.731, 2.770	2.751, 2.843
Cl	2.926, 3.048	2.729, 2.770	2.749, 2.846
CF_3	2.913, 3.056	2.731, 2.772	2.746, 2.853
CN	2.904, 3.061	2.730, 2.759	2.739, 2.845
NO_2	2.895, 3.068	2.731, 2.772	2.735, 2.867

[a] Values *x, y* are for **3** and **4**. [b] HF/6-311+G(d, p).

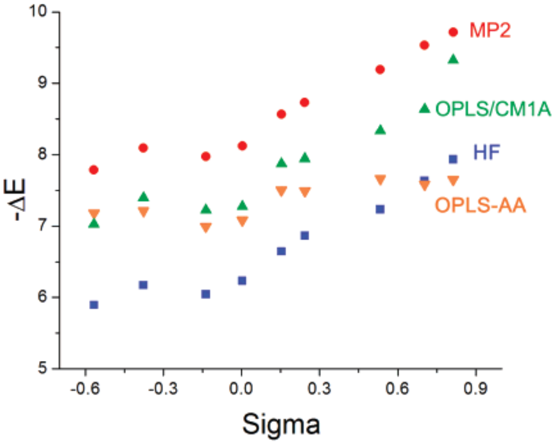


Figure 3. Computed interaction energies for the *p*-XPhO-H...OH_2 complexes **3** vs σ(X).

The substituent effects on the hydrogen-bond strengths are substantial with a 2 kcal/mol range from both the HF and MP2 calculations for the complexes **3**. In view of the constancy of the OPLS-AA partial charges for the phenolic hydroxyl group, it is not surprising that the range for Δ*E* is compressed to 0.7 kcal/mol and the ordering of the values is poor. There is also negligible variation in the OO distances in Table 4 with OPLS-AA, while the HF results show a reduction of the hydrogen-bond lengths by 0.06 Å in going from X = NH_2 to NO_2 for the complexes **3**. In contrast, use of the 1.14*CM1A charges nicely corrects the problems with the interaction energies and yields absolute values roughly midway between the HF and MP2 results (Table 3 and Figure 3). The level of accord was not anticipated, but it suggests that the polarization of the charge distributions by the CM1A method is remarkably accurate within the context of the simple point charge model for the force fields (single atom-centered partial charges). The hydrogen-bond lengths with OPLS/CM1A also decrease with increasing strength, though the range is less than from the HF optimizations (Table 4). The absolute hydrogen-bond lengths are 0.15−0.20 Å shorter from the force fields than from the ab initio calculations, which is normal for fixed-charge force fields that are intended for condensed-phase simulations.[7,8,13,16]

Table 5. Computed Interaction Energies (−Δ*E*, kcal/mol) for Complexes **4**

X	HF[a]	MP2[a]	OPLS-AA	OPLS/CM1A
NH_2	3.99	5.42	5.95	5.51
CH_3	3.74	5.19	6.05	5.23
OH	3.88	5.48	5.93	4.92
H	3.60	5.05	5.98	5.16
F	3.55	5.20	5.72	4.80
Cl	3.39	5.13	5.73	4.73
CF_3	3.22	5.11	5.65	4.44
CN	3.11	4.94	5.85	4.35
NO_2	3.00	4.90	5.65	3.87
mue	1.66	(0)	0.68	0.49

[a] As in Table 3.

Table 6. Interaction Energies (−Δ*E*) and Enthalpies (−Δ*H*) for Complexes **5** (kcal/mol)

X	HF[a]	MP2[a]	OPLS-AA	OPLS/CM1A	−Δ*H*, exptl[b]
NH_2	17.01	23.49	15.39	14.98	
CH_3	17.60	24.38	14.25	15.69	24.1
OH	18.11	24.70	15.39	16.25	
H	18.34	24.95	14.90	16.06	24.5
F	20.94	27.62	17.64	19.13	26.4
Cl	22.17	28.61	17.66	19.46	28.1
CF_3	24.24	31.79	18.87	21.34	
CN	26.80	33.84	18.44	23.38	33.6
NO_2	28.19	34.93	18.93	26.77	
mue	6.77	(0)	11.43	9.03	

[a] As in Table 3. [b] Reference 25.

XPhOH−Water Complexes 4. Turning to the complexes **4** with the phenol as the hydrogen-bond acceptor, the trends for hydrogen-bond strengths and lengths from the ab initio calculations are now opposite with electron-withdrawing substituents weakening the basicity of the phenolic oxygen. Thus the MP2 results for **3** range from −7.8 to −9.7 kcal/mol in going from X = NH_2 to NO_2, while the corresponding values for **4** are −5.4 to −4.9 kcal/mol (Table 5), i.e., opposite in trend, much weaker, and in a narrower range. Qualitatively similar results are obtained from the HF calculations, though the range for the complexes **4** is 1.0 kcal/mol. As before, the OPLS-AA results are too invariant, while the OPLS/CM1A model is successful in showing the weakening of the hydrogen bond for **4** with increasing electron-withdrawing character for the substituent X.

A key point from Tables 3 and 5 is the increasing gap between the substituted phenol's ability to act as a hydrogen bond donor and acceptor with increasing electron-withdrawing character for X. E.g., for *p*-cyanophenol as donor and acceptor the difference in hydrogen-bond strengths is 4.6 kcal/mol from the MP2 results and 4.3 kcal/mol with OPLS/CM1A, while the difference is only 1.7 kcal/mol from the OPLS-AA calculations. It is clear that (a) such modulation of hydrogen-bonding ability is important for proper description of intermolecular interactions, and (b) its accurate description requires methodology that allows polarization of the charge distributions. It is also apparent from Figure 2 and Tables 3 and 5 that the hydrogen-bond strengths are very sensitive to small changes in the atomic charges. The variations for the hydroxyl oxygen and hydrogen are only ca. 0.01 e with OPLS/CM1A; the variation for the *ipso* carbon is actually much greater, 0.1 e. For phenol itself, if the OPLS-AA charges for the hydroxyl oxygen and hydrogen are changed by 0.01 e, the strength of the hydrogen-bond for the complex **3** changes by ca. ±0.3 kal/mol in the expected manner. This sensitivity is well-known and has always been a challenge in the development of force fields.[7,8]

XPhOH−Cl⁻ Complexes 5. Naturally, the substituent effects are much enhanced for the complexes with chloride ion, **5** (Table 6). It is noted that the OPLS chloride ion parameters ($q = -1.0$ e, $\sigma = 4.02$ Å, $\epsilon = 0.71$ kcal/mol) that were used here are from a recent, comprehensive study of the hydration of halide and alkali ions.[24] For the complexes with the substituted phenols, the ranges for the interaction energies are −17 to −28 (HF), −23 to −35 (MP2), −14 to −19 (OPLS-AA), and −15 to −27 kcal/mol (OPLS/CM1A). Some experimental data from high-pressure mass spectrometry are also available for complexes **5**, as listed in Table 6.[25] For anion−molecule complexes like these, conversion of the computed electronic energy change Δ*E* to Δ*H* at 298 K involves a correction of ca. +0.9 kcal/mol.[26] It is apparent that the MP2 results are in close accord with the experimental data, while the HF and force-field results significantly underestimate the hydrogen-bond strengths. However, the OPLS/CM1A approach again does much better than OPLS-AA for the magnitude, pattern, and range for the interaction energies. In this case, polarization of the substituted phenol by the chloride ion can be expected to be significant, and the fixed-charge OPLS-AA and OPLS/CM1A models are both inadequate. This is the case in spite of the fact that the optimal interaction energy for Cl^- with a TIP4P water molecule of −13.0 kcal/mol is in good agreement with the best available estimates.[24] The larger phenols are more polarizable than a water molecule.

For proper treatment of such strong ion−molecule interactions, it is accepted that a fully polarizable force field is required.[7,11] Thus, we have been exploring the addition of inducible dipoles to the OPLS models. A simple approach has been taken by which an inducible dipole can be added to non-hydrogen atoms. Furthermore, the electric field that determines the inducible dipoles is only computed from the permanent charges (eq 1), and the total polarization energy is given by the usual formula, eq 2. The key approximation

$$\vec{\mu}_i = \alpha_i \vec{\mathrm{E}}_i^{\,0} \tag{1}$$

$$E_{\mathrm{pol}} = -\,(^1/_2)\sum_i \vec{\mu}_i \bullet \vec{E}_i^{\,0} \tag{2}$$

is that the induced dipoles do not contribute to the electric field, which simplifies the computations since an iterative solution for the dipoles is not required. Addition of the induced dipoles to OPLS-AA and OPLS/CM1A yields OPLS-AAP and OPLS/CM1AP. The implementations are residue-based in that the electric field at an atom is determined by the charges on all other atoms not in the same residue. For the complex **5**, the substituted phenol and chloride ion are treated as separate residues, so the induced dipoles for the phenol are only determined by the field from the chloride ion. The same polarization model has been used

Table 7. Computed Interaction Energies ($-\Delta E$, kcal/mol) for Complexes **5** Using Polarizable Force Fields and Optimized O−Cl Distances (Å)

X	$-\Delta E$ OPLS-AAP	$-\Delta E$ OPLS/CM1AP	r(O−Cl) HF[a]	r(O−Cl) OPLS-AAP	r(O−Cl) OPLS/CM1AP
NH_2	20.48	19.93	3.181	3.119	3.146
CH_3	19.32	20.74	3.164	3.124	3.136
OH	20.59	21.40	3.164	3.111	3.129
H	19.95	21.08	3.154	3.121	3.133
F	22.83	24.35	3.145	3.111	3.120
Cl	22.91	24.76	3.125	3.113	3.118
CF_3	24.16	26.76	3.098	3.109	3.106
CN	23.69	28.89	3.080	3.112	3.101
NO_2	24.17	32.47	3.059	3.109	3.082
mue[b]	6.25	3.77			

[a] HF/6-311+G(d, p). [b] Mean unsigned error to MP2 ΔE in Table 6.

by others,[27,28] and it performed well in a previous study of ours for reproducing solvent effects for the gauche/anti equilibrium for 1,2-dichloroethane in multiple solvents and the free energy of solvation of water in cyclohexane.[29]

Modest parameter optimization has been carried out for the polarizabilities α to reproduce gas-phase complexation energies (MP2/6-311G(d, p)) for ca. 30 ion−molecule complexes focusing on cation-π interactions. This led to setting $\alpha_i = 1.0$ Å^3 for carbon and $\alpha_i = 1.5$ Å^3 for heteroatoms. With these choices, the phenol−chloride ion complexes were optimized yielding the results in Table 7. Inclusion of the induced dipoles is found to enhance the interaction energies by 5−6 kcal/mol. The hydrogen-bond lengths are also shortened by ca. 0.1 Å to yield the values that are listed in Table 7. The agreement between the ab initio and OPLS/CM1AP results is certainly respectable, while the OPLS-AAP results still suffer from the underlying problems with the invariant partial charges in the OPLS-AA model. For the phenol−water complexes, addition of the inducible dipoles to the force fields strengthens the hydrogen bonds uniformly by 0.4−0.5 kcal/mol and shortens them by ca. 0.02 Å.

Conclusion

Substituent effects on the interaction energies for complexes of phenols with water and chloride ion have been investigated. For the complexes with water, the OPLS/CM1A model was found to yield good reproduction of ab initio results, while the OPLS-AA force field with invariant partial charges for the hydroxyl group compresses the substituent effects. For the complexes with chloride ion, the interaction energies and substituent effects are much magnified, and the need for explicit treatment of the intermolecular polarization energy is apparent. Addition of inducible dipoles on non-hydrogen atoms was found to enhance the interaction energies by 5−6 kcal/mol. The resultant OPLS/CM1AP model performed well and warrants further investigation. It is emphasized that the ability to predict accurately substituent effects on intermolecular interactions is central to key applications of molecular modeling, for example, in the design of drugs, materials, and catalysts.

Acknowledgment. Gratitude is expressed to the National Science Foundation (CHE-0446920) and National Institutes of Health (GM32136) for support.

Supporting Information Available: Ab initio energies for the substituted phenols and complexes **3**−**5**. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 663−667.

(2) Ruiz-Caro, J.; Basavapathruni, A.; Kim, J. T.; Wang, L.; Bailey, C. M.; Anderson, K. S.; Hamilton, A. D.; Jorgensen, W. L. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 668−671.

(3) Ren, J.; Esnouf, R. M.; Hopkins, A. L.; Warren, J.; Balzarini, J.; Stuart, D. I.; Stammers, D. K. *Biochemistry* **1998**, *37*, 14394−14403.

(4) Blagović, M. U.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2003**, *125*, 6016−6017.

(5) Das, K.; Clark, A. D., Jr.; Lewi, P. J.; Heeres, J.; de Jonge, M. R.; Koymans, L. M. H.; Vinkers, H. M.; Daeyaert, F.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Ye, H.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; de Béthune, M.-P.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A. J.; Arnold, E. *J. Med. Chem.* **2004**, *47*, 2550−2560.

(6) Gohlke, H.; Klebe, G. *Angew. Chem. Int. Ed.* **2002**, *41*, 2644−2676.

(7) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27−85.

(8) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci U.S.A.* **2005**, *102*, 6665−6670.

(9) Thakur, V. V.; Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5664−5667.

(10) Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 15372−15373.

(11) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, *18*, 89−146.

(12) Smith, M. B.; March, J. *March's Advanced Organic Chemistry*, 5th ed.; Wiley: New York, 2001; Chapter 9.

(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(14) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87−110.

(15) Blagović, M. U.; Morales, de Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322−1332.

(16) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Phys. Chem.* **1983**, *79*, 926−935.

(17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai,

H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(18) Curtiss, L. A.; Frurip, D. J.; Blander, M. *J. Chem. Phys.* **1979**, *71*, 2703−2711.

(19) (a) Frisch, M. J.; Del Bene, J. E.; Binkley, J. S.; Schaefer, H. F., III *J. Chem. Phys.* **1986**, *84*, 2279−2289. (b) Halkier, A.; Koch, H.; Jorgensen, P.; Christiansen, O.; Beck, Nielsen, I. M.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 150−157.

(20) Feller, D.; Feyereisen, M. W. *J. Comput. Chem.* **1993**, *14*, 1027−1035.

(21) (a) Jorgensen, W. L.; Nguyen, T. B. *J. Comput. Chem.* **1993**, *13*, 195−205. (b) Jorgensen, W. L.; Laird, E. R.; Nguyen, T. B.; Tirado-Rives, J. *J. Comput. Chem.* **1993**, *13*, 206−215.

(22) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689−1700.

(23) Price, D. J.: Brooks, C. L., III *J. Comput. Chem.* **2005**, *26*, 1529−1541.

(24) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 1499−1509.

(25) (a) Cumming, J. B.; French, M. A.; Kebarle, P. *J. Am. Chem. Soc.* **1977**, *99*, 6999−7003. (b) Paul, G. J. C.; Kebarle, P. *Can. J. Chem.* **1990**, *68*, 2070−2077.

(26) Gao, J.; Garner, D. S.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1986**, *108*, 4784−4790.

(27) King, G.; Warshel, A. *J. Chem. Phys.* **1990**, *93*, 8682−8692.

(28) Straatsma, T. P.; McCammon, J. A. *Chem. Phys. Lett.* **1991**, *177*, 433−440.

(29) Jorgensen, W. L.; McDonald, N. A.; Selmi, M.; Rablen, P. R. *J. Am. Chem. Soc.* **1995**, *117*, 11809−11810.

CT7001754

JCTC Journal of Chemical Theory and Computation

Many-Body Polarization, a Cause of Asymmetric Solvation of Ions and Quadrupoles

Anders Öhrn* and Gunnar Karlström

Department of Theoretical Chemistry, Chemical Center, P.O. Box 124, S-221 00 Lund, Sweden

Received January 22, 2007

Abstract: Three models are used to study the effect of many-body polarization in the solvation of non-dipolar molecules and ions in water. Two of the models are very simplified and are used to show a number of basic principles of correlation of solvent degrees of freedom and asymmetric solvent structures. These principles are used to interpret results from the third model: an accurate simulation of para-benzoquinone (PBQ) in aqueous solution with a combined quantum chemical statistical mechanical solvent model with an explicit solvent. It is found that nonzero polarizability of PBQ introduces correlation in the solvent degrees of freedom through the many-body nature of the polarization. The fluctuating electric field from the solvent on the solute increases in magnitude with the correlation. Solvent effects are hence modified. This correlation is not described within the mean-field approximation. In practice, the correlation will show up as an increased probability for asymmetric solvation of the molecule.

1. Introduction

The solvation of molecules in liquid solvent or large organic assemblies, such as proteins or micelles, constitutes a large and important part of chemistry. Most chemistry, after all, takes place in an environment. Along improving computers and quantum chemical methods, our comprehension and ability to predict properties of molecular matter have been taken to new levels of detail, accuracy, and size. Useful as this may be, to be able to pinpoint and characterize the features relevant to the question of the particular system and thus get a better understanding, nontrivial simplifications are needed by definition. This article is meant to attain good understanding of a molecular system in aqueous solution studied in a previous article with an accurate model.[1] We wish to establish the relevant aspect of the system for a property that was found in the results.

The system is para-benzoquinone (PBQ) surrounded by water at room temperature. For this system, it is found that the solvent structure in the vicinity of the two carbonyl groups is correlated in such a way that asymmetric structures are favored. On the basis of a qualitative comparison with studies on solvation of ions and their affinity to surfaces, it was suggested that the many-body polarization of PBQ was the cause of this observation.[2−6] To test and refine this statement, we will use three different models, all involving different simplifications, in order to properly investigate the problem.

The results obtained and presented below give credence to the claim that many-body polarization indeed is the relevant aspect for understanding the correlation and asymmetry that is observed. PBQ in aqueous solution is, however, not the only system for which polarization can have this influence. Rather, we argue that it can be of importance to a much wider set of solvation problems, especially in polar environments. The many-body effects that are found to be of importance are disregarded in a mean-field approximation. The molecular nature of the solvent and a more detailed statistical mechanical treatment have to be considered in order to account for these effects.

2. Models and Results

Three models are used to address the question of this article. The first two are very simple models that do not treat the problem in its full complexity. They will instead unambiguously demonstrate simple relations. These relations are then used to analyze the results of the third model, which is a

* Corresponding author e-mail: anders.ohrn@teokem.lu.se

10.1021/ct700022b CCC: $37.00 © 2007 American Chemical Society
Published on Web 04/13/2007

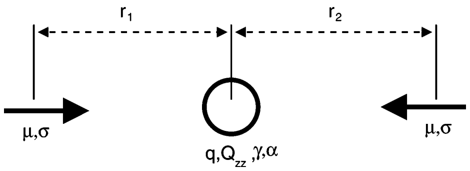


Figure 1. One-dimensional model system A. The central particle can have charge, quadrupole, and polarizability; the peripheral particles have dipoles of equal magnitude, but of opposite direction.

realistic simulation of PBQ in aqueous solution at room temperature.

2.1. Polarizable Trimer. The one-dimensional model system A is depicted in Figure 1. It consists of three particles: one central polarizable particle with either charge, q, or quadrupole, Q_{zz}, and two peripheral particles with dipoles of equal magnitude, $|\mu|$, but of opposite orientation. The potential energy of model system A is

$$U(r_1, r_2) = \left(\frac{\sigma+\gamma}{2r_1}\right)^{12} + \left(\frac{\sigma+\gamma}{2r_2}\right)^{12} + \left(\frac{\sigma}{r_1+r_2}\right)^{12} + \frac{2|\mu|^2}{(r_1+r_2)^3} + 3\frac{Q_{zz}|\mu|}{r_1^{\,4}} + 3\frac{Q_{zz}|\mu|}{r_2^{\,4}} + \frac{q|\mu|}{r_1^{\,2}} + \frac{q|\mu|}{r_2^{\,2}} - 2|\mu|^2\alpha\left(\frac{1}{r_1^{\,3}} - \frac{1}{r_2^{\,3}}\right)^2 \quad (1)$$

The first three terms are the Lennard-Jones repulsion between all particles; parameters σ and γ denote the sizes of the peripheral and the central particles, respectively. The fourth term is the repulsive dipole−dipole interaction between the two peripheral particles. The next four terms will all be attractive in the present application and derive from electrostatic pair interactions between peripheral and central particles. The final term is the induction energy from the polarizability, α, on the central particle. Correlation is obtained if the joint-probability function for r_1 and r_2, $p(r_1,r_2)$, cannot be written as a product of two functions that only depends on either variable. Fundamental results of statistical mechanics give that $p(r_1,r_2) = e^{-U(r_1,r_2)/kT}$; hence, the statement on correlation can be reformulated as correlation is obtained if there are terms in the potential energy that cannot be written as a sum of two terms that only depends on either variable. Hence, from eq 1, it is seen that only three terms can contribute to correlation between r_1 and r_2, namely, the two repulsive terms between the peripheral particles and the polarization term.

A contour plot of the potential in eq 1 is shown in Figure 2. The central particle is charged and either nonpolarizable (upper half) or polarizable (lower half). The parameters in eq 1 are $q = -1e$, $|\mu| = 2.54$ D, $\sigma = \gamma = 1.59$ Å, and α = 0.0 or 3.0 Å^3. They are set to roughly approximate two water molecules in aqueous solution interacting with a chloride ion in aqueous solution; observe that the construction of the system implies that the individual values of γ and σ are unimportant; only their sum will be of any significance. A symmetric configuration, with the peripheral particles close to the central particle, is the minimum of the potential, both with and without polarizability. The attractive and long-ranged charge−dipole pair terms are the reason for this behavior. Only slight differences are seen in the potential plot between the nonpolarizable and polarizable systems. On the right side of Figure 2, the contour plot of the partial derivative $\partial U(r_1,r_2)/\partial r_1$ in an arbitrary unit is shown for the two systems. The contour lines of the nonpolarizable system are almost parallel with the r_2 axis. This implies that the dependence of $U(r_1,r_2)$ on r_1 is almost independent of the value on r_2 (and vice versa from the symmetry of model system A). The slight dependence comes only from the dipole−dipole term, since the polarizability is zero and the Lennard-Jones term between the peripheral particles is at these distances effectively equal to zero. The polarizable system, however, shows a different behavior: the lines are denser at large values of r_2, and at a decreasing value of r_2, the contour lines are far from parallel with the axis. The polarization term in the potential thus introduces a much larger coupling between the two degrees of freedom.

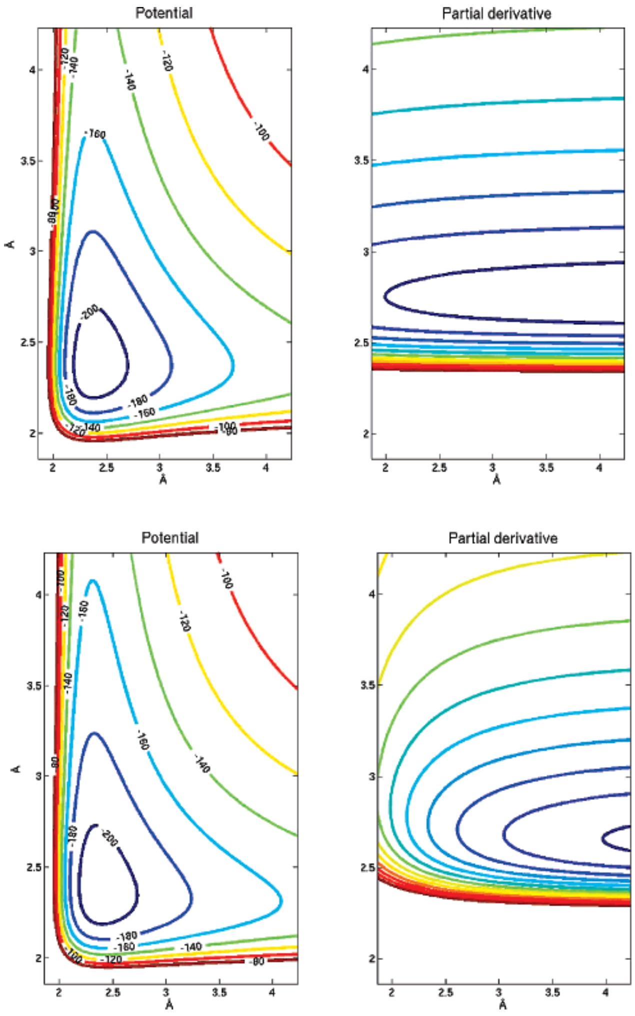


Figure 2. Contour plots of potential (eq 1) for the trimer in kilojoules per mole (left) and the partial derivate along one axis (right). The upper two figures are for a nonpolarizable central particle with charge; the lower ones are for a polarizable central particle with charge.

To get quantitative information about this coupling, the statistical mechanical problem defined by $U(r_1,r_2)$ is solved at $T = 300$ K. Since the potential is so simple, that problem can be solved essentially exactly by numerical integration.

Table 1. Asymmetry Properties for Ionic Simple Trimer; ρ Is Unitless, $\sqrt{\langle \mathscr{E}^2 \rangle}$ in MV/cm, and Both $\langle |r_1 - r_2| \rangle$ and $\langle r_x \rangle$ in Å; μ in D and α in Å^3

μ, q	property	α = 0.0	α = 1.48	α = 2.96	α = 4.45	α = 5.93
1.02, −1.0	ρ	−0.004	−0.018	−0.033	−0.048	−0.063
	$\sqrt{\langle \mathscr{E}^2 \rangle}$	8.64	8.78	8.92	9.07	9.23
	$\langle \|r_1 - r_2\| \rangle$	0.201	0.205	0.209	0.213	0.217
	$\langle r_x \rangle$	2.679	2.679	2.680	2.681	2.683
2.54, −1.0		−0.010	−0.061	−0.117	−0.180	−0.249
		15.69	16.55	17.56	18.80	20.35
		0.091	0.096	0.102	0.110	0.120
		2.397	2.398	2.399	2.400	2.402
4.06, −1.0		−0.016	−0.121	−0.249	−0.408	−0.560
		21.87	24.36	27.99	34.09	50.20
		0.065	0.073	0.084	0.103	0.154
		2.285	2.286	2.287	2.290	2.301
1.02, −2.0		−0.002	−0.011	−0.020	−0.024	−0.029
		6.65	6.71	6.77	6.81	6.84
		0.106	0.107	0.108	0.108	0.109
		2.449	2.449	2.450	2.450	2.450
2.54, −2.0		−0.005	−0.037	−0.071	−0.088	−0.107
		12.44	12.85	13.30	13.54	13.80
		0.053	0.055	0.057	0.058	0.059
		2.220	2.221	2.221	2.221	2.221
4.06, −2.0		−0.008	−0.072	−0.144	−0.184	−0.227
		17.33	18.48	19.90	20.75	21.71
		0.038	0.041	0.044	0.046	0.048
		2.118	2.119	2.119	2.119	2.119

Table 2. Asymmetry Properties for Quadrupolar Simple Trimer; ρ Is Unitless, $\sqrt{\langle \mathscr{E}^2 \rangle}$ in MV/cm, and Both $\langle |r_1 - r_2| \rangle$ and $\langle r_x \rangle$ in Å; μ in D and α in Å^3 and Quadrupole in D·Å

μ,Q_{zz}	property	α = 0.0	α = 4.45	α = 8.89	α = 14.8	α = 26.7
1.02, −23.8	ρ	−0.004	−0.007	−0.011	−0.016	−0.026
	$\sqrt{\langle \mathscr{E}^2 \rangle}$	2.69	2.70	2.71	2.72	2.74
	$\langle \|r_1 - r_2\| \rangle$	1.845	1.851	1.857	1.865	1.881
	$\langle r_1 \rangle$	6.404	6.404	6.404	6.405	6.405
2.54, −23.8		−0.006	−0.018	−0.029	−0.045	−0.075
		4.92	5.02	5.13	5.29	5.65
		0.418	0.431	0.444	0.464	0.514
		4.586	4.591	4.597	4.605	4.626
4.06, −23.8		−0.007	−0.027	−0.047	−0.075	−0.128
		5.74	5.87	6.01	6.22	6.76
		0.199	0.204	0.209	0.217	0.238
		4.208	4.209	4.210	4.212	4.218
1.02, −37.2		−0.003	−0.007	−0.010	−0.015	−0.025
		2.72	2.73	2.74	2.75	2.78
		1.192	1.198	1.205	1.213	1.231
		5.401	5.404	5.406	5.409	5.415
2.54, −37.2		−0.003	−0.010	−0.018	−0.028	−0.050
		3.58	3.61	3.64	3.69	3.78
		0.200	0.202	0.203	0.206	0.212
		4.212	4.212	4.213	4.213	4.215
4.06, −37.2		−0.004	−0.017	−0.30	−0.048	−0.087
		4.64	4.70	4.77	4.86	5.06
		0.123	0.124	0.126	0.129	0.134
		3.941	3.941	3.942	3.942	3.943

Four different quantities are computed: (i) The correlation coefficient

$$\rho(r_1,r_2) = \frac{\langle r_1 r_2 \rangle - \langle r_1 \rangle \langle r_2 \rangle}{\sqrt{(\langle {r_1}^2 \rangle - \langle r_1 \rangle^2)(\langle {r_2}^2 \rangle - \langle r_2 \rangle^2)}} \tag{2}$$

where

$$\langle {r_x}^y \rangle = \frac{1}{\mathscr{Z}} \int \int {r_x}^y \, \mathrm{e}^{U(r_1,r_2)/kT} \, \mathrm{d}r_2 \, \mathrm{d}r_1$$

and $\mathscr{Z}$ is the partition function is determined. If the degrees of freedom are uncoupled, $\rho(r_1,r_2) = 0$; negative coupling (one large, the other small) leads to a negative correlation coefficient, but not below −1, and positive coupling leads to positive values, at most 1. (ii) The square-average electric field on the central particle is also obtained: $\sqrt{\langle \mathscr{E}^2 \rangle}$. It measures how large the electric perturbation from the peripheral particles are, which in a completely symmetric configuration is zero due to cancellation. (iii) The average absolute difference between r_1 and r_2 is computed: $\langle |r_1 - r_2| \rangle$. To interpret this quantity, (iv) the average separations $\langle r_1 \rangle$ and $\langle r_2 \rangle$ are also needed. In Table 1, the quantities i−iv are reported for different dipoles, polarizabilities, and magnitudes of charge. Three noteworthy relations are found: (i) Irrespective of charge or dipole magnitude, an increase in polarizability of the central particle gives rise to an increase of the magnitude of $\rho(r_1,r_2)$, the electric field, $\langle |r_1 - r_2| \rangle$, and $\langle r_x \rangle$. Negative correlation is an effect of asymmetric structures being favored by the increased electric field $\mathscr{E}$ on the polarizable particle in such structures. If only pair interactions are considered, the dependence of the field could be falsely attributed to stronger interactions from an increased polarizability and hence to a structure with the peripheral particles closer to the central particle. That this is a false argument is seen from $\langle r_x \rangle$ increasing slightly rather than decreasing, so the electric field will not become larger on account of a smaller denominator in the expression for the electric field. (ii) $\sqrt{\langle \mathscr{E}^2 \rangle}$ decreases with an increased charge on the central particle, despite the simultaneous decrease in $\langle r_x \rangle$. The reason is that the importance of the pair terms in the potential increases, and they favor a symmetric configuration, where the electric field is zero. The dependence of the correlation coefficient on the charge confirms this explanation. (iii) With a larger dipole magnitude, the correlation increases for a fixed polarizability, and also, with larger magnitude, the increase of the correlation with polarizability is greater than for the system with a smaller magnitude. This fits well with the prediction that the last term in the potential in eq 1 is the one that mainly determines the degree of correlation.

In Table 2, results from a system with a quadrupole instead of a charge are collected. The Lennard-Jones repulsion is also modified to $\sigma = \gamma = 2.75$ Å, and the quadrupole moment and polarizability are set to qualitatively represent the PBQ−water system. The same relations are found as above, with the only exception being that, for the smallest dipole moment, the electric field increases upon an increased quadrupole moment; this deviation is explained by the correlation being almost unaffected by this modification, while the pair terms see to it that the system gets tighter

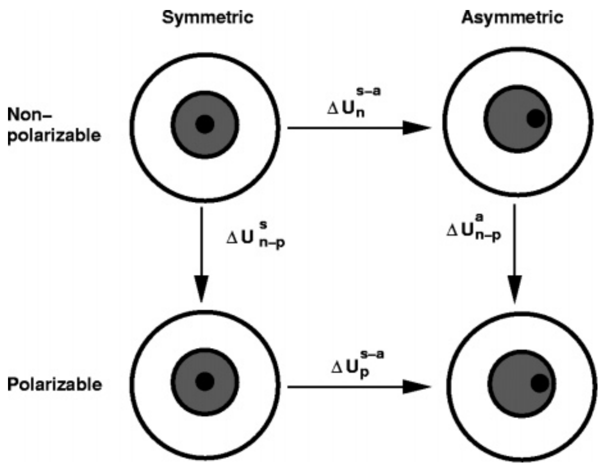


Figure 3. Two-state model system B. All solvent matter included, but separated into two regions: close to the solute particle and far away from the solute particle; the system has two states: symmetric solvation and asymmetric solvation; included in the figure are also the transition energies between the different states.

(see $\langle r_x \rangle$), and that gives a greater field on the central particle. The magnitudes are, however, distinctly different for the quadrupolar system compared to the charged one: the correlation is smaller, and while the increased correlation could lead to a doubling of the field for the charged system, the field is at most increased 20% going from the two extremes in polarizability in Table 2. One reason for this is the weaker pair interactions in the quadrupolar system: With weak interactions, entropy will drive the system to a loose state with the peripheral particles on average far away, which in turn leads to a small field and thus a small correlating energy term in the potential. However, as seen in the charged system with the transition from a monovalent to a divalent system, if entropy is too small, the symmetric energy-minimum configuration will become dominant and that way reduce correlation. Another reason for the small correlation can be that the quadrupolar particle is bigger and the polarization is described with a point polarizability in the middle of the particle, which of course leads to smaller fields; a better description of the polarization of PBQ could change the quantities. To conclude, the two types of model system A have established a few principles and showed that they are the same for charged and quadrupolar systems, while possibly quite different in degree.

2.2. Two-State Many-Body Solvation Model. Model system B is a purely qualitative model for bulk solvation and includes all solvent matter but uses a very simple two-state description of the system, see Figure 3 for illustration. The system is in either a symmetric solvation state or an asymmetric solvation state, and the solvent is divided into two regions, close (gray area in Figure 3) and far away (white area) from the solute (black area). The interaction between the solute and solvent in the former region disrupts what otherwise would be the preferred solvent structure with optimal interactions between the two solvent regions. This implies that, by somehow weakening the solute−solvent interaction, favorable interactions in the entire system can in fact increase (compare this with some explanations of hydrophobic attraction).[7−9] Therefore, a transition from the symmetric to asymmetric state, which changes the solute−solvent interaction in the gray region, can be both favorable and unfavorable; which situation that applies is mainly determined by the balance between solute−solvent and solvent−solvent pair interactions. Another feature of model system B is that any transition from the nonpolarizable to polarizable state is favorable. However, it is assumed, on the basis of the results from model system A, that this transition in the asymmetric state is more favorable than that in the symmetric state, that is, $\Delta U_{n-p}^{a} \leq \Delta U_{n-p}^{s} \leq 0$. Three cases can be distinguished:

(i) If $\Delta U_n^{s-a} \geq 0$ and $\Delta U_n^{s-a} \geq \Delta U_{n-p}^{s} - \Delta U_{n-p}^{a}$, then $\Delta U_p^{s-a} \geq 0$; in other words, in both the nonpolarizable and polarizable states, the symmetric solvation is more probable than the asymmetric solvation.

(ii) If $\Delta U_n^{s-a} \leq 0$, then $\Delta U_p^{s-a} \leq 0$, or in other words, if already the nonpolarizable state favors the asymmetric solvation, the polarizable state will also do so.

(iii) If $\Delta U_n^{s-a} \geq 0$ and $\Delta U_n^{s-a} \leq \Delta U_{n-p}^{s} - \Delta U_{n-p}^{a}$, then $\Delta U_p^{s-a} \leq 0$, which means that the introduction of polarizability will turn the system from being preferably symmetrically to asymmetrically solvated.

Model system B shows that, with the entire solvent (or a sufficient amount, at least) in the treatment, other factors related to the balance between the different interactions in the system become important in understanding the structure around the solute and hence the solvent effects. The qualitative nature of the model and its unrealistic account of entropy precludes any predictions for specific systems, however. Hence, model system B is only a thought experiment to warn against interpretations based only on the solvent in contact with the solute, as in the preceding model system A.

2.3. Explicit Many-Body Solvation. An accurate description of a solvation phenomenon requires that both features of model systems A and B be taken into account, that is, a plausible description of the interactions and the thermodynamics of the system, and that a sufficient portion of the solvent−solvent interactions be accounted for to adequately describe their indirect effect on the solute−solvent interaction. An obstacle, on the conceptual level, is that, once both features are combined, it is not as easy to analyze the system as in model systems A and B, and this results in observations that are not easy to trace. This also makes predictions based only on simple physical relations of the constituents more difficult to make. It may, however, be that the simultaneous introduction of both features invalidates questions about causation of a nature as precise as that in the previous two models. A system in its full complexity may very well entangle the causes and therefore, on a fundamental level, rule out clear statements on cause and effect in the present state of the theoretical development. Before this discussion is continued in a section below, a simulation of a solute−solvent system in its full complexity is done, both with a polarizable and a nonpolarizable solute. This provides one particular realistic system from which limited generalization can be made.

The easiest system would be a monatomic ion in a polar solvent. These systems have already been extensively studied with Monte Carlo and molecular dynamics simulations, and their results will be discussed from the present perspective below.[2−6,10−16] As noted in the Introduction, it has recently been established that similar questions on solvent structure and correlation are meaningful also for neutral solutes.[1] The neutral and non-dipolar PBQ in aqueous solution is chosen as the model system.

The model used is the combined quantum chemical statistical mechanical solvent model called QMSTAT.[17] In the present simulation, a Hartree−Fock (HF) wave function is used for PBQ since only the electronic ground state of PBQ is relevant; an extension of QMSTAT for excited states has been formulated and was used in the previous study on PBQ and its electronic spectrum.[1,18] The model solves the quantum chemical problem in a truncated natural molecular orbital basis. The model is thus compact, and the small dimension of the Fock matrix as well as the ability to store all two-electron integrals in memory leads to a single calculation being a relatively easy task. In the subsequent Monte Carlo simulation, it is then tenable to solve a quantum chemical problem in each step. Therefore, the combined quantum chemical statistical mechanical problem can be solved with a so-called hybrid approach which enables the statistical error to be made arbitrarily small, as compared to the more common but approximate sequential approach. Since a key aspect in the present study is the polarization of the solute and its consequences on statistical solvent properties, QMSTAT is a suitable model since both polarization and statistical mechanics are treated well. Details of QMSTAT are available elsewhere, and below only the particular aspects for PBQ are presented.[17,18] The Møller−Plesset optimized structure is used for PBQ (same as in ref 1).[19] An atomic natural orbital (ANO) basis set is used for all orbital calculations; contractions are C,O 4s3p2d; H 2s1p.[20] The natural orbitals are constructed from diagonalization of an average density matrix; the different density matrices adding up to the average comes from HF calculations with the same set of perturbations as in the inhomogeneous basis in ref 1. For the nonpolarizable system, only occupied orbitals are included in the basis; hence, the electronic wave function has no flexibility and is frozen in its gas-phase form. For the polarizable state, 42 orbitals are included in total, which retains almost all polarizability that the full ANO-basis set gives. Solute−solvent dispersion interaction is parametrized as in ref 1, and the repulsive solute−solvent pseudo-potential parameters are $d = -0.32$ and $\beta_4 = 2.5$ (see ref 18 for relevant equations). A total of 150 explicit polarizable water molecules are included as the solvent, and a nonperiodic boundary condition using the image-charge approximation is added.[21−23] To rule out statistical uncertainties for the observed properties, special attention is paid to the convergence of the Monte Carlo simulation and the statistical significance of computed quantities; this technical discussion is put in Appendix A. Statistical properties are computed from simulations of 4.2×10^6 Monte Carlo steps where every 100th configuration is sampled. All quantum chemical calculations are done with the quantum chemical software package MOLCAS, which also is the platform for the development of QMSTAT.[24,25]

Table 3. Quantities and 99.9% Confidence Intervals from Simulation on PBQ, with and without Polarization[a]

	no polarization		with polarization	
	average	conf. int.	average	conf. int.
$\rho(r_1,r_2)$	−0.032	(−0.047,−0.016)	−0.166	(−0.181,−0.151)
$\langle \lvert r_1 - r_2 \rvert \rangle$	0.316	(0.312,0.319)	0.313	(0.309,0.317)
$\rho(\phi(r_1),\phi(r_2))$	0.043	(0.027,0.060)	−0.114	(−0.130,−0.097)
$^1/_2\langle \phi(r_1) + \phi(r_1) \rangle$	0.380	(0.375,0.384)	0.512	(0.507,0.516)
$\sigma(^1/_2(\phi(r_1) + \phi(r_2)))$	0.284	(0.270,0.297)	0.296	(0.278,0.312)

[a] Distances in Å and potentials in V.

In this complex system, an analysis of asymmetry and correlation in the solvent structure is more difficult. No clear-cut choice of quantities to characterize the degree of asymmetry is evident to us. To achieve a close correspondence between the results of this model and the results of model system A, however, five different quantities are chosen. (i) The correlation coefficient (eq 2) for r_1 and r_2 is computed, where r_1 and r_2 are defined as the shortest distance between a hydrogen atom in the solvent and the oppositely located oxygen atoms of PBQ. This corresponds to $\rho(r_1,r_2)$ in model system A. (ii) The average difference $\langle |r_1 - r_2| \rangle$, which corresponds to the same type of quantity in model system A, is obtained. (iii) Next, the correlation coefficient between $\phi(x_1)$ and $\phi(x_2)$, where $\phi(x_1)$ is the electric potential from the solvent at one of the oxygen atoms in PBQ, is determined; $\phi(x_2)$ is the corresponding quantity at the other oxygen atom. This corresponds in part to $\sqrt{\langle \mathcal{E}^2 \rangle}$ in model system A, since it measures the correlation of the electric perturbation on the solute. The magnitude of the electric field in the middle of PBQ, which superficially has more in common with $\sqrt{\langle \mathcal{E}^2 \rangle}$, is a less adequate measure since it implicitly assumes that a polarizability in the center of mass best characterizes the polarization of PBQ in aqueous solution, which hardly is a valid approximation to the polarization in QMSTAT. (iv) Finally, the average value of $\phi(x_1)$ and $\phi(x_2)$ and (v) their standard deviation are evaluated. Although they do not measure the degree of asymmetry or correlation, they show what effect the polarization will have on the magnitude of the electric perturbation from the solvent on the solute, which in part also corresponds to $\sqrt{\langle \mathcal{E}^2 \rangle}$ in model system A. The quantities are reported in Table 3 with 99.9% confidence intervals obtained with the bootstrap method (see Appendix A).

In Table 3, it is seen that both correlation coefficients i and iii are significantly different between nonpolarizable and polarizable PBQ. For the nonpolarizable PBQ, a faint correlation is found, while polarizable PBQ has a negative correlation between both r_1 and r_2, as well as $\phi(x_1)$ and $\phi(x_2)$. The average of $|r_1 - r_2|$ is not significantly different between the two system, however. The reason for this is that two effects cancel: on the one hand, the more negative correlation in the polarizable state increases the difference; on the other hand, the shorter separation between the solute and solvent in the same state decreases the difference; see Figure 4 for the radial distribution functions which prove the latter statement. Further, the solvent electric potential on the oxygen atoms is significantly larger in the polarizable state. The standard deviation, however, has no significant

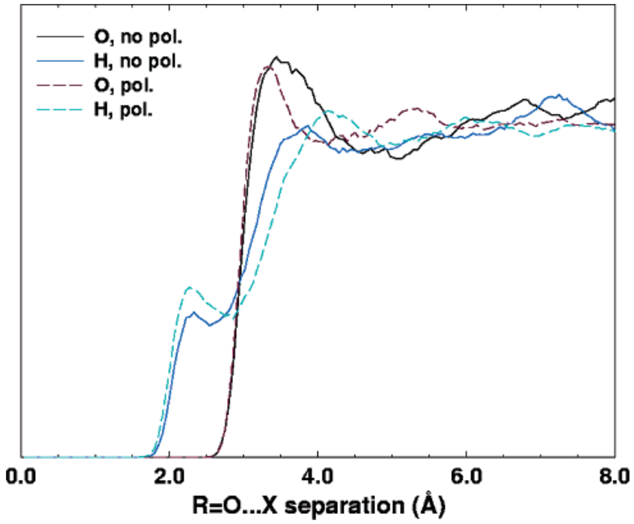


Figure 4. Radial distribution functions around the oxygen atoms for nonpolarizable and polarizable PBQ in an aqueous solution.

difference between the two systems. The reason the average is larger will to some extent be explained by the increased asymmetry in the polarizable state. That is probably not the entire reason, though. An additional contribution is likely to come from the increased reaction field from the polarizable solvent when the solute polarizes.

3. Discussion

Asymmetric solvent configurations are themselves nothing novel: thermal fluctuations, or entropy, will always, at nonzero temperatures, lead to configurations outside of the, possibly, symmetric energy minimum; see, for example, $\sqrt{\langle \mathscr{E}^2 \rangle}$ in Tables 1 and 2 at zero polarizability where the nonzero value comes only from the aforementioned fluctuations and not from the correlation discussed above. Still, the average over all configurations may be symmetric. This observation leads to a critique of the mean-field approximation , which is the basis for the widespread continuum solvent models. There, the solute interacts with the field from the average solvent configuration, which as observed above may be a symmetric one with zero electric field. If the particle solvated in the dielectric cavity is polarizable, an attractive term will be missing because fluctuations are missing. This critique has been formalized elsewhere, and corrections to continuum solvent models have been proposed.[26−28] Other problems with the continuum models are discussed by de Vries et al.[29]

Model system A, above, shows that the thermal molecular fluctuations can be correlated. The polarizability will not only interact favorably with the fluctuation field coming from independent random fluctuations (i.e., the electric field at $\alpha = 0$ in model system A) but will couple the solvent degrees of freedom and enhance the magnitude of the fluctuating electric field on the solute. To put it differently, model system A shows that polarization can introduce a bias for asymmetric solvent configurations and, by that, increase the magnitude of the fluctuations.

Results from simulations of monatomic and lately also polyatomic ions in aqueous solution have in some cases been interpreted in similar terms. Bulk simulations by Carignano et al. of ions with variable polarizability, but with fixed sizes, have shown that the solvation environment tends to get more asymmetric with increasing polarizability for a fixed size.[10] Other simulations of highly polarizable anions in bulk show that the structure in the closest hydration shell is less symmetric than in less polarizable ions, although other effects are not always ruled out.[11−13] Wilson and Madden have also argued that layered structures for certain simple ionic compounds are effects of anion polarizability and their affinity for asymmetric environments.[30,31] Results from simulations on the distribution of ions between bulk and air/water interfaces have also been shown to be dependent on the polarizability of the ionic solute.[2−6] The air/water interface can be seen as an extreme asymmetric environment. Thus, the polarizability of the ion, in the same fashion as in model system A, increases the probability for the interface to be populated as compared to the less asymmetric bulk environment. But clearly, other factors will have an influence, as the simple thought experiment in model system B shows. Hrobárik et al. give a lucid example with several tetraalkylammonium cations with different alkyl chain lengths: the different surface propensities of the ions are rationalized by the increasing hydrophobicity of the ion with increasing chain length.[6] And there are studies on monatomic ions that establish and emphasize the importance of the balance between solute−solvent and solvent−solvent interactions also for these systems.[4,11,14−16,32] A recent experiment also found that surface affinity correlates most strongly with ion size, not polarizability.[33] As a final remark, however, we observe that the polarizability of molecules in solution, and anions in particular, is a property that depends on the environment and is consequently not easy to unequivocally assign.[34−39]

PBQ is neutral but has a significant quadrupole moment due to the polar carbonyl groups. Interactions with quadrupoles can be large, and to consider them insignificant is in many applications quite wrong.[40−43] Also, as shown with model system A, the same principles that apply to the ionic system with respect to correlation of the solvent degrees of freedom by polarization apply also to the quadrupolar system. Further, we established above in a detailed simulation that there is a significant difference in the correlation with and without polarizability on PBQ. Together, these two results strongly suggest that it is the polarization of PBQ that couples the solvent degrees of freedom on opposite sides of the solute molecule by the same simple mechanism that operates in model system A. As shown with that model system, however, a polarizability is not enough to cause correlation; there has to be favorable solute−solvent pair terms that order the solvent adjacent to the solute for a significant fluctuation field to appear on the solute. In the most weakly interacting case for model system A, entropy almost manages to dissociate the solute−solvent system. In the simulation of PBQ, packing effects will, however, make the space close to PBQ occupied at all times; instead, other degrees of freedom can be used to contain the hydrophobic solute so that the solvent−solvent interactions are minimally perturbed. Dominance of such configurations implies a small fluctuation field on the solute. Therefore, the hydrogen bonds between water and the carbonyl oxygen atom (primarily a pair term) have an indirect but important effect on the correlation.

It is important to observe, though, that the hypothesis only deals with why the correlation exists and its implications on the fluctuation field. The details from which the actual magnitude of the fluctuation field could be computed are not dealt with. In model system A, it is shown that the correlation and the electric field on the solute have the same type of dependence on the polarizability no matter what charge or quadrupole the central particle has. But the magnitude of the field depends on these central particle properties for reasons discussed above. For a complex system such as PBQ in aqueous solution, several specific features of the system will contribute to this, and we do not propose to have a quantitative theory for this intricate problem.

Solvent effects on chemical reactions, absorption and fluorescence spectra, and many other relevant processes are effects of the electric perturbation the solvent exerts on the solute.[44] From the perspective of a mean-field theory of the solvent, this perturbation will be small for molecules with no dipole and be of the same symmetry as the solute molecule. As previously discussed, this will lead to the neglect of a stabilizing fluctuation term. Another effect is that symmetry-breaking terms that should appear in the Hamiltonian are lost. For example, in the study of nonlinear optics, quadrupolar (or higher) chromophores are often used, and their electronic ground and, more often, excited states can be quite labile to symmetry-breaking terms in the Hamiltonian and through them give rise to distinct modifications compared to gas-phase or other nonpolar surroundings.[45−47] The excited state from the one-photon excitation $n \rightarrow \pi^*$ in PBQ is in fact near-degenerate, and therefore the solvated state undergoes a large mixing because there is a significant symmetry-broken term in the effective Hamiltonian from the solvent.[1] Zijlstra et al. also show how drastic the effects of solvent-induced symmetry breaking are on excited ethylene.[48] As shown here, the magnitude of this symmetry-breaking perturbation in these and similar systems can be even larger than expected since correlations caused by the polarizability and the discrete molecular nature of the solvent are likely to exist.

4. Conclusions

Starting from a very simple model of a non-dipolar molecule in a polar surrounding, and going to a simulation of great detail, we have found support for the hypothesis that the many-body nature of the polarization of the non-dipolar solute couples the solvent degrees of freedom in such a way that asymmetric solvent configurations are favored. In asymmetric configurations, the electric perturbation from the solvent on the solute is larger, which leads to an increase in magnitude of the fluctuating field, implied by the thermal fluctuations of the solvent, with an increase of the polarizability of the solute. However, the pair interactions between the solute and solvent must also be included in the explanation since they have an indirect effect on the correlation and the fluctuations. The mechanism we propose is valid for the solvation of both ions and quadrupolar molecules, and we comment on previous studies of ion solvation and the surface affinity of ions. Other situations where this can be of relevance is in understanding solvent effects on multipolar molecules. The present study highlights the intricacy of the fluctuating electric field a molecule experiences in a solvent.

Acknowledgment. Financial support from the Swedish Research Council within the Linné-project Organizing Molecular Matter is acknowledged. Computer time at Lunarc, Center for Scientific and Technical Computing, Lund University is acknowledged.

Appendix A

To obtain statistically certain quantities for the QMSTAT simulation of PBQ in water, convergence diagnostics and bootstrap confidence intervals are used. Both methods are described below.

In any Monte Carlo simulation of a complex system, knowing when a balanced sampling of the configuration space has been achieved is difficult. This problem is especially critical when free-energy barriers exist in the system. Brooks and Gelman have proposed a simple convergence diagnostic, which has become popular in applications of Monte Carlo statistical techniques in medicine.[49] Other more advanced diagnostics exist, but the Brook−Gelman diagnostic (BGD) is judged to be of sufficient accuracy for the present application and also has a feature that fits well with our simulation approach.[50]

To obtain the BGD, (i) *N* parallel Monte Carlo simulations with different initial configurations are run for *m* steps each. (ii) Some measure (vide infra) of variance is computed between the different chains of configurations as well as within the chains. If these measures differ significantly, that tells that the individual chains have not been sampled over a sufficient space, and *m* should be increased. (iii) When between and within measures are of similar value, convergence is likely to have been reached, although it does not guarantee convergence

Two different measures of variance are used, both proposed by Brooks and Gelman (observe that a different notation is used in ref 49). First, the second moments:

$$\hat{B}_2 = \frac{1}{Nm - 1}\sum_{i,j}^{N,m}(x_{ij} - x_{..})^2$$

$$\hat{W}_2 = \frac{1}{N(m - 1)}\sum_{i,j}^{N,m}(x_{ij} - x_{i\cdot})^2$$

where x_{ij} is the *j*th element in the *i*th chain, and a dot in the indices means that that index has been averaged. Another measure is

$$\hat{L}_B = |l_l^{\mathrm{tot}} - l_r^{\mathrm{tot}}|$$

$$\hat{L}_W = \frac{1}{N}\sum_{i}^{N} |l_l^i - l_r^i|$$

where l_l^i and l_r^i are the estimated confidence interval limits for $\{x_{ij}\}_{j=1,\cdots m}$ for a given chain *i*, and l_l^{tot} and l_r^{tot} are the

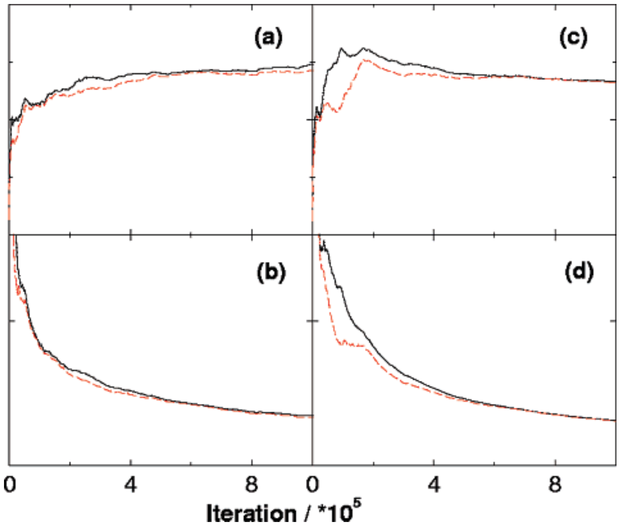


Figure 5. (a) The second moment of the minimum oxygen−hydrogen atom distance on one side of nonpolarizable PBQ. Part c is like a, only that PBQ is polarizable. Diagrams b and d show the confidence interval width for nonpolarizable and polarizable PBQ, respectively. All black lines are the between measure; all red dashed lines are the within measures.

same for the entire data set. Observe that $\hat{L}_B \sim 1/\sqrt{m}$ as $m \rightarrow \infty$; the same goes for $\hat{L}_W$. In Figure 5, the progression of the measures as longer individual Monte Carlo chains are sampled is shown; four parallel chains have been used. The sampled quantities x_{ij} are the shortest distance between a specific oxygen atom on PBQ and the hydrogen atoms of the solvent. As can be seen, the measures have converged to being very close to each other in both cases. Other quantities x_{ij} of the simulation show the same behavior. This suggests (but does not prove) that convergence has been reached and that the total data set is a balanced sample of the relevant configuration space. Observe that the equilibration period (called burn-in period in ref 49) is not included in Figure 5.

The bootstrap method is a resampling method constructed by Efron.[51] It is a method that can be used to solve statistical problems of a very diverse nature, even when the knowledge of the probability distribution is incomplete. For the present study, it is used to construct confidence intervals in Table 3, an application for which bootstrap is suitable.[52] A premise for this method is that the samples are collected from independent distributions. In a Monte Carlo simulation, this is not fulfilled. However, the bootstrap method can still be used if the collected sample is a balanced sample of the configuration space, or in other words, where the set of sampled points appears as if they were obtained independently. Therefore, to apply the bootstrap method properly, convergence has to be reached in the sense described above in relation with BGD. Once this is established, the nonparametric percentile bootstrap method is applied to construct confidence intervals for the correlation coefficients and the other variables. A description of this method can be found in advanced textbooks on statistical inference or on resampling methods.

References

(1) Öhrn, A.; Aquilante, F. *Phys. Chem. Chem. Phys.* **2007**, *9*, 470−480.

(2) Jungwirth, P.; Tobias, D. J. *J. Phys. Chem. B* **2002**, *106*, 6361−6373.

(3) Karlström, G.; Hagberg, D. *J. Phys. Chem. B* **2002**, *106*, 11585−11592.

(4) Hagberg, D.; Brdarski, S.; Karlström, G. *J. Phys. Chem. B* **2005**, *109*, 4111−4117.

(5) Jungwirth, P.; Tobias, D. J. *Chem. Rev.* **2006**, *106*, 1259−1281.

(6) Hrobárik, T.; Vrbka, L.; Jungwirth, P. *Biophys. Chem.* **2006**, *124*, 238−242.

(7) Israelachvili, J. N. *Intermolecular and Surface Forces*, 2nd ed.; Academic Press: London, Great Britain, 1992.

(8) Forsman, J.; Jönsson, B.; Woodward, C. E. *J. Phys. Chem.* **1996**, *100*, 15005−15010.

(9) Meyer, E. E.; Rosenberg, K. J.; Israelachvili, J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15739−15746.

(10) Carignano, M. A.; Karlström, G.; Linse, P. *J. Phys. Chem. B* **1997**, *101*, 1142−1147.

(11) Raugei, S.; Klein, M. L. *J. Chem. Phys.* **2002**, *116*, 196−202.

(12) Rajamani, S.; Ghosh, T.; Garde, S. *J. Chem. Phys.* **2004**, *120*, 4457−4466.

(13) Tofteberg, T.; Öhrn, A.; Karlström, G. *Chem. Phys. Lett.* **2006**, *429*, 436−439.

(14) Perera, L.; Berkowitz, M. L. *J. Chem. Phys.* **1992**, *96*, 8288−8294.

(15) Stuart, S. J.; Berne, B. J. *J. Phys. Chem.* **1996**, *100*, 11934−11943.

(16) Grossfield, A. *J. Chem. Phys.* **2005**, *122*, 024506.

(17) Moriarty, N. W.; Karlström, G. *J. Phys. Chem.* **1996**, *100*, 17791−17796.

(18) Öhrn, A.; Karlström, G. *Mol. Phys.* **2006**, *104*, 3087−3099.

(19) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618−622.

(20) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. *Theor. Chim. Acta* **1995**, *90*, 87−114.

(21) Wallqvist, A.; Ahlström, P.; Karlström, G. *J. Phys. Chem.* **1990**, *94*, 1649−1656.

(22) Friedman, H. L. *Mol. Phys.* **1975**, *29*, 1533−139.

(23) Wallqvist, A. *Mol. Simul.* **1993**, *10*, 13−17.

(24) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222−239.

(25) Veryazov, V.; Widmark, P.-O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2004**, *100*, 626−635.

(26) Baur, M. E.; Nicol, M. *J. Chem. Phys.* **1966**, *44*, 3337−3343.

(27) Karlström, G.; Halle, B. *J. Chem. Phys.* **1993**, *99*, 8056−8062.

(28) Sánchez, M. L.; Martín, M. E.; Galván, I. F.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Phys. Chem. B* **2002**, *106*, 4813−4817.

(29) de Vries, A. H.; van Duijnen, P. T.; Juffer, A. H. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1993**, *27*, 451−466.

(30) Wilson, M.; Madden, P. A. *J. Phys.: Condens. Matter* **1994**, *6*, 159−170.

(31) Stone, A. J. *The Theory of Intermolecular Forces*, 1st ed.; Oxford University Press: Oxford, Great Britain, 1996.

(32) Frank, H. S.; Wen, W.-Y. *Discuss. Faraday Soc.* **1957**, *24*, 133−140.

(33) Cheng, J.; Vecitis, C. D.; Hoffmann, M. R.; Colussi, A. J. *J. Phys. Chem. B* **2006**, *110*, 25598−25602.

(34) Mayer, J. E.; Mayer, M. G. *Phys. Rev.* **1933**, *43*, 605−611.

(35) Pyper, N. C.; Pike, C. G.; Edwards, P. P. *Mol. Phys.* **1992**, *76*, 353−372.

(36) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 9903−9906.

(37) Öhrn, A.; Karlström, G. *J. Phys. Chem. B* **2004**, *108*, 8452−8459.

(38) Krishtal, A.; Senet, P.; Yang, M.; van Alsenoy, C. *J. Chem. Phys.* **2006**, *125*, 034312.

(39) Heaton, R. J.; Madden, P. A.; Clark, S. J.; Jahn, S. *J. Chem. Phys.* **2006**, *125*, 144104.

(40) Buckingham, A. D.; Fowler, P. W. *J. Chem. Phys.* **1983**, *79*, 6426−6428.

(41) Hurst, G. J. B.; Fowler, P. W.; Stone, A. J.; Buckingham, A. D. *Int. J. Quantum Chem.* **1986**, *29*, 1223−1239.

(42) Moriarty, N. W.; Karlström, G. *J. Chem. Phys.* **1997**, *106*, 6470−6474.

(43) Zhao, X. C.; Johnson, J. K. *Mol. Simul.* **2005**, *31*, 1−10.

(44) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd ed.; Wiley-VCH: Weinheim, Germany, 2003.

(45) Droumaguet, C. L.; Mongin, O.; Werts, M. H. V.; Blanchard-Desce, M. *Chem. Commun.* **2005**, 2802−2804.

(46) Terenziani, F.; Painelli, A.; Katan, C.; Charlot, M.; Blanchard-Desce, M. *J. Am. Chem. Soc.* **2006**, *128*, 15742−15755.

(47) Bidault, S.; Brasselet, S.; Zyss, J.; Maury, O.; Bozec, H. L. *J. Chem. Phys.* **2007**, *126*, 034312.

(48) Zijlstra, R. W. J.; Grozema, F. C.; Swart, M.; Feringa, B. L.; van Duijnen, P. T. *J. Phys. Chem. A* **2001**, *105*, 3583−3590.

(49) Brooks, S. P.; Gelman, A. *J. Comput. Graph. Stat.* **1998**, *7*, 434−455.

(50) Brooks, S. P.; Roberts, G. O. *Stat. Comput.* **1998**, *8*, 319−335.

(51) Efron, B. *Ann. Stat.* **1979**, *7*, 1−26.

(52) Efron, B.; Tibshirani, R. *Stat. Sci.* **1986**, *1*, 54−75.

CT700022B

JCTC Journal of Chemical Theory and Computation

The Effect of Polarizability for Understanding the Molecular Structure of Aqueous Interfaces

Collin D. Wick,[†] I-Feng W. Kuo,[‡] Christopher J. Mundy,*[,†] and Liem X. Dang*[,†]

Pacific Northwest National Laboratory, Richland, Washington 99352, and Lawrence Livermore National Laboratory, Livermore, California 94550

Received April 26, 2007

Abstract: A review is presented on recent progress of the application of molecular dynamics simulation methods with the inclusion of polarizability for the understanding of aqueous interfaces. Comparisons among a variety of models, including those based on density functional theory of the neat air−water interface, are given. These results are used to describe the effect of polarizability on modeling the microscopic structure of the neat air−water interface, including comparisons with recent spectroscopic studies. Also, the understanding of the contribution of polarization to the electrostatic potential across the air−water interface is elucidated. Finally, the importance of polarizability for understanding anion transfer across an organic−water interface is shown.

Introduction

Aqueous interfaces are ubiquitous in nature and pose characteristics that affect countless biological, atmospheric, pharmaceutical, and industrial processes. These processes are dependent on the molecular-level details of these interfaces and are manifested in enhanced or depleted molecular activity and reaction rates at interfaces, detergent agents, membrane permeability, and molecular uptake in aqueous aerosols. Because of this, there is a strong effort to understand the molecular-level properties of these interfaces. This understanding is beginning to form due, in part, to the introduction of polarizability in the molecular models used to study aqueous interfaces. Polarizability has been found to be of the highest importance for the realization that some anions have a propensity for the interface.[1,2] However, the importance of polarizable interactions for understanding the properties of neat air−water interfaces is not comprehensive. In fact, while there is some indication of the importance of polarizability for the determination of thermodynamic properties at the air−water interface,[3] there is also some indication that polarizability is of secondary importance for air−water interfacial properties.[4,5]

In the past few years, there has been a large amount of surface-sensitive spectroscopic techniques dedicated to studying the air−water interface.[6−12] The vibrational sum frequency generation spectroscopic technique and the emerging area of X-ray techniques applied to liquid−vapor interfaces are elucidating significant details of the molecular structure of the air−water interface.[6,8−11,13] Experimental findings include the characterizations of both a single donor (a free O−H vibration) and acceptor-only (two free O−H stretches) hydrogen-bond species at the air−water interface, and thus fewer on average hydrogen bonds for interfacial waters than for bulk ones.[14] Because of the heterogeneous nature of the interfacial region, it can be easily justified that the hydrogen-bond populations and degree of hydrogen bonding will differ from their bulk values. However, the dependence of these populations on the interaction potential and the ability to understand and agree with spectroscopic determinations of interfacial hydrogen bonding are still a topic of debate.[9]

Recent X-ray absorption fine structure (EXAFS) experiments found another interesting feature, namely, that there is an expansion in the average water oxygen−oxygen distances at the air−water interface when compared with the bulk.[7] A following computational study of the air−water interface found no expansion using a variety of classical force fields but did find that with Car−Parrinello molecular dynamics (CPMD), using density functional theory (DFT)

* Corresponding author e-mail: chris.mundy@pnl.gov (C.J.M.); liem.dang@pnl.gov. (L.X.D.).

[†] Pacific Northwest National Laboratory.

[‡] Lawrence Livermore National Laboratory.

10.1021/ct700098z CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/27/2007

with a BLYP exchange and correlation functional, surface expansion at the air−water interface was observed.[15,16] One may question as to what features are necessary in a classical molecular model to capture this experimentally observed surface relaxation.

The inclusion of polarizability may be the key for the observation of surface relaxation at the air−water interface. Two of the most common ways to account for polarizability for rigid water models are using the fluctuating charge (FQ)[17] technique and including explicit point polarizabilities. The important distinction between explicitly polarizable and FQ models is that, for a polarizable model, a dipole is induced at one or more point polarizabilities on the basis of the local electric field. For FQ water models, the local electric field induces a change in the charge distribution between the hydrogens and the oxygen or other nonatomic interaction sites keeping an overall neutral molecule. Both techniques are designed to mimic charge reorganization in a water molecule in response to its solvation environment.

Another way to characterize interfaces is to determine the electrostatic potential (EP) across them.[18] The electrostatic potential can be used to characterize the distribution of electrostatic charge and thus the molecular structure at an interface. Although the empirical potentials cannot capture the true potential due to the nuclear charge and electrons, the value of the surface potential appears to be insensitive to the type of empirical interaction potential (viz., fix charge or polarizable).[19] With the inclusion of polarizability, the effect of specific molecular structures and orientations can be separated from effects due to rearrangement of the charge in a molecule. However, the effects of a smeared charge distribution cannot be easily dismissed. It has been shown that, for a simple Gaussian model of charge smearing, the degree of smearing as determined by the width of the Gaussian can have dramatic effects on the value of the surface potential.[20] Understanding the effect of polarization and a realistic charge distribution can be a major factor in interpreting electrostatic potential measurements.

While polarizability has been found to be paramount for understanding anions at air−water interfaces, only recently has polarizability been used to understand ions at organic−water interfaces.[21] With an organic (in this case CCl_4) present at the interface with water, the interfacial properties are different than at an air−water interface.[22] With these different interfacial properties, understanding if the effect of polarizability for organic−water interfaces is similar to that for ion transfer across air−water interfaces is of importance.

This paper is organized as follows. The next section gives details for some simulations carried out for this work. The Results and Discussion section gives a comparison of a variety of molecular models for understanding the air−water interface, followed by a discussion as to the relevance of polarizability to understanding interfacial electrostatic potentials. Then, the free energy profile of a polarizable hydronium molecule across an air−water interface is shown. Next, a comparison of the free energy profile for iodide across organic−water interfaces with and without polarizable interactions is given. Finally, a summary and conclusions are given.

Models and Simulation Details

Classical Simulations of Pure Water. Classical molecular dynamics (MD) simulations were carried out utilizing the rigid four-site TIP4P,[23] rigid four-site Dang−Chang[24] (D−C), and flexible three-site SPC−FW[25] water models. The TIP4P and D−C water models are rigid with four interaction sites. All models contain a single Lennard-Jones interaction site located on the oxygen atomic position, and the SPC−FW model has a negative charge located at the oxygen position. All models have two hydrogen atomic sites with positive charges, and the TIP4P and D−C models have an additional *m* site located along their oxygen−hydrogen bisectors. For the TIP4P and D−C models, the *m* site contains a negative charge, but the D−C model has an additional point polarizability located on it. The point polarizability allows the formation of induced dipoles in response to the local electric field. Induced dipoles were evaluated by a self-consistent iterative procedure, which is described in detail elsewhere.[24] A potential truncation of 9 Å was employed for short-ranged interactions, and the particle mesh Ewald summation technique was used to handle long-ranged electrostatics.[26] For the SPC−FW model, since it is flexible, the RESPA algorithm was used with multiple time steps,[27] with a time step of 1 fs for intermolecular interactions and a 0.01 fs time step for bonded interactions.

A total of 1000 water molecules were set up in boxes in slab geometry with periodic liquid containing water molecules in the *x* and *y* directions, and elongated in the *z* direction, giving dimensions of 30 Å (*x*) × 30 Å (*y*) × 100 Å (*z*). The amount of air volume was approximately double the liquid volume for these simulations. Data were collected in a 500 ps production run for the D−C and SPC−FQ water models, and a 1 ns production run was carried out for TIP4P, both after extensive equilibration. The temperature was kept constant at 298 K with the Berendsen thermostat for the TIP4P and D−C models,[28] and the SHAKE algorithm was used to keep the molecules rigid.[29] The SPC−FW model had its temperature kept constant with a Nose−Hoover chains thermostat with one chain for each atom.[30]

Car−Parrinello Molecular Dynamics of Neat Aqueous Liquid−Vapor Interface. The details for the −CPMD simulations are described in detail elsewhere,[15,16,31] and only a brief overview is given here. The CPMD simulations perform DFT-based calculations with the BLYP exchange and correlation functional.[32,33] The system was set up in slab geometry with dimensions 15 Å (*x*) × 15 Å (*y*) × 71.44 Å (*z*) and 216 water molecules. A total of 10 ps of equilibration was carried out, and the results were obtained over 4 ps.

Results and Discussion

Density Profiles. The density as a function of the *z* coordinate is given in Figure 1 for the D−C, TIP4P, and BLYP simulation results. The density profiles were fit to a hyperbolic tangent to determine the Gibbs dividing surface (GDS) and to elucidate the interfacial width (δ):

$$\rho(z) = \frac{1}{2}(\rho_l + \rho_v) - \frac{1}{2}(\rho_l - \rho_v) \tanh\left(\frac{z - z_{\mathrm{GDS}}}{\delta}\right) \qquad (1)$$

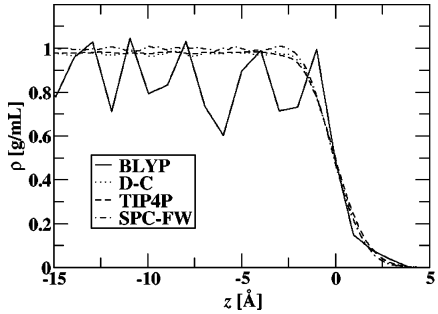


Figure 1. Density profiles for the simulation results for BLYP, TIP4P, D−C, and SPC−FW. Zero in the *z* axis represents the GDS for all figures.

where ρ_l and ρ_v are the average liquid and gas densities, respectively. Table 1 gives the average liquid densities and interfacial widths of the tested water models along with previously determined results[15] for the TIP4P-POL2[34] and TIP4P-FQ[17] water models. The TIP4P-POL2 and TIP4P-FQ models are four-site water models, similar to D−C and TIP4P, but are FQ models instead of using point polarizabilities. While the densities of the TIP4P and D−C water models are indistinguishable, the interfacial length of the D−C water model is smaller than that of TIP4P. The interfacial length for the SPC−FW model is similar to that of D−C, and the interfacial widths for the FQ models are the greatest. The BLYP simulations are dominated by noise, resulting in an icelike profile. However, this is only an artifact of the spatial and temporal sampling in the common procedure for computing density profiles. In a previous study, we computed the Voronoi polyhedra for liquid water averaged over time.[15,16] This procedure only relies on the continuous particle positions and was shown to give identical fluctuations to those obtained with classical simulations. In the same study, the short-time rotational dynamics of the water molecules at the surface and in bulk obtained with classical empirical and DFT interaction potentials were compared.[15] It was found that the time scale of the librational dynamics was nearly identical between models, indicating the presence of a fluid state. However, it is still clear from examining the radial distribution functions obtained with BLYP in the interior regions of the interface that an overstructured water is yielded that is consistent with recent DFT calculations on bulk liquid water.[35−39] There is still considerable speculation as to the exact cause of the observed overstructuring obtained with DFT interaction potentials (e.g., system size, basis set, functionals, and quantum effects). A recent study has shown that utilizing BLYP in the complete basis set limit can reduce the amount of overstructuring.[35] Another DFT study has shown that the use of hybrid density functionals containing exact exchange can also reduce the overstructuring.[36] One should be reminded that all of the aforementioned studies on the overstructuring of liquid water as determined by the radial distribution function were performed at constant volume. The BLYP interface was not constrained to be at 1 g/cm^3, leading to the calculated density being less than 1 g/cm^3 (see Figure 1). To investigate whether this is a result of poor sampling or simulation protocol, extensive Monte Carlo (both Gibbs' ensemble and NPT) studies were conducted to map out the liquid−vapor coexistence of liquid water utilizing DFT interaction potentials.[40−43] These studies have all concluded that the density of liquid water at 298 K and 1 atm is less than 1 g/cm^3, in good agreement with the results obtained in the interior of the liquid−vapor interface. Furthermore, Monte Carlo studies using different functionals and basis sets have been completed, yielding the same qualitative conclusions that DFT interaction potentials yield: a density of water that is less than 1 g/cm^3.[40] From these results, it is not clear how polarizability specifically affects the air−water interfacial width, δ. One should be reminded that the evaluation of δ using the BLYP trajectory was obtained by giving all points in the density profile the same weight.[15] Thus, statistics will play a significant role in this number, and it is more instructive to look at a variety of structural and electronic properties in order to synthesize a coherent picture of the effects of polarization on interfacial properties.

Dipole Distributions. The dipole distributions for the D−C, SPC−FW, and BLYP simulations are given in Figure 2, with the average bulk dipole, along with the average dipole at the GDS for a variety of water models given in Table 1. For all polarizable models and BLYP, the dipole decreases somewhat from the bulk to the GDS and drops off to much lower values outside the GDS. The experimental value of 2.9 ± 0.6 for bulk water[44] is in agreement with all of the models shown, except TIP4P, which is outside this range. BLYP has the greatest decrease in dipole from the bulk to the GDS. Because DFT interaction potentials do not contain dispersion , all of the long range interaction is governed by electrostatics. Thus, the large drop in dipole moment in the vicinity of the interface will give rise to a dramatic loss in the interaction energy, which may account for the surface expansion seen in DFT models of the aqueous liquid−vapor interface. For the classical force fields, the TIP4P-POL model has the smallest drop, while the TIP4P-FQ model has the largest drop (D−C is in between them). Apparently, the type of technique used to model charge rearrangement does not significantly affect the change in water dipole as it approaches the interface. It should be noted that, while flexible water models have significantly different dipoles in the gas and liquid phases,[25] there is very little difference between the bulk and the interface in the molecular dipole for SPC−FW, which is at odds with the DFT interaction potentials.

Water Electrostatic Potential. The EP from atomic charges ($\Delta\phi_q(z)$) can be determined from the integral of the electric field from some reference point in the vapor (z_0) across the air−water interface into the water bulk.[20,45]

$$\Delta\phi_q(z) = \phi_q(z) - \phi_q(z_0) = \int_{z_0}^{z} E_q(z')\,\mathrm{d}z' \qquad (2)$$

The electric field due to fixed charges (E_q) is determined from the integral of charge density as a function of position ($\rho_q(z')$):

$$E_z(z) = \frac{1}{\epsilon_0}\int_{z_0}^{z} \langle \rho_q(z')\rangle\,\mathrm{d}z' \qquad (3)$$

where ϵ_0 is the permittivity of the vacuum and the brackets denote an ensemble average for a liquid slab of 0.5 Å width.

Table 1. Interfacial Widths (δ) and Total Dipole Moments in the Water Bulk and at the GDS for Various Water Molecules

	BLYP	D−C	TIP4P	SPC−FW	TIP4P-POL2[a]	TIP4P-FQ[a]
δ (Å)	0.78	1.45	1.56	1.45	1.782	1.575
ρ_l (g/cm^3)	0.857	0.98	0.98	1.00	0.995	1.007
$\langle\mu_{\mathrm{Bulk}}(D)\rangle$	3.02	2.74	2.18	2.39	2.48	2.64
$\langle\mu_{\mathrm{GDS}}(D)\rangle$	2.6	2.53	2.18	2.39	2.38	2.41

[a] Results taken from ref 15.

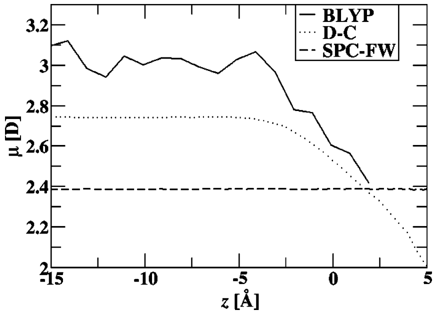


Figure 2. Average dipole as a function of *z* position for models described in Figure 1.

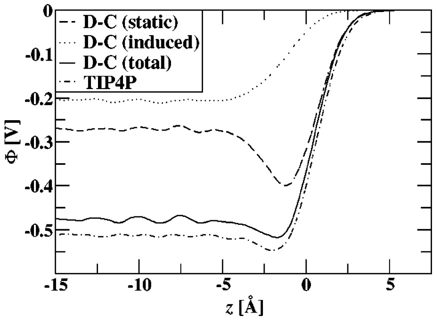


Figure 3. Electrostatic potentials across the air−water interface for the TIP4P and Dang−Chang (D−C in figure) water models, including contributions from static charges and induced dipoles for Dang−Chang.

Equation 2 gives the total electrostatic potential for the TIP4P water model. For polarizable models, such as D−C, an additional contribution comes from the induced dipoles:[19]

$$\Delta\phi_{\mu}^{\mathrm{ind}}(z) = \phi_{\mu}^{\mathrm{ind}}(z) - \phi_{\mu}^{\mathrm{ind}}(z_0) = \frac{1}{\epsilon_0}\int_{z_0}^{z} \langle\rho_{\mu}^{\mathrm{ind}}(z')\rangle\, \mathrm{d}z' \quad (4)$$

where $\rho_{\mu}^{\mathrm{ind}}$ is the induced dipole density. The EPs from static charges and induced dipoles for the TIP4P and D−C molecular models are given in Figure 3. The total EPs for both classical models are quite similar, around −0.5 V, with TIP4P being slightly greater in magnitude. Experimental values suggest that the surface potential for neat water is likely positive,[46] in disagreement with the results here. Wilson et al. found that smearing the charges in a Gaussian distribution results in an increase in surface potential to positive values,[20] which, if applied to the results here, could result in positive surface potential values. The EP for DFT BLYP simulations are underway and will directly address the effects of charge transfer and smeared charge distribution on the calculated surface potential.

The agreement with TIP4P and D−C, along with a large number of classical potentials giving similar EP values,[45] suggests that polarizability has little effect on the total EP if the bulk-phase properties are similar. For the D−C model, though, the EP is distributed among static charges and induced dipoles. The orientation of the TIP4P and D−C models with respect to the surface normal are related to their static EPs. When the static EP decreases from left to right, the water hydrogens are pointing toward the water bulk, and when the EP increases, they are pointing primarily toward the vapor. In the region between 0 and 5 Å from the GDS, the two models' static EPs are nearly identical, showing a similar orientation. Where the models differ significantly in static EP, though, is in the region between 0 and −5 Å from the GDS. In this region, both models show a general decrease in static EP, but the D−C model shows this to a much greater degree. This corresponds to D−C waters orienting their hydrogens in this region toward the water bulk to a much greater degree than those of TIP4P. It should be noted that this orientation of the water dipoles is consistent with second harmonic generation results.[12]

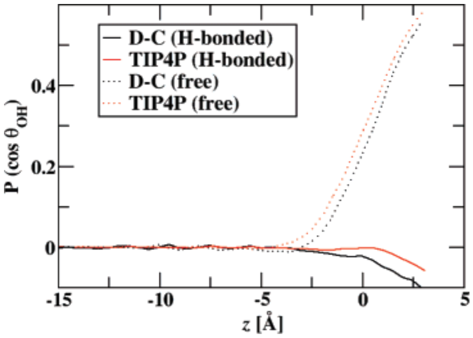


Figure 4. Average oxygen−hydrogen angle with the surface normal, with positive values corresponding to hydrogens pointing away from the water center of mass for the models in Figure 1. The lack of statistics for the BLYP run make a direct comparison between the models to be difficult and inconclusive. However, P(cos θ_{OH}) tended to be positive for both free and H-bonded cases.

Interfacial Water Orientation. To better elucidate the orientation of interfacial water molecules, the distribution of the angle the water oxygen−hydrogen vector forms with respect to the surface normal is given in Figure 4 for both hydrogen-bonded and non-hydrogen-bonded (free) hydrogens. The criteria for a hydrogen bond are described in the next section. The first point of interest is the fact that the free hydrogen orientations are very similar between the D−C and TIP4P models, showing very strong orientation of the free hydrogen toward the vapor, in agreement with many experimental observations.[8−11] There is a noticeable difference between the two models in that the point where the free hydrogen points toward the interface for TIP4P is shifted slightly more toward the interior than for D−C. The most pronounced difference between the two models, though, is

Table 2. Hydrogen-Bond Populations for the Water Bulk and Interface for Models Tested

BLYP[a] TIP4P D−C	0D		1D		2D	
	bulk	interface	bulk	interface	bulk	interface
0A	0.8	3.5	2.9	8.3	2.1	2.3
	1.1	2.6	5.5	9.3	3.4	3.3
	0.8	3.4	3.6	7.6	2.4	2.4
1A	3.5	8.4	19.3	34.2	19.8	14.8
	3.7	6.4	21.1	27.6	17	14.2
	2.8	6.1	17.7	26.0	17.5	14.8
2A	2.2	2.2	18.3	13.8	30.8	12.5
	3.2	3.4	21.3	18.4	22.0	14.0
	2.4	2.6	20.7	17.9	30.0	18.2

[a] Results taken from ref 15.

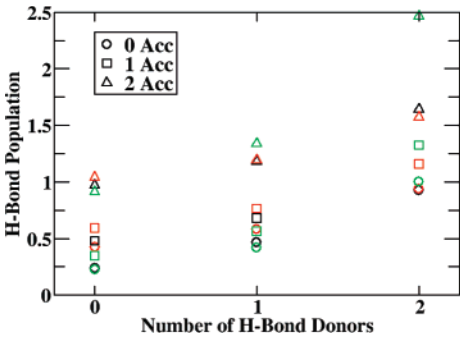


Figure 5. Ratio of bulk to interfacial hydrogen-bond populations for the D−C (black), TIP4P (red), and BLYP (green) results as a function of the number of hydrogen-bond donors and acceptors.

present with the hydrogens that are involved in H bonds. In the region of −2.5 Å and greater, the D−C model clearly shows a greater orientation of its H-bonded hydrogens toward the liquid interior. This is similar to the observation shown in the electrostatic potentials of the two models. The strong decrease in the D−C EP with respect to TIP4P in Figure 3 between −5 and 0 Å is shown to be the result of a combination of a decrease in the propensity for a non-H-bonded hydrogen to point toward the vapor along with an increase in the propensity for an H-bonded hydrogen to point toward the interior.

Hydrogen-Bond Populations. The hydrogen-bond populations in the water bulk and at the interface are given in Table 2. The criteria for hydrogen bonding are a combination of an intermolecular oxygen−hydrogen distance less than 2.27 Å and an oxygen−hydrogen−oxygen angle greater than 150°. Previous studies found that the qualitative trends between the interface and the bulk are similar between these criteria and many others.[15] The interfacial region defined here is considered to be 2*d* from the GDS for TIP4P and D−C. Since the *d* value for the BLYP simulations was much smaller than those for the other two systems, a value of 1.61 Å (same as a previous paper with BLYP)[15] was used for this study to be similar to the other two. To make better comparisons between the different simulation results, Figure 5 gives the ratio of bulk to interfacial hydrogen-bond populations for D−C, TIP4P, and BLYP. It should be noted that the symbol for one donor and two acceptors for the D−C model (black square in right column) is overlapped by the result for TIP4P (red square). The first noticeable trend is that, for most cases, the ratio for D−C is shifted toward the BLYP results from the TIP4P (i.e., the D−C ratio is closer to the BLYP ratio for most cases). The ratios for all entrees are largest for the TIP4P water model except the case with two donors and two acceptors, in which TIP4P is the smallest. From these results, it can be inferred that the inclusion of polarizability decreases the number of fully coordinated hydrogen-bonding waters at the interface. However, the overall population trends in the water bulk are independent of the type of interaction potential.

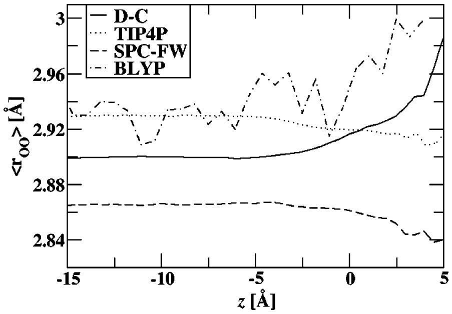


Figure 6. Average first solvation shell oxygen−oxygen distance for water as a function of position.

Table 3. Average Oxygen−Oxygen Distance in the Water Bulk and between the GDS and 2δ from the GDS[a]

model	$\langle r_{OO}\rangle_{bulk}$	$\langle r_{OO}\rangle_{interface}$
BLYP	2.93 Å	2.96 Å
TIP4P	2.930 Å	2.922 Å
D−C	2.900 Å	2.909 Å
SPC−FW	2.866 Å	2.863 Å
TIP4P-POL2[b]	2.96 Å	2.93 Å
TIP4P-FQ[b]	2.99 Å	2.98 Å

[a] Uncertainties are smaller than the last digit reported. [b] Taken from ref 15.

Surface Relaxation. One interesting feature that has been recently observed experimentally using the EXAFS technique is that the oxygen−oxygen distance expands at the interface with respect to the bulk.[7] The concept of surface relaxation is not new and is studied extensively in the solid-state physics community where surface relaxation effects are known to be due to a charge rearrangement of unsatisfied bonds at the solid−vapor interface. Quantifying surface relaxation in a disordered system is much more difficult. The only reporting of this quantity using computational models, to our knowledge, showed that surface relaxation at the neat liquid−vapor interface has not been observed with any classical force fields, including FQ models. However, as previously mentioned, surface relaxation was observed using DFT interaction potentials in conjunction with the BLYP exchange and correlation functional.[15,16] Here, we present the running average oxygen−oxygen distance (r_{OO}) as a function of the position for the models tested in this review (Figure 6). Table 3 gives the average value at the bulk and interface for models

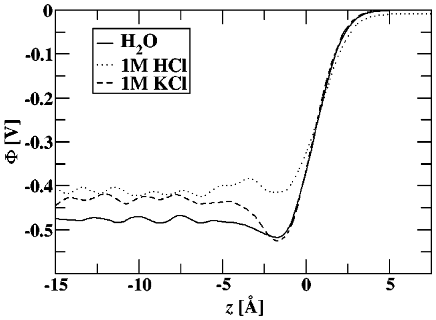


Figure 7. Electrostatic potentials using polarizable models for water, 1 M KCl, and 1 M HCl.

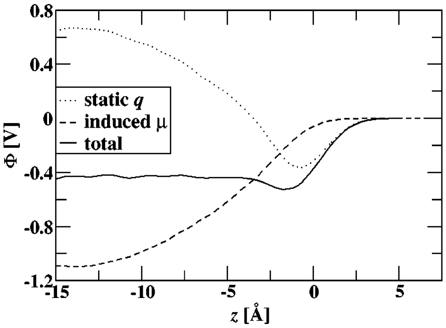


Figure 8. Decomposition of electrostatic potential into contributions from static charge and induced dipoles.

considered in the review, which is to be compared to the data in ref 15. All water models show a contraction at the interface, with the exception of the D−C model and the BLYP results. It is interesting that the D−C model provides an outward expansion that is qualitatively similar to BLYP and experimental results, unlike all the other models tested. The values shown in Table 3 for BLYP and D−C show only a very small increase in r_{OO} corresponding to 1% and 0.3%, respectively, at the GDS. This is much lower than the experimental expansion of 5.9%.[7] However, Figure 6 shows that, outside the GDS, further expansion of the r_{OO} distances occur, leading to increases of 2.4% and 2.9% at 5 Å for BLYP and D−C, respectively, closer to experimental results. In order to make quantitative contact with experimental results, the calculation of the surface versus bulk EXAFS spectra needs to be computed. This is work that is currently underway using representative configurations from the D−C and DFT-BLYP interface calculation in conjunction with the FEFF code to compute the EXAFS spectra. It should be noted that two of the models, the TIP4P-FQ and SPC−FW, do have versions that include polarizability.[47,48]

Electrostatic Potentials For Salt and Acid Solutions. The simulated EPs for 1 M KCl[49] and 1 M HCl solutions with polarizable models were determined. The 1 M HCl solution used 48 classical polarizable hydronium ions,[50] 48 polarizable chloride ions,[51] and 1000 D−C water molecules. These EP results were obtained from 1 ns of simulation time. The total EPs for pure water, 1 M KCl, and 1 M HCl solutions are given in Figure 7. The addition of KCl salt increases the surface potential, in agreement with experimental observations.[18]

The decomposition of the EP into contributions from static charges and induced dipoles is given in Figure 8. The static EP drops originally due to dangling hydrogens from the water molecules, as is the case for pure water, followed by a significant increase in static EP. This increase in static EP is due to the anisotropic pairing of KCl at the interface. The computed density profiles for the 1 M KCl salt solutions confirmed this, by showing the higher anion concentration near the GDS (not shown).[49] Also, it showed an increase in K^+ density between −5 and −7.5 Å from the GDS, just next to the region where Cl^- density is greater than K^+ density. This double layer creates a dipole at the surface pointing toward the gas phase, which contributes negatively to the electric field and positively to the static EP from the vapor to the liquid. The induced dipole EP works against the static EP, being significantly negative in value. The result is that the total EP is negative, but more positive than for pure water. It should be noted that if the total EP was used as a gauge to understand ion pairing at the interface, it would significantly underestimate the true amount of ion pairing, since it does not take into account the effect of induced polarization.

The computed surface potential for 1 M HCl is also included in Figure 7. Upon examining the results, there are several observations that are in order: (1) The shift in the surface potential of 1 M HCl is larger than the corresponding 1 M KCl shift, which is consistent with experimental results.[18] (2) This larger shift is probably due, in part, to the presence of the hydronium ions at the interface. This observation is demonstrated in the snapshots taken from MD simulations shown in Figure 9.

To bring insight into hydronium interfacial activity, its free energy profile is determined using the constrained molecular dynamics potential of mean force (PMF) technique. The PMF technique drags a molecule across an interface, constraining the molecule position and liquid center of mass. The force acting between the constrained liquid and molecule is recorded as a function of the *z* position, yielding a free energy profile across the interface:

$$\Delta F(z_s) = F(z_s) - F_0 = \int_{z_0}^{z_s} \langle f_z(\zeta) \rangle \, d\zeta \qquad (5)$$

For this work, a single hydronium ion was dragged in 1 Å increments across an air−water interface with 1000 water molecules. Figure 10 gives the free energy profile as a function of the position for the hydronium across the air−water interface. As conjectured above, the PMF shows a free energy minimum at the interface, showing a propensity for the hydronium for the air−water interface, in agreement with recent nonpolarizable simulation results[52] and experimental[53] results.

Ion Transfer Across Organic−Water Interfaces. A recent study of the transfer of iodide across the organic−water interface compared the free energy profile with polarizable and nonpolarizable models.[54] The simulations with polarizable models used the D−C water model,[24] a polarizable CCl_4 model,[22] and a polarizable iodide.[51] The simulations with nonpolarizable models included the TIP4P water model,[23] OPLS CCl_4 model,[55] and a nonpolarizable iodide.[19] The free energies for the polarizable and nonpo-

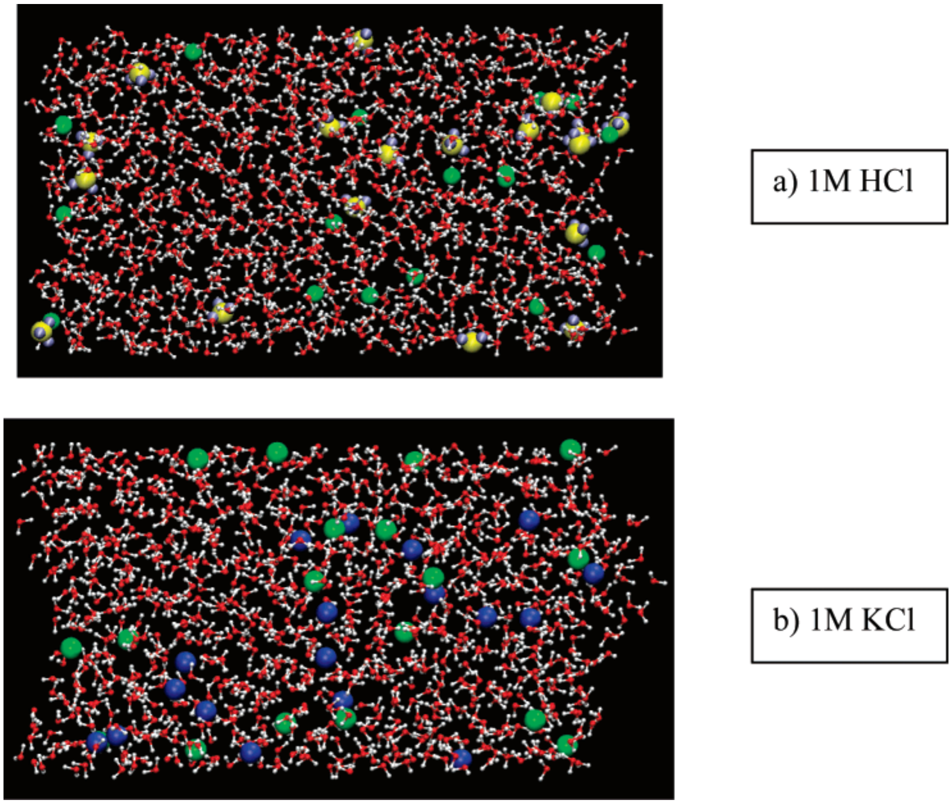


Figure 9. Snapshots taken from MD simulations of 1 M KCl and 1 M HCl.

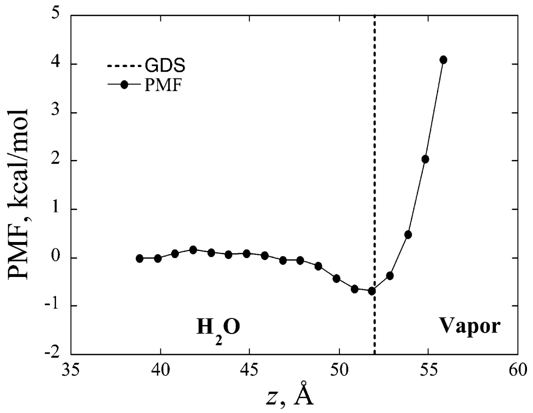


Figure 10. Free energy for transferring a hydronium ion across the air−water interface with polarizable potential models.

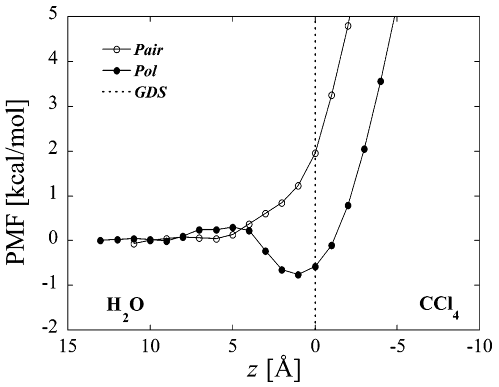


Figure 11. Free energy profile of the transfer of iodide across the $H_2O−CCl_4$ interface for polarizable (pol) and nonpolarizable (non-pol) models.

larizable models using the PMF technique are shown in Figure 11. There is a clear free energy minimum for the simulations with the polarizable model between −2.5 and 0 Å of the GDS, which is not present with the nonpolarizable model. This minimum in the free energy at the water interface that was only present when using polarizability is slightly shallower than that calculated for the air−water interface.[19] What is clear, though, is that the inclusion of polarizability is paramount for the understanding of ion transport across organic−water interfaces, just as it was found for the air−water interface.

Conclusions

We presented a review on the recent progress of the application of molecular dynamics simulation methods, including which polarizable potential models were used, to describe interactions among species, and how they affect a variety of chemical and physical processes at interfaces. It was found that polarizability played an important role for determining the molecular structure and orientation at neat air−water interfaces, including observing surface relaxation at the air−water interface. To our knowledge, only BLYP and Dang−Chang have been shown to result in an expansion at the air−water interface, but it should be stated that other models, especially those with polarizability, would likely show this also. In addition, the effect of polarizability on the understanding of electrostatic potential across the air−water interface, and how it is influenced by the addition of KCl salt and HCl acid, is important. Finally, only with the inclusion of polarizability, the free energy profile of iodide was shown to have a minimum at the organic−water interface.

Acknowledgment. This work was performed at Pacific Northwest National Laboratory (PNNL) under the auspices of the Division of Chemical Sciences, Office of Basic Energy Sciences, U.S. Department of Energy. PNNL is operated by Battelle.

References

(1) Jungwirth, P.; Tobias, D. J. *Chem. Rev.* **2006**, *106*, 1259−1281.

(2) Chang, T. M.; Dang, L. X. *Chem. Rev.* **2006**, *106*, 1305−1322.

(3) Rivera, J. L.; Starr, F. W.; Paricaud, P.; Cummings, P. T. *J. Chem. Phys.* **2006**, *125*.

(4) Motakabbir, K. A.; Berkowitz, M. L. *Chem. Phys. Lett.* **1991**, *176*, 61−66.

(5) Saturo, I.; Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2007**, *126*, 124505 (1−13).

(6) Wilson, K. R.; Rude, B. S.; Catalano, T.; Schaller, R. D.; Tobin, J. G.; Co, D. T.; Saykally, R. J. *J. Phys. Chem. B* **2001**, *105*, 3346−3349.

(7) Wilson, K. R.; Schaller, R. D.; Co, D. T.; Saykally, R. J.; Rude, B. S.; Catalano, T.; Bozek, J. D. *J. Chem. Phys.* **2002**, *117*, 7738−7744.

(8) Richmond, G. L. *Chem. Rev.* **2002**, *102*, 2693−2724.

(9) Gopalakrishnan, S.; Liu, D. F.; Allen, H. C.; Kuo, M.; Shultz, M. J. *Chem. Rev.* **2006**, *106*, 1155−1175.

(10) Du, Q.; Superfine, R.; Freysz, E.; Shen, Y. R. *Phys. Rev. Lett.* **1993**, *70*, 2313−2316.

(11) Shultz, M. J.; Baldelli, S.; Schnitzer, C.; Simonelli, D. *J. Phys. Chem. B* **2002**, *106*, 5313−5324.

(12) Kemnitz, K.; Bhattacharyya, K.; Hicks, J. M.; Pinto, G. R.; Eisenthal, K. B.; Heinz, T. F. *Chem. Phys. Lett.* **1986**, *131*, 285−290.

(13) Ghosal, S.; Hemminger, J. C.; Bluhm, H.; Mun, B. S.; Hebenstreit, E. L. D.; Ketteler, G.; Ogletree, D. F.; Requejo, F. G.; Salmeron, M. *Science* **2005**, *307*, 563−566.

(14) Raymond, E. A.; Richmond, G. L. *J. Phys. Chem. B* **2004**, *108*, 5051−5059.

(15) Kuo, I. F. W.; Mundy, C. J.; Eggimann, B. L.; McGrath, M. J.; Siepmann, J. I.; Chen, B.; Vieceli, J.; Tobias, D. J. *J. Phys. Chem. B* **2006**, *110*, 3738−3746.

(16) Kuo, I. F. W.; Mundy, C. J. *Science* **2004**, *303*, 658−660.

(17) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141−6156.

(18) Randles, J. E. B. *Phys. Chem. Liq.* **1977**, *7*, 107−179.

(19) Dang, L. X.; Chang, T. M. *J. Phys. Chem. B* **2002**, *106*, 235−238.

(20) Wilson, M. A.; Pohorille, A.; Pratt, L. R. *J. Chem. Phys.* **1988**, *88*, 3281−3285.

(21) Wick, C. D.; Dang, L. X. *J. Phys. Chem. B* **2006**, *110*, 6824−6831.

(22) Chang, T. M.; Dang, L. X. *J. Chem. Phys.* **1996**, *104*, 6772−6783.

(23) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(24) Dang, L. X.; Chang, T. M. *J. Chem. Phys.* **1997**, *106*, 8149−8159.

(25) Wu, Y. J.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*.

(26) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(27) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990−2001.

(28) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(29) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(30) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635−2643.

(31) Mundy, C. J.; Kuo, I. F. W. *Chem. Rev.* **2006**, *106*, 1282−1304.

(32) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098−3100.

(33) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(34) Chen, B.; Xing, J. H.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 2391−2401.

(35) Lee, H. S.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*.

(36) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I. F. W.; Mundy, C. J. *J. Phys. Chem. B* **2006**, *110*, 3685−3691.

(37) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300−311.

(38) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*.

(39) Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990−12998.

(40) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J. *Mol. Phys.* **2006**, *104*, 3619−3626.

(41) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *J. Phys. Chem. A* **2006**, *110*, 640−646.

(42) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *ChemPhysChem* **2005**, *6*, 1894−1901.

(43) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J.; VandeVondele, J.; Sprik, M.; Hutter, E.; Mohamed, F.; Krack, M.; Parrinello, M. *Comput. Phys. Commun.* **2005**, *169*, 289−294.

(44) Badyal, Y. S.; Saboungi, M. L.; Price, D. L.; Shastri, S. D.; Haeffner, D. R.; Soper, A. K. *J. Chem. Phys.* **2000**, *112*, 9206−9208.

(45) Sokhan, V. P.; Tildesley, D. J. *Mol. Phys.* **1997**, *92*, 625−640.

(46) Paluch, M. *Adv. Colloid Interface Sci.* **2000**, *84*, 27−45.

(47) Jeon, J.; Lefohn, A. E.; Voth, G. A. *J. Chem. Phys.* **2003**, *118*, 7504−7518.

(48) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237−2251.

(49) Wick, C. D.; Dang, L. X.; Jungwirth, P. *J. Chem. Phys.* **2006**, *125*.

(50) Dang, L. X. *J. Chem. Phys.* **2003**, *119*, 6351−6353.

(51) Dang, L. X. *J. Phys. Chem. B* **2002**, *106*, 10388−10394.

(52) Petersen, M. K.; Iyengar, S. S.; Day, T. J. F.; Voth, G. A. *J. Phys. Chem. B* **2004**, *108*, 14804−14806.

(53) Petersen, P. B.; Saykally, R. J. *J. Phys. Chem. B* **2005**, *109*, 7976−7980.

(54) Wick, C. D.; Dang, L. X. *J. Chem. Phys.* **2007**, *126*, 134702 (1−4).

(55) Duffy, E. M.; Severance, D. L.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1992**, *114*, 7535−7542.

CT700098Z

JCTC Journal of Chemical Theory and Computation

Self-Consistent Reaction Field Model for Aqueous and Nonaqueous Solutions Based on Accurate Polarized Partial Charges

Aleksandr V. Marenich, Ryan M. Olson, Casey P. Kelly, Christopher J. Cramer,* and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota,
207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431

Received June 10, 2007

Abstract: A new universal continuum solvation model (where "universal" denotes applicable to all solvents), called SM8, is presented. It is an implicit solvation model, also called a continuum solvation model, and it improves on earlier SM*x* universal solvation models by including free energies of solvation of ions in nonaqueous media in the parametrization. SM8 is applicable to any charged or uncharged solute composed of H, C, N, O, F, Si, P, S, Cl, and/or Br in any solvent or liquid medium for which a few key descriptors are known, in particular dielectric constant, refractive index, bulk surface tension, and acidity and basicity parameters. It does not require the user to assign molecular-mechanics types to an atom or group; all parameters are unique and continuous functions of geometry. It may be used with any level of electronic structure theory as long as accurate partial charges can be computed for that level of theory; we recommend using it with self-consistently polarized Charge Model 4 or other self-consistently polarized class IV charges, in which case analytic gradients are available. The model separates the observable solvation free energy into two components: the long-range bulk electrostatic contribution arising from a self-consistent reaction field treatment using the generalized Born approximation for electrostatics is augmented by the non-bulk-electrostatic contribution arising from short-range interactions between the solute and solvent molecules in the first solvation shell. The cavities for the bulk electrostatics calculation are defined by superpositions of nuclear-centered spheres whose sizes are determined by intrinsic atomic Coulomb radii. The radii used for aqueous solution are the same as parametrized previously for the SM6 aqueous solvation model, and the radii for nonaqueous solution are parametrized by a training set of 220 bare ions and 21 clustered ions in acetonitrile, methanol, and dimethyl sulfoxide. The non-bulk-electrostatic terms are proportional to the solvent-accessible surface areas of the atoms of the solute and have been parametrized using solvation free energies for a training set of 2346 solvation free energies for 318 neutral solutes in 90 nonaqueous solvents and water and 143 transfer free energies for 93 neutral solutes between water and 15 organic solvents. The model is tested with three density functionals and with four basis sets: 6-31+G(d,p), 6-31+G(d), 6-31G(d), and MIDI!6D. The SM8 model achieves mean unsigned errors of 0.5−0.8 kcal/mol in the solvation free energies of tested neutrals and mean unsigned errors of 2.2−7.0 kcal/mol for ions. The model outperforms the earlier SM5.43R and SM7 universal solvation models as well as the default Polarizable Continuum Model (PCM) implemented in *Gaussian 98/03*, the Conductor-like PCM as implemented in *GAMESS*, *Jaguar*'s continuum model based on numerical solution of the Poisson equation, and the GCOSMO model implemented in *NWChem*.

1. Introduction

Realistic solvation models must include long-range electrostatic polarization effects, which decrease as R^{-4} (R is the distance between the solute and a given solvent molecule), shorter-range polarization effects, and shorter-range non-electrostatic effects such as cavitation, dispersion, and solvent structural effects (CDS), the latter including both hydrogen bonding and exchange repulsion effects.[1−6] These effects can be treated either in terms of explicit (atomistic) solvent or

* Corresponding author e-mail: cramer@chem.umn.edu (C.J.C.) and truhlar@umn.edu (D.G.T.).

10.1021/ct7001418 CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/04/2007

implicit solvent, where the latter is usually represented by a continuous (also called continuum) medium characterized by both macroscopic properties, such as dielectric constant and bulk surface tension, and microscopic properties, such as polarizability and effective solvent radius. The "electrostatic" effect may be described as the electric polarization of the solvent by the polar or nonuniform charge distribution of the solute, and it also includes the effect of the self-consistent distortion of the solute by the polarized solvent. Although some efforts have been made[4,5] to treat nonelectrostatic terms (at least, dispersion) self-consistently (in so-called direct reaction field methods), a much more common assumption in self-consistent solvation models is that the solute charge distribution polarizes due to the electrostatic effects but not due to the nonelectrostatic ones. Thus the solute properties depend on the way that these effects are separated. Unfortunately though, there is no unique way to separate electrostatic effects from solvent structural effects. The ambiguity in current models is well illustrated by a comparison, a few years ago, of three successful aqueous solvation models based on different assumptions and model parameters.[7] For typical solutes (nitroethane, acetone, acetonitrile, benzaldehyde, and tagged water), the average difference between models of predicted standard free energies of solvation is 0.7 kcal/mol, whereas for the same cases the average difference from experiment is 0.6 kcal/mol, and the average difference between models of the electrostatic component is 2.1 kcal/mol.[7] Clearly the nonelectrostatic terms have been parametrized in a way that compensates for the differences in electrostatics.

Solvation models are usually parametrized and/or validated in terms of their ability to predict free energies of solvation, and implicit solvation models approximate such free energies as a sum of electrostatic and nonelectrostatic effects without cross terms. However, because the cross terms are not negligible, there is no unambiguous way to sort out the electrostatic and nonelectrostatic components of free energy. In fact the only possible separation of the free energy changes into components that are state functions (and hence independent of path) is the separation into enthalpy and entropy contributions[8,9] with a further possible separation, usually of little interest, of enthalpy into internal energy and work of compression.

A given separation of the free energy of solvation into electrostatic and nonelectrostatic contributions may therefore be associated with a particular implicit path for thermodynamic integration, and some paths may have more predictive power for modeling than others do.[9] One particularly relevant issue in this regard is that the magnitudes of solvation free energies of ions are much larger than those of neutral solutes and are dominated by large electrostatic contributions. Therefore a parametrization that is carried out in such a way that free energies of solvation of ions are accurate must be doing a good job of modeling electrostatics. By using the same parameters for neutrals one might also achieve a physical estimation of the electrostatics for cases where electrostatic and nonelectrostatic terms are comparable. The nonelectrostatic contribution can then be defined as the difference between the experimentally accessible and path-independent total free energy of solvation and the modeled electrostatic contribution.

The most important parameters for modeling the electrostatics are the atomic radii; we call the radii used in the bulk electrostatics calculation the Coulomb radii (to distinguish them from van der Waals radii or covalent radii and from the radii used in the nonelectrostatic calculation). In the SM*x* series of solvation models ($x = 1,2,...,8$),[3,10−12] the Coulomb radii are calculated by a dielectric descreening approximation[2] from a set of intrinsic atomic Coulomb radii, and it has been our usual practice to optimize these intrinsic atomic Coulomb radii in calculations on ions, then fix these parameters and optimize the nonelectrostatic terms on data for neutrals. There have, however, been two flies in the ointment.

The first problem is that most ionic solvation data have involved an uncertainty related to the partition of the free energy of solvation of a salt or Brønsted acid into separate contributions associated with the cation and the anion because only their sum is well defined in classical thermodynamics.[13] This is resolved by molecular statistical mechanics by determining one absolute ionic solvation free energy, traditionally that of the proton.[14] However, there have been controversies about the value of that key quantity. Recent work, though, has largely eliminated these uncertainties,[15−17] and this enabled us to make a large database of ionic free energies of aqueous solvation.[17] These data were used[17] to test the performance of 13 solvation models for aqueous solvation energies of ions (see ref 18 for details), and the best performance was found for the SM6 solvation parameters[18] with the mPW1PW density functional[19] (also called MPW25, mPW0, and mPW1PW91), the Charge Model 4 (CM4),[18] and the 6-31G(d)[20] basis set.

A second problem though is that implicit solvation models have not been well studied for nonaqueous ionic solvation. One reason is the complication of ion pairing in media with low dielectric constants. Even in dilute solutions in more strongly solvating nonaqueous media, where ionic pairing may be neglected in calculating solvation free energies, although it is not totally absent,[21−24] there is a paucity of data, and until recently there has been no accurate determination of an absolute single-ion solvation energy, which is required, just as in water, to anchor the separate cationic and anionic scales. Recently, the absolute solvation free energy of the proton has been determined in methanol, acetonitrile, and dimethyl sulfoxide (DMSO),[25] and these absolute values can be used to determine databases of ionic solvation data in all three solvents. This now allows us to extend to nonaqueous solutions the strategy of adjusting Coulomb radii to fit ionic data and using the electrostatic model thusly parametrized even for neutrals where non-bulk-electrostatic effects are comparable to electrostatic ones.

In the present article we parametrize a new solvation model for both aqueous and nonaqueous solvents by using the Coulomb radii of SM6 for water and by parametrizing new Coulomb radii for nonaqueous solvents with the new database. The new model is called SM8. Note that SM7[26] denotes an unpublished universal solvation model in which the SM6 Coulomb radii are used in both aqueous and nonaqueous media (the modeling strategy of employing the same Coulomb radii in all media was also used in SM5.2,[27] SM5.4,[28,29] SM5.42,[30−32] SM5C,[33] and SM5.43[34]); however, SM7 yielded some large errors for nonaqueous free energies of solvation of a subset of the ions.

One potential approach that could be used to parametrize the SM8 solvation model would be to develop individual sets of parameters for each solvent. For example, an adequate amount of experimental data exists in solvents like 1-octanol and hexadecane so that developing reasonably accurate sets of solvation parameters in these solvents would be possible. Indeed, two earlier SM*x* models for hexadecane[28] and for chloroform[35] used this approach. However, a major disadvantage that is associated with following this approach is that adequate experimental data do not exist in most other organic solvents, so that developing separate sets of solvation parameters in these solvents is not practical.

To circumvent this problem, a series of universal SM*x* models that can be applied to any solvent has been developed.[18,26,30−34,36−39] In the universal SM*x* models the solvation parameters are functions of a small set of solvent descriptors that are transferable to *any* condensed-phase medium. In this way, a single set of solvation parameters can be developed against a training set that includes experimental data in all solvents, including those solvents for which very little data exist.

In the most recent previously published universal solvation model, SM5.43,[34,39] the solvation parameters were optimized against a training set of data that contained 2237 solvation free energies for 295 solutes in 91 different solvents, including water, 79 transfer free energies between water and 12 organic solvents for an additional 51 solutes, and 47 aqueous solvation free energies for 47 ionic solutes. No experimental data for ionic solutes in nonaqueous solvents were included in this training set. Furthermore the aqueous ion data set used for SM5.43 is smaller than that used for SM6, which was parametrized only for aqueous solution. In this article, an updated version of the SM5.43 neutral training set and a new parametrization strategy involving a smaller number of parameters will be used to develop a new universal solvation model called SM8. It is especially noteworthy that the parametrization will involve new single-ion solvation free energies in acetonitrile, DMSO, and methanol.

All SM*x* solvation models except SM5C are based on discrete partial atomic charges, whereas the SM5C solvation model[33] was based on the continuous electronic density $\rho(\mathbf{r})$ where $\mathbf{r}$ denotes a point in space. The partial atomic charges in models SM1−SM5.2 were obtained by Mulliken[40] population analysis (which yields class II[41] charges), and those in models SM5.4, SM5.42, and SM5.43 were obtained by class IV charge models CM1,[41] CM2,[42] and CM3,[43] respectively. The SM8 model, like SM6[18] and SM7,[26] will be parametrized for the CM4[18] class IV charge model.

We have already mentioned one key respect in which SM8 differs from all previous SM*x* models, namely it involves intrinsic Coulomb radii adjusted to improve the solvation energies of ions in nonaqueous media. A second key difference is the catholicity of the parametrization. In solvation models SM1−SM7, there was a separate set of solvation parameters for each electronic structure level, for example, separate sets for AM1, HF/6-31G(d), mPW1PW/6-31G(d), and mPW1PW/6-31+G(d,p) where AM1 denotes the Austin Model 1 semiempirical molecular orbital theory, HF denotes ab initio Hartree Fock theory, and 6-31+G(d,p)[20] is a basis set. The main reason for carrying out separate parametrizations is that the partial charges depend to some extent on the electronic structure level, and the parameters must be consistent with the partial atomic charges. However, this is true to a much greater extent for Mulliken[40] or Löwdin[44−47] charges than for class IV charges. Therefore in SM8 we will develop only a single set of solvation parameters that is designed to be used with any level of electronic structure theory that supports either the CM4 charge model or other comparably accurate charges. For example, it gives similar accuracy with any class IV charges, and we will show that it can also be used, although somewhat less accurately, with charges from population analysis. Although partial atomic charges are not physical observables, they can still be considered accurate within a given model context if they vary physically with molecular geometry and environment and can be used to accurately reproduce observables such as dipole moments or if they can be derived consistently and realistically from accurate experimental data, for instance, from the dipole moment of a diatomic molecule. The parameters of CM1 and CM2 depend on the specifics of the electronic structure level, but the parameters of CM3 are more general. They can be used with either HF theory or density functional theory (DFT), and they depend only on the basis set and the fraction of HF exchange. CM3 is also parametrized[48] for the self-consistent charge density-functional tight-binding model.[49] The parameters of CM4 are even more general and depend only on basis set. Thus CM3 and CM4 are parametrized for all density functionals (including hybrid ones with any amount of HF exchange), which means that SM8 can use class IV charges with all density functionals. Currently, CM4 parameter sets are available for the MIDI!6D[50,51] basis set and for Pople's[20] 6-31G(d), 6-31+G(d), and 6-31+G(d,p) basis sets. Additional CM4 parameters are under development and will be published soon.[52]

2. Description of the SM8 Universal Model

In the SM*x* models, the standard-state free energy of solvation $\Delta G_\mathrm{S}^\mathrm{o}$ is partitioned according to

$$\Delta G_\mathrm{S}^\mathrm{o} = \Delta E_\mathrm{E} + \Delta E_\mathrm{N} + \Delta G_\mathrm{conc}^\mathrm{o} + G_\mathrm{p} + G_\mathrm{CDS} \qquad (1)$$

where ΔE_E is the change in the solute's internal electronic (E) energy in moving from the gas phase to the liquid phase at the same geometry, ΔE_N is the change in the solute's internal energy due to changes in the equilibrium nuclear (N) positions in the solute that accompany the solvation process, $\Delta G_\mathrm{conc}^\mathrm{o}$ (which is also called[25,39,53] the free energy of liberation or $\Delta G_\mathrm{lib}^\mathrm{o}$) accounts for the concentration change between the gas-phase and the liquid-phase standard states, G_P is the polarization free energy, and G_CDS is the component of the free energy that is nominally associated with cavitation, dispersion, and solvent structure. Following the notation used in previous SM*x* models, the sums $\Delta E_E + G_P$ and $\Delta E_E + \Delta E_N + G_P$ will be referred to as ΔG_{EP} and ΔG_{ENP}, respectively. Since the same concentration (1 mol/L) is used in both the gaseous and solution phases, ΔG^o_{conc} is 0.[53,54] (If we used a gas-phase standard state of 1 atm, ΔG^o_{conc} would be +1.9 kcal/mol.) All calculations reported here are based on gas-phase geometries (although the present model can be used to optimize geometries in the liquid phase[55]), so ΔE_N is assumed to be 0 in this article, although not in the model in general. Since all free energies in this article are standard free energies, we will omit the standard-state modifier in most of the text for brevity.

The ΔG_{EP} contribution to the total solvation free energy is calculated from a self-consistent molecular orbital calculation,[30,56,57] where the generalized Born approximation[2,3,58−61] is used to calculate the polarization contribution to the total free energy according to

$$G_P = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\sum_{k,k'} q_k \gamma_{kk'} q_{k'} \tag{2}$$

In the above equation, the summations go over atoms k in the solute, ϵ is the dielectric constant of the solvent, q_k is the partial atomic charge of atom k, and $\gamma_{kk'}$ is a Coulomb integral involving atoms k and k'.

As in the most recent previously published SM*x* solvation model, SM6,[18] the solvation parameters presented in this work are based on polarization free energies computed by eq 2 using CM4 partial atomic charges self-consistently polarized in solution. In CM4, the partial atomic charges are functions of the partial atomic charges obtained from a Löwdin population analysis[44−47] or a redistributed Löwdin population analysis (RLPA),[62] the gas-phase or liquid-phase Mayer bond orders,[63−65] and a set of atomic-number-dependent empirical parameters. These parameters have been optimized in earlier work and were chosen so as to minimize the errors between accurate gas-phase dipole moments and the dipole moments computed using gas-phase CM4 partial atomic charges. CM4 differs from the previous CM*x* model, CM3,[43,48,66−68] in that for hydrocarbons, CM4 is designed to accurately reproduce the partial atomic charges obtained from the Optimized Potentials for Liquid Simulations (OPLS) force field.[69] (Many of the hydrocarbons, e.g., ethane, in the CM4 training set do not have permanent dipole moments.)

One of the reasons it is preferable to optimize the parameters contained in the SM*x* solvation models using polarization free energies computed with class IV partial charges (as in CM4 and earlier CM*x* models[41−43,48,66−68,70]) is because these types of charge models are usually able to remove many of the systematic errors, in particular basis set dependence, that are present in partial atomic charges obtained from Mulliken,[40] Löwdin,[44−47] and redistributed Löwdin[62] population analyses. This helps to properly shift the focus of the modeling effort toward the various components of the solvation process. In addition, CM4 charges yield more accurate electrostatic potentials than population analysis charges, and this makes the solvation models more physical. It is worth noting that partial atomic charges obtained from any method can be used in eq 2 to compute polarization free energies (see, for example, ref 71). However, one should be aware that in many cases, using different charge models can lead to very different partial atomic charges for a given molecule (and hence polarization free energies).[72] Because of this, it is recommended that, whenever possible, one should use the SM8 solvation parameters with CM4 partial atomic charges or with other charge models that have been validated to give partial atomic charges similar to those of CM4.

The Coulomb integrals $\gamma_{kk'}$ are calculated according to ref 2

$$\gamma_{kk'} = [R^2_{kk'} + \alpha_k\alpha_{k'}\ \exp(-R^2_{kk'}/d\alpha_k\alpha_{k'})]^{-1/2} \tag{3}$$

where $R_{kk'}$ is the distance between atoms k and k', and α_k is the effective Born radius of atom k, which is described below. In the above equation, d is an empirical constant that is usually set equal to 4 (this value was originally proposed by Still et al.,[2] because, for intermediate values of $R_{kk'}$, it gives polarization free energies that are close to those predicted using the classical equation for a dipolar sphere embedded in a dielectric medium), although during the development of SM6 and some earlier solvation models, it was found that optimizing this parameter increased the accuracy. In SM*x* models prior to SM6, when d was not set equal to 4 for all k and k',[27,29,30,32−35,38,39,73−75] it was always set equal to 4 or 3.9 except for the case where one of the atoms k and k' is carbon and the other is hydrogen. In SM6, d was made independent of k and k', and it was optimized to the value of 3.7. We also used that value in SM7, and we will also use it in SM8.

The Born radius is calculated by[76]

$$\alpha_k = \left(\frac{1}{R'} + \int_{\rho_{Z_k}'}^{R'} \frac{A_k(\mathbf{R},r,\{\rho_{Z'}\})}{4\pi r^4}\,\mathrm{d}r\right)^{-1} \tag{4}$$

where R' is the radius of the sphere centered on atom k that completely engulfs all other spheres centered on the other atoms of the solute, and $A_k(\mathbf{R},r,\{\rho_Z\})$ is the exposed area[76] of a sphere of radius r that is centered on atom k. This area depends on the geometry of the solute, $\mathbf{R}$, and the radii of the spheres centered on all the other atoms in the solute. The radii of these spheres are given by a set of intrinsic Coulomb radii ρ_{Z_k} that depend on the atomic number Z_k of the atom k.

The final term on the right-hand-side of eq 1, G_{CDS}, is the first-solvation-shell contribution to the solvation free energy. Examples of first-solvation-shell effects include, but are not limited to, cavitation (C), dispersion (D), and structural (S) effects of solvent molecules in the first solvation shell. In SM8, G_{CDS} is given by

$$G_{CDS} = \sum_k^{\text{atoms}} \sigma_k A_k(\mathbf{R},\{R_{Z_k} + r_s\}) + \sigma^{[M]} \sum_k^{\text{atoms}} A_k(\mathbf{R},\{R_{Z_k} + r_s\}) \tag{5}$$

where σ_k and $\sigma^{[M]}$ are the atomic and the molecular surface tensions of atom k, respectively, and A_k is the solvent-accessible surface area (SASA)[77,78] of atom k. The SASA depends on the geometry $\mathbf{R}$, the set $\{R_{Z_k}\}$ of all atomic van der Waals radii, and the solvent radius r_s, which is added to each of the atomic van der Waals radii. Adding a nonzero value for solvent radius to the atomic radii defines the spheres that are used to compute the SASA of a given solute.[76] Notice that the van der Waals radii used in the SASA calculation are not the same as the intrinsic Coulomb radii used in eq 4; in fact we use Bondi's values[79] for the van der Waals radii.

The atomic surface tensions are given by

$$\sigma_k = \tilde{\sigma}_{Z_k} + \sum_{k'}^{\text{atoms}} \tilde{\sigma}_{Z_k Z_{k'}} T_k(\{Z_{k'}, R_{kk'}\}) \tag{6}$$

where $\tilde{\sigma}_Z$ is an atomic-number-specific parameter, $\tilde{\sigma}_{ZZ'}$ is a parameter that depends on the atomic numbers of atoms k and k', and $T_k(\{Z_{k'},R_{k,k'}\})$ is a geometry-dependent switching function called a cutoff tanh, or COT; this function is described in a previous publication.[18] For H, C, N, O, F, P, S, Cl, and Br, SM8 uses the same functional forms T_k as does SM6. The atomic surface tensions for these atoms were also presented in the previous publication.[18] For SM8, an additional atomic surface tension for Si was added; this atomic surface tension is set equal to the atomic-number-specific parameter for Si (i.e., this atomic surface tension does not include any COT functions)

$$\sigma_Z|_{Z=14} = \tilde{\sigma}_Z \tag{7}$$

As in previous SMx universal solvation models, in SM8 the atomic surface tensions are made to depend on the solvent by making the parameters $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ functions of a set of solvent descriptors. This dependence is given by

$$\tilde{\sigma}_i = \tilde{\sigma}_i^{[n]} n + \tilde{\sigma}_i^{[\alpha]} \alpha + \tilde{\sigma}_i^{[\beta]} \beta \tag{8}$$

where $\tilde{\sigma}_i$ is either $\tilde{\sigma}_Z$ or $\tilde{\sigma}_{ZZ'}$, n is the refractive index of the solvent at room temperature (which is conventionally taken as 293 K for this quantity), α is Abraham's[80−83] hydrogen bond acidity parameter of the solvent (which Abraham denotes as $\Sigma\alpha_2$), β is Abraham's hydrogen bond basicity parameter of the solvent (which Abraham denotes as $\Sigma\beta_2$), and $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, and $\tilde{\sigma}_i^{[\beta]}$ are empirical parameters that depend on i. (Note that Abraham developed $\Sigma\alpha_2$ and $\Sigma\beta_2$ as solute descriptors, but we use them as solvent descriptors.) Besides making the atomic surface tensions depend on the solvent through the use of eq 8, SM8 also uses a molecular surface tension that is multiplied by the total SASA of the given solute (see eq 5; the total SASA of the solute is equal to the sum of the SASAs of each of the individual atoms in the solute). The molecular surface tension is also a function of solvent descriptors, and it is given by

$$\sigma^{[\text{M}]} = \tilde{\sigma}^{[\gamma]}\left(\frac{\gamma}{\gamma_o}\right) + \tilde{\sigma}^{[\phi^2]}\phi^2 + \tilde{\sigma}^{[\psi^2]}\psi^2 + \tilde{\sigma}^{[\beta^2]}\beta^2 \tag{9}$$

where γ is the macroscopic surface tension of the solvent at air/solvent interface at 298.15 K expressed in cal mol^{-1}Å^{-2} (note that 1 dyn/cm = 1.43932 cal mol^{-1}Å^{-2}), γ_o = 1 cal mol^{-1}Å^{-2}, ϕ^2 is the square of the fraction of solvent atoms that are aromatic carbon atoms (carbon aromaticity), ψ^2 is the square of the fraction of solvent atoms that are F, Cl, or Br (electronegative halogenicity), β^2 is the square of Abraham's hydrogen bond basicity parameter of the solvent, and $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$ are empirical parameters that are independent of the solute.

The chosen solvent descriptors are physically meaningful.[27,75] For example, the refractive index n is a measure of solvent's polarizability, which in turn is related to dispersion interactions of the solvent. The acidity and basicity parameters are related to the solvent's ability to donate and accept hydrogen bonds, respectively. The solvent's macroscopic surface tension represents the energy required for cavitation (creation of a surface) in the solvent. The aromaticity and electronegative halogenicity factors are used to account for systematic differences in intermolecular interactions in aromatic solvents and solvents containing electronegative halogen atoms.

In SM8 as well as in previous universal solvation models water is treated as a special solvent that is given its own set of $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ values, so that eq 8 is not needed for water. Also, for water, the molecular surface tension $\sigma^{[\text{M}]}$ is set equal to zero. Thus, when SM8 is used to compute solvation free energies in aqueous solvent, eq 5 reduces to

$$G_{\text{CDS,water}} = \sum_{k}^{\text{atoms}} \sigma_k A_k(\mathbf{R}, \{R_{Z_k} + r_s\}) \tag{10}$$

where $\tilde{\sigma}_Z$ or $\tilde{\sigma}_{ZZ'}$ used in eq 6 to obtain σ_k are simply numbers that do not depend on solvent descriptors.

3. Parameters to be Optimized

During the development of SM6,[18] the parametrization effort was focused on two types of parameters: (1) the atomic radii used in eq 4 and (2) the parameters $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ used in eq 10. For previous universal SMx models, it has been shown that using solvent-independent values for the intrinsic Coulomb radii, the van der Waals radii, and the solvent radius r_s leads to relatively accurate solvation free energies in both water and nonaqueous solvents; that is, it was assumed that the solvent dependence of the solvation free energy is entirely contained in $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$. This assumption is too restrictive for the present work, and we will allow the intrinsic atomic Coulomb radii to depend on solvent in nonaqueous solvents while retaining the SM6 values in water. Then the parameters to be determined are a solvent-dependent set of intrinsic atomic Coulomb radii for nonaqueous solvents and the full set of atomic surface tensions, namely $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, and $\tilde{\sigma}_i^{[\beta]}$ that appear in eq 8 and $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$ that appear in eq 9. The $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ parameters for water also are to be optimized as part of this work. The symbols $\tilde{\sigma}_i^{[\text{water}]}$ will be used to denote the $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ parameters that are optimized specifically for water, where i is either Z or ZZ'. For Si, which was not included in the SM5.43R[34,39] or the SM6 parametrizations, Bondi's value[79] of 2.10 Å will be used in eq 4 for water and also to compute the SASA for all solvents; this is the same value for the atomic radius that was used by a previous universal SMx model that included Si.[75]

4. SM8 Universal Model Training Set

Standard States. In this article, all gas-phase free energies use a standard-state gas-phase pressure of 1 atm. All solvation free energies in the present article are tabulated for the gas-phase solute having a standard state of an ideal gas at a gas-phase concentration of 1 mol/L and for the liquid-phase solute being dissolved in an ideal solution at a liquid-phase concentration of 1 mol/L. Transfer free energies between water and organic solvents use a standard state for which the concentration is equal in both phases.

Experimental Data for Neutrals. The present training set begins with the portions of the SM5.43 and SM6 training sets that contain experimental aqueous solvation free energies of neutral solutes. This subset contains aqueous solvation free energies for 273 neutral solutes (including the water dimer) containing one or more of the elements H, C, N, O, F, P, S, Cl, or Br. To this subset of data was added the experimental aqueous solvation free energy of tetramethylsilane, which was taken from an earlier training set.[75]

The nonaqueous portion of the SM5.43 training set[34,39] contains 1980 solvation free energies for 263 solutes containing one or more of the elements H, C, N, O, F, P, S, Cl, or Br, in 90 organic solvents. This training set also contains 79 transfer free energies between water and 12 organic solvents, which are a subset of the 90 organic solvents. These transfer free energies were determined directly from experimental partition coefficients according to

$$\Delta G^{\mathrm{o}}_{\mathrm{o/w}} = -2.303RT \log P_{\mathrm{o/w}} \tag{11}$$

where $\Delta G^{\mathrm{o}}_{\mathrm{o/w}}$ is the standard-state free energy associated with transferring the solute from the aqueous phase w to the organic phase o, and $P_{\mathrm{o/w}}$ is the corresponding partition coefficient, which is given by

$$P_{\mathrm{o/w}} = \frac{[\mathrm{solute}]_{\mathrm{o}}}{[\mathrm{solute}]_{\mathrm{w}}} \tag{12}$$

where $[\mathrm{solute}]_{\mathrm{o}}$ is the equilibrium concentration of the solute in the organic phase, and $[\mathrm{solute}]_{\mathrm{w}}$ is the equilibrium concentration of the solute in the aqueous phase. Transfer free energy data are included in this as well as several previous SM*x* training sets,[33,34,39,75] because for many solutes the experimental data that are required to determine the solvation free energy between the gas and liquid phases are not available. Thus, if one were restricted to considering only solvation free energies, many solutes containing important functionality would not be well represented (or not represented at all) in the training set. It is worth noting that many of the solvation free energies in the SM5.43 training set are derived from experimental partition coefficients and experimental aqueous solvation free energies, that is

$$\Delta G^{\mathrm{o}}_{\mathrm{o/a}} = \Delta G^{\mathrm{o}}_{\mathrm{w/a}} - 2.303RT \log P_{\mathrm{o/w}} \tag{13}$$

where $\Delta G^{\mathrm{o}}_{\mathrm{o/a}}$ is the solvation free energy in the organic solvent o (a denotes the gas phase, or air), $\Delta G^{\mathrm{o}}_{\mathrm{w/a}}$ is the aqueous solvation free energy, and $P_{\mathrm{o/w}}$ is the partition coefficient.

Before incorporating the experimental data taken from the SM5.43 training set into the current training set, several updates and corrections were made to these data. For nitromethane, the SM5.43 training set contains solvation free energies in carbon tetrachloride and cyclohexane as well as transfer free energies between water and carbon tetrachloride and water and cyclohexane. The two transfer free energies are redundant and were removed. Using an experimental value for the aqueous solvation free energy of nitromethane,[18] the transfer free energy of nitromethane between octanol and water was converted to the solvation free energy of nitromethane in octanol using eq 13. Similarly, the transfer free energies of 3,5-dimethylpyridine between benzene and water, 4-ethylpyridine between octanol and water, γ-butyrolactone between octanol and water, pyrrole between chloroform and water, octanol and water, and cyclohexane and water, and quinoline between chloroform and water, octanol and water, and cyclohexane and water were converted to solvation free energies in the above organic solvents using eq 13 and experimental values[18,84] for the aqueous solvation free energies of these solutes. During the development of SM6, the experimental value for the aqueous solvation free energy of hydrazine was updated from −9.30 kcal/mol to −6.26 kcal/mol. In the SM5.43 training set, experimental values for the solvation free energies of hydrazine in benzene, octanol, diethyl ether, and chloroform were determined using experimental partition coefficients and an experimental value of −9.30 kcal/mol for the aqueous solvation free energy of hydrazine in eq 13; these solvation free energies were adjusted to reflect the above update in the aqueous solvation free energy of hydrazine. Finally, in the SM5.43 training set, the experimental value for the transfer free energy of phenylurea between chloroform and water is incorrectly listed as −0.86 kcal/mol. The current training set uses the correct value, which is +0.86 kcal/mol.[85]

To the above subset of data, experimental partition coefficients[85] and experimental aqueous solvation free energies were used to add 80 solvation free energies in organic solvents for the following 14 solutes: hydrogen peroxide, urea, benzamide, methylhydrazine, 2-methylaniline, 3-methylaniline, 4-methylaniline, *N*-methylaniline, *N*-methyl-2-pyrrolidinone, 2-pyrrolidinone, formamide, *N*,*N*-dimethylformamide, *N*-methylformamide, and *N*,*N*-dimethylacetamide. Sixty-three relative solvation free energies between water and organic solvents for 31 solutes, the majority of which contain amide groups, were also added. Finally, experimental values for the solvation free energy of tetramethylsilane in hexadecane and in octanol and experimental values for the transfer free energies of 13 other solutes containing Si between water and octanol were added. These experimental data were taken from an earlier training set.[75]

Combining all of the data from above results in a total of 2346 solvation free energies for 318 neutral solutes in 91 solvents (including water) and 143 transfer free energies for 93 neutral solutes between water and 15 organic solvents. These data will be referred to as the SM8 universal model neutral training set. Note that this training set does not contain any ionic solutes. Experimental data for ionic solutes will be discussed below.

Table 1. 90 Nonaqueous Solvents in the SM8 Neutral Training Set[a]

acetic acid	*1,2-dibromoethane*	methoxyethanol
acetonitrile*	*dibutyl ether*	methylene chloride*
acetophenone	*o*-dichlorobenzene	methylformamide
aniline*	*1,2-dichloroethane**	4-methyl-2-pentanone
anisole	*diethyl ether**	2-methylpyridine
*benzene**	diisopropyl ether	*nitrobenzene*
benzonitrile	*N*,*N*′-dimethylacetamide	nitroethane
benzyl alcohol	*N*,*N*′-dimethylformamide	nitromethane*
bromobenzene	2,6-dimethylpyridine	*o*-nitrotoluene
bromoethane	dimethyl sulfoxide*	nonane
bromoform	dodecane	nonanol
bromooctane	ethanol*	octane
1-*butanol*	ethoxybenzene	*octanol*
2-butanol	*ethyl acetate*	pentadecane
butanone	ethylbenzene	pentane
butyl acetate	fluorobenzene	pentanol
n-butylbenzene	1-fluoro-*n*-octane	perfluorobenzene
sec-butylbenzene	*heptane**	phenyl ether
t-butylbenzene	heptanol	propanol
carbon disulfide	hexadecane	pyridine
*carbon tetrachloride**	hexadecyl iodide	tetrachloroethene
*chlorobenzene**	*hexane*	tetrahydrofuran*
*chloroform**	hexanol	tetrahydrothiophene dioxide
chlorohexane	iodobenzene	tetralin
m-cresol	isobutanol	toluene*
*cyclohexane**	isooctane	tributylphosphate
cyclohexanone	isopropanol	triethylamine
decalin (mixture)	isopropylbenzene	1,2,4-trimethylbenzene
decane	*p*-isopropyltoluene	undecane
decanol	mesitylene	xylene (mixture)

[a] Methanol is not included in this training set because there are no data for neutral solutes in methanol. The asterisk denotes the nonaqueous solvents presently available with the default solvation model implemented in *Gaussian 03* in addition to methanol and water. The names of 15 solvents for which we used solvent−water transfer free energies are italicized.

Table 1 shows the 90 organic solvents used in the SM8 parametrization. A table containing experimental values for the 2346 solvation free energies and 143 transfer free energies contained in the SM8 universal model neutral training set is given in the Supporting Information. Also included in the Supporting Information are calculated values obtained using the SM8 model described below.

Solvation Free Energies of Ions in Water. The current ion training set for ions in aqueous solution was explained in the article where SM6 was parametrized.[18] In particular we use the data set called the selectively clustered set. In this set, there are 112 ions; 81 of these are unclustered and 31 are clustered with a single water molecule each (these ions are not included in unclustered form). The criterion for whether to cluster an ion is that it is clustered if it contains three or fewer atoms, or if the partial charge on any oxygen is more negative than the partial charge on oxygen in water, or if the ion is an oxonium or ammonium cation. The rationale for this criterion is explained in the SM6 paper.[18]

Single-Ion Solvation Free Energies of Unclustered Ions in Acetonitrile, DMSO, and Methanol. In previous work, the cluster pair approximation was used to estimate the absolute solvation free energy of the proton in acetonitrile, DMSO, and methanol.[25] These values can be combined with experimental or calculated data to determine absolute solvation free energies of other ionic solutes in these solvents. For example, using thermochemical cycle 1 (illustrated in Scheme 1) the absolute solvation free energy of the cation BH^+ can be written as

Scheme 1. Thermochemical Cycle Relating the Solvation Free Energy of BH^+ to the Gas-Phase Basicity (GB) of the Base B

$$\begin{tikzcd}
BH^+(g) \arrow[r, "\Delta G_g^o(BH^+) = GB(B)"] \arrow[d, "\Delta G_S^o(BH^+)"] & B(g) \arrow[d, "\Delta G_S^o(B)"] & + & H^+(g) \arrow[d, "\Delta G_S^o(H^+)"] \\
BH^+(s) \arrow[r, "2.303RT\mathrm{p}K_a(BH^+)"] & B(s) & + & H^+(s)
\end{tikzcd}$$

Scheme 2. Thermochemical Cycle Relating the Solvation Free Energy of A^- to the Gas-Phase Acidity of the Acid AH

$$\begin{tikzcd}
AH(g) \arrow[r, "\Delta G_g^o(AH)"] \arrow[d, "\Delta G_S^o(AH)"] & A^-(g) \arrow[d, "\Delta G_S^o(A^-)"] & + & H^+(g) \arrow[d, "\Delta G_S^o(H^+)"] \\
AH(s) \arrow[r, "2.303RT\mathrm{p}K_a(AH)"] & A^-(s) & + & H^+(s)
\end{tikzcd}$$

$$\Delta G_S^o(BH^+) = \Delta G_g^o(BH^+) + \Delta G_S^o(B) - 2.303RT\mathrm{p}K_a(BH^+) + \Delta G_S^o(H^+) + \Delta G^{1\mathrm{atm}\rightarrow 1\mathrm{M}} \quad (14)$$

where $\Delta G_g^o(BH^+)$ is the gas-phase acidity of BH^+, which is equal to

$$\Delta G_g^o(BH^+) = G^o(B) + G^o(H^+) - G^o(BH^+) \quad (15)$$

$\Delta G_S^o(B)$ is the solvation free energy of the neutral species B, pK_a is the negative common logarithm of the solution-phase acid dissociation constant of BH^+, and $\Delta G^{1\mathrm{atm}\rightarrow 1\mathrm{M}}$ is the free energy change associated with moving from a gas-phase pressure of 1 atm to a liquid-phase concentration of 1 M. Similarly, thermochemical cycle 2 (illustrated in Scheme 2) can be used to write the absolute solvation free energy of the anion A^-

$$\Delta G_S^o(A^-) = -\Delta G_g^o(AH) + \Delta G_S^o(AH) + 2.303RT\mathrm{p}K_a(AH) - \Delta G_S^o(H^+) - \Delta G^{1\mathrm{atm}\rightarrow 1\mathrm{M}} \quad (16)$$

In previous work, the above equations were combined with experimental pK_a values, gas-phase acidities, and aqueous solvation free energies of neutral species in order to determine single-ion solvation free energies in aqueous solution. However, for the solvents acetonitrile, DMSO, and methanol, finding solutes for which experimental pK_a values, gas-phase acidities, *and* solvation free energies (of the neutral species) exist is challenging. In particular, for those solutes where experimental pK_a values in one of the solvents above and gas-phase acidities are available, experimental solvation free energies in the given solvent are typically not available.

Table 2. Reference Free Energies of Solvation of Selected Ions (kcal/mol)[a]

neutral molecule (AH or B)	charge	ΔG_g^o	ΔG_s^o (neutral)	pK_a	ΔG_s^o (ion)
			Acetonitrile		
ammonia	+1	195.7	−4.29	16.5[90]	−89.3
pyridine	+1	214.7	−6.34	12.3[89,90]	−66.7
acetic acid	−1	341.4	−6.04	22.3[98,106,107]	−58.8
phenol	−1	342.9	−7.20	27.0[97,113]	−55.1
			DMSO		
ammonia	+1	195.7	−3.95	10.5[114]	−93.9
pyridine	+1	214.7	−5.71	3.5[114]	−67.2
acetic acid	−1	341.4	−5.95	12.3[114]	−59.2
phenol	−1	342.9	−7.22	18.0[114]	−54.2
			Methanol		
ammonia	+1	195.7	−5.05	10.8[125]	−85.6
pyridine	+1	214.7	−6.57	5.4[125,126]	−60.8
acetic acid[b]	−1	341.4	−6.25	9.7[110,125,128]	−72.9
phenol	−1	342.9	−7.60	14.4[125,131,139]	−69.3

[a] The free energies of solvation for all ions and the auxiliary data are listed in the Supporting Information. For cations BH^+, ΔG_g^o is the gas-phase acidity of BH^+ equal to the gas-phase basicity[87] of the neutral base B (see eq 14 and Scheme 1). For anions A^-, ΔG_g^o is the gas-phase acidity[88] of the neutral acid AH (see eq 16 and Scheme 2). The free energies of solvation for neutral species ΔG_s^o(neutral) are calculated at the SM7/mPW1PW/6-31G(d) level of theory.[26] The values of pK_a(BH^+) and pK_a(AH) are reference data. In case of multiple references, the pK_a values are averaged over the references. The absolute free energies of ions ΔG_s^o(ion) are based on the following values[25] for the absolute solvation free energy of the proton in the three solvents: −260.2 (acetonitrile), −273.3 (DMSO), and −263.5 (methanol) kcal/mol. [b] See also refs 130, 131, and 133 for pK_a.

The SM8 universal solvent model as well as earlier SM*x* universal solvent models can predict solvation free energies of neutral solutes in nonaqueous media to an accuracy of ∼0.6 kcal/mol. Thus, calculated instead of experimental values can be used for the solvation free energies of neutral solutes in the above equations to obtain relatively reasonable estimates of the solvation free energies of single ions. To do this, a data set[86] of experimental gas-phase acidities[87,88] and experimental pK_a values[89−139] in acetonitrile, DMSO, and methanol was created. For each of the species, solvation free energies were calculated at the SM7/*m*PW1PW/6-31G(d) level of theory.[26] Using these experimental and calculated data, single-ion solvation free energies were determined in acetonitrile, DMSO, and methanol. The resulting free energies for all the ions and the auxiliary data used to determine them are listed in the Supporting Information. Table 2 shows examples of these data for a few typical ions in the three solvents.

Clustered Ions in Nonaqueous Solvents. For all four cations in DMSO and for any ion in acetonitrile and DMSO that contains one or more halogen atoms, we calculated solvation free energies of clustered ions (with a single solvent molecule in the cluster) by using precisely the same procedure as used previously[18,25] to obtain solvation free energies of clustered ions in water. This procedure is illustrated in cycle 3 (see Scheme 3) that relates the solvation free energy of a clustered ion to the gas-phase binding free energy of the cluster. The binding free energies were calculated at the B97-1[140]/MG3S[141] level of theory using *Gaussian 03*,[142] except for the clusters of Cl^- and Br^-, for which the experimental binding free energies were available in standard reference data.[143] The molecular geometries for all of the clusters were optimized, and conformational analysis was carried out by calculating harmonic frequencies to verify the nature of minima and by searching to find the global minimum conformations in the gas phase. The primary reason for adding the clustered-ion data was to increase the number of data to achieve a more robust fit. The resulting solvation free energies of clusters are given in Table 3. The molecular structures of a few typical clusters are depicted in Figure 1. The Cartesian coordinates corresponding to the B97-1/MG3S optimized global minima for all of the nonaqueous clusters used in this work are given in the Supporting Information.

Scheme 3. Thermochemical Cycle Relating the Solvation Free Energy of an Ionic Solute $M^{\pm}$ to the Solvation Free Energy of the Solvent−Solute Cluster $S\bullet M^{\pm}$

$$\begin{tikzcd}
M^{\pm}(g) \arrow[d, "\Delta G_S^o(M^{\pm})"] & + & S(g) \arrow[d, "\Delta G_S^o(S)"] \arrow[r, "\Delta G_g^o(BE)"] & S\bullet M^{\pm}(g) \arrow[d, "\Delta G_S^o(S\bullet M^{\pm})"] \\
M^{\pm}(s) & + & S(s) \arrow[r, "zero"] & S\bullet M^{\pm}(s)
\end{tikzcd}$$

Table 3. Solvation Free Energies of Solvent−Solute Clusters Used in Optimization of the SM8 Coulomb Radii[a]

neutral molecule (AH or B)	ion	ΔG_g^o (BE)[b]	ΔG_s^o (bare ion)[c]	ΔG_s^o (cluster)
		Acetonitrile Clusters		
hydrochloric acid	Cl^-	−9.5[d]	−62.4	−55.9
hydrobromic acid	Br^-	−8.7[d]	−59.3	−53.6
trifluoroacetic acid	$CF_3CO_2{}^-$	−6.3	−45.6	−42.2
3-trifluoromethylphenol	$CF_3C_6H_4O^-$	−5.0	−46.9	−44.8
chloroacetic acid	$CH_2ClCO_2{}^-$	−7.2	−54.6	−50.4
2-chlorobenzoic acid	$C_6H_4ClCO_2{}^-$	−5.7	−53.5	−50.8
3-chlorophenol	$C_6H_4ClO^-$	−5.8	−50.6	−47.7
dichloroacetic acid	$CHCl_2CO_2{}^-$	−5.7	−51.2	−48.5
3,4,5-trichlorophenol	$C_6H_2Cl_3O^-$	−4.0	−43.8	−42.7
		DMSO Clusters		
ammonia	$NH_4{}^+$	−29.1[e]	−93.9	−70.6
aniline	$C_6H_5NH_3{}^+$	−18.9[e]	−79.8	−66.7
methylamine	$CH_3NH_3{}^+$	−23.1[e]	−82.4	−65.0
pyridine	$C_5H_5NH^+$	−18.3[e]	−67.2	−54.6
hydrochloric acid	Cl^-	−12.5[d]	−62.7	−55.9
hydrobromic acid	Br^-	−10.9[d]	−57.8	−52.6
2,2,2-trifluoroethanol	$CF_3CH_2O^-$	−11.3	−56.1	−50.6
trifluoroacetamide	CF_3NHCO^-	−5.9	−49.2	−49.1
trifluoroacetic acid	$CF_3CO_2{}^-$	−5.7	−45.0	−45.0
2-chlorobenzoic acid	$C_6H_4ClCO_2{}^-$	−6.4	−53.6	−52.9
4-chlorobenzoic acid	$C_6H_4ClCO_2{}^-$	−7.0	−52.6	−51.3
dichloroacetic acid	$CHCl_2CO_2{}^-$	−5.6	−49.2	−49.3

[a] All entries (in kcal/mol) are given for 298.15 K. [b] Solvent−solute binding free energies (BE), calculated in this work at the B97-1/MG3S level of theory, unless indicated otherwise. [c] Taken from Tables S1−S3 in the Supporting Information (part II). [d] Experimental energies.[143] [e] B97-1/MG3S energies.[25]

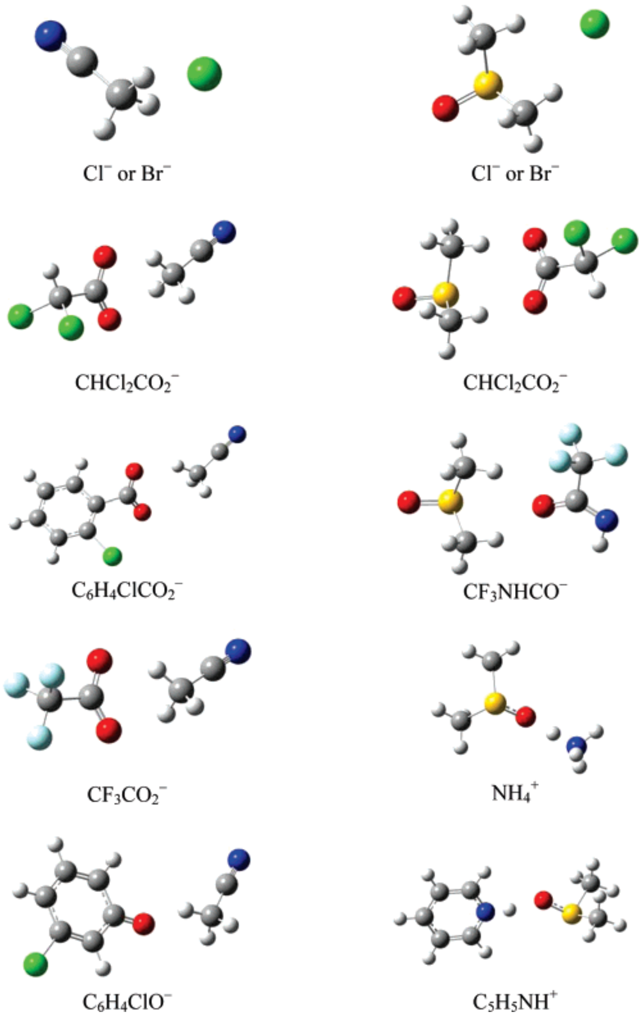


Figure 1. Clusters of selected ions with acetonitrile and DMSO.

Molecular Geometries of Solutes. All computed solvation free energies in this article are based on rigid, gas-phase geometries. The molecular geometries of all unclustered neutral and ionic solutes are optimized at the *m*PW1PW/MIDI![50,51,144] level of electronic structure theory.

The use of gas-phase optimized geometries allows us to save computational time without any significant loss of accuracy in the SM8 parametrization. Indeed, because most solutes considered here prefer similar conformations and structures in the gas phase and solution the difference in solvation free energy between using gas-phase geometries and using liquid-phase geometries is smaller than the mean error of the model in many cases. Having obtained the parameters with such a training set, they can be used more broadly in further applications, for instance for the liquid-phase geometry optimization when the solute's geometry is expected to change significantly upon passing a solute from the gas phase to solution.

5. Parametrization

The first step was to parametrize the intrinsic Coulomb radii. The intrinsic Coulomb radii for aqueous solution are frozen at the values that were optimized for the SM6 model in previous work.[18] The intrinsic Coulomb radii for nonaqueous solution were optimized using the solvation free energies of ions in acetonitrile, DMSO, and methanol in Tables 2 and S1−S3 in part II of the Supporting Information. A number of schemes were tested in which the intrinsic Coulomb radii of various sets of atoms were optimized as functions of the solvent hydrogen-bond acidity α and hydrogen-bond basicity *β*. The values of these solvent descriptors for the four solvents including water in which we have ionic data are listed in Table 4. After considerable trial and error we concluded that there was no reason to change any of the intrinsic Coulomb radii from their water values in methanol and no reason to change the carbon or nitrogen intrinsic Coulomb radii in any solvent. We also found that for the three nonaqueous solvents tested, making the Coulomb radii functions of *β* had little effect on the overall accuracy of the model. Thus we settled on the scheme

Table 4. Solvent Acidity and Basicity Descriptors for the Four Solvents with Ionic Data[a]

solvent	α	*β*
acetonitrile	0.07	0.32
DMSO	0	0.88
methanol	0.43	0.47
water	0.82	0.35

[a] α is Abraham's[80−83] hydrogen bond acidity parameter (which Abraham denotes as $\Sigma\alpha_2$), and *β* is Abraham's hydrogen bond basicity parameter (which Abraham denotes as $\Sigma\beta_2$).

$$\rho_Z = \begin{cases} \rho_Z(\text{water}) & \alpha \geq 0.43 \\ \rho_Z(\text{water}) + a(0.43 - \alpha) & \alpha < 0.43 \end{cases} \tag{17}$$

with *a* as a parameter, and we constrained *a* to zero for $Z = 6$ and 7. There is not enough data to optimize *a* for $Z = 15$, so it was set equal to the value for $Z = 16$. The optimized radii are given in Table 5 where they are compared to some previous SM*x* intrinsic Coulomb radii[18,26,30−32,34] and to the van der Waals radii of Bondi. (The other columns in this table will be explained below.) The free energies of solvation for ions in acetonitrile and DMSO calculated by SM8 and SM7 are compared to the corresponding reference data in Tables S4 and S5 in part II of the Supporting Information.

A technical point should be mentioned here. The optimum intrinsic Coulomb radii actually depend slightly on the atomic surface tensions (whereas the atomic surface tensions depend strongly on the intrinsic Coulomb radii). Thus we optimized the intrinsic Coulomb radii for ions with SM7 atomic surface tensions,[26] then determined a first round of SM8 surface tensions using neutral data, then reoptimized the intrinsic Coulomb radii for ions with these atomic and molecular surface tensions, and then found the final atomic surface tensions by a final round of calculations on neutrals.

To begin the parametrization, of the atomic surface tensions, ΔG_{EP} values were calculated for all of the solutes in the SM8 universal model training set for which solvation free energies are available (total of 2346 calculations). The ΔG_{EP} values for all of the solutes in the training set for which transfer free energies are available, in water, and in the organic solvent to which the transfer free energy refers, were

Table 5. Intrinsic Coulomb Radii (Å) of Various Models and Bondi's van der Waals Radii (Å)[a]

		SM8									
atom	Z	*a*	ρ_Z(water)	ρ_Z(DMSO)	SM6/SM7[b]	SM5.43[c]	SM5.42[d]	C-PCM GAMESS	PB Jaguar	GCOSMO NWChem	Bondi[e]
H	1	−0.52	1.02	0.80	1.02	0.79	0.91	1.44	1.15	1.20	1.20
C	6	0	1.57	1.57	1.57	1.81	1.78	2.04	1.90	1.50	1.70
N	7	0	1.61	1.61	1.61	1.66	1.92	1.92	1.60	1.50	1.55
O	8	1.54	1.52	2.18	1.52	1.63	1.60	1.80	1.60	1.40	1.52
F	9	2.69	1.47	2.63	1.47	1.58	1.50	1.62	1.68	1.35	1.47
Si	14	0	2.10	2.10	2.10	2.10	2.10	2.40	2.15	1.17	2.10
P	15	0.77	1.80	2.13	1.80	2.01	2.27	2.28	2.07	1.80	1.80
S	16	0.77	2.12	2.45	2.12	2.22	1.98	2.22	1.90	1.75	1.80
Cl	17	1.42	2.02	2.63	2.02	2.28	2.13	2.17	1.97	1.70	1.75
Br	35	0.59	2.60	2.85	2.60	2.38	2.31	2.34	2.10	1.80	1.85

[a] The SM8 parameters *a* (eq 17) were optimized for H, O, F, S, Cl, and Br. The value of *a*(P) was fixed at the *a*(S) value. The SMx model radii are compared to the default values of radii used in the Conductor-like Polarizable Continuum Model (C-PCM/GAMESS) as implemented in *GAMESS*, in *Jaguar*'s Poisson−Boltzmann self-consistent reaction field solver (PB/Jaguar), and in the generalized Conductor-like screening model implemented in *NWChem* (GCOSMO/NWChem). [b] References 18 and 26. [c] Reference 34. [d] References 30−32 and 75. [e] Reference 79.

also calculated (a total of 286 calculations). A locally modified version[145] of the *Gaussian 03*[142] electronic structure package was used to carry out the above calculations. The above calculations also gave the computed COT functions for each molecule in the training set as well as the SASAs for each atom in each molecule in the training set (the COT functions and SASAs are independent of the solvent).

Optimizing the parameters for nonaqueous solvents, $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$, $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$, and the parameters for water, $\tilde{\sigma}_i^{[\text{water}]}$, involves minimizing the following error function

$$\chi = \sum_{j=1}^{4} \sum_{J=1}^{2489} |\Delta G_S^o(\text{expt},J) - \Delta G_{EP}(j,J) - G_{CDS}(j,J)| \tag{18}$$

where the first summation is over four levels of electronic structure theory (in particular, mPW1PW[19] with four different basis sets: MIDI!6D,[50,51] 6-31G(d),[20] 6-31+G(d),[20] and 6-31+G(d,p)[20]), and the second summation is over all data points in the neutral training set (2346 solvation free energies plus 143 transfer free energies), and $\Delta G_S^o(\text{expt},J)$ is the experimental solvation or transfer free energy. For solvation free energies, $\Delta G_{EP}(j,J)$ and $G_{CDS}(j,J)$ can be calculated directly with the solvation model, whereas for transfer free energies, two separate solvation model calculations are required, that is,

$$\Delta G_{\text{EP, transfer}} = \Delta G_{\text{EP,organic}} - \Delta G_{\text{EP,water}} \tag{19}$$

$$G_{\text{CDS,transfer}} = G_{\text{CDS,organic}} - G_{\text{CDS,water}} \tag{20}$$

where $\Delta G_{\text{EP,organic}}$ and $\Delta G_{\text{EP,water}}$ are calculated in the same way, except that different values are used for the dielectric constant in eq 2, and $G_{\text{CDS,organic}}$ and $G_{\text{CDS,water}}$ are computed using eqs 5 and 10, respectively. Note that because transfer free energies depend on both the aqueous solvation free energy and the solvation free energy in the organic solvent (eq 13), the parameters for nonaqueous solvents and the parameters for water must be optimized simultaneously.

The optimization of the above parameters was carried out in three stages. First, the $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$, and $\tilde{\sigma}_i^{[\text{water}]}$ parameters for atoms involving at most H, C, N, and O and the $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$ parameters were optimized against data for molecules containing H, C, N, and/or O. Next, these parameters were frozen, then the $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$, and $\tilde{\sigma}_i^{[\text{water}]}$ parameters for atoms involving the elements F, S, Cl, and Br were optimized against data for molecules containing H, C, N, and/or O, plus F, S, Cl, and/or Br. Finally, these parameters were frozen, and then the $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$, and $\tilde{\sigma}_i^{[\text{water}]}$ parameters for atoms involving P or Si were optimized against molecules containing P or Si.

SM8 uses the same functional forms for the atomic surface tensions as SM6, which contains 25 different $\tilde{\sigma}_i^{[\text{water}]}$ values. Thus, SM8 contains these 25 parameters for water, plus 75 $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$ parameters, the 4 $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$ parameters, and the 4 $\tilde{\sigma}_{\text{Si}}^{[\text{water}]}$, $\sigma_{\text{Si}}^{[n]}$, $\sigma_{\text{Si}}^{[\alpha]}$, and $\sigma_{\text{Si}}^{[\beta]}$ parameters for silicon that were introduced as part of this work (108 parameters in all). However, as demonstrated by the performance of previous universal SM*x* models, it is not necessary (or desirable) to use all of these parameters. In an earlier paper,[27] a set of rules was adopted for determining which parameters to include, which are as follows: (1) If a parameter affects less than two different solutes, it is set to zero. (2) If using a parameter does not improve the mean unsigned error for the affected solutes by at least 0.1 kcal/mol, the parameter is set to zero. (3) Any surface tension coefficient that is not set to zero by either of these rules is retained.

For SM8 we instead used a different approach. In order to determine which parameters to retain in the SM8 parametrization, we used an approach based on statistical significance. First, *all* 108 parameters were optimized using the three-step procedure outlined above. Any parameter with a value greater than 1000 cal mol^{-1} Å^{-2} was removed, then the remaining parameters were reoptimized. In addition to being very large in value, all of the parameters that were removed in this stage of the optimization also had low values for the statistical significance. Next, the parameter with the least amount of statistical significance was removed, and then all of the remaining parameters were reoptimized. This step was repeated, until the statistical significance associated with each parameter was greater or equal to 95%. The only

Table 6. Surface Tension Model Parameters for SM8[a]

i	$\tilde{\sigma}_i^{[\text{water}]}$	$\tilde{\sigma}_i^{[n]}$	$\tilde{\sigma}_i^{[\alpha]}$	$\tilde{\sigma}_i^{[\beta]}$
H	58.93	22.02		
C	91.53	59.79	19.30	75.66
H, C	−81.35	−66.35		
C, C	−70.57	−63.62		−54.83
O	−97.68	−20.89	71.43	−142.42
H, O	−123.51	−78.77		
O, C	164.72	−11.64	127.68	134.04
O, O	86.92		122.12	−58.17
N	47.91	57.33	−120.41	
H, N	−118.50	−50.59		
C, N		−89.40	297.26	−105.62
N, C	−52.45	−3.88	−48.74	
N, C(2)	−194.27		−453.99	
N, C(3)	69.80			161.49
O, N	190.74			243.64
F	32.13			
Cl		−25.74		
Br	−19.92	−37.89		
S	−38.08	−49.29		
O, P	163.44		271.82	
Si		−72.87		
	$\tilde{\sigma}^{[\gamma]} = 0.21$[b]	$\tilde{\sigma}^{[\phi^2]} = -2.79$[b]	$\tilde{\sigma}^{[\psi^2]} = -8.46$[b]	$\tilde{\sigma}^{[\beta^2]} = 2.51$[b]

[a] The units of $\tilde{\sigma}$ are cal mol^{-1}Å^{-2}. The 54 CDS coefficients were optimized by fitting theoretical electrostatic solvation energies to the corresponding reference data that contained 274 aqueous free energies, 2072 nonaqueous free energies, and 143 transfer free energies. [b] These quantities are multiplied by the total solvent accessible surface area of a solute molecule.

exceptions were made for the two parameters $\tilde{\sigma}_{\text{Si}}^{[n]}$ and $\tilde{\sigma}_{\text{Br}}^{[\text{water}]}$, which were retained despite having statistical significances of 93% and 91%, respectively. In all, only 54 of the original 108 parameters were retained, compared to the 75 parameters that are used by SM5.43. The final set of parameters obtained using the procedure described above is listed in Table 6. Note that in this table, only 21 types of surface tension parameters are listed, even though there are 26 possible types. This is because for $\tilde{\sigma}_{\text{C,C(2)}}$, $\tilde{\sigma}_{\text{H,S}}$, $\tilde{\sigma}_{\text{S,S}}$, $\tilde{\sigma}_{\text{P}}$, and $\tilde{\sigma}_{\text{S,P}}$ the final $\tilde{\sigma}_i^{[\text{water}]}$, $\tilde{\sigma}_i^{[n]}$, $\sigma_i^{[\alpha]}$, and $\sigma_i^{[\beta]}$ values that result from following the above procedure are all equal to zero.

Thus we vary only 64 independent parameters, consisting of the 10 parameters a in eq 17 (see also Table 5) and 54 CDS parameters in eqs 5 and 10 (Table 6). These parameters are fit to 2730 reference free energies, which consist of 2346 solvation free energies for 318 neutral solutes in 90 nonaqueous solvents and water, 143 transfer free energies for 93 neutral solutes between water and 15 organic solvents, and 241 ionic free energies for 220 bare ions and 21 clustered ions in acetonitrile, dimethyl sulfoxide, and methanol. Thus there are more than 42 data per parameter, and this large ratio contributes to the robustness of the SM8 model. We refer the reader to results of the cross-validation procedure performed for the earlier SM5.43 model (that uses more parameters than SM8 and thus is formally less robust than SM8) by random removal of 25% of the data from the SM5.43 training set.[34] In these tests, the solvation free energies of solutes not used to train the SM5.43 model were predicted with only slightly increased mean unsigned errors (0.47 vs 0.42 kcal/mol for aqueous and 0.52 vs 0.50 kcal/mol for organic neutral data).

6. Performance for Neutrals

Table 7 gives a breakdown of the errors in calculated aqueous solvation free energies by solute class. In Tables 8 and 9, the errors are broken down by solute class for calculated solvation free energies in nonaqueous solvents and for calculated transfer free energies, respectively. These tables show the results not only for the four levels of electronic structure theory used in the parametrization but also for two other density functional levels, namely B3LYP[146−148]/6-31G-(d) and M06-2X[149]/6-31G(d). Tables 7−9 show generally good agreement with experiment across both solute classes and electronic structure levels. The mean errors for electronic structure levels not included in the parametrization are not systematically worse (and in fact are often better) than those for the four levels used in parametrization.

Table 10 provides an overall summary of the performance of SM8 for neutral data and a comparison to SM7 and SM5.43 (which is called SM5.43R in the tables because in the SM5.43 model, the convention had been to append "R" if a gas-phase geometry was used; we will continue to say SM5.43 in the text). The performance of SM8 is quite comparable to the performance of SM5.43 and SM7 for neutral solvation.

Table 10 also shows (in parentheses) the mean unsigned errors we obtain if the sum over j in eq 18 is restricted to a single term, yielding a new set of atomic surface tensions for each level of electronic structure theory. The results are typically better by only 0.01−0.04 kcal/mol. We are willing to accept the slightly larger errors obtained with the catholic parameters, and so we are not publishing the individually optimized parameters.

The SM8 errors in solvation and transfer free energies for a few selected solvents, for which we have the most abundant solute data, are listed in Table 11. The errors over all solvents are given in the Supporting Information. The errors are small, lying within the uncertainty of experimental data. We found no statistically significant correlation between the errors and solvent values of hydrogen bond acidity or basicity parameters or other solvent properties.

7. Performance for Ions in Acetonitrile, DMSO, Methanol, and Water

Using the solvation free energies listed in Tables 2 and S1−S3 in part II of the Supporting Information as well as the earlier SM6 selectively clustered aqueous ion set, the performance of SM8 was tested for predicting solvation free energies of ions. Note that the data in Table 3 were used in parametrization but are not included in the error analyses of this section. See Tables S4 and S5 in part II of the Supporting Information for the error analyses including the clusters.

Table 12 shows the mean unsigned errors in SM8 solvation free energies of ions in the four solvents with three density functionals and four basis sets. As was mentioned above for neutral solutes, the overall performance of the SM8 model for ions only slightly depends on the change of electronic

Table 7. Mean Unsigned Errors (kcal/mol) in Aqueous Solvation Free Energies Calculated Using SM8, by Solute Class[a]

		mPW1PW				B3LYP	M06-2X
solute class	*N*	MIDI!6D	6-31G(d)	6-31+G(d)	6-31+G(d,p)	6-31G(d)	6-31G(d)
H_2,NH_3,H_2O,$(H_2O)_2$	4	1.07	1.59	1.81	1.63	1.43	1.37
unbranched alkanes	8	0.88	0.88	1.05	1.01	0.82	0.83
branched alkanes	5	0.82	0.79	0.91	0.90	0.73	0.74
cycloalkanes	5	0.74	0.74	0.81	0.78	0.63	0.66
alkenes	9	0.43	0.45	0.59	0.53	0.35	0.36
alkynes	5	0.44	0.36	0.29	0.39	0.59	0.53
arenes	8	0.27	0.31	0.80	0.60	0.37	0.24
alcohols	12	0.63	0.41	0.40	0.46	0.50	0.59
phenols	4	0.94	0.53	0.47	0.58	0.98	0.79
ethers	12	0.45	0.46	0.57	0.61	0.46	0.62
aldehydes	6	0.74	0.46	0.39	0.45	0.36	0.38
ketones	12	0.50	0.36	0.56	0.57	0.28	0.52
carboxylic acids	5	0.45	0.44	1.07	1.04	0.59	0.86
esters	13	0.60	0.43	0.41	0.40	0.26	0.15
peroxides	3	0.43	0.14	0.26	0.22	0.13	0.12
bifunctional H,C,O compounds	5	0.76	0.45	0.40	0.47	0.62	0.94
aliphatic amines	15	0.75	0.61	0.57	0.59	0.60	0.60
anilines	7	0.61	0.41	0.26	0.53	0.92	0.79
aromatic N-heterocycles (1 N)	10	0.15	0.20	0.54	0.47	0.53	0.66
aromatic N-heterocycles (2 Ns)	2	1.29	0.43	0.45	0.47	0.35	0.34
nitriles	4	0.65	0.34	1.07	1.05	0.39	1.03
hydrazines	3	1.22	0.99	0.97	0.89	0.86	0.85
bifunctional H,C,N compounds	3	0.46	0.55	0.39	0.83	0.69	0.63
amides	4	1.00	0.71	0.99	1.05	0.85	1.12
ureas	2	0.52	0.41	1.04	0.52	0.29	0.41
thymines (uracils)	1	1.18	1.76	0.23	0.77	1.61	0.68
nitrohydrocarbons	7	0.74	0.30	0.42	0.45	0.32	0.41
bifunctional H,C,N,O compounds	3	0.72	0.30	0.16	0.22	0.04	0.23
fluoroalkanes	5	1.00	0.58	0.32	0.31	0.55	0.29
fluoroarenes	1	0.11	0.28	0.80	0.71	0.00	0.08
chloroalkanes	13	0.30	0.31	0.46	0.53	0.28	0.27
chloroalkenes	6	0.57	0.59	0.44	0.52	0.66	0.70
chloroarenes	8	0.29	0.41	0.68	0.85	0.22	0.20
bromoalkanes	9	0.18	0.17	0.35	0.35	0.16	0.15
bromoalkenes	1	0.14	0.06	0.09	0.06	0.17	0.23
bromoarenes	4	0.27	0.41	0.65	0.53	0.16	0.22
multihalogen hydrocarbons	12	0.48	0.29	0.31	0.33	0.28	0.27
halogenated bifunctional compounds	9	1.39	1.12	1.77	1.77	1.18	1.48
thiols	4	0.71	0.60	0.50	0.43	0.74	0.72
sulfides	5	0.79	0.80	0.59	0.51	0.92	0.94
disulfides	2	0.16	0.09	0.49	0.65	0.20	0.22
sulfur heterocycles	1	0.24	0.24	0.27	0.11	0.50	0.41
halogenated sulfur compounds	2	1.26	1.70	0.84	1.08	1.48	1.71
phosphorus compounds	14	1.21	1.48	1.55	1.58	1.50	1.62
silicon compounds	1	0.33	0.23	0.27	0.27	0.31	0.17
all neutral data	274	0.63	0.55	0.66	0.67	0.57	0.62

[a] All the solvation free energies were obtained using the SM8 model parameters. *N* is the number of data in a given solute class.

structure level. Again, the errors for B3LYP/6-31G(d) and M06-2X/6-31G(d), which were not included in the SM8 parametrization, are not systematically larger than for those levels of theory that were used in parametrization.

Experimental solvation free energies ΔG_S^o for neutral solutes in the SM8 training set vary from −14.1 kcal/mol (for chrysene in hexadecane) to 4.3 kcal/mol (for octafluoropropane in water) with the average value (averaged over all 2346 data) equal to −4.8 kcal/mol. Experimental solvation free energies ΔG_S^o for ions in acetonitrile, DMSO, methanol, and water vary from −110.3 kcal/mol (for aqueous H_3O^+) to −36.0 kcal/mol (for 2,4-dinitrophenoxide anion in acetonitrile) with the average value (averaged over all 332 data) equal to −65.0 kcal/mol. The SM8 model predicts these average values of ΔG_S^o quite precisely: indeed, −4.9 kcal/mol for neutrals and −66.0 kcal/mol for ions. The average SM8 values of ΔG_{ENP} are −2.1 kcal/mol for neutral solutes and −64.3 kcal/mol for ions. (Recall that ΔG_{ENP} is approximated as ΔG_{EP} in the present article. All SM8 results given in this paragraph are calculated at the mPW1PW/6-

Table 8. Mean Unsigned Errors (kcal/mol) in Nonaqueous Solvation Free Energies Calculated Using SM8, by Solute Class[a]

		mPW1PW				B3LYP	M06-2X
solute class	*N*	MIDI!6D	6-31G(d)	6-31+G(d)	6-31+G(d,p)	6-31G(d)	6-31G(d)
H_2, NH_3, H_2O, $(H_2O)_2$	29	1.60	1.81	2.00	2.03	1.73	1.72
unbranched alkanes	85	0.45	0.45	0.41	0.42	0.45	0.45
branched alkanes	7	0.39	0.41	0.39	0.39	0.41	0.41
cycloalkanes	13	0.47	0.47	0.49	0.48	0.42	0.43
alkenes	18	0.41	0.40	0.36	0.37	0.42	0.42
alkynes	9	0.52	0.47	0.39	0.46	0.57	0.55
arenes	134	0.44	0.50	0.83	0.71	0.35	0.38
alcohols	272	0.38	0.39	0.39	0.39	0.38	0.38
phenols	109	0.75	0.59	0.52	0.59	0.81	0.72
ethers	87	0.68	0.67	0.67	0.68	0.69	0.71
aldehydes	32	0.69	0.68	0.76	0.77	0.63	0.60
ketones	195	0.86	0.77	0.75	0.80	0.70	0.51
carboxylic acids	120	0.51	0.54	0.75	0.73	0.58	0.68
esters, including lactones[b]	243	0.44	0.42	0.48	0.50	0.44	0.47
peroxides	17	0.58	0.60	0.60	0.60	0.58	0.59
bifunctional H,C,O compounds	24	1.37	1.22	1.07	1.10	1.30	1.42
aliphatic amines	154	0.43	0.41	0.40	0.40	0.42	0.43
anilines	61	0.38	0.36	0.37	0.36	0.48	0.43
aromatic N-heterocycles (1 N)	52	0.62	0.61	0.59	0.60	0.64	0.65
aromatic N-heterocycles (2 Ns)	8	0.46	0.58	0.84	1.15	0.81	0.94
nitriles	20	0.70	0.58	0.75	0.75	0.54	0.51
hydrazines	5	1.30	1.26	1.26	1.24	1.26	1.27
bifunctional H,C,N compounds	2	0.79	1.02	0.77	0.81	0.80	0.80
amides	26	0.69	0.60	0.65	0.67	0.70	0.83
ureas	7	1.14	0.85	0.59	0.93	1.02	1.10
lactams	4	0.69	0.79	0.84	0.89	0.88	0.95
thymines (uracils)	1	0.67	0.97	0.19	0.22	0.78	0.42
nitrohydrocarbons	86	0.77	0.56	0.43	0.44	0.56	0.53
bifunctional H,C,N,O compounds	3	0.72	0.67	0.83	0.97	0.77	0.80
fluoroalkanes	5	0.86	0.72	0.54	0.54	0.70	0.61
fluoroarenes	11	0.54	0.60	0.76	0.73	0.55	0.57
chloroalkanes	26	0.44	0.51	0.59	0.63	0.48	0.43
chloroalkenes	15	0.69	0.72	0.58	0.62	0.75	0.77
chloroarenes	31	0.31	0.33	0.42	0.39	0.34	0.32
bromoalkanes	21	0.55	0.54	0.61	0.61	0.51	0.48
bromoalkenes	2	0.09	0.08	0.18	0.16	0.10	0.11
bromoarenes	16	0.41	0.49	0.63	0.58	0.35	0.39
multihalogen hydrocarbons	14	0.46	0.39	0.42	0.42	0.39	0.35
halogenated bifunctional compounds	37	1.14	0.99	1.13	1.18	1.14	1.11
thiols	10	0.30	0.23	0.18	0.17	0.35	0.31
sulfides	13	0.88	0.91	0.94	0.92	0.87	0.89
disulfides	4	0.38	0.41	0.37	0.36	0.47	0.47
sulfurheterocycles	4	0.63	0.63	0.29	0.39	0.80	0.74
sulfoxides	1	0.33	0.92	0.92	0.96	0.76	0.40
phosphorus compounds	37	1.43	1.63	1.61	1.63	1.67	1.73
silicon compounds	2	1.96	1.53	1.92	1.92	1.46	1.57
all neutral data	2072	0.60	0.57	0.61	0.62	0.58	0.57

[a] All the solvation free energies were obtained using the SM8 model parameters. *N* is the number of data in a given solute class. [b] Five lactones and 238 other esters.

31G(d)/CM4 level.) The ΔG_{EP} results indicate that the average signed value of non-bulk-electrostatic contributions in the free energy of solvation ΔG_S^o is approximately the same for both neutral and ionic solutes (−2.8 kcal/mol for neutrals and −1.7 kcal/mol for ions; the smaller absolute value for ions is understandable in that the typical ion in our data set is smaller than the typical neutral solute). The average absolute value of G_{CDS}, i.e. $\langle|G_{CDS}|\rangle$, is 3.0 kcal/mol for neutrals and 2.3 kcal/mol for ions, whereas $\langle|\Delta G_S^o|\rangle$ is 5.0 kcal/mol for neutrals and 66.0 kcal/mol for ions, and $\langle|\Delta G_{EP}|\rangle$ is the same as $|\langle\Delta G_{EP}\rangle|$ since ΔG_{EP} is intrinsically negative. In the generalized Born approximation, ΔG_{EP} need not be negative (as it should be), but there are only 7 positive values (out of 2346 neutral solvation free energies in our SM8 data set), and the largest—hexadecane in hexadecane—is only +0.15 kcal/mol.

Table 9. Mean Unsigned Errors (kcal/mol) in Transfer Free Energies between Water and Organic Solvents Calculated Using SM8, by Solute Class[a]

		mPW1PW				B3LYP	M06-2X
solute class	*N*	MIDI!6D	6-31G(d)	6-31+G(d)	6-31+G(d,p)	6-31G(d)	6-31G(d)
lactones	10	1.27	1.03	0.96	0.96	1.05	0.87
aromatic N-heterocycles	6	0.43	0.44	0.39	0.34	0.37	0.33
bifunctional H,C,N compounds	2	0.79	0.91	0.74	0.74	0.89	0.81
amides	13	0.96	0.79	0.98	0.77	0.88	1.14
ureas	11	0.30	0.35	0.32	0.27	0.34	0.27
lactams	4	1.66	1.60	1.72	1.73	1.62	1.74
thymines and uracils	12	0.78	0.98	0.70	0.73	1.03	0.60
bifunctional H,C,N,O compounds	5	0.45	0.60	0.47	0.46	0.64	0.53
halogenated bifunctional compounds	39	0.77	0.83	0.69	0.66	0.84	0.65
sulfur compounds (with no P)	19	0.42	0.47	0.54	0.53	0.47	0.49
phosphorus compounds	9	0.69	1.15	0.55	0.55	1.21	1.44
silicon compounds	13	0.82	0.81	0.76	0.78	0.80	0.81
all neutral data	143	0.74	0.78	0.70	0.66	0.80	0.74

[a] All the solvation free energies were obtained using the SM8 model parameters. *N* is the number of data in a given solute class.

Table 10. Mean Unsigned Errors (kcal/mol) in Solvation Free Energies of Neutral Solutes Using CM4

			MUE		
model[a]	DFT method	basis	aqueous[b]	organic[c]	transfer[d]
SM8-CM4	mPW1PW	MIDI!6D	0.63 (0.58)	0.60 (0.60)	0.74 (0.70)
SM8-CM4	mPW1PW	6-31G(d)	0.55 (0.55)	0.57 (0.57)	0.78 (0.68)
SM8-CM4	mPW1PW	6-31+G(d)	0.66 (0.63)	0.61 (0.60)	0.70 (0.69)
SM8-CM4	mPW1PW	6-31+G(d,p)	0.67 (0.63)	0.62 (0.61)	0.66 (0.68)
SM8-CM4	B3LYP	6-31G(d)	0.57 (0.55)	0.58 (0.57)	0.80 (0.70)
SM8-CM4	M06-2X	6-31G(d)	0.62 (0.55)	0.57 (0.54)	0.74 (0.68)
SM7[e]	mPW1PW	6-31G(d)	0.53	0.61	0.70
SM5.43R[f]	mPW1PW	6-31G(d)	0.55	0.61	1.02

[a] The CDS contributions in the SM8 free energies of solvation were found using the 54 parameters ($\tilde{\sigma}_i$) of the SM8 model presented in Table 6 with the exception of the numbers in the parentheses that were obtained by optimization of the 54 parameters $\tilde{\sigma}_i$ for each specified basis set and density functional. [b] Two hundred seventy-four data. [c] Two thousand seventy-two data. [d] One hundred forty-three data. [e] The SM7 model[26] uses 56 parameters $\tilde{\sigma}_i$ and the CM4 charge model. [f] The SM5.43R model[34] uses 75 parameters $\tilde{\sigma}_i$ and the CM3 charge model.

Table 11. Errors (kcal/mol) in Solvation and Transfer Free Energies for Neutrals Calculated at the SM8/mPW1PW/6-31G(d) Level of Theory, by Solvent[a]

solvent	*N*	α	*β*	MSE[b]	MUE[c]
benzene	75	0.00	0.14	0.21	0.65
carbon tetrachloride	78	0.00	0.00	−0.45	0.62
chlorobenzene	38	0.00	0.07	−0.36	0.54
chloroform	105	0.15	0.02	0.11	0.77
cyclohexane	91	0.00	0.00	0.05	0.49
decane	39	0.00	0.00	−0.18	0.41
dichloroethane	38	0.10	0.11	0.30	0.59
diethyl ether	67	0.00	0.41	0.05	0.71
heptane	66	0.00	0.00	−0.02	0.36
hexadecane	190	0.00	0.00	−0.03	0.50
hexane	59	0.00	0.00	−0.08	0.49
isooctane	32	0.00	0.00	−0.37	0.48
octane	38	0.00	0.00	−0.09	0.39
octanol	206	0.37	0.48	−0.10	0.66
octanol−water transfer	90	0.37	0.48	0.04	0.65
toluene	51	0.00	0.14	0.08	0.45
water	274	0.82	0.35	−0.06	0.55
xylene	48	0.00	0.16	0.11	0.45

[a] *N* is the number of neutral solute data in a given solvent. Only the solvents with *N* > 30 are listed here. See the Supporting Information for all data. α is Abraham's[80−83] hydrogen bond acidity parameter (which Abraham denotes as $\Sigma\alpha_2$), and *β* is Abraham's hydrogen bond basicity parameter (which Abraham denotes as $\Sigma\beta_2$). [b] Mean signed error. [c] Mean unsigned error.

8. Performance of Other Continuum Models

In addition to SM8, SM7, and SM5.43, we tested the performance of several other implicit solvent models, in particular five models that serve as default solvation models in five popular quantum-chemical program packages: (1) the Integral Equation Formalism Polarizable Continuum Model[150,151] of *Gaussian 03*,[142] namely IEF-PCM/G03;[152−155] (2) the dielectric version[150,151,156] of PCM (D-PCM/G98) as implemented in *Gaussian 98*;[157] (3) the Conductor-like PCM model[150,151,158−164] in *GAMESS* (C-PCM/GAMESS);[165−167] (4) *Jaguar*'s Poisson−Boltzmann (PB) self-consistent reaction field solver (PB/Jaguar);[168−170] and (5) the Generalized Conductor-like Screening Model (GCOSMO) as implemented in *NWChem* (GCOSMO/NWChem).[171]

For nonelectrostatic contributions, we accept the defaults of these program packages. Thus the *Gaussian* PCM calculations include not only electrostatics but also cavitation, dispersion, and repulsion, as explained in the original references.[154,156] In contrast, the default in *GAMESS*[165−167] and *NWChem*[171] is to only include electrostatics. In *Jaguar*,[170] the default involves only electrostatics for the nonaqueous solvents but both electrostatics and nonelectrostatic terms[169] for the aqueous model.

There are various ways to implement the Conductor-like Screening Model (COSMO) algorithm,[33,160,161,172−175] and the various later implementations should not be confused with the original COSMO method of Klamt and Schüürmann[160] or with the Conductor-like Screening Model for Real Solvents (COSMO-RS)[176,177] that provides a current enhanced version of the COSMO method.[160] Analysis of the performance of the original COSMO method[160] or COSMO-RS[176] or the performance of GCOSMO with the radii optimized by Stefanovich and Truong[174] is beyond the scope of the present study. By GCOSMO/NWChem we refer to the default implementation of the COSMO method in the

Table 12. Mean Unsigned Errors (kcal/mol) in Ionic Solvation Free Energies Calculated Using SM8[a]

		mPW1PW				B3LYP	M06-2X
solute class	*N*	MIDI!6D	6-31G(d)	6-31+G(d)	6-31+G(d,p)	6-31G(d)	6-31G(d)
				Acetonitrile			
H,C,N,O cations[b]	36	6.3	6.4	6.7	6.6	6.5	6.5
S cations[c]	3	15.9	16.1	16.6	16.6	16.2	16.1
all cations	39	7.0	7.2	7.4	7.4	7.2	7.2
H,C,N,O anions[b]	19	3.5	4.3	5.5	5.5	3.9	4.3
F,Cl,Br,S anions[c]	11	2.6	3.2	4.3	4.4	3.0	3.3
all anions	30	3.1	3.9	5.1	5.1	3.6	3.9
all ions	69	5.3	5.8	6.4	6.4	5.6	5.8
				DMSO			
H,C,N,O cations[b]	4	1.8	1.7	2.0	2.2	1.7	1.8
all cations	4	1.8	1.7	2.0	2.2	1.7	1.8
H,C,N,O anions[b]	52	7.8	8.3	8.9	8.8	7.9	8.0
F,Cl,Br,S anions[c]	15	3.2	3.9	4.6	4.7	3.7	4.0
all anions	67	6.8	7.3	7.9	7.9	6.9	7.1
all ions	71	6.5	7.0	7.6	7.6	6.6	6.8
				Methanol			
H,C,N,O cations[b]	26	2.0	2.0	2.3	2.3	2.1	2.1
Cl,Br cations[c]	3	0.3	0.7	1.3	1.7	0.4	0.7
all cations	29	1.8	1.9	2.2	2.3	1.9	2.0
H,C,N,O anions[b]	36	2.1	2.3	3.4	3.5	2.2	2.3
F,Cl,Br anions[c]	15	2.0	2.3	4.3	4.3	2.1	2.4
all anions	51	2.1	2.3	3.6	3.7	2.2	2.4
all ions	80	2.0	2.2	3.1	3.2	2.1	2.2
				Water[d]			
H,C,N,O cations[b]	48	2.8	2.7	3.4	3.4	2.7	2.7
Cl,S cations[c]	4	2.3	2.3	2.7	2.9	2.1	2.4
all cations	52	2.7	2.7	3.3	3.4	2.7	2.7
H,C,N,O anions[b]	43	4.9	4.0	3.5	3.4	4.3	4.3
F,Cl,Br,S anions[c]	17	3.4	3.0	2.5	2.5	3.1	3.1
all anions	60	4.5	3.7	3.2	3.2	4.0	3.9
all ions	112	3.7	3.2	3.3	3.3	3.4	3.4

[a] The solvation free energies obtained using the SM8 model parameters. [b] Ions containing no elements other than H, C, N, or O. [c] Ions containing any of the listed elements in addition to H, C, N, or O. [d] One hundred twelve selectively clustered ions from the SM6 model training set as defined in ref 18.

NWChem-version 4.7 computer package.[171] This implementation uses the atomic radii in Table 5 with the generalized COSMO[161,172,173] (GCOSMO) dielectric screening factor for the conductor-like surface charge. Note that the radii used by default in *NWChem*, as given in Table 5, differ from the values given by the *NWChem* manual. Note also that the default radius of the silicon atom (1.17 Å) is equal to the covalent Si radius given in Table 7-13 in ref 178 that is much smaller than Bondi's van der Waals atomic radius for Si, $R(\mathrm{Si}) = 2.10$ Å.[79] Nevertheless we used the default radii given in Table 5, and we also accepted all other program defaults.

The radii used for IEF-PCM/G03, D-PCM/G98, C-PCM/GAMESS, and PB/Jaguar electrostatic calculations also require further discussion. The PCM methods in *Gaussian* were tested both with atomic radii and group radii; in the latter case one treats certain groups consisting of an atom and its covalently attached hydrogens as a pseudoatom (called a united atom) in forming the cavity. We used three different schemes for assigning atomic or group radii in the IEF-PCM/G03 calculation. First we used the United-Atom Hartree−Fock (UAHF) scheme[179] that is the recommended method for predicting solvation free energies with PCM according to the *Gaussian 03* manual.[142] We also tested IEF-PCM with the UA0 and Bondi schemes. (We note that although UAHF is the recommended scheme in the *Gaussian 03* manual for using with the Hartree−Fock method or DFT, UA0 is the default scheme.) With the UA0 scheme (also called the "united atoms topological model") in *Gaussian 03*, one sometimes needs to use the "sphereonh=N" option to place an individual sphere on a hydrogen that *Gaussian* recognizes as having more than one bond. In particular this is required for the anion of 2-hydroxybenzoic acid and for all of the solute-water clusters used in the set of selectively clustered ions. The D-PCM/G98 model was tested with UAHF group radii. The values of intrinsic atomic radii used for cavity construction with C-PCM/GAMESS and PB/Jaguar are listed in Table 5. The *Jaguar* program uses the atomic radii of Table 5 only for calculation of nonaqueous solvation free energies, whereas for calculation of aqueous solvation energies it employs atomic radii that depend on typing certain functional groups in a solute molecule.[169] The boundary between the solute and solvent used by PB/Jaguar is the so-called molecular surface,[168] which depends on the

Table 13. Errors (kcal/mol) in Ionic Solvation Free Energies Calculated Using Various Solvent Models[a]

					IEF-PCM/G03				D-PCM/G98	C-PCM**	PB**	GCOSMO**
solute class	N	SM8	SM7	SM5.43R	UA0	UAHF	Bondi	UAHF*	UAHF*	GAMESS	Jaguar	NWChem
						Acetonitrile						
MSE (cations)	39	5.1	6.6	4.2	18.7	24.2	12.7	23.8	23.4	14.6	7.3	−3.2
MSE (anions)	30	−3.9	−13.7	−10.1	2.2	1.0	−9.1	−1.0	−1.4	−6.9	−12.4	−22.7
MUE (cations)	39	7.2	6.6	5.6	18.7	24.2	12.7	23.8	23.4	14.6	7.3	4.6
MUE (anions)	30	3.9	13.7	10.1	3.4	2.7	9.1	2.0	2.4	6.9	12.4	22.7
						DMSO						
MSE (cations)	4	−1.3	5.0	0.4	15.7	23.7	6.9	23.5	23.5	15.0	5.0	0.4
MSE (anions)	67	−7.0	−14.3	−10.1	−3.5	−1.5	−12.2	−3.1	−2.4	−6.4	−13.2[b]	−22.1
MUE (cations)	4	1.7	5.0	2.1	15.7	23.7	6.9	23.5	23.5	15.0	5.0	1.4
MUE (anions)	67	7.3	14.3	10.1	5.0	4.9	12.2	5.7	4.5	6.6	13.2[b]	22.1
						Methanol						
MSE (cations)	29	−1.0	−1.2	−4.0	7.5	5.0	0.3	4.3	4.5	8.0	0.1	−10.3
MSE (anions)	51	−1.5	−1.4	3.0	7.3	1.7	−0.4	−1.6	−4.3	6.2	0.5	−9.7
MUE (cations)	29	1.9	1.9	4.5	7.5	5.1	2.0	4.5	4.6	8.0	2.1	10.3
MUE (anions)	51	2.3	2.2	3.6	7.8	3.0	3.3	2.5	4.3	6.4	3.2	10.0
						Water[c]						
MSE (cations)	52	1.0	1.0	−0.2	10.9	5.8	2.8	5.3	5.8	7.7	2.4	−10.8
MSE (anions)	60	1.8	2.0	6.4	13.7	6.2	3.8	4.1	4.5	8.9	3.0	−6.9[d]
MUE (cations)	52	2.7	2.8	5.2	10.9	6.2	3.7	5.7	6.1	7.7	3.1	11.0
MUE (anions)	60	3.7	3.8	6.7	13.7	10.7	5.5	8.9	5.4	8.9	4.8	7.0[d]

[a] N is the number of data in a given solute class. MSE/MUE refers to mean signed/unsigned error. The SM*x* models are described in the text. IEF-PCM/G03 was used with the following methods for assigning atomic or group radii: the united-atom universal force field topological model (UA0), the united-atom Hartree−Fock model (UAHF), and the Bondi atomic radii (Bondi). D-PCM/G98 is the dielectric version of PCM implemented in *Gaussian 98* with using the UAHF radii. The calculations were performed at the mPW1PW/6-31G(d) level of theory, except for the calculations marked by the asterisks: they used the Hartree−Fock method (*) and B3LYP (**). [b] No data were obtained for 3-hydroxybenzoic acid (anion). The total count is reduced to $N - 1$. [c] One hundred twelve selectively clustered ions from the SM6 model training set as defined in ref 18. [d] No data were obtained for hydroperoxyl radical (anion). The total count is reduced to $N - 1$.

solvent probe radius, which was set, following the instructions in the manual,[170] to 1.4 Å for water and calculated from the solvent density by assuming a packing fraction of 0.5 for other solvents. The values listed in Table 5 for C-PCM/GAMESS are the radii used in the C-PCM/GAMESS electrostatic calculations. Note that the user would have to input values a factor of 1.2 smaller than those in the table since in the cavity construction algorithm, *GAMESS* multiplies the input values by 1.2.

To test all of the continuum models we used the same data on 332 ions in acetonitrile, DMSO, methanol, and water as described above. However, the test of neutrals was performed only for those 17 solvents (including acetonitrile, DMSO, and water) that are available for IEF-PCM/G03 in *Gaussian 03* (actually *Gaussian 03* supports 21 solvents, but we have neutral data in only 17 of them; see Table 1). Note that, for example, methanol is available in *Gaussian 03*, but we have no neutral data for this solvent. All the calculations for these comparisons were carried out with the 6-31G(d) basis. The mPW1PW density functional[19] was used with the SM8, SM7, SM5.43, and IEF-PCM/G03 models. However mPW1PW was not available with C-PCM/GAMESS, PB/Jaguar, and GCOSMO/NWChem. In these three cases we employed B3LYP[146−148] instead of mPW1PW. The IEF-PCM/G03 and D-PCM/G98 models with the UAHF scheme for assigning atomic radii were also tested using the Hartree−Fock method because the parameters contained in the UAHF model were originally optimized for the HF/6-31G(d) level of theory.[179]

Before turning to the results, we comment on the standard states used by the various program packages. All programs tested use a gas-phase standard state of 1 mol/L, and all results presented in the paper use this standard state.

Table 13 shows the mean signed and unsigned errors between calculated and experimental solvation free energies of anions and cations for each of the implicit solvent models mentioned above, in acetonitrile, DMSO, methanol, and water (see the Supporting Information for more details). For ions in methanol and water, SM8 and SM7 give nearly identical average errors, and both are more accurate than the other models tested. PB/Jaguar also gives quite accurate (but still inferior to SM8) predictions for ions in methanol and water. For acetonitrile and DMSO, the performance of each of these models is highly dependent on whether the solute is an anion or cation. For SM7 and SM5.43, the calculated solvation free energies in acetonitrile and DMSO are significantly more accurate for cations than for anions. The opposite occurs when IEF-PCM/G03 or D-PCM/G98 (that yields similar results to those from IEF-PCM/G03) is used with either UA0 or UAHF radii; any of these models is able to predict solvation free energies of anions in acetonitrile and DMSO fairly accurately, whereas for cations, any of these PCM models gives mean unsigned errors of over 15 kcal/mol! One can observe that in many cases the Bondi scheme in conjunction with IEF-PCM/G03 can provide more accurate predictions than the united atom models. The GCOSMO/NWChem model gives errors for cations in acetonitrile and DMSO that are smaller than

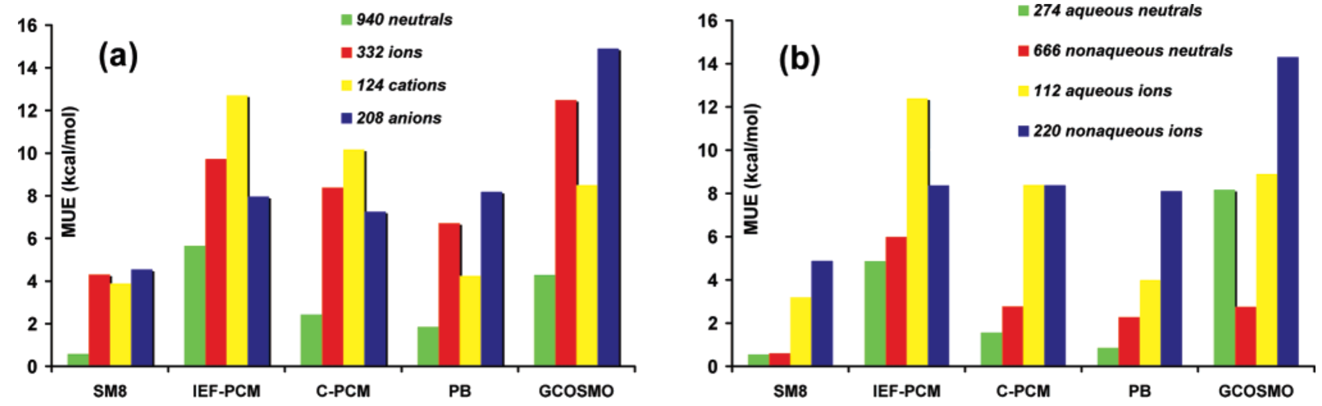


Figure 2. Mean unsigned errors (MUEs) in solvation free energies of neutral and ionic solutes calculated using SM8 and other continuum models including IEF-PCM/G03 with the UA0 model for assigning atomic or group radii, C-PCM/GAMESS, PB/Jaguar, and GCOSMO/NWChem. B3LYP was used with *GAMESS*, *Jaguar*, and *NWChem*, because mPW1PW was unavailable. The calculation was done only for 18 solvents, which are available for IEF-PCM/G03, including acetonitrile (ions and neutral solutes), DMSO (ions and neutral solutes), methanol (only ions), water (ions and neutral solutes), and an additional 14 organic solvents from the SM8 neutral training set (Table 1; see also footnotes in Tables 13 and 14). (a) MUEs are given for ions and neutrals in all of the 18 solvents. (b) MUEs for solutes in aqueous solutions are compared to MUEs for solutes in nonaqueous solutions.

Table 14. Errors (kcal/mol) in Solvation Free Energies Calculated Using Various Solvent Models[a]

method	aqueous neutrals[b]		organic neutrals[c]		ions[d]	
	MSE	MUE	MSE	MUE	MSE	MUE
SM8	−0.06	0.55	−0.02	0.61	−1.02	4.31
SM7	−0.07	0.53	−0.11	0.59	−3.09	6.59
SM5.43R	0.00	0.55	0.10	0.67	−1.18	6.60
IEF-PCM/UA0	4.86	4.87	5.94	5.99	7.45	9.73
IEF-PCM/UAHF	0.61	1.18	3.88	3.94	5.63	8.15
C-PCM/GAMESS	−0.65	1.57	2.62	2.78	4.45	8.39
PB/Jaguar	0.22	0.86	1.69	2.28	−1.86[e]	6.72[e]
GCOSMO	−8.17[f]	8.17[f]	−2.12	2.76	−12.21[g]	12.49[g]
3PM	0.00	2.65	0.00	1.49	0.00	8.60

[a] MSE/MUE refers to mean signed/unsigned error. IEF-PCM/G03 was used with the united-atom universal force field topological model (UA0) and the united-atom Hartree−Fock model (UAHF) for assigning group radii. The SM*x* and IEF-PCM calculations were performed at the mPW1PW/6-31G(d) level; the other calculations were performed at the B3LYP/6-31G(d) level. 3PM refers to the three-parameter model described in the text. [b] Two hundred seventy-four neutral data, unless indicated otherwise. [c] Six hundred sixty-six neutral data in 16 nonaqueous solvents available with IEF-PCM/G03. [d] Three hundred thirty-two data (unless indicated otherwise) for ions in acetonitrile, DMSO, methanol, and water. [e] No data were obtained for 3-hydroxybenzoic acid (anion) in DMSO. The total count is reduced to 331. [f] No data were obtained for 11 phosphorus-containing compounds and tetramethylsilane. The total count is reduced to 262. [g] No data were obtained for hydroperoxyl radical (anion) in water. The total count is reduced to 331.

Table 15. Mean Unsigned Errors (kcal/mol) in Solvation Free Energies Calculated with the mPW1PW Density Functional and with Class II and Class IV Partial Atomic Charges[a]

model	basis	neutral data			ions[e]
		aqueous[b]	organic[c]	transfer[d]	
SM8-LPA	MIDI!6D	1.26	0.83	0.81	4.21
SM8-LPA	6-31G(d)	1.93	1.40	0.78	5.60
SM8-RLPA	6-31+G(d)	2.18	1.51	0.93	6.33
SM8-RLPA	6-31+G(d,p)	1.39	0.89	0.90	5.54
SM8-CM4	MIDI!6D	0.63	0.60	0.74	4.20
SM8-CM4	6-31G(d)	0.55	0.57	0.78	4.31
SM8-CM4	6-31+G(d)	0.66	0.61	0.70	4.79
SM8-CM4	6-31+G(d,p)	0.67	0.62	0.66	4.81

[a] LPA denotes Löwdin population analysis, and RLPA denotes redistributed Löwdin population analysis; population analysis yields class II charges as defined in ref 41. CM4 denotes charge model 4, which yields class IV charges, also defined in ref 41. [b] Two hundred seventy-four data. [c] Two thousand seventy-two data. [d] One hundred forty-three data. [e] Three hundred thirty-two data.

those obtained by SM8, but the errors for anions exceed 20 kcal/mol. Similarly, SM7 and SM5.43 tend to overestimate the solvation free energies of anions in acetonitrile and DMSO. Thus SM8 has a much better average performance in acetonitrile than any other model in Table 13.

The errors in predicting the solvation free energies of neutral solutes by different models are listed in Table 14. We limited our calculations on neutral solutes only to testing the following models at the DFT/6-31G(d) level of theory: IEF-PCM/G03/UA0, IEF-PCM/G03/UAHF, C-PCM/GAMESS, PB/Jaguar, and GCOSMO/NWChem. Again, the SM*x* models provide much more accurate predictions of experimental free energies of solvation than any of these models. In particular, GCOSMO/NWChem gives an unacceptably large overestimate of aqueous neutral data (up to 8 kcal/mol on average). The most accurate non-SM*x* model tested is PB/Jaguar. However even in this case the error for the neutral solutes in organic solvents is 4.5 times larger than obtained with the SM8 model. Figure 2 complements the analysis of various continuum solvation models presented in Tables 13 and 14 and shows again that the newly developed SM8 solvation model significantly outperforms the most popular implicit solvent models that are widely used in modeling condensed media.

We close this section by evaluating one more solvation model, which we call the three-parameter model (3PM). The 3PM predicts that all neutral solvation free energies in aqueous solution are −2.99 kcal/mol, all neutral solvation

Table 16. Mean Unsigned Errors (kcal/mol) in Solvation Free Energies Calculated Using SM*x* and Non-SM*x* Implicit Solvent Models[a]

					IEF-PCM/G03							
solute class	*N*	SM8	SM7	SM5.43R	UA0	UAHF	Bondi	UAHF*	D-PCM/G98 UAHF*	C-PCM** GAMESS	PB** Jaguar	GCOSMO** NWChem
all neutrals	940	0.59	0.57	0.64	5.66	3.14				2.43	1.86	4.29[b]
all ions	332	4.31	6.59	6.60	9.73	8.15	7.08	7.67	7.15	8.39	6.72[c]	12.49[d]
all cations	124	3.90	10.19	5.06	12.71	12.17	6.24	11.69	11.75	10.18	4.25	8.51
all anions	208	4.55	8.22	7.53	7.97	5.79	7.64	5.30	4.41	7.26	8.19[c]	14.90[d]
aqueous neutrals	274	0.55	0.53	0.55	4.87	1.18				1.57	0.86	8.17[b]
nonaqeous neutrals	666	0.61	0.59	0.67	5.99	3.94				2.78	2.28	2.76
aqueous ions	112	3.24	3.31	6.00	12.43	8.61	4.64	7.43	5.73	8.36	4.03	8.85[d]
nonaqueous ions	220	4.88	8.26	6.90	8.37	7.93	8.34	7.81	7.89	8.38	8.11[c]	14.31

[a] *N* is the number of data in a given solute class. The SM*x* models are described in the text. IEF-PCM/G03 was used with the following methods for assigning atomic or group radii: the united-atom universal force field topological model (UA0), the united-atom Hartree−Fock model (UAHF), and the Bondi atomic radii (Bondi). D-PCM/G98 is the dielectric version of PCM implemented in *Gaussian 98* with using the UAHF radii. The calculations were performed at the mPW1PW/6-31G(d) level of theory, except for the calculations marked by the asterisks: they used the Hartree−Fock method (*) and B3LYP (**). [b] No data were obtained for 11 phosphorus-containing compounds and tetramethylsilane. The total count is reduced to *N* − 12. [c] No data were obtained for 3-hydroxybenzoic acid (anion). The total count is reduced to *N* − 1. [d] No data were obtained for hydroperoxyl radical (anion). The total count is reduced to *N* − 1.

free energies in organic solvents are −5.38 kcal/mol, and all ionic solvation free energies are −65.0 kcal/mol; these are the average experimental values averaged over 274, 666, 332 data, respectively (Table 14). The mean unsigned error of the 3PM is 2.7 kcal/mol for neutrals in water, 1.5 kcal/mol for neutrals in nonaqueous solvents, and 8.6 kcal/mol for ions. Comparison to the MUE column in Table 14 shows, somewhat disappointingly, that the non-SM*x* models outperform the 3PM in only 6 out of 15 possible cases.

9. Using Other Charge Models

Although in this work the performance of SM8 has only been illustrated for six electronic structure levels, experience with SM6[18] and the MPW*X* series[34,39] (where *X* denotes a fraction of Hartree−Fock exchange) shows that the SM*x* models can be used with any density functional or with the Hartree−Fock approximation as long as one uses class IV charges. SM8 can also be used with other kinds of charges. One can expect the most reliable results if the user validates that the charge model chosen gives partial atomic charges that are reasonably similar to CM4 charges. Table 15 and Tables S8−S11 in part II of the Supporting Information show examples of using SM8 with other charge models; the results are less accurate than with CM4 charges, but even with the less accurate class II charges based on Löwdin or redistributed population analysis the average errors of SM8 are smaller than the errors of many non-SM*x* model listed in Table 14.

10. Summary and Concluding Remarks

Using experimental pK_a values in acetonitrile, DMSO, and methanol, experimental gas-phase acidities, accurate values for the absolute solvation free energy of the proton, and solvation free energies of neutral solutes that were computed using SM7, a data set of single-ion solvation free energies in the three solvents above was assembled. Using these data and data assembled previously for solvation free energies of ions in water, solvation free energies of neutrals in water and 90 nonaqueous solvents, and transfer free energies of neutrals from water to 15 nonaqueous solvents, a new universal implicit solvent model called SM8 has been developed for predicting solvation free energies of neutral and ionic solutes in water and in nonaqueous solvents. For nonaqueous solvents, SM8 uses a small set of solvent descriptors that characterize the properties of the solvent.

Like several previous universal SM*x* models, SM8 gives solvation free energies of neutral solutes that are typically within ∼0.6 kcal/mol of the experimental value, despite using fewer parameters than the earlier models. For ionic solvation the present models provide considerable improvement over all previous methods. Since new ionic data are used to obtain a physical partitioning of the solvation energy into bulk electrostatic and non-bulk-electrostatic components, and self-consistently polarized charge distributions are used to calculate the bulk electrostatic contributions, we expect that not only the solvation free energies but also the charge distributions and properties of the dissolved molecules should be well represented. Thus the present model can be used with confidence to calculate partition coefficients (e.g., Henry's Law constants, octanol/water partition coefficients, etc),[54] solubilities,[180] vapor pressures,[34,181] liquid-phase geometries of neutral and charged species (including transition state species),[55] and, when combined with gas-phase acid dissociation free energies, liquid-phase pK_a constants.[182] These properties can be calculated in any solvent that can be characterized by the solvent descriptors used by the SM8 model.

The key descriptors used by SM8 such as dielectric constant, Abraham's hydrogen bond acidity and basicity parameters, refractive index, and macroscopic surface tension at an air/solvent interface are tabulated in the literature for almost all possible organic solvents, and that is the primary sense in which our model is universal. However the applicability of SM8 as presented here is still limited to room temperature. (An extension to variable temperature for aqueous solutions is essentially complete,[183] and it will be submitted soon.). Although the SM8 training set includes solvation free energies of solutes only in pure aqueous and organic solvents but not in mixtures of solvents, the SM8 model can also be applied to complex "solvents" such as membranes,[184] interfaces,[185] or mixtures, provided effective values for the solvent descriptors are available or can be obtained.

Table 16 summarizes the comparison of the present model to the previously published universal solvation model from our group (SM5.43), to an unpublished model based on using atomic radii optimized for water in all solvents (SM7), and to several solvation models from popular computer packages. For neutral solutes in aqueous solution, the mean unsigned error of SM8 is 0.55 kcal/mol, whereas the errors in the five non-SM*x* models we tested are 0.9−8.2 kcal/mol. For solvation of neutrals in nonaqueous solvents the mean unsigned error of SM8 increases to 0.61 kcal/mol, whereas the errors in the five non-SM*x* model we tested are 2.3−6.0 kcal/mol. For ions, SM8 gives a mean unsigned error of 4.3 kcal/mol, whereas the errors in the eight non-SM*x* models we tested are 6.7−12.5 kcal/mol.

Acknowledgment. This work was supported by the National Institutes of Health training grant for Neurophysical-computational Sciences, by the U.S. Army Research Office under the Multidisciplinary Research Program of the University Research Initiative through grant no. DAAD19-02-1-0176, by the National Science Foundation (grant CHE06-10183 and grant CHE07-04974), and by the Office of Naval Research under grant no. 00014-05-01-0538.

Supporting Information Available: Two thousand three hundred forty-six reference solvation free energies and 143 reference transfer energies for neutral solutes in the SM8 training set; reference free energies for 112 selectively clustered ions in water; and 220 unclustered ions and 21 ionic clusters in acetonitrile, DMSO, and methanol (part I) and reference pK_a constants, gas-phase acidity and basicity values, and solvation free energies of neutral species used for evaluation of the reference solvation free energies of the corresponding ions in acetonitrile, DMSO, and methanol; SM7 and SM8 calculated free energies of nonaqueous ions; errors in solvation and transfer free energies calculated by SM8 using the class IV CM4 charges and the class II charges based on Löwdin or redistributed Löwdin population analyses; errors in solvation free energies of neutral and ionic solutes calculated using SM*x* and non-SM*x* implicit solvent models, by solvent and by solute class; and the Cartesian coordinates corresponding to the B97-1/MG3S optimized global minima for nonaqueous clusters (part II). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Floris, F.; Tomasi, J. *J. Comput. Chem*. **1989**, *10*, 616.

(2) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc*. **1990**, *112*, 6127.

(3) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc*. **1991**, *113*, 8305.

(4) Tomasi, J.; Persico, M. *Chem. Rev*. **1994**, *94*, 2027.

(5) Cramer, C. J.; Truhlar, D. G. *Chem. Rev*. **1999**, *99*, 2161.

(6) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev*. **2005**, *105*, 2999.

(7) Curutchet, C.; Cramer, C. J.; Truhlar, D. G.; Ruiz-López, M. F.; Rinaldi, D.; Orozco, M.; Luque, F. J. *J. Comput. Chem*. **2003**, *24*, 284.

(8) Mark, A. E.; van Gunsteren, W. F. *J. Mol. Biol*. **1994**, *240*, 167.

(9) Smith, P. E.; van Gunsteren, W. F. *J. Phys. Chem*. **1994**, *98*, 13735.

(10) Giesen, D. J.; Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. In *Computational Thermochemistry*; Irikura, K., Frurip, D. J., Eds.; ACS Symposium Series 677; American Chemical Society: Washington, DC, 1998; p 285.

(11) Hawkins, G. D.; Zhu, T.; Li, J.; Chambers, C. C.; Giesen, D. J.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. In *Combined Quantum Mechanical and Molecular Mechanical Methods*; Gao, J., Thompson, M. A., Eds.; ACS Symposium Series 712; American Chemical Society: Washington, DC, 1998; p 201.

(12) Cramer, C. J.; Truhlar, D. G. In *Trends and Perspectives in Modern Computational Science*; Maroulis, G., Simos, T. E., Eds.; Lecture Series on Computer and Computational Sciences 6; Brill/VSP: Leiden, 2006; p 112.

(13) Klotz, I. M.; Rosenberg, R. M. *Chemical Thermodynamics: Basic Theory and Methods*, 5th ed.; Wiley: New York, 1994; p 459.

(14) Lewis, G. N.; Randall, M.; Pitzer, K. S.; Brewer, L. *Thermodynamics*, 2nd ed.; McGraw-Hill: New York, 1961; p 399.

(15) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V. *J. Phys. Chem. A* **1998**, *102*, 7787.

(16) Camaioni, D. M.; Schwerdtfeger, C. A. *J. Phys. Chem. A* **2005**, *109*, 10795.

(17) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066.

(18) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput*. **2005**, *1*, 1133.

(19) Adamo, C.; Barone, V. *J. Chem. Phys*. **1998**, *108*, 664.

(20) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(21) Fuoss, R. M.; Accascina, F. *Electrolytic Conductance*; Interscience: New York, 1959.

(22) Szwarc, M. *Acc. Chem. Res*. **1969**, *2*, 87.

(23) Mayer, U. *Coord. Chem. Rev*. **1976**, *21*, 159.

(24) Krell, M.; Symons, M. C. R.; Barthel, J. *J. Chem. Soc., Faraday Trans. 1* **1987**, *83*, 3419.

(25) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 408.

(26) (a) Kelly, C. P. Ph.D. Thesis, University of Minnesota: Minneapolis, 2007. (b) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Unpublished.

(27) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257.

(28) Giesen, D. J.; Storer, J. W.; Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc*. **1995**, *117*, 1057.

(29) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem*. **1996**, *100*, 16385.

(30) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys*. **1998**, *109*, 9117; errata: **1999**, *111*, 5624 and **2000**, *113*, 3930.

(31) Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett*. **1998**, *288*, 293.

(32) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc*. **1999**, *103*, *9*.

(33) Dolney, D. M.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem*. **2000**, *21*, 340.

(34) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532.

(35) Giesen, D. J.; Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1997**, *101*, 2061.

(36) Giesen, D. J.; Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1997**, *98*, 85.

(37) Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1998**, *63*, 4305.

(38) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 2178.

(39) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2005**, *113*, 107.

(40) Mulliken, R. S. *J. Chem. Phys*. **1955**, *23*, 1833.

(41) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

(42) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.

(43) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 10707.

(44) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365.

(45) Golebiewski, A.; Rzeszowska, E. *Acta Phys. Pol., A* **1974**, *45*, 563.

(46) Baker, J. *Theor. Chim. Acta* **1985**, *68*, 221.

(47) Kar, T.; Sannigrahi, A. B.; Mukherjee, D. C. *J. Mol. Struct.: THEOCHEM* **1987**, *153*, 93.

(48) Kalinowski, J. A.; Lesyng, B.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2545.

(49) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Status Solidi B* **2000**, *217*, 41.

(50) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(51) Li, J.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc*. **1998**, *99*, 192.

(52) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.*, **2007**, *6*, 2046−2054.

(53) Ben-Naim, A. *Solvation Thermodynamics*; Plenum: New York, 1987.

(54) Cramer, C. J.; Truhlar, D. G. In *Free Energy Calculations in Rational Drug Design*; Reddy, M. R., Erion, M. D., Eds.; Kluwer/Plenum: New York, 2001; p 63.

(55) Zhu, T.; Li, J.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 5503.

(56) Tapia, O. In *Quantum Theory of Chemical Reactions*; Daudel, R., Pullman, A., Salem, L., Veillard, A., Eds.; Reidel: Dordrecht, 1980; p 25ff.

(57) Cramer, C. J.; Truhlar, D. G. In *Solvent Effects and Chemical Reactivity*; Tapia, O., Bertrán, J., Eds.; Kluwer: Dordrecht, 1996; p 1.

(58) Hoijtink, G. J.; de Boer, E.; van der Meij, P. H.; Weijland, W. P. *Recl. Trav. Chim. Pays-Bas Belg.* **1956**, *75*, 487.

(59) Peradejordi, F. *Cahiers Phys.* **1963**, *17*, 393.

(60) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(61) Cramer, C. J.; Truhlar, D. G. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 6, p 1.

(62) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *PhysChemComm* **2002**, *5*, 117.

(63) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.

(64) Mayer, I. *Chem. Phys. Lett.* **1985**, *117*, 396.

(65) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 73.

(66) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291.

(67) Brom, J. M.; Schmitz, B. J.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 6483.

(68) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2005**, *113*, 133.

(69) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(70) Li, J.; Williams, B.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 724.

(71) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264.

(72) Martin, F.; Zipse, H. *J. Comput. Chem.* **2004**, *26*, 97.

(73) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824.

(74) Winget, P.; Cramer, C. J.; Truhlar, D. G. *Environ. Sci. Technol.* **2000**, *34*, 4733.

(75) Winget, P.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 5160.

(76) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1995**, *16*, 422.

(77) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(78) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

(79) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(80) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. *J. Chem. Soc., Perkin Trans. 2* **1989**, 699.

(81) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.

(82) Abraham, M. H. *J. Phys. Org. Chem.* **1993**, *6*, 660.

(83) Abraham, M. H. In *Quantitative Treatment of Solute/Solvent Interactions*; Theoretical and Computational Chemistry Series Vol. 1; Politzer, P., Murray, J. S., Eds.; Elsevier: Amsterdam, 1994; p 83.

(84) *Physical/Chemical Property Database (PHYSPROP);* SRC Environmental Science Center: Syracuse, NY, 1994.

(85) Leo, A. J. *Masterfile from MedChem Software*; BioByte Corp.: Claremont, CA, 1994.

(86) Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. G.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database version 3.0*; University of Minnesota: Minneapolis, MN 55455-0431, 2007.

(87) Hunter, E. P. L.; Lias, S. G. *J. Phys. Chem. Ref. Data* **1998**, *27*, 413.

(88) Lias, S. G.; Bartness, J. E.; Liebman, J. F.; Holmes, J. L.; Levin, R. D.; Mallard, W. G. Ion Energetics Data. In *NIST Chemistry WebBook*; NIST Standard Reference Database Number 69; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, March 2003.

(89) Chantooni, M. K., Jr.; Kolthoff, I. M. *J. Am. Chem. Soc.* **1968**, *90*, 3005.

(90) Coetzee, J. F.; Padmanabhan, G. R. *J. Am. Chem. Soc.* **1965**, *87*, 5005.

(91) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1973**, *95*, 8539.

(92) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1973**, *95*, 4768.

(93) Kolthoff, I. M.; Chantooni, M. K., Jr.; Bhowmik, S. *Anal. Chem.* **1967**, *39*, 1627.

(94) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1968**, *90*, 3320.

(95) Beltrame, P.; Gelli, G.; Loi, A. *Gazz. Chim. Ital.* **1980**, *110*, 491.

(96) Chantooni, M. K., Jr.; Kolthoff, I. M. *J. Phys. Chem.* **1974**, *78*, 839.

(97) Coetzee, J. F.; Padmanabhan, G. R. *J. Phys. Chem.* **1965**, *69*, 3193.

(98) Jasinski, T.; El-Harakany, A. A.; Halaka, F. G.; Sadek, H. *Croat. Chem. Acta* **1978**, *51*, 1.

(99) Kolthoff, I. M.; Chantooni, M. K., Jr.; Bhowmik, S. *J. Am. Chem. Soc.* **1966**, *88*, 5430.

(100) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Phys. Chem.* **1966**, *70*, 856.

(101) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1970**, *92*, 7025.

(102) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1971**, *93*, 3843.

(103) Chantooni, M. K., Jr.; Kolthoff, I. M. *J. Phys. Chem.* **1976**, *80*, 1306.

(104) Chantooni, M. K., Jr.; Kolthoff, I. M. *J. Phys. Chem.* **1975**, *79*, 1176.

(105) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1969**, *91*, 4621.

(106) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1975**, *97*, 1376.

(107) Kolthoff, I. M.; Chantooni, M. K., Jr.; Bhowmik, S. *J. Am. Chem. Soc.* **1968**, *90*, 23.

(108) Ludwig, M.; Pytela, O.; Vecera, M. *Collect. Czech. Chem. Commun.* **1984**, *49*, 2593.

(109) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1976**, *98*, 5063.

(110) Chantooni, M. K., Jr.; Kolthoff, I. M. *Anal. Chem.* **1979**, *51*, 133.

(111) Kolthoff, I. M.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1965**, *87*, 4428.

(112) Kolthoff, I. M.; Bruckenstein, S.; Chantooni, M. K., Jr. *J. Am. Chem. Soc.* **1961**, *83*, 3927.

(113) Kolthoff, I. M.; Chantooni, M. K., Jr.; Bhowmik, S. *Anal. Chem.* **1967**, *39*, 315.

(114) Bordwell, F. G. *Acc. Chem. Res.* **1988**, *21*, 456.

(115) Jasinski, T.; Stefaniuk, K. *Chem. Anal. (Warsaw)* **1965**, *10*, 211.

(116) Ritchie, C. D.; Uschold, R. E. *J. Am. Chem. Soc.* **1968**, *90*, 2821.

(117) Matthews, W. S.; Bares, J. E.; Bartmess, J. E.; Bordwell, F. G.; Cornforth, F. J.; Drucker, G. E.; Margolin, Z.; McCallum, R. J.; McCollum, G. J.; Vanier, N. R. *J. Am. Chem. Soc.* **1975**, *97*, 7006.

(118) Bordwell, F. G.; Algrim, D. J. *J. Am. Chem. Soc.* **1988**, *110*, 2964.

(119) Clare, B. W.; Cook, D.; Ko, E. C. F.; Mac, Y. C.; Parker, A. J. *J. Am. Chem. Soc.* **1966**, *88*, 1911.

(120) Courtot-Coupez, J.; Le Demezet, M. *Bull. Soc. Chim. Fr.* **1969**, 1033.

(121) Kolthoff, I. M.; Reddy, T. B. *Inorg. Chem.* **1962**, *1*, 189.

(122) Olmstead, W. N.; Margolin, Z.; Bordwell, F. G. *J. Org. Chem.* **1980**, *45*, 3295.

(123) Ritchie, C. D.; Uschold, R. E. *J. Am. Chem. Soc.* **1967**, *89*, 1721.

(124) Bordwell, F. G.; McCallum, R. J.; Olmstead, W. N. *J. Org. Chem.* **1984**, *49*, 1424.

(125) Rived, F.; Rosés, M.; Bosch, E. *Anal. Chim. Acta* **1998**, *374*, 309.

(126) Rochester, C. H. *J. Chem. Soc. B* **1967**, 33.

(127) Ritchie, C. D.; Heffley, P. D. *J. Am. Chem. Soc.* **1965**, *87*, 5402.

(128) Juillard, J. *Bull. Soc. Chim. Fr.* **1966**, 1727.

(129) Juillard, J.; Dondon, M.-L. *Bull. Soc. Chim. Fr.* **1963**, 2535.

(130) Konovalov, O. M. *Zh. Fiz. Khim.* **1965**, *39*, 693.

(131) Bolton, P. D.; Rochester, C. H.; Rossall, B. *Trans. Faraday Soc.* **1970**, *66*, 1348.

(132) Juillard, J. Ph.D. Thesis, University of Clermont-Ferrand: Clermont-Ferrand, France, 1967.

(133) Shedlovsky, T.; Kay, R. L. *J. Phys. Chem.* **1956**, *60*, 151.

(134) Kolthoff, I. M.; Lingane, J. J.; Larson, W. D. *J. Am. Chem. Soc.* **1938**, *60*, 2512.

(135) Leung, C. S.; Grunwald, E. *J. Phys. Chem.* **1970**, *74*, 696.

(136) Charlot, G.; Tremillon, B. *Chemical Reactions in Solvents and Melts*, 1st Engl. ed.; Pergamon Press: New York, 1969; p 278.

(137) Izmailov, N. A.; Chernyi, V. S.; Spivak, L. L. *Zh. Fiz. Khim.* **1963**, *37*, 822.

(138) Mason, R. B.; Kilpatrick, M. *J. Am. Chem. Soc.* **1937**, *59*, 572.

(139) Rochester, C. H. *Trans. Faraday Soc.* **1966**, *62*, 355.

(140) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(141) Fast, P. L.; Sánchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.

(142) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revisions C.01, C.02, and D.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(143) Meot-Ner, M. M.; Lias, S. G. Binding Energies Between Ions and Molecules, and the Thermochemistry of Cluster Ions. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, March 2003.

(144) Thompson, J. D.; Winget, P.; Truhlar, D. G. *PhysChemComm* **2001**, *16*, 1.

(145) Olson, R. M.; Marenich, A. V.; Chamberlin, A. C.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Hawkins, G. D.; Winget, P.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J. *MN-GSM, version 2007-beta*; University of Minnesota: Minneapolis, MN, 55455-0431, 2007.

(146) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(147) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(148) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(149) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc*. In press.

(150) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(151) Miertuš, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239.

(152) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(153) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.

(154) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253.

(155) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43.

(156) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1996**, *255*, 327.

(157) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J. W.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A.11*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(158) Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 71.

(159) Li, H.; Jensen, J. H. *J. Comput. Chem.* **2004**, *25*, 1449.

(160) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(161) Truong, T. N.; Stefanovich, E. V. *Chem. Phys. Lett.* **1995**, *240*, 253.

(162) Baldridge, K.; Klamt, A. *J. Chem. Phys.* **1997**, *106*, 6622.

(163) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(164) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.

(165) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem.* **1993**, *14*, 1347.

(166) Gordon, M. S.; Schmidt, M. W. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 1167.

(167) *GAMESS computer package, version 7 SEP 2006 (R6)*; Iowa State University: Ames, IA, 2006. http://www.msg.ameslab.gov/GAMESS/GAMESS.html (accessed Feb 2007).

(168) Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M.; Goddard, W. A., III; Honig, B. *J. Am. Chem. Soc.* **1994**, *116*, 11875.

(169) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775.

(170) *Jaguar 6.5*, *Release 112*; Schrödinger, Inc.: Portland, OR, 2005.

(171) Bylaska, E. J.; de Jong, W. A.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Aprà, E.; Windus, T. L.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, R. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.7*; Pacific Northwest National Laboratory: Richland, WA, 2006.

(172) Truong, T. N.; Stefanovich, E. V. *J. Phys. Chem.* **1995**, *99*, 14700.

(173) Truong, T. N.; Stefanovich, E. V. *J. Chem. Phys.* **1995**, *103*, 3709.

(174) Stefanovich, E. V.; Truong, T. N. *Chem. Phys. Lett.* **1995**, *244*, 65.

(175) Truong, T. N.; Nguyen, U. N.; Stefanovich, E. V. *Int. J. Quantum Chem.* **1996**, *60*, 1615.

(176) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224.

(177) Klamt, A.; Eckert, F. *Fluid Phase Equilib.* **2000**, *172*, 43.

(178) Pauling, L. *The Nature of The Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, NY, 1960.

(179) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(180) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **2003**, *119*, 1661.

(181) Winget, P.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2000**, *104*, 4726.

(182) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 2493.

(183) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. To be published.

(184) Chambers, C. C.; Giesen, D. J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G.; Vaes, W. H. J. In *Rational Drug Design*; Truhlar, D. G., Howe, W. J., Hopfinger, A. J., Blaney, J. M., Dammkoehler, R. A., Eds.; Springer: New York, 1999; p 51.

(185) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2004**, *108*, 12882.

CT7001418

JCTC Journal of Chemical Theory and Computation

Polarizable Force Fields: History, Test Cases, and Prospects

Arieh Warshel,* Mitsunori Kato, and Andrei V. Pisliakov

*University of Southern California, 418 SGM Building, 3620 McClintock Avenue,
Los Angeles, California 90089-1062*

Received May 28, 2007

Abstract: A consistent treatment of electrostatic energies is arguably the most important requirement for the realistic modeling of biological systems. An important part of electrostatic modeling is the ability to account for the polarizability of the simulated system. This can be done both macroscopically and microscopically, but the use of macroscopic models may lead to conceptual traps, which do not exist in the microscopic treatments. The present work describes the development of microscopic polarizable force fields starting with the introduction of these powerful tools and following some of the subsequent developments in the field. Special effort has been made to review a wide range of applications and emphasize cases when the use of polarizable force fields is important. Finally, a brief perspective is given on the future of this rapidly growing field.

1. The Emergence of Polarizable Force Fields

Electrostatic effects, and solvation effects in particular, play a major role in determining the energetics and dynamics of charge transfer and related processes in solution (e.g. refs 1−3). Such effects also play a crucial role in determining the function of macromolecules (e.g. refs 4−13). Thus, the ability to quantify electrostatic interactions is essential for the quantitative description both of processes in solution and for structure−function correlation studies of proteins (e.g. ref 5). However, accomplishing this task has been quite challenging for both microscopic and macroscopic approaches (for reviews see e.g. refs 6−13).

Here, we will focus on one crucial aspect of the microscopic modeling of electrostatic energies, namely, the treatment of electronic polarizability. We will start by presenting some of the historical background of this rapidly growing field. We will then move to key examples and finally to a discussion of the prospects of the field.

The idea that matter can be represented by induced dipoles goes back to the early literature on electrostatics. However, the rationalization of the proper description of microscopic polarization and the replacement of electronic polarization by classical polarizable induced diploes is more recent. In fact, most textbooks treat the energetics of polarizable matter in a macroscopic way whose relationship to the microscopic world is not clear. For example, according to the well-established macroscopic theory (e.g. refs 14 and 15), one can express the energy of a polarizable volume element by

$$W = -\frac{1}{2}\mathbf{P}\mathbf{E}_0 = -\frac{1}{2}\alpha\mathbf{E}_0{}^2 \tag{1a}$$

where **P** is the induced polarization, $\mathbf{E}_0$ is the macroscopic field, and α is the corresponding polarizability. However, the validity of such a treatment in microscopic systems may look less clear to a chemist who comes from the molecular atomistic background, where it is known that the interaction between a charge and the induced dipole of a single atom in a collection of atoms is given by $W = -\boldsymbol{\mu}\boldsymbol{\xi}_0 = -\alpha\boldsymbol{\xi}_0{}^2$ (where ξ is the microscopic field on the atom). Thus, the origin of the (1/2) factor is not obvious. This point can be verified by trying to ask a physics or electrical engineering professor how the factor 1/2 in microscopic systems is obtained. The typical answer usually involves the well-known $\int Q\mathrm{d}Q = Q_0{}^2/2$ macroscopic integral,[15] or arguments about the linear response nature of matter, but it will not satisfy those who

* Corresponding author e-mail: warshel@usc.edu.

10.1021/ct700127w CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/25/2007

insist on a molecular explanation. In fact, the microscopic relationship for a collection of charges and induced dipoles is[16]

$$W = -\sum_{i,j} Q_i(\boldsymbol{\mu}_{\mathbf{j}}\cdot\mathbf{r}_{\mathbf{ij}})/{r_{ij}}^3 + \sum_{j>j'} \boldsymbol{\mu}_{\mathbf{j}}[\nabla(\boldsymbol{\mu}_{\mathbf{j'}}\mathbf{r}_{\mathbf{j'j}})/{r_{j'j}}^3] + \frac{1}{2}\sum_j \alpha_j|\boldsymbol{\xi}_{\mathbf{j}}|^2 \tag{1b}$$

where the first term comes from the interaction of the charges *i* with the dipoles *j*, the second term from the interaction of the dipoles *j* and *j′*, and the third term from the energy that must be spent in distorting the electron cloud of the atom to create the induced dipoles. This energy cost can be verified by using a model that views the electron as being attached to the nuclear core by a spring or by actual quantum mechanical calculations which consider an atom in an external field. At any rate, we can rewrite eq 1b as[16]

$$W = -\frac{1}{2}\sum_{i,j} Q_i(\boldsymbol{\mu}_{\mathbf{j}}\cdot\mathbf{r}_{\mathbf{ij}})/{r_{ij}}^3 \tag{1c}$$

or in other words (see also ref 9)

$$W = -\frac{1}{2}\sum_j \boldsymbol{\mu}_{\mathbf{j}}\cdot\boldsymbol{\xi}_{\mathbf{j}}^{\mathbf{0}} \tag{1d}$$

where $\boldsymbol{\xi}_{\mathbf{j}}^{\mathbf{0}}$ is the field on the *j*th dipole from the charges in the system. This field does not include the field from the other dipoles; that leads, however, to the actual value of $\boldsymbol{\mu}_{\mathbf{j}}$. The above derivation has not appeared, to the best of our knowledge, in the early macroscopic literature.

Similar problems arise when one tries to consider other features of polarizable matter in a microscopic way by starting from macroscopically based textbooks. Here, one becomes puzzled about the nature of the dielectric constant of small molecular size volume elements, and the problem can only be resolved by microscopic treatments, as was done in section 1 of ref 9.

The problem may become even more profound when one tries to solve time-dependent problems in polarizable matter by starting from a macroscopic perspective (see for example the controversy about nonequilibrium effects,[17,18] which could be easily resolved microscopically by using, for example, a polarizable empirical valence bond (EVB) type model). The conceptual difficulties with the macroscopic picture (and the corresponding dielectric behavior) of the polarizable (nonpolar) medium disappear once one takes a fully microscopic treatment of a collection of induced dipoles into account. Such a microscopic derivation has been presented in refs 9 and 16. Classical treatments of electronic polarizability of isolated molecules emerged in the early 1970s[19] in addition to quantum mechanical treatments of isolated molecules in electric fields.[20,21] As far as classical treatments are concerned, the work of Applequist and co-workers[19] has provided a classical way of evaluating the polarization of an isolated molecule in the gas phase by an external electric field. Although this has been an important advance in the field, it was neither developed into an approach for calculations of the energy of interacting molecules nor for a tool in force field studies.

Classical microscopic treatments of the energetics of induced dipoles for solutions and large molecules emerged only in the mid 1970s. In particular, a preliminary attempt to study dielectric effects in nonpolar environments was reported by Hopfinger,[22] who placed a methyl group between two charges. However, this study overlooked the fact that most of the dielectric effects come from the molecules around the charges rather than between them. Thus, the first physically consistent microscopic study of dielectric effects in nonpolar environments was reported by Warshel and Levitt (WL),[16] who simulated the electrostatic environment in lysozyme by a classical polarizable force field and represented the effect of the surrounding solvent by a grid of Langevin-type dipoles. Similar approaches were used for other proteins (e.g. ref 23) and for polarizable grids of dipoles (e.g. ref 24). Alder and co-workers[25,26] subsequently used a polarizable model for simulations of charges and dipoles in nonpolar solvents. Thus, the use of polarizable force fields dates back to the work of Warshel and Levitt,[16] who introduced this approach as a general way of capturing the effect of electronic polarization and the corresponding dielectric constant in protein modeling. This was done using both iterative and noniterative approaches. Subsequent early instructive studies include those reported in refs 27 and 28.

The use of polarizable force fields became an integral part of the simulations in our group,[29,30] and we analyzed its effect on electrostatic modeling in many subsequent studies.[9,31,32] The general realization that the effect of induced dipoles is important has been relatively slow (some workers initially argued that this cannot be an important effect[33]), but it is now widely appreciated.

Recent works have advanced the use of polarizable models to many force field programs and also refined the accuracy of such models.[34−43] Furthermore, the use of polarizable models in simulations has progressed significantly, and many studies have implemented polarizable water models.[27,39,40,44−47] The general advances in the development of polarizable force fields will be described by other workers in this issue, including detailed descriptions of specific implementations and their differences and similarities to earlier models.

Although we leave it up to other workers to describe their specific implementations, we would like to comment on the fact that the inclusion of induced dipoles allows one to transfer gas-phase ab initio potentials to condensed phases. That is, Wallqvist and Karlstrom[48] have shown that it is possible to represent the gas-phase potential of a water dimer by a potential surface that includes classical induced dipoles. A further crucial step was done by Kuwajima and Warshel[44] who demonstrated that a polarizable potential that was fitted to an ab initio potential of a water dimer can be directly transferred to condensed phases and reproduces, for example, the many-body effect of water molecules on the dipole moment of each water molecule in condensed phases.

We would also like to clarify that in contrast to the possible implications from a recent study,[46] the KW model is a quite consistent model, and its inability to reproduce the exact gas-phase dimer spectra properties is entirely due to the fact that the MCY gas-phase ab initio potential available at that time[49] was not perfect (the MCY and KW potentials give identical gas-phase results as verified by Saykally and co-workers[50]). The point of the KW paper was to show how to transfer ab initio potentials to solution and not how to improve ab initio calculations.

To conclude this section, it might be useful to re-emphasize that a general-purpose polarizable force field program has been already available as early as 1975. It was originally implemented in the program used in ref 16 which of course provided all the relevant parameters. Subsequently, it was implemented in the POLARIS and ENZYMIX programs.[29,51] A detailed description of the program, the parameters, and the performance is given in ref 29. Several versions of the polarizable force field have been used both in simplified PDLD studies (e.g. refs 16, 24, and 30) and in MD simulations starting with ref 52 as well as countless subsequent studies by our group. Thus, claims that such programs were only recently developed are not useful.

2. Calibration of Cation Force Fields Using Binding Energies to Valinomycin

The most crucial need for a polarizable force field is probably in the treatment of ions and ionized groups. To demonstrate this point, we will describe a recent calibration study, which was aimed at refining force field parameters for studies of ion channels. We start this section by pointing out that one of the most important factors in any reliable study of the selectivity of biological ion channels is the accuracy of the parameters that describe the solvation of the ions by water and by the protein environment.[53] In view of the challenges of obtaining converging results in ion channels studies, it is obviously important to reduce any errors associated with the accuracy of the force field. The calibration of force field parameters can be done by using results from high level ab initio calculations of simple systems in the gas phase. Unfortunately, those parameters do not always give proper results in a condensed phase. Therefore, it is a reasonable approach to adjust force field parameters to reproduce experimental hydration energies (e.g. refs 54−56). An improved agreement for highly charged ions can be obtained by specialized approaches (e.g. ref 57). At any rate, regardless of the procedures used, it is absolutely crucial to validate and refine the parameters by comparing calculated and observed solvation energies in proteins and solutions. The problem is, however, that convergence errors in the protein active site can be larger than the “errors” in the force field parameters. Moreover, since it is trivial to reproduce the solvation in water by adjusting the force field parameters, it is important to use in the refinement process additional information which reflects the difference between the solvation of the cation in the protein and in water. In our view, the best strategy is to compare the “solvation” energy of the cations in water and in macrocycles. Of course, requiring that the resulting force field will also reproduce ab initio results can augment this type of treatment. At any rate, we describe below a systematic force field calibration by calculations of cation solvation energies in water and in a system that contains the key groups of the cation binding sites. In our view, valinomycin is an excellent system for the validation of cation parameters because it is relatively simple (cyclododecadecipeptide), the solvation of its polar groups is closely related to the corresponding solvation in proteins, and it shows cation binding selectivity.[58] The relative simplicity of valinomycin is crucial since it allows for proper convergence, which is hard to obtain in studies of cations binding to proteins.

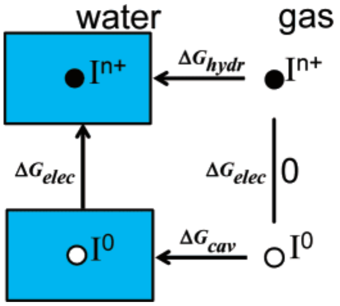


Figure 1. The thermodynamic cycle used for calculations of absolute solvation energies.

Force field parameters for the cations were obtained for both polarizable and nonpolarizable force fields and were first adjusted to reproduce experimental hydration free energies.[54−56] These were then validated by comparing calculated relative binding energies (to valinomycin) with the corresponding experimental values.[59,60] The calculations of the hydration energy were based on the thermodynamic cycle described in Figure 1. This cycle divides the hydration energy into two contributions, the electrostatic and the cavitation energy, using

$$\Delta G_{\mathrm{hydr}}(I^{n+}) = \Delta G_{\mathrm{elec}}(I^0 \rightarrow I^{n+}) + \Delta G_{\mathrm{cav}} \tag{2}$$

where I^0 and I^{n+} are the uncharged and ionized state of a cation respectively, and ΔG_{cav} is the free energy of solvation of the uncharged cation. The electrostatic contribution, ΔG_{elec}, was calculated by the adiabatic charging (AC) free energy perturbation (FEP) approach[1,61] using

$$V_m(\lambda_m) = V_0(1 - \lambda_m) + V_1\lambda_m \tag{3}$$

$$\exp\{-\Delta G(\lambda_m \rightarrow \lambda_{m+1})\beta\} = \langle \exp\{-(V_{\lambda_{m+1}} - V_{\lambda_m})\}\beta\rangle_{V_{\lambda_m}} \tag{4}$$

$$\Delta G_{V_0 \rightarrow V_1} = \sum_{m=0}^{n-1} \Delta\Delta G_{\lambda_m \rightarrow \lambda_{m+1}} \tag{5}$$

where V_0 is the potential where the charge of the cation is zero, V_1 is the potential where the charge of the cation is +1 or +2 depending on the cation type and λ_m are mapping windows between V_0 and V_1. Typically, 51 windows were used with a 5 ps simulation time and 1 fs time steps.

The force field potential for the interaction between the cation and other atoms was defined by

$$V_{I-W} = \sum_j (A_I A_j r_{Ij}{}^{-12} - B_I B_j r_{Ij}{}^{-6}) + \sum_j C Q_I q_j / r_{Ij} + U_{\mathrm{ind}}(\mathbf{r}) \tag{6}$$

where I represents a cation, j represents other atoms, A_i and B_i are the vdW parameters for the given atom, Q_I and q_j are the charges (or residual charges) of the ion and the jth solvent atom, while C is 332. The charges are given in atomic units,

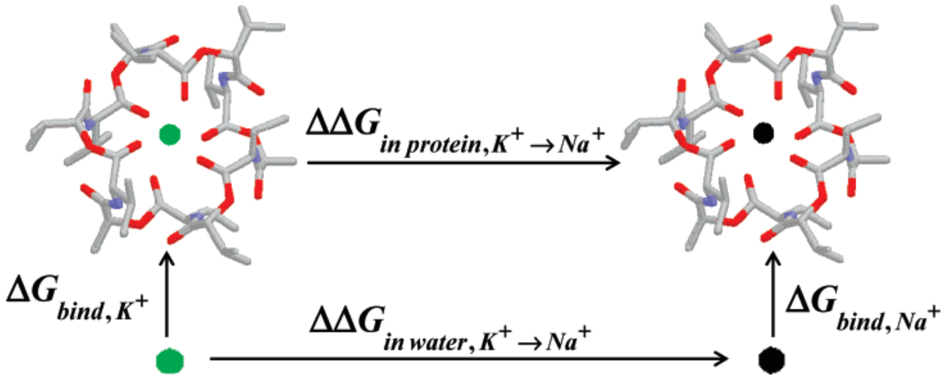


Figure 2. The thermodynamic cycle used for evaluation of the relative binding energy of sodium and potassium to valinomycin.

the distance in Å, and the energy in kcal/mol. The cavitation energy (the nonelectrostatic contribution ΔG_{cav}) was calculated by a FEP treatment, in which V_1 was defined as the potential where the vdW parameters A and B of the cation are at their values in eq 6 and V_0 is the potential where A and B are set to zero.

After calibrating the solvation energy in water, we moved to the next step of evaluating the free energy of binding of the cations to valinomycin (the "protein"). In principle, we could evaluate the energetics of the absolute binding energies using the thermodynamic cycle of Figure 2. However, in the present case, we focus on the relative binding energies. These relative binding energies were obtained by taking the difference of the free energies to transform the cation in valinomycin (surrounded by water) and in bulk water. For example, for the K^+ and Na^+ pair we used

$$\Delta\Delta G_{\mathrm{bind,K^+\rightarrow Na^+}} = \Delta G_{\mathrm{bind,Na^+}} - \Delta G_{\mathrm{bind,K^+}} = \Delta\Delta G_{\mathrm{protein,K^+\rightarrow Na^+}} - \Delta\Delta G_{\mathrm{water,K^+\rightarrow Na^+}} \quad (7)$$

The mutation of the cations was done by an AC FEP procedure using 51 windows of 5 ps with 1 fs time steps. The refined parameters and the corresponding hydration energies are summarized in Table 1, and the results for monovalent ions are also given in Figure 3. As seen from the table, we obtained very reasonable results for both the nonpolarizable and polarizable force fields. In fact, a better agreement for the divalent ions can be easily obtained by using six center dummy atom models for the ion (e.g., refs 57 and 62). At any rate, optimized parameters were then used to evaluate the relative binding free energies of cations to valinomycin and the calculated results are summarized in Figure 4. As seen from the figure, we obtained reasonable results for the binding of monovalent ions (Figure 4a) to valinomycin for both the polarizable and nonpolarizable force fields, although the order of the binding selectivity of cations was not always correct. This is clearly satisfactory considering the 1 kcal/mol error range of the parametrization for the hydration energies. However, in the case of the divalent ions (Figure 4b) the polarizable model gives significantly better results than the nonpolarizable model. More specifically, both the polarizable and nonpolarizable force fields give reasonable results in (A), while in (B) only the polarizable force field does (e.g., the deviations in the case of $Sr^{2+} \rightarrow Ca^{2+}$ are around 4 kcal/mol).

Table 1. Cation vdW Parameters and Solvation Energies Calculated with Nonpolarizable (A) and Polarizable (B) Force Fields[a]

	vdW parameters		hydration energy (kcal/mol)		
cation	A	B	$\Delta G_{\mathrm{hydr,calc}}$	$\Delta G_{\mathrm{hydr,expt}}$	$\Delta\Delta G_{\mathrm{(expt-calc)}}$
			(A) Nonpolarizable Force Fields		
Na^+	94	3.89	−98	−98.2	−0.2
K^+	333	4.35	−80.2	−80.6	−0.4
Rb^+	508	4.64	−74.7	−75.5	−0.8
Cs^+	892	5.44	−68.9	−67.8	1.1
Ca^{2+}	205	18.82	−378.4	−380.8	−2.4
Sr^{2+}	470	20.54	−345	−345.9	−0.9
Ba^{2+}	1045	24.13	−312.2	−315.1	−2.9
			(B) Polarizable Force Fields		
Na^+	47	3.89	−97.8	−98.2	−0.4
K^+	205	4.35	−79.7	−80.6	−0.9
Rb^+	318	4.64	−75.9	−75.5	0.4
Cs^+	655	5.44	−68.7	−67.8	0.9
Ca^{2+}	85	18.82	−381.3	−380.8	0.5
Sr^{2+}	242	20.54	−344.8	−345.9	−1.1
Ba^{2+}	668	24.13	−314.4	−315.1	−0.7

[a] The parameters for the solvent and the protein are the standard MOLARIS parameters.[29]

At any rate, the most important conclusion of the present study is that we can easily fit parameters that reproduce the observable solvation energy in water by both polarizable and nonpolarizable models. The advantage of polarizable models only becomes apparent when we move from water to other environments and even then (if we deal with ions that are in contact with water) only in the case of divalent ions.

3. General Applications of Polarizable Force Fields

This section will cover a wide range of examples of the application of polarizable force fields to different systems, focusing mainly on contributions from our lab. In each case, we will emphasize the importance of the use of polarizable force fields relative to the problems associated with other factors (e.g., convergence effects).

3.1. Calibration and Examination by Studies of Solvation Energies of Small Molecules. The modeling of a biological process can be helped enormously by calibrating the calculations or the conceptual considerations relative to the observed (or estimated) solvation free energy of the

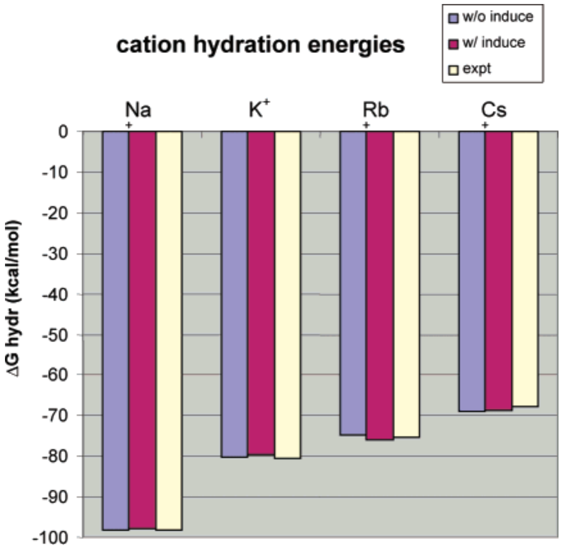


Figure 3. Cation hydration energies obtained after the parametrization. The white bars show the experimental hydration energies, while blue and red bars show the calculated hydration energies with nonpolarizable and polarizable force fields, respectively.

relevant reacting system in aqueous solution (e.g. refs 1 and 9). This is true with regards to enzymatic reactions where the catalytic effect is defined relative to the corresponding solution reaction and, of course, for calculations of ligand binding processes where one has to compare the solvation energy of the ligand in the protein site with the corresponding solvation energy in solution. Early attempts to estimate solvation energies (e.g. refs 63 and 64) were based on the use of the Born or Onsager models with an arbitrary cavity radius. The first attempts to move toward quantitative evaluations of solvation energies can be divided into two branches. One direction involved attempts to examine the interaction between the solute and a single solvent molecule (e.g. ref 65) quantum mechanically. The other direction, which turned out to be more successful, involves the realization that quantitative evaluation of solvation free energies requires parametrization of the solute−solvent van der Waals interaction in a complete solute−solvent system[24] and evaluation of the interaction between the solute and many (rather than one) solvent molecules. Although such an empirical approach was initially considered by the quantum mechanical community as having "too many parameters", it was eventually realized that having an atom-solvent parameter for each type of the solute atoms is the key requirement in any quantitative semiempirical solvation model.

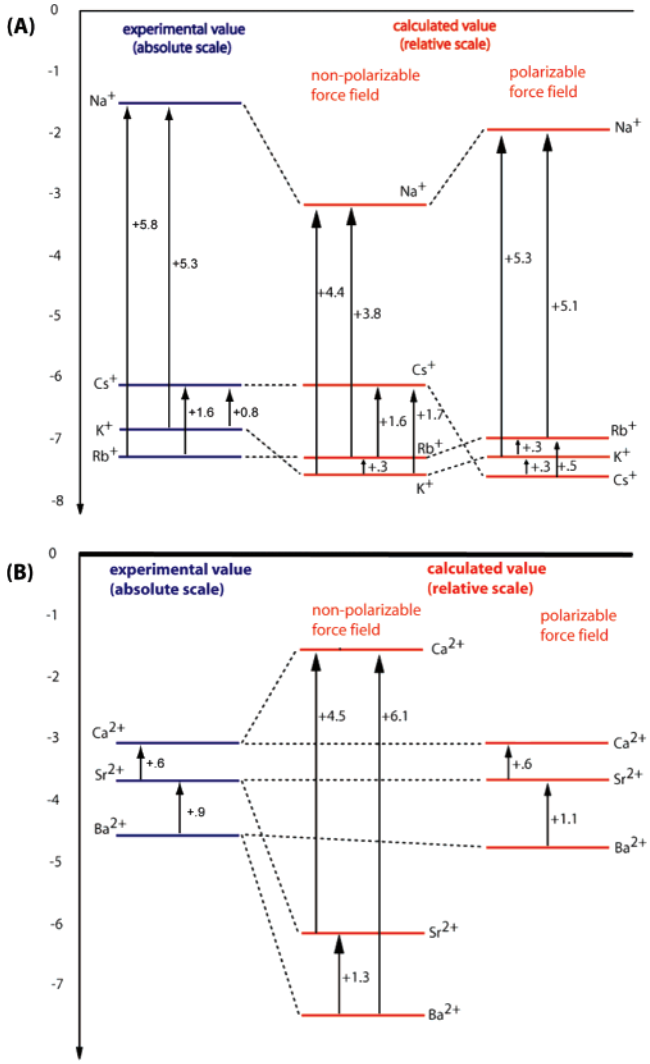

(A)
experimental value (absolute scale)
calculated value (relative scale)
non-polarizable force field
polarizable force field
0
-1
-2
-3
-4
-5
-6
-7
-8
Na+
Cs+
K+
Rb+
+5.8
+5.3
+1.6
+0.8
+4.4
+3.8
+1.6
+1.7
+.3
+5.3
+5.1
+.3
+.3
+.5
(B)
experimental value (absolute scale)
calculated value (relative scale)
non-polarizable force field
polarizable force field
Ca2+
Sr2+
Ba2+
+.6
+.9
+4.5
+6.1
+1.3
+.6
+1.1

Figure 4. The relative free energies (in kcal/mol) for the binding of cations to valinomycin. The experimental values are shown in blue, while the calculated values are shown in red. The figure gives the results for monovalent (A) and divalent (B) ions. The experimental binding energies are given by reporting the corresponding absolute values, while the calculated values are given as relative energies (e.g., K^+ relative to Na^+). As seen from the figure, both the polarizable and nonpolarizable force fields give reasonable results in (A), while in (B) only the polarizable force field does (e.g., the deviations in the case of $Sr^{2+} \rightarrow Ca^{2+}$ are around 4 kcal/mol).

In our view, the successes of calculations of solvation energies of small molecules in solution with a parametrized potential (e.g. refs 24 and 66−69) are very important but, in some respect, obvious. That is, in such cases the environment is uniform, and the solvation free energy is related to the effective atomic radius in a simple way. Thus, reasonable parametrization can usually be accomplished (e.g., see section 3.1 as well as refs 29 and 68). However, the ability to reproduce solvation energies in solution is not a guarantee for reasonable results for the solvation energies of charged ligands in proteins. This issue will be addressed in subsequent sections. At any rate, since it is always possible to fit parameters that reproduce the solvation of a given molecule, the issue here is whether the use of a polarizable model improves the agreement between the calculated and observed solvation energies in a series of related molecules (where we cannot freely adjust the van der Waals parameters). Some interesting studies along this line were done with the amine series,[69−71] although it is not clear whether the actual agreement was improved by the use of a polarizable model. It is possible that the difficulties in fitting reflect charge transfer to the solvent that has not been accounted for in the models used. Here, the best strategy should probably involve calculations of solvation in small clusters by both ab initio and force field models followed by adjustment of the force field parameters to reproduce both the solvation in the cluster and in the bulk (e.g. ref 62). Such an approach should allow separation of the charge transfer and inductive effects. At any rate, the parameters obtained by calibration on solvation in solution should be validated when moving to the protein site as was done in the studies described in section 2.

3.2. Evaluation of pK_as of Ionizable Residues in Proteins. Ionizable residues in proteins play a major role in most biological processes including enzymatic reactions, proton pumps, and protein stability. This role involves both the interaction between the ionizable groups and the energetics of the ionization process. Thus, the ability to calculate pK_as of ionizable groups in proteins is crucial in attempts to correlate the structure and function of proteins and to validate different models for electrostatic energies in proteins.[9]

Calculations of pK_as by all-atom FEP approaches have been reported in a surprisingly small number of cases (e.g. refs 52, 72, and 73). Recent works include studies of the pK_a of metal-bound water molecules[74] and proton transfer in proteins[75] as well as functionally important groups (e.g. refs 76−78). All-atom LRA calculations were also reported.[79,80] In only a few cases was any attempt made to actually estimate the error range in these calculations (e.g. ref 81). It appears that the error range of the all-atom models is still somewhat disappointing, although the inclusion of proper long-range treatments and induced dipoles leads to some improvement.[72,79] As far as the effect of induced dipoles is concerned, we would like to clarify that all of the early PDLD studies of pK_as in proteins included explicitly induced dipoles and explored the role of the induced energy (e.g. ref 9). Similarly, most all-atom studies of pK_as in proteins by our group included the use of a polarizable force field.[79] The effect of induced dipoles appeared to be important mainly in the case of ionizable groups in protein interiors (e.g. ref 82).

3.3. Redox Potential of Proteins and Electron Transport Processes. Electron transport processes are involved in key energy transduction processes in living systems (most notably, photosynthesis). Such processes involve changes in the charges of the donor and acceptor involved and are thus controlled by the electrostatic energies of the corresponding charges and the reorganization energies involved in the charge-transfer process. Here, the challenge is to evaluate the redox energies and the reorganization energies using the relevant protein structure. Probably the first attempt to address this problem was reported by Kassner,[83] who represented the protein as a nonpolar sphere. The idea that such a model can be used for analyzing redox properties held on for a long time (see discussion in refs 84−90 and in ref 91). However, the use of the microscopic PDLD model,[92,93] with its self-consistent polarizability treatment, has shown that the evaluation of redox potentials must take the protein permanent dipoles and the penetration of water molecules into account. The role of the protein permanent dipoles has been most clearly established in subsequent studies of iron−sulfur proteins.[94,95] Another interesting factor is the effect of ionized groups on redox potentials. PB studies of redox proteins have progressed significantly since the early studies that considered the protein as a nonpolar sphere (see above). These studies (e.g. refs 84, 87, and 96−98) started to reflect a gradual recognition of the importance of the protein permanent dipoles, although some confusion still exists (see discussion in refs 84 and 90). The realization of the importance of the protein permanent dipoles could not be accomplished in a convincing way without accounting for the effect of the induced dipoles, which has been done in many of the above studies. Microscopic estimates of protein reorganization energies have been reported[31,99,100] and were used very effectively in studies of the rate constants of biological electron transport. This also includes studies of the nuclear quantum mechanical effect associated with the fluctuations of the protein polar groups (for review see ref 101). As far as the role of induced dipoles is concerned, probably the most systematic study to date has been reported by Muegge et al.[99] who explored the dielectric effect in cytochrome *c* for microscopic, semimacroscopic, and macroscopic models. The inclusion of induced dipoles has also been shown to be crucial in studies of photosynthetic systems,[31,101,102] where the correct mechanism was first elucidated theoretically[102] rather than experimentally.

3.4. Electrostatic Effects in Ligand Binding to Proteins. A reliable evaluation of the free energy of ligand binding can potentially play a major role in designing effective drugs against various diseases (e.g. ref 103). Here, there is an interplay between electrostatic, hydrophobic, and steric effects, but accurate estimates of the relevant electrostatic contributions are still crucial. In principle, it is possible to evaluate binding free energies by performing FEP calculations and 'mutating' the ligand to 'nothing' in water and in the protein active site. This approach, however, encounters major convergence problems, and, at present, the reported results are disappointing with the exception of cases of very small ligands. Alternatively, in simple cases one could study the effect of small 'mutations' of the given ligand,[104] for example, a replacement of NH_2 by OH. However, when one is interested in the absolute binding of medium-size ligands, it is essential to use simpler approaches. Perhaps the most useful alternative is offered by the LRA approach augmented by estimates of the nonelectrostatic effects. That is, the LRA approach is particularly effective in calculating the electrostatic contribution to the binding energy.[105,106] With this approximation one can express the binding energy as

$$\Delta G_{\mathrm{bind}} = \frac{1}{2}[\langle U^{\mathrm{p}}_{\mathrm{elec,l}}\rangle_{\mathrm{l}} + \langle U^{\mathrm{p}}_{\mathrm{elec,l}}\rangle_{\mathrm{l'}} - \langle U^{\mathrm{w}}_{\mathrm{elec,l}}\rangle_{\mathrm{l}} - \langle U^{\mathrm{w}}_{\mathrm{elec,l}}\rangle_{\mathrm{l'}}] + \Delta G^{\mathrm{nonelec}}_{\mathrm{bind}} \quad (8)$$

where $U^{\mathrm{p}}_{\mathrm{elec,l}}$ is the electrostatic contribution for the interaction between the ligand and its surroundings, p and w designate the protein and water, respectively, and l and l′ designate the ligand in its actual charged form and the 'nonpolar' ligand (where all the residual charges are set to zero), respectively. In this expression, the terms $\langle U_{\mathrm{elec,l}} - U_{\mathrm{elec,l'}}\rangle$ are replaced by $\langle U_{\mathrm{elec,l}}\rangle$ since $U_{\mathrm{elec,l'}} = 0$. Now, the evaluation of the nonelectrostatic contribution $\Delta G^{\mathrm{nonelec}}_{\mathrm{bind}}$ is still very challenging, since these contributions might not follow the LRA. A useful option, which was used in refs 105 and 106, is to estimate the contributions to the binding free energy from hydrophobic effects, van der Waals, and water penetration by the PDLD approach. Another powerful option is the so-called linear interaction energy (LIE) approach.[67] This approach starts from the LRA approximation for the electrostatic contribution but neglects the $\langle U_{\mathrm{elec,l}}\rangle_{\mathrm{l}'}$ terms. The binding energy is then expressed as

$$\Delta G_{\mathrm{bind}} \approx \alpha[\langle U^{\mathrm{p}}_{\mathrm{elec,l}}\rangle_{\mathrm{l}} - \langle U^{\mathrm{w}}_{\mathrm{elec,l}}\rangle_{\mathrm{l}}] + \beta[\langle U^{\mathrm{p}}_{\mathrm{vdW,l}}\rangle_{\mathrm{l}} - \langle U^{\mathrm{w}}_{\mathrm{vdW,l}}\rangle_{\mathrm{l}}] \tag{9}$$

where α is a constant that is around 1/2 in many cases, and β is an empirical parameter that scales the vdW component of the protein−ligand interaction. A careful analysis of the relationship between the LRA and LIE approaches as well as the origin of the α and β parameters is given in refs 106 and 107.

As far as the effect of induced dipoles is concerned, it seems to us that we are probably not yet at a stage where the inclusion of induced dipoles makes a major difference in binding calculations of neutral molecules, since the convergence problems are still larger than the errors associated with the implicit inclusion of the induced dipoles in the parametrization procedure. However, some of our binding studies did include polarizable force field.[108]

3.5. Enzyme Catalysis. The elucidation of the origin of the catalytic power of enzymes is a subject of big practical and fundamental importance.[1,109−111] The introduction of combined quantum mechanical/molecular mechanics (QM/MM) computational models (e.g. refs 16, 109, and 111−117) provided a way to quantify the main factors that allow enzymes to reduce the activation free energies of the corresponding reactions. QM/MM studies, including those conducted by the empirical valence bond (EVB) method,[1] provided compelling support to the proposal[118] that the electrostatic effects of preorganized active sites play a major role in stabilizing the transition states of enzymatic reactions.[119] In fact, there is now a growing appreciation of this view (e.g. refs 120 and 121). Simulation approaches that focused on the electrostatic aspects of enzyme catalysis (i.e., the difference between the stabilization in the enzyme and in solution) appear to give much more quantitative results than those which focused on the quantum mechanical aspects of the problem but overlooked the proper treatment of long-range effects (see discussion in ref 122). Apparently, some problems can be effectively treated even by PB approaches (see, e.g., ref 123) without considering quantum mechanical issues. Interestingly, evaluation of the activation free energies of enzymatic reactions appeared to be simpler, in terms of the stability of the corresponding results, than other types of electrostatic calculations such as binding free energies (see discussion in ref 124). This advantage has been exploited for a long time in EVB studies (see, e.g., ref 109) and is now being reflected in molecular orbital QM/MM studies (e.g. refs 111, 114, and 125).

Our studies of enzymatic reactions have included explicit treatments of induced dipoles since the initial QM/MM study.[16] In some cases it appeared that one can capture the entire catalytic effect without the use of induced dipoles as long as the focus is on the difference between the reaction in water and the protein active site. However, the inclusion of induced dipoles in simulations of enzymatic reactions has clearly been important in terms of gaining confidence about the importance of electrostatic effects in enzyme catalysis.

3.6. Ion Channels. The control of ion permeation by transmembrane channels underlies many important biological functions (e.g. ref 126). Quantifying the factors that determine ion selectivity by ion channels is a basic problem in protein electrostatics that turns out to be a truly challenging task (e.g. refs 58 and 127). The primary problem is the evaluation of the free energy profile for transferring the given ion from water to the given position in the channel. It is also essential to evaluate the interaction between the conducted ions in the channel if the ion current involves a multi-ion process.[77] Early studies of ion channels focused on the energetics of ions in the gramicidine channel.[32,128] The first microscopic study of this system (or for that matter of any other ion channel) that included all the electrostatic elements of the system (including channel residual charges, channel induced dipoles, solvent, and membrane) explicitly was reported by Åqvist and Warshel.[32] The "solvation" free energy of the system was explored by both the PDLD model and by FEP calculations. The inclusion of the induced dipoles was criticized in ref 33 although the same authors later argued that inclusion of induced dipoles is very important (e.g. ref 129).

The solution of the structure of the KcsA potassium channel[130] provided a model for real biological channels and a major challenge for simulation approaches. Some early studies majorly overestimated the barriers for ion transport (e.g. refs 131 and 132), and the first reasonable results were obtained by the FEP calculations of Åqvist and Luzhkov.[133] These calculations involved the LRF long-range treatment and the SCSSA boundary conditions that probably helped in obtaining reliable results. Microscopic attempts to obtain the selectivity difference between K^+ and Na^+ were also reported.[134] However, these attempts did not evaluate the activation barriers for the two different ions and thus could not be used in evaluating the difference in the corresponding currents. Furthermore, attempts to evaluate the so-called potential of mean force (PMF) for ion penetration, that have the appearance of truly rigorous approaches, have not succeeded in reproducing the actual PMF for moving the ions from water to the channel (see discussion in ref 77).

Our studies of the KcsA potassium channel[53,77] have focused on the evaluation of the selectivity of the ion channel while at the same time using a realistic protein model. It was found that the convergence problems can be overcome in calculations of the energies of the ion binding but become too serious in studies of the activation barriers. Thus, we focused on the use of the semimacroscopic PDLD/S-LRA model combined with Brownian dynamics. However, our studies also involved FEP all-atom calculations of the ion binding using the parameters refined in the procedure described in section 2. These studies also explored the effect of induced dipoles but concluded that in the case of monovalent ions it is reasonable to use nonpolarizable models in view of the fact that the convergence errors are probably larger (at present) than the errors associated with neglect of the induced effects (considering the fact that the parameters are adjusted accordingly).

3.7. Proton Transport. The discovery of aquaporins and their remarkable role in conducting water molecules through cell membranes has attracted major interest in recent years (e.g. refs 135−137). One of the important questions that has been raised is the origin of the blockage of protons by the aquaporin channels. This issue has been[138] and is continuing to be a major field of interest in the biophysical community.[139−148] Early studies (e.g. refs 139 and 143) suggested that this blockage is due to water orientational effects that disrupt the Grotthuss mechanism.[149−151] However, recent works[140,142,144,145,148,152] came to the conclusion that this is due to the electrostatic barrier, in agreement with our general proposal[153,154] which argued that PTR in proteins is controlled by electrostatic barriers.[155]

Assuming that the above point is generally accepted, we can move to our main subject (which remains quite controversial), namely, the origin of the electrostatic barrier and its magnitude. The controversy reflects significant misunderstanding as well as the diverse background of workers in the field and in some cases even unfamiliarity with the progress in electrostatic calculations. Some authors have attributed the barrier to special structural elements[140,142] and, in particular, to the so-called NPA motif,[138,142,148,152] to the ionized residues,[148] and /or to the helix dipoles.[139,144] On the other hand, Burykin and Warshel (BW) concluded that although the electrostatic barrier reflects all the electrostatic contributions of the channel (polar and nonpolar groups), the barrier will remain very high even when these contributions are removed. The different views can be summarized by a schematic drawing of Figure 2 in ref 155, which presents crucial modifications and clarifications (see below) of a similar illustration that was presented before in ref 144.

At any rate, a recent study[155] examined the origin of the barrier for PTR in aquaporin by semimacroscopic and microscopic calculations and explored the effect of different factors. This study confirmed the BW conclusion and clarified the problems with some of the alternative approaches (e.g., not allowing the protein to relax in Poisson−Boltzmann studies).

Overall, it has been demonstrated that the barrier for PTR in proteins, in general, and in aquaporin, in particular, is determined by the overwhelming reduction in solvation energy upon moving from water to the protein, and this can be modulated by specific electrostatic interactions. The barrier can be eliminated only when the sum of the electrostatic contributions from the protein permanent dipoles, the induced dipoles, and the charges is as large as the solvation in water.

Since the reduction in solvation plays such an important role in PTR in proteins, it is quite obvious that proper microscopic studies of such processes should involve the use of polarizable force fields. In fact, the EVB method[1,156] (that is arguably the most effective current model for treating PTR in a full atomistic way) has included induced dipoles in many of our studies of PT in proteins.[157] Similarly, the adaptation of the EVB by Voth and co-workers has also recently emphasized the need for using polarizable models.[158]

3.8. Helix Macrodipoles versus Localized Molecular Dipoles. The idea that the macroscopic dipoles of alpha helices provide critical electrostatic contribution[159,160] has gained significant popularity and appeared in many proposals. The general acceptance of this idea and the corresponding estimates (see below) are, in fact, a reflection of a superficial attitude. That is, we have here a case where the idea that microscopic dipoles (e.g., hydrogen bonds and carbonyls) play a major role in protein electrostatics[9,118] is replaced by a problematic idea that the source of large electrostatic effects is macrodipoles. The main reason for the acceptance of the helix dipole idea (except the structural appeal of this proposal) is the use of incorrect dielectric concepts. That is, estimates of large helix dipole effects[160−164] involve a major underestimation of the corresponding dielectric constant and the customary tendency to avoid proper validation studies. In more detail, almost none of the attempts to estimate the magnitude of the helix dipole effect have tried to verify this estimate by using the same model in calculations of relevant observables (e.g., pK_a shift and enzyme catalysis). The first quantitative estimate of the effect of the helix dipole[165] established that the actual effect is due to the first few microscopic dipoles at the end of the helix and not to the helix macrodipole. It was also predicted that neutralizing the end of the helix by an opposing charge would have a very small effect. This prediction was confirmed experimentally.[166]

One of the most dramatic recent examples of the need for proper consideration of the helix dipole effect has been provided by the KcsA K^+ channel. The study of ref 167 used PB calculations with $\epsilon_p = 2$ and obtained an extremely large effect from the helix dipoles on the stabilization of the K^+ ion in the central cavity (∼ −20 kcal/mol). However, a recent study[53] that used a proper LRA procedure in the framework of the PDLD/S-LRA approach gave a much smaller effect of the helix macrodipole (see Figure 12). Basically, the use of $\epsilon_p = 2$ overestimates the effect of the helix dipole by a factor of 3, and the effect is rather localized on the first few residues. A similar problem occurred with the analysis of the helix dipole in aquaporin where, as stated in section 3.7, it has been suggested that the barrier for PTR is due to the helix dipole.[144,145] However, the careful analysis of ref 155 indicated that the helix macrodipole (or more precisely, its end) only contributes about 4 kcal/mol to the overall barrier. Finally, it is important to note that recent experimental attempts to "neutralize" the effect of the macrodipole in KcsA[168] has confirmed our earlier predictions, as summarized in Figure 5.

The inclusion of induced dipoles either explicitly[165] or implicitly[155] has been a crucial part of the examination of the helix dipole idea, because, in this case, the dielectric effect reduces the helix dipole effect. However, in this respect it is important to point out a misunderstanding that repeatedly appears in some incomplete quantum mechanical studies. There were ab initio attempts to describe the cooperative electrostatic effects, namely, the interaction between charges and collection of amino acids (e.g. refs 169 and 170). These studies concluded that nonadditive effects increase the contribution of the helix dipole and may thus be crucial in enzyme action. Unfortunately, these findings reflect the

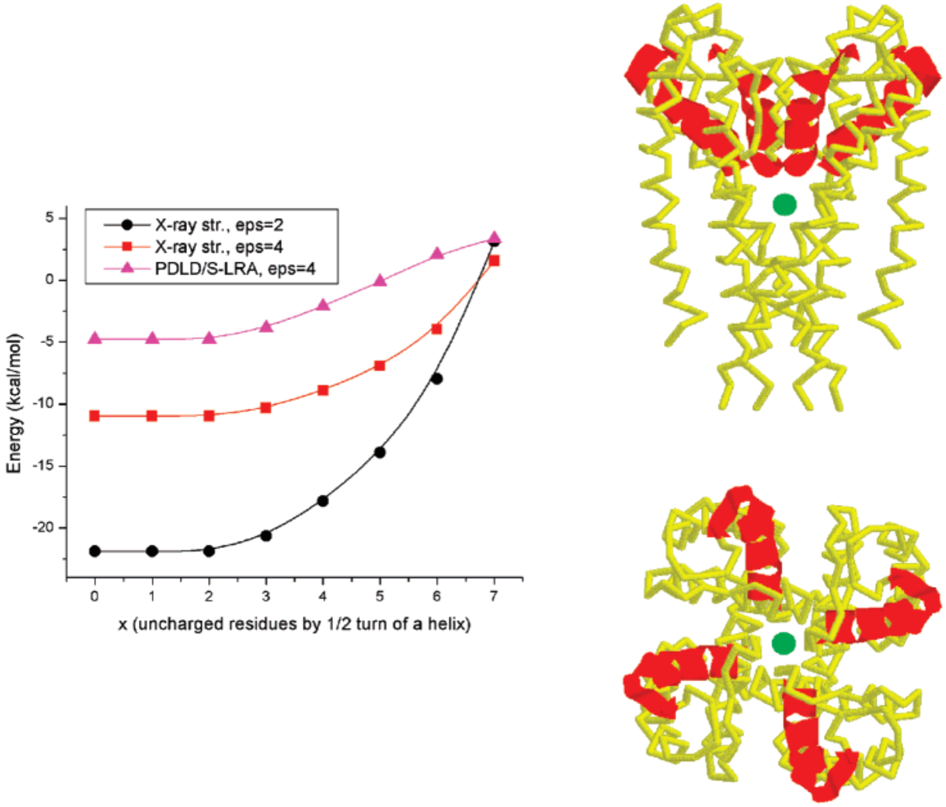


Figure 5. Examination of the effect of the helix dipoles of the KcsA ion channel (upper panel) on a K^+ ion on the central cavity. The lower panel presents the contribution of the residues in the four helices as a function of the dielectric treatment used. It is shown that the use of $\epsilon_p = 2$ drastically overestimates the contribution of the macrodipoles, which is evaluated more quantitatively with the PDLD/S-LRA treatment.

artifact of considering an isolated helix without its surroundings. In this case, the use of a polarizable model (there is no need for any quantum mechanical treatment) demonstrates that the inductive effect *enhances* the interaction. The problem is, however, that most of the dielectric effect comes for the medium around the helix and not from the polarizable matter within the helix (the same is true for the interaction between charges). Thus the effect of the helix dipole is reduced by about one-half due to nonadditive inductive effects when the surrounding is properly included. This fact can be easily verified even in the ab initio studies by embedding the charge and the helix in a polarizable medium.

4. Concluding Remarks

Almost all biological processes are controlled or modulated by electrostatic effects. Thus, the key for quantitative structure−function correlation is the ability to perform accurate electrostatic calculations. Apparently, despite a clear increase in the recognition of the importance of electrostatic effects, there are still significant problems with accepting the need for discriminative validation studies and understanding the relationship between microscopic and macroscopic calculations (see discussion in ref 6).

Nevertheless, one of the issues that is now widely appreciated is the need for polarizable models. This realization is demonstrated by the recent development of many polarizable force fields. However, in some cases we might be overemphasizing the importance of induced dipoles and unjustified in the belief that the reliability problems will disappear once we improve our force field (overlooking convergence issues and other problems).

Despite the advances of polarizable models, there is still a lack of appreciation of simple models that can capture most of the effect of the induced dipoles. For example, in the case of induced dipoles (where the dielectric is small), the noniterative model of WL[16] is very effective, but such models have not been used by the most research groups, with the exception of its adaptation by refs 171 and 172. Similarly, as far as interaction between charges is concerned, it has not been widely realized that the use of Coulomb's law with a dielectric of two is an extremely good approximation even at very close distances (see Figure 13 in ref 9).

Quantum mechanical examinations of the nonadditive effect of induced dipoles are very useful. However, some of these studies have reached incorrect physical conclusions by overlooking hints from simpler approaches. An example is the idea that induced dipoles increase the effect of the helix dipole (see section 3.8). Nevertheless, consistent quantum mechanical studies with QM/MM inclusion of the rest of the environment should be extremely useful in separating the effect of the induced dipoles from the charge-transfer effects.

In conclusion, polarizable force fields offer a practical and effective way of capturing the nonadditive effect of induced dipoles. It is strongly recommended to use such force fields in studies of the charge energetics of protein interiors and in any case where permanent polarization does not account for most of the simulated effect.

Acknowledgment. This work was supported by the NIH grants GM40283, GM24492, and R37GM21422 and the NSF grant MCB-0003872. We thank Dr. Fred Lee and Professor George Eisenman for their initial contribution to the valinomycin study.

References

(1) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons: New York, 1991.

(2) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027−2094.

(3) *Structure and Reactivity in Aqueous Solution*; Cramer, C. J., Truhlar, D. G., Eds.; American Chemical Society: Washington, DC, 1994; Vol. 568.

(4) Perutz, M. F. *Science* **1978**, *201*, 1187−1191.

(5) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284−290.

(6) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647−1676.

(7) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211−217.

(8) Simonson, T. *Rep. Prog. Phys.* **2003**, *66*, 737−787.

(9) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283−421.

(10) Sharp, K. A.; Honig, B. *Ann. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301−332.

(11) Linderstrom-Lang, K. *C. R. Trav. Lab. Carlsberg* **1924**, *15*, 1−29.

(12) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.

(13) Nakamura, H. *Quart. Rev. Biophys.* **1996**, *29*, 1−90.

(14) Jackson, J. D. *Classical Electrodynamics*; John Wiley & Sons: New York, 1999.

(15) Boettcher, C. J. F. *Theory of Electric Polarization*; Elsevier: Amsterdam, The Netherlands, 1973.

(16) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.

(17) Gehlen, J. N.; Chandler, D.; Kim, H. J.; Hynes, J. T. *J. Phys. Chem.* **1992**, *96*, 1748−1753.

(18) Kim, H. J.; Hynes, J. T. *J. Chem. Phys.* **1990**, *93*, 5194−5210.

(19) Applequist, J.; Carl, J. R.; Fung, K. K. *J. Am. Chem. Soc.* **1972**, *94*, 2952−2960.

(20) Gready, J. E.; Bacskay, G. B.; Hush, N. S. *Chem. Phys.* **1977**, *24*, 333−341.

(21) Metzger, R. M. *J. Chem. Phys.* **1981**, *74*, 3444−3457.

(22) Hopfinger, A. J. In *Peptides, Polypetides and Proteins*; Blout, E. R., Goodman, F. A. B. M., Lotan, N., Eds.; Wiley: New York, 1974; pp 77−78.

(23) Warshel, A. *Biochemistry* **1981**, *20*, 3167−3177.

(24) Warshel, A. *J. Phys. Chem.* **1979**, *83*, 1640−1650.

(25) Pollock, E. L.; Alder, B. J. *Phys. Rev. Lett.* **1977**, *39*, 299−302.

(26) Pollock, E. L.; Alder, B. J.; Pratt, L. R. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 49−51.

(27) Barnes, P.; Finney, J. L.; Nicholas, J. D.; Quinn, J. E. *Nature* **1979**, *282*, 459−464.

(28) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341−350.

(29) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161−185.

(30) Russell, S. T.; Warshel, A. *J. Mol. Biol.* **1985**, *185*, 389−404.

(31) Alden, R. G.; Parson, W. W.; Chu, Z. T.; Warshel, A. *J. Am. Chem. Soc.* **1995**, *117*, 12284−12298.

(32) Åqvist, J.; Warshel, A. *Biophys. J.* **1989**, *56*, 171−182.

(33) Roux, B.; Karplus, M. *J. Am. Chem. Soc.* **1993**, *115*, 3250−3260.

(34) Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7829−7834.

(35) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236−242.

(36) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048−1057.

(37) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515−1531.

(38) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1−15.

(39) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497−1506.

(40) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237−2251.

(41) Saint-Martin, H.; Hernandez-Cobos, J.; Bernal-Uruchurtu, M. I.; Ortega-Blake, I.; Berendsen, H. J. C. *J. Chem. Phys.* **2000**, *113*, 10899−10912.

(42) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185−5197.

(43) Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221−234.

(44) Kuwajima, S.; Warshel, A. *J. Phys. Chem.* **1990**, *94*, 460−466.

(45) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910−8922.

(46) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *Science* **2007**, *315*, 1249−1252.

(47) Donchev, A. G.; Galkin, N. G.; Illarionov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8613−8617.

(48) Wallqvist, A.; Karlstrom, G. *Chem. Scr.* **1989**, *29A*, 131−137.

(49) Matsuoka, O.; Clementi, E.; Yoshimine, M. *J. Chem. Phys.* **1976**, *64*, 1351−1361.

(50) Fellers, R. S.; Braly, L. B.; Saykally, R. J.; Leforestier, C. *J. Chem. Phys.* **1999**, *110*, 6306−6318.

(51) Warshel, A.; Creighton, S. In *Computer Simulation of Biomolecular Systems*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM: Leiden, The Netherlands, 1989; pp 120−138.

(52) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368−8372.

(53) Burykin, A.; Kato, M.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 412−426.

(54) Burgess, M. A. *Metal Ions in Solution*; Ellis Horwood: Chichester, U.K., 1978.

(55) Pauling, L. *The Nature of the Chemical Bond;* Cornell University Press: Ithaca, NY, 1960.

(56) Magini, M. In *Ions and Molecules in Solution*; Tanaka, N., Ohtaki, H., Tamamushi, R., Eds.; Elsevier: Amsterdam, 1983; p 97.

(57) Åqvist, J.; Warshel, A. *J. Am. Chem. Soc.* **1990**, *112*, 2860−2868.

(58) Eisenman, G.; Alvarez, O. *J. Membr. Biol.* **1991**, *119*, 109−132.

(59) Ovchinnikov, Y. A.; Ivanov, V. T.; Shkrob, A. M. *Membrane-active Complexones, BBA Library Vol. 12;* Elsevier: New York, 1974.

(60) Grell, E.; Funck, T.; Eggers, F. *Membranes* **1975**, *3*, 1−126.

(61) Valleau, J. P.; Torrie, G. M. In *Modern Theoretical Chemistry;* Plenum Press: New York, 1977; Vol. 5, pp 169−194.

(62) Oelschlaeger, P.; Klahn, M.; Beard, W. A.; Wilson, S. H.; Warshel, A. *J. Mol. Biol.* **2007**, *366*, 687−701.

(63) Kosower, E. M. *Introduction to Physical Organic Chemistry*; Wiley: New York, 1968.

(64) Mataga, N.; Kubota, T. *Molecular Interactions and Electronic Spectra*; M. Dekker: New York, 1970.

(65) Pullman, A.; Pullman, B. *Quart. Rev. Biol.* **1975**, *7*, 506−566.

(66) Grossfield, A.; Ren, P. Y.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671−15682.

(67) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512−9521.

(68) Warshel, A.; Chu, Z. T. *Structure and Reactivity in Aqueous Solution. Characterization of Chemical and Biological Systems*; Cramer, C. J., Truhlar, D. G., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 1994; pp 72−93.

(69) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827−4836.

(70) Ding, Y. B.; Bernardo, D. N.; Kroghjespersen, K.; Levy, R. M. *J. Phys. Chem.* **1995**, *99*, 11575−11583.

(71) Morgantini, P. Y.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 6057.

(72) Saito, M. *J. Phys. Chem.* **1995**, *99*, 17043−17048.

(73) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167−4180.

(74) Fothergill, M.; Goodman, M. F.; Petruska, J.; Warshel, A. *J. Am. Chem. Soc.* **1995**, *117*, 11619−11627.

(75) Åqvist, J.; Fothergill, M. *J. Biol. Chem.* **1996**, *271*, 10010−10016.

(76) Luzhkov, V. B.; Åqvist, J. *Biochim. Biophys. Acta* **2000**, *1481*, 360−70.

(77) Burykin, A.; Schutz, C. N.; Villa, J.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 265−280.

(78) Schutz, C. N.; Warshel, A. *Proteins* **2004**, *55*, 711−723.

(79) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458−4472.

(80) Delbuono, G. S.; Figueirido, F. E.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 85−97.

(81) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100−3107.

(82) Kato, M.; Warshel, A. *J. Phys. Chem. B* **2006**, *110*, 11566−11570.

(83) Harvey, S.; Hoekstra, P. *J. Phys. Chem.* **1972**, *76*, 2987−2994.

(84) Zhou, H.-X. *J. Biol. Inorg. Chem.* **1997**, *2*, 109−113.

(85) Bertini, I.; Gori-Savellini, G.; Luchinat, C. *J. Biol. Inorg. Chem.* **1997**, *2*, 114−118.

(86) Mauk, A. G.; Moore, G. R. *J. Biol. Inorg. Chem.* **1997**, *2*, 119−125.

(87) Gunner, M. R.; Alexov, E.; Torres, E.; Lipovaca, S. *J. Biol. Inorg. Chem.* **1997**, *2*, 126−134.

(88) Naray-Szabo, G. *J. Biol. Inorg. Chem.* **1997**, *2*, 135−138.

(89) Armstrong, F. A. *J. Biol. Inorg. Chem.* **1997**, *2*, 139−142.

(90) Warshel, A.; Papazyan, A.; Muegge, I. *J. Biol. Inorg. Chem.* **1997**, *2*, 143−152.

(91) Rogers, N. K.; Moore, G. R. *FEBS Lett.* **1988**, *228*, 69−73.

(92) Churg, A. K.; Weiss, R. M.; Warshel, A.; Takano, T. *J. Phys. Chem.* **1983**, *87*, 1683−1694.

(93) Churg, A. K.; Warshel, A. *Biochemistry* **1986**, *25*, 1675−1681.

(94) Stephens, P. J.; Jollie, D. R.; Warshel, A. *Chem. Rev.* **1996**, *96*, 2491−2513.

(95) Swartz, P. D.; Beck, B. W.; Ichiye, T. *Biophys. J.* **1996**, *71*, 2958−2969.

(96) Rabenstein, B.; Ullmann, G. M.; Knapp, E. W. *Biochemistry* **1998**, *37*, 2488−2495.

(97) Noodleman, L.; Lovell, T.; Liu, T. Q.; Himo, F.; Torres, R. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 259−273.

(98) Teixeira, V. H.; Soares, C. M.; Baptista, A. M. *J. Biol. Inorg. Chem.* **2002**, *7*, 200−216.

(99) Muegge, I.; Qi, P. X.; Wand, A. J.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 825−836.

(100) Yelle, R. B.; Ichiye, T. *J. Phys. Chem. B* **1997**, *101*, 4127−4135.

(101) Warshel, A.; Parson, W. W. *Q. Rev. Biophys.* **2001**, *34*, 563−670.

(102) Creighton, S.; Hwang, J.-K.; Warshel, A.; Parson, W. W.; Norris, J. *Biochemistry* **1988**, *27*, 774−781.

(103) Muegge, I.; Rarey, M. In *Reviews in computational chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH, John Wiley and Sons: New York, 2001; Vol. 17, pp 1−60.

(104) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.

(105) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215−228.

(106) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393−407.

(107) Brandsdal, B.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. *Adv. Prot. Chem.* **2003**, *66*, 123−158.

(108) Florian, J.; Goodman, M. F.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 5739−5753.

(109) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425−443.

(110) Fersht, A. *Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman and Company: New York, 1999.

(111) Field, M. *J. Comput. Chem.* **2002**, *23*, 48−58.

(112) Bash, P. A.; Field, M. J.; Davenport, R. C.; Petsko, G. A.; Ringe, D.; Karplus, M. *Biochemistry* **1991**, *30*, 5826−5832.

(113) Hartsough, D. S.; Merz, K. M., Jr. *J. Phys. Chem.* **1995**, *99*, 11266−11275.

(114) Alhambra, C.; Gao, J.; Corchado, J. C.; Villà, J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1999**, *121*, 2253−2258.

(115) Marti, S. A., J.; Moliner, V.; Silla, E.; Tunon, I.; Bertran, J. *Theor. Chem. Acc.* **2001**, *3*, 207−212.

(116) Mulholland, A. J.; Grant, G. H.; Richards, W. G. *Protein Eng.* **1993**, *6*, 133−147.

(117) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718−730.

(118) Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5250−5254.

(119) Warshel, A. *J. Biol. Chem.* **1998**, *273*, 27035−27038.

(120) Roca, M.; Marti, S.; Andres, J.; Moliner, V.; Tunon, M.; Bertran, J.; Williams, A. H. *J. Am. Chem. Soc.* **2003**, *125*, 7726−7737.

(121) Cannon, W.; Benkovic, S. *J. Biol. Chem.* **1998**, *273*, 26257−60.

(122) Náray-Szabó, G.; Fuxreiter, M.; Warshel, A. In *Computational Approaches to Biochemical Reactivity*; Náray-Szabó, G., Warshel, A., Eds.; Kluwer Academic Publishers: Dordrecht, 1997; pp 237−293.

(123) van Beek, J.; Callender, R.; Gunner, M. R. *Biophys. J.* **1997**, *72*, 619−626.

(124) Warshel, A.; Villá, J.; Štrajbl, M.; Florián, J. *Biochemistry* **2000**, *39*, 14728−14738.

(125) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569−585.

(126) Hille, B. *Ion Channels of Excitable Membranes*, 3rd ed.; Sinauer Assoc.: Sunderland, MA, 2001.

(127) Eisenman, G.; Horn, R. *J. Membr. Biol.* **1983**, *50*, 1025−1034.

(128) Jordan, P. C. *J. Phys. Chem.* **1987**, *91*, 6582−6591.

(129) Allen, T. W.; Andersen, O. S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117−122.

(130) Morals-Cabral, J. H.; Zhou, Y.; MacKinnon, R. *Nature* **2001**, *414*, 37−42.

(131) Shrivastava, I. H.; Sansom, M. S. P. *Biophys. J.* **2000**, *78*, 557−570.

(132) Allen, T. W.; Kuyucak, S.; Chung, S. H. *Biophys. J.* **1999**, *77*, 2502−2516.

(133) Åqvist, J.; Luzhkov, V. *Nature* **2000**, *404*, 881−884.

(134) Luzhkov, V.; Åqvist, J. *Biochim. Biophys. Acta* **2001**, *36446*, 1−9.

(135) Agre, P.; Kozono, D. *FEBS Lett.* **2003**, *555*, 72−78.

(136) Murata, K.; Mitsuoka, K.; Hirai, T.; Walz, T.; Agre, P.; Heymann, J. B.; Engel, A.; Fujiyoshi, Y. *Nature* **2000**, *407*, 599−605.

(137) Sui, H.; Han, B.-G.; Lee, J. K.; Walian, P.; Jap, B. K. *Nature* **2001**, *414*, 872−878.

(138) Yarnell, A. *Chem. Eng. News* **2004**, *82*, 42−44.

(139) de Groot, B.; Grubmuller, H. *Science* **2001**, *294*, 2353−2357.

(140) de Groot, B. L.; Frigato, T.; Helms, V.; Grubmuller, H. *J. Mol. Biol.* **2003**, *333*, 279−293.

(141) Decoursey, T. E. *Physiol. Rev.* **2003**, *83*, 475−579.

(142) Jensen, M. O.; Tajkhorshid, E.; Schulten, K. *Biophys. J.* **2003**, *85*, 2884−2899.

(143) Tajkhorshid, E.; Nollert, P.; Jensen, M.; Miercke, L.; Stroud, R. M.; Schulten, K. *Science* **2002**, *296*, 525−530.

(144) de Groot, B. L.; Grubmuller, H. *Curr. Opin. Struct. Biol.* **2005**, *15*, 176−183.

(145) Chakrabarti, N.; Roux, B.; Pomes, R. *J. Mol. Biol.* **2004**, *343*, 493−510.

(146) Burykin, A.; Warshel, A. *FEBS Lett.* **2004**, *570*, 41−46.

(147) Burykin, A.; Warshel, A. *Biophys. J.* **2003**, *85*, 3696−3706.

(148) Miloshevsky, G. V.; Jordan, P. C. *Biophys. J.* **2004**, *87*, 3690−3702.

(149) Agmon, N. *Chem. Phys. Lett.* **1995**, *244*, 456−462.

(150) Eigen, M. *Angew. Chem. Int. Ed.* **1964**, *3*, 157−164.

(151) Zundel, G.; Fritcsh, J. Elsevier: Amsterdam, 1986; Vol. 2, Chapter 2.

(152) Ilan, B.; Tajkhorshid, E.; Schulten, K.; Voth, G. A. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 223−228.

(153) Sham, Y.; Muegge, I.; Warshel, A. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 484−500.

(154) Warshel, A. *Photochem. Photobiol.* **1979**, *30*, 285−290.

(155) Kato, M.; Pisliakov, A. V.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2006**, *64*, 829−844.

(156) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523−2544.

(157) Warshel, A.; Russell, S. *J. Am. Chem. Soc.* **1986**, *108*, 6569−6579.

(158) Lefohn, A. E.; Ovchinnikov, M.; Voth, G. A. *J. Phys. Chem. B* **2001**, *105*, 6628−6637.

(159) Wada, A. *Adv. Biophys.* **1976**, *9*, 1−63.

(160) Hol, W. G. J.; Duijnen, P. T. V.; Berendson, H. J. C. *Nature* **1978**, *273*, 443−446.

(161) Roux, B.; Berneche, S.; Im, W. *Biochemistry* **2000**, *39* (44), 13295−13306.

(162) Daggett, V. D.; Kollman, P. A.; Kuntz, I. D. *Chem. Scr.* **1989**, *29A*, 205−215.

(163) Gilson, M.; Honig, B. *Proteins: Struct., Funct., Genet.* **1988**, *3*, 32−52.

(164) van Duijnen, P. T.; Thole, B. T.; Hol, W. G. J. *Biophys. Chem.* **1979**, *9*, 273−280.

(165) Åqvist, J.; Luecke, H.; Quiocho, F. A.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2026−2030.

(166) Lodi, P. J.; Knowles, J. R. *Biochemistry* **1993**, *32*, 4338−4343.

(167) Roux, B.; MacKinnon, R. *Science* **1999**, *285*, 100−102.

(168) Chatelain, F. C.; Alagem, N.; Xu, Q.; Pancaroglu, R.; Reuveny, E.; Minor, D. L. *Neuron* **2005**, *47*, 833−843.

(169) Guo, H.; Salahub, D. R. *Angew. Chem. Int. Ed.* **1998**, *37*, 2985−2990.

(170) van Duijnen, P. T.; Thole, B. T. *Biopolymers* **1982**, *21*, 1749−1761.

(171) Straatsma, T. P.; McCammon, J. A. *Chem. Phys. Lett.* **1990**, *167*, 252−254.

(172) Jorgensen, W. L.; McDonald, N. A.; Selmi, M.; Rablen, P. R. *J. Am. Chem. Soc.* **1995**, *117*, 11809−11810.

CT700127W

JCTC Journal of Chemical Theory and Computation

Charge Model 4 and Intramolecular Charge Polarization

Ryan M. Olson, Aleksandr V. Marenich, Christopher J. Cramer,* and
Donald G. Truhlar*

*Department of Chemistry and Supercomputing Institute, University of Minnesota,
207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431*

Received June 27, 2007

Abstract: Partial atomic charges provide the most widely used model for molecular charge polarization, and Charge Model 4 (CM4) is designed to provide partial atomic charges that correspond to an accurate charge distribution, even though they may be calculated with polarized double-ζ basis sets with any density functional. Here we extend CM4 to six additional basis sets, and we present a model (CM4M) that is individually optimized for the M06 suite of density functionals for ten basis sets. These charge models yield class IV partial atomic charges by mapping from those obtained with Löwdin or redistributed Löwdin population analyses of density functional electronic charge distributions. CM4M/M06-2X/6-31G(d)//M06-2X/6-31+G(d,p) partial atomic charges are calculated for ethylene, CH_nCl_{4-n} ($n = 0-4$), benzene, nitrobenzene, phenol, and fluoromethanol and used to discuss gas-phase polarization effects.

1. Introduction

Molecular polarization is an important aspect of molecular structure, stability, and reactivity; it accounts for the nonuniform distribution of electrons within a molecule and for changes in this distribution due to various interactions. Qualitative theories of molecular polarization are often used to interpret structure and reactivity. The present article concerns polarization effects within single gas-phase molecules, which may be considered to be the starting point for all discussions of polarization.

The degree to which molecular polarization is present in a molecule is called polarity. One measure of polarity is the dipole moment; however, dipole moments are only a single measure of a molecule’s polarity, and dipole moments alone are insufficient to describe the charge distributions within a molecule. Partial atomic charges provide a description of polarity that is intermediate between giving the full electronic charge distribution and giving only the dipole moment. Partial atomic charges are not physical observables because they lack a unique definition that is associated with a quantum mechanical operator, such as the dipole moment operator or the electrostatic potential operator.

The variations in the partial atomic charges with respect to changes in the chemical environment, such as substitution, complexation, or solvation, are key polarization effects that can be quantified with partial charge models. Partial atomic charges are also used in molecular mechanics force fields[1−3] and for calculating the electrostatic contribution to the free energy of solvation using the generalized Born approximation.[4−7]

Numerous methods have been proposed for assigning partial atomic charges. These methods may be assigned to four distinct classes.[8] Class I charges are based on concepts from classical physics and are not based on quantum mechanical calculations. Class II charges are based on a reasonable partitioning of the electron density from a quantum mechanical wave function into atomic populations. Examples of Class II charges are the charges obtained by Mulliken population analysis,[9] Löwdin population analysis,[10] natural population analysis (NPA),[11] Hirshfeld population analysis,[12] atomic polar tensor population analysis,[13] and the population analysis proposed by Bader and co-workers.[14] Class III charges are partial atomic charges constrained to reproduce calculated physical observables such as electrostatic potentials and dipole moments. Schemes such as ChElP[15]/ChElPG,[16] electrostatic interaction energy (ESIE) fitting,[17] and those proposed by Kollman and co-workers[18,19] are examples of Class III charges. Second-generation elec-

* Corresponding author e-mail: truhlar@umn.edu (D.G.T.) and cramer@chem.umn.edu (C.J.C.).

10.1021/ct7001607 CCC: $37.00
Published on Web 10/25/2007

trostatic fitting algorithms such as RESP[20] include restraints to tame unphysical conformational dependences that sometimes occur[21,22] in electrostatic fitting. Finally, Class IV charges[8] are defined as charges that accurately reproduce or predict either charge-dependent experimental observables or well-defined observables obtained by well converged quantum mechanical calculations.

A series of Class IV charge models[7,8,23−26] has been developed for molecular orbital theory and density functional theory (DFT), including ab initio Hartree−Fock (HF) theory and hybrid DFT as special cases. These development efforts led to the recently proposed Charge Model 4 (CM4).[7] Class IV charge models have been designed to map Class II charges obtained from population analysis to accurately reproduce experimental (i.e., accurate) dipole moments. Dipole moments govern the electrostatic potential at long range. By parametrizing the models to reproduce the dipole moments of small, monofunctional molecules, we hope to obtain the correct bond polarity in both small and large molecules and thus to obtain realistic representations of the higher-order multipole moments as well as dipole moments in multifunctional molecules. The parametrized charge models simultaneously correct for the incompleteness of the one-electron basis set and the imperfect treatment of the electron correlation, and therefore the resulting partial atomic charges do not depend strongly on the density functional and one-electron basis set used to obtain the population analysis charges that serve as input to the mappings. Using a simple functional form for the mapping, the CM4 model provides an accurate, efficient, and stable means of assigning partial atomic charges.

The CM1 charge model[8] was developed only for neglect-of-diatomic-differential-overlap theory, but CM2,[23−25] CM3,[26] and CM4[7] may be used with ab initio HF theory and DFT. In this article, we extended the CM4 model so that it can be used with any basis set from for which we previously parametrized a CMx model (x = 2, 3, or 4). These basis sets include the following: 6-31G(d),[27−31] 6-31+G(d),[32] 6-31+G(d,p),[33] MIDI!,[34−36] MIDI!6D,[34−36] DZVP,[37] and cc-pVDZ.[38] The general CM4 model was also extended to include the following additional basis sets: 6-31G(d,p),[30,31,39] 6-31B(d),[40] and 6-31B(d,p).[40] The parameters of the CM4 model for a given basis set are defined to be functions only of the percentage of Hartree−Fock exchange associated with the density functional, and thus they may be used with any exchange-correlation functional. However, somewhat higher accuracy can be obtained by parametrizing for a specific density functional. With this in mind, in this article we specifically optimize a set of parameters for use with the M06 suite[41−43] of functionals (M06, M06-2X, M06-L, and M06-HF); this model will be referred to as the CM4M model. The M06-2X and CM4M methods are then used to discuss polarization effects in a representative set of small molecules.

2. CM4 Model

2.1. Theory. CM4M is a special case of CM4, so we need only to explain the equations for CM4. As in CM2[23−25] and CM3,[26] the charges for the CM4 model are mapped from Class II charges obtained using population analysis by the following formula

$$q_k = q_k^0 + \sum_{k \neq k'} T_{kk'}(B_{kk'}) \tag{1}$$

where q_k is the resulting CM4 charge on atom k, q_k^0 is the input Class II partial atomic charge, and $T_{kk'}$ is a quadratic function of the Mayer bond order[44−46] ($B_{kk'}$):

$$T_{kk'}(B_{kk'}) = (D_{Z_k Z_{k'}} + C_{Zk} Z_{k'} B_{kk'})B_{kk'} \tag{2}$$

The CM4 parameters are the values of $C_{Z_k Z_{k'}}$ and $D_{Z_k Z_{k'}}$; these parameters depend on the choice of the Class II charges used to generate the initial q_k^0 charges, the density functional, and the one-electron basis set. The CM4 parameters are optimized such that the errors in charge-dependent observables calculated from them are minimized. The method for determining the CM4 parameters is discussed in section 2.4.

Löwdin population analysis (LPA) was chosen as the Class II charge model to generate initial charges for one-electron basis sets without diffuse functions, while redistributed Löwdin population analysis[47] (RLPA) was chosen for use with basis sets containing diffuse functions. In a recent study,[47] the dipole moments predicted by Löwdin charges were found to be more accurate than those predicted by Mulliken analysis. Furthermore, redistributed Löwdin population analysis (RLPA) was shown to lead to lower errors in dipole moments and more stable charges than either Löwdin or Mulliken population analysis when the one-electron basis set contains diffuse functions. In the absence of diffuse functions, RLPA charges are equivalent to LPA charges. We note that LPA charges have been shown[48,49] to depend on the orientation of the molecule with respect to a fixed coordinate system when Cartesian basis functions with angular quantum numbers greater than 1 are employed. Table 1 shows the average and standard deviation of CM4M and LPA charges for phenol over ten random rotations using the 6-31G(d) basis set. The LPA (and derived CM4M) charges vary by a chemically insignificant amount so that we conclude that LPA and RLPA Class II charges are a reliable and stable set of input charges for the CM4 mapping.

Table 1. Average and Standard Deviation (stdev) of Löwdin and CM4M Charges of Phenol over 10 Random Rotations Using M06-2X/6-31G(d)//M06-2X/6-31+G(d,p)[a]

	CM4M		Löwdin	
	average	stdev	average	stdev
C1	0.130	0.001	0.106	0.001
C2, (*ortho*)	−0.109	0.001	−0.193	0.001
C3, C5 (*meta*)	−0.066	0.001	−0.150	0.001
C4 (*para*)	−0.105	0.001	−0.189	0.001
C6 (*ortho*)	−0.139	0.002	−0.223	0.001
H7 (*ortho*)	0.090	0.001	0.174	0.001
H8, H10 (*meta*)	0.081	0.001	0.165	0.001
H9 (*para*)	0.080	0.001	0.164	0.000
H11 (*ortho*)	0.076	0.001	0.160	0.001
O12	−0.389	0.001	−0.396	0.001
H13	0.336	0.001	0.366	0.001

[a] Refer to Figure 1 for atom labels.

Table 2. Parameters Defining the CM4 and CM4M Models[a]

parameter	$C_{ZZ'}$	$D_{ZZ'}$	occurrences[b]
H−C		1	234
H−N		2	61
H−O		2	31
H−Si		4	22
H−P		5	25
H−S		3	14
Li−C		6	9
Li−N		6	2
Li−O		6	4
Li−F		6	1
Li−S		6	2
Li−Cl		6	2
C−N		2	149
C−O	2	2	157
C−F		3	111
C−Si		4	10
C−P		6	23
C−S		3	58
C−Cl		3	69
C−Br		3	20
N−O		2	22
N−P		6	1
O−Si	5	5	12
O−P	6	6	24
O−S		3	13
F−Si		5	17
F−P		6	9
Si−Cl		5	18
P−S	6	6	9
P−Cl		6	9

[a] Columns 2 and 3 denote at which stage in the optimization process each parameter was optimized. [b] Number of interactions in the molecules in the parametrization where the Mayer bond order between the atom pairs was greater than 0.20.

Table 3. CM4M Parameters Optimized for the 6-31G(d) Basis Set for the M06 Series of Density Functionals

	M06-L	M06	M06-2X	M06-HF
		$C_{ZZ'}$		
C−O	0.054	0.055	0.058	0.058
O−Si	−0.063	−0.061	−0.066	−0.069
O−P	−0.094	−0.093	−0.093	−0.091
P−S	−0.045	−0.047	−0.047	−0.042
		$D_{ZZ'}$		
H−C	−0.090	−0.091	−0.091	−0.099
H−N	0.031	0.036	0.039	0.045
H−O	−0.041	−0.039	−0.037	−0.036
H−Si	0.019	0.011	0.011	0.019
H−P	0.080	0.070	0.064	0.053
H−S	−0.007	−0.004	−0.002	0.000
Li−C	0.448	0.459	0.472	0.499
Li−N	0.661	0.667	0.695	0.726
Li−O	0.681	0.681	0.719	0.752
Li−F	0.605	0.608	0.615	0.628
Li−S	0.542	0.538	0.539	0.546
Li−Cl	0.594	0.584	0.587	0.587
C−N	0.086	0.086	0.092	0.094
C−O	−0.019	−0.029	−0.030	−0.034
C−F	0.033	0.022	0.024	0.014
C−Si	−0.029	−0.030	−0.023	−0.013
C−P	0.130	0.135	0.136	0.141
C−S	0.141	0.140	0.139	0.137
C−Cl	0.094	0.096	0.100	0.105
C−Br	0.073	0.069	0.059	0.041
N−O	−0.011	−0.020	−0.027	−0.052
N−P	−0.005	−0.003	−0.008	−0.009
O−Si	0.134	0.135	0.145	0.161
O−P	0.244	0.254	0.255	0.263
O−S	0.111	0.123	0.131	0.155
F−Si	0.075	0.084	0.077	0.078
F−P	0.176	0.187	0.181	0.181
Si−Cl	0.020	0.021	0.018	0.011
P−S	0.030	0.036	0.034	0.027
P−Cl	−0.088	−0.086	−0.083	−0.074

2.2. Density Functionals. In previous work, the CM2 parameters were defined as functions of both the method used for the treatment of electron correlation *and* the one-electron basis set. The parameters of the more recent CM3 and CM4 models depend only on the percentage (X) of Hartree−Fock exchange used by the functional and on the one-electron basis set. CM4 parameters are determined by fitting $C_{ZZ'}$ and $D_{ZZ'}$ as a quadratic function of X, for example

$$P^{[X]}_{ZZ'} = b_{ZZ'} + \sum_{i=1}^{1\text{or}2} X^i m^{[i]}_{ZZ'} \tag{3}$$

where P is either C or D for values of $C_{ZZ'}$ and $D_{ZZ'}$ optimized at $X = 0$, 25, 42.8, 60.6, and 99.9 using the mPW1PW*X* functional[50,51] as described in ref 23. The middle values of X used for the mPW1PW*X* functionals correspond to named functionals, mPW1PW[50] ($X = 25$, also called mPW1PW91, mPW0, and MPW25), MPW1K[52] ($X = 42.8$), and MPW1KK[26] ($X = 60.6$), while the limits of $X = 0$ and $X = 99.9$ ensure a smooth fit over the entire range of X. In this work we extend the CM4 model to the following basis sets: MIDI!, 6-31G(d,p), 6-31B(d), 6-31B(d,p), DZVP, and cc-pVDZ.

The CM4 parameters are intended to be compatible with both current and future density functionals; however, the errors in charge-dependent observables can be further reduced if one optimizes the CM4 parameters for specific functionals. As an example, the optimal set of CM4 parameters for a new M06 suite of functionals[41−43] was determined. This model will be referred to as CM4M.

2.3. Basis Sets. CM4 and CM4M parameters were obtained for all basis sets used in previous CM*x* models, as itemized in the Introduction. Both the MIDI! and cc-pVDZ basis sets are defined to use spherical-harmonic *d*-functions, i.e., five *d*-functions are used instead of six Cartesian *d* functions. The remaining basis sets are all defined to use Cartesian *d* functions. The valence/core and polarization functions defined by Binning et al.[31] were used to define 6-31G basis functions for bromine, and the diffuse *s* and *p* functions (exponent = 0.035) for bromine were those defined by Winget and co-workers.[26] The 6-31B basis sets are not defined for Br, so we used the 6-31G definition for bromine in 6-31B calculations.

2.4. Parametrization. The method for determining the CM4 parameters has been described previously.[7] The CM4

Table 4. CM4 Parameters at Fixed Values of Hartree−Fock Exchange ($X = 0$, 25, 42.8, 60.6, 99.9) and the Quadratic Coefficients ($m^{[2]}_{ZZ'}$, $m^{[1]}_{ZZ'}$, $b_{ZZ'}$) Which Define the CM4 Parameters for All Other Values of X

	0	25	42.8	60.6	99.9	$m^{[2]}_{ZZ'}$	$m^{[1]}_{ZZ'}$	$b_{ZZ'}$
				$C_{ZZ'}$				
C−O	0.052	0.054	0.055	0.056	0.056	−0.006	0.010	0.052
O−Si	−0.059	−0.062	−0.064	−0.065	−0.067	0.006	−0.013	−0.059
O−P	−0.089	−0.090	−0.090	−0.091	−0.095	−0.005	0.000	−0.089
P−S	−0.041	−0.049	−0.055	−0.064	−0.085	−0.018	−0.027	−0.041
				$D_{ZZ'}$				
H−C	−0.094	−0.097	−0.099	−0.102	−0.106	0.000	−0.013	−0.094
H−N	0.041	0.035	0.031	0.027	0.017	0.000	−0.024	0.041
H−O	−0.027	−0.035	−0.041	−0.047	−0.060	0.000	−0.033	−0.027
H−Si	−0.003	0.006	0.013	0.019	0.031	0.000	0.034	−0.002
H−P	0.049	0.057	0.063	0.068	0.080	0.000	0.030	0.050
H−S	−0.011	−0.009	−0.007	−0.006	−0.003	0.000	0.007	−0.011
Li−C	0.473	0.472	0.473	0.475	0.483	0.018	−0.007	0.473
Li−N	0.677	0.689	0.700	0.713	0.751	0.036	0.037	0.677
Li−O	0.676	0.692	0.706	0.723	0.772	0.045	0.050	0.676
Li−F	0.595	0.608	0.620	0.634	0.675	0.039	0.041	0.595
Li−S	0.540	0.542	0.544	0.547	0.554	0.007	0.007	0.540
Li−Cl	0.576	0.590	0.601	0.613	0.640	0.009	0.056	0.576
C−N	0.095	0.090	0.086	0.082	0.072	−0.004	−0.019	0.095
C−O	−0.004	−0.021	−0.032	−0.043	−0.065	0.008	−0.069	−0.004
C−F	0.060	0.033	0.014	−0.004	−0.045	0.000	−0.106	0.060
C−Si	−0.043	−0.033	−0.026	−0.020	−0.006	0.000	0.037	−0.043
C−P	0.127	0.131	0.134	0.136	0.140	−0.005	0.019	0.127
C−S	0.140	0.138	0.137	0.136	0.132	−0.002	−0.005	0.140
C−Cl	0.106	0.101	0.097	0.093	0.085	0.000	−0.021	0.106
C−Br	0.066	0.059	0.054	0.049	0.037	0.000	−0.029	0.066
N−O	0.008	−0.017	−0.032	−0.046	−0.078	0.012	−0.096	0.007
N−P	−0.017	−0.011	−0.006	−0.002	0.009	0.000	0.026	−0.017
O−Si	0.105	0.130	0.148	0.166	0.203	0.000	0.098	0.106
O−P	0.220	0.241	0.256	0.272	0.310	0.000	0.090	0.219
O−S	0.091	0.119	0.140	0.160	0.206	0.000	0.116	0.090
F−Si	0.028	0.064	0.090	0.117	0.177	0.000	0.149	0.027
F−P	0.131	0.167	0.192	0.217	0.272	0.000	0.141	0.131
Si−Cl	0.039	0.025	0.016	0.007	−0.013	0.000	−0.052	0.039
P−S	0.035	0.037	0.041	0.047	0.063	0.023	0.006	0.035
P−Cl	−0.066	−0.078	−0.085	−0.093	−0.109	0.000	−0.043	−0.067

parametrization scheme is identical to the method used[26] in the development of CM3 parameters with one exception, namely that the CM4 D_{HC} parameters describing the polarity of the C−H bond were fit to the partial charges from the OPLS force field model[53] for a series of 19 hydrocarbons, whereas the CM3 D_{HC} parameters were fit to adjust the partial charges on ethylene and benzene to preselected values. The resulting CM4 partial atomic charges predict less polar C−H bonds than the previous CM3 model, as will be discussed in section 3.2.1.

The list of parameters optimized for the CM4 model is given in Table 2. The first step in fitting the parameters is to obtain the Mayer bond order matrix and the set of LPA and/or RLPA partial atomic charges for each of the 416 molecular geometries in the training set. The training set[26] consists of 19 hydrocarbon molecules and 397 conformational isomers of 386 unique molecules.

Table 2 also describes the order in which the parameters were optimized and the number of atom−atom interactions affected significantly by each parameter during the optimization step. For this purpose, a significant interaction is defined as a bond order greater than 0.20. The choice of 0.20 was chosen as the bond order cutoff value to report the number of significant interactions, but since CM4 charges are continuous functions of bond order even for bond orders lower than this, the use of this cutoff value for Table 2 has no effect on the calculations. The Mayer bond order is a function of the one-electron basis set and the level of theory employed; thus the values in Table 2 are exact for M06-2X/6-31G(d), whereas for all other methods and basis sets, the values in this table are only approximate.

As previously mentioned, the first parameter to be optimized was the D_{HC} parameter. This was accomplished by minimizing the error function (χ) of the D_{HC} parameter

$$\chi^{[D_{HC}]} = \sum_{k}^{\text{atoms}} (q_k^{\text{CM4}} - q_k^{\text{OPLS}})^2 \qquad (4)$$

over the set of all the atoms in the 19 molecules of the C−H training set.

Table 5. Mean Unsigned Errors (in Debyes) for CM4M Predicted Dipole Moments Using the M06 Suite of Density Functionals and the 6-31G(d) Basis Set

compounds	no.[a]	M06-L	M06	M06-2X	M06-HF
inorganics	10	0.24	0.23	0.23	0.23
alcohols, phenol	13	0.12	0.13	0.12	0.12
ethers	11	0.13	0.12	0.12	0.13
aldehydes	5	0.22	0.22	0.17	0.13
ketones	11	0.18	0.17	0.17	0.17
carboxylic acids	9	0.15	0.18	0.20	0.23
esters	6	0.24	0.22	0.18	0.15
other C, H, O	12	0.21	0.19	0.20	0.20
aliphatic amines	13	0.17	0.19	0.20	0.22
aromatic nitrogen	11	0.23	0.22	0.21	0.19
nitriles	12	0.18	0.19	0.17	0.18
imines	6	0.34	0.32	0.34	0.37
other CHN	14	0.13	0.12	0.12	0.16
amides	17	0.15	0.17	0.16	0.17
nitrohydrocarbons	5	0.14	0.13	0.11	0.17
bifunctional HCNO	11	0.19	0.21	0.22	0.22
HCNO polar	162	0.18	0.18	0.18	0.19
F contaning	39	0.15	0.14	0.13	0.13
Cl contaning	33	0.12	0.11	0.11	0.10
Br contaning	14	0.13	0.11	0.13	0.13
halogenated bifunctionals	23	0.20	0.18	0.18	0.17
thiols	8	0.12	0.11	0.11	0.10
sulfides, disulfides	9	0.23	0.24	0.24	0.23
other sulfur	23	0.41	0.40	0.40	0.40
phosphorus	10	0.35	0.34	0.34	0.34
multifunctional P	13	0.30	0.28	0.27	0.28
S and P containing	7	0.20	0.20	0.15	0.12
CH and Si	9	0.12	0.12	0.12	0.13
CHO and Si	9	0.33	0.33	0.32	0.30
CH, Si, and halogen	18	0.40	0.40	0.41	0.45
lithium compounds	16	0.22	0.20	0.19	0.18
CM3 training set	397	0.21	0.20	0.20	0.20

[a] Number of occurrences of various functional groups in the training set.

Table 6. Mean Unsigned Errors (in Debyes) for CM4 Dipole Moments Using the mPW Exchange Functional and the PW91 Correlation Functional with Various Percentages *X* of Hartree−Fock Exchange and the 6-31G(d) Basis Set

		X =				
compounds	no.[a]	0	25	42.8	60.6	99.9
inorganics	10	0.24	0.23	0.23	0.24	0.27
alcohols, phenol	13	0.12	0.12	0.12	0.12	0.13
ethers	11	0.14	0.13	0.12	0.12	0.11
aldehydes	5	0.22	0.19	0.18	0.16	0.14
ketones	11	0.19	0.17	0.16	0.15	0.13
carboxylic acids	9	0.14	0.18	0.20	0.22	0.24
esters	6	0.20	0.20	0.20	0.20	0.20
other CHO	12	0.23	0.20	0.19	0.18	0.17
aliphatic amines	13	0.19	0.19	0.18	0.18	0.17
aromatic nitrogen	11	0.23	0.22	0.22	0.23	0.25
nitriles	12	0.19	0.18	0.18	0.18	0.18
imines	6	0.32	0.32	0.32	0.32	0.32
other CHN	14	0.15	0.12	0.11	0.11	0.14
amides	17	0.13	0.15	0.17	0.18	0.22
nitrohydrocarbons	5	0.12	0.12	0.12	0.14	0.18
bifunctional HCNO	11	0.22	0.20	0.20	0.20	0.20
HCNO polar	162	0.18	0.18	0.17	0.18	0.18
F contaning	39	0.16	0.15	0.14	0.14	0.14
Cl contaning	33	0.12	0.11	0.11	0.10	0.10
Br contaning	14	0.14	0.13	0.13	0.12	0.10
halogenated bifunctionals	23	0.20	0.18	0.17	0.17	0.17
thiols	8	0.11	0.11	0.12	0.13	0.17
sulfides, disulfides	9	0.21	0.23	0.25	0.27	0.32
other sulfur	23	0.38	0.41	0.43	0.45	0.51
phosphorus	10	0.32	0.34	0.36	0.37	0.40
multifunctional P	13	0.27	0.27	0.28	0.28	0.29
S and P containing	7	0.16	0.15	0.15	0.14	0.15
CH and Si	9	0.13	0.13	0.13	0.13	0.13
CHO and Si	9	0.34	0.33	0.32	0.32	0.30
CH, Si, and halogen	18	0.40	0.40	0.41	0.41	0.42
lithium compounds	16	0.20	0.19	0.20	0.20	0.21
CM3 training set	397	0.20	0.20	0.20	0.20	0.21

[a] Number of occurrences of various functional groups in the training set.

The remaining parameters were divided into five disjoint groups, labeled 2−6 in Table 2. The parameters for each group were optimized in a stepwise manner such that the parameters for previously optimized groups were held fixed. For each group the parameters were optimized to minimize the sum of the squares of the deviations of dipole moments calculated from CM4 charges from a set of target dipole moments, which were either experimental dipole moments or dipole moments calculated from one-electron expectation values of the full electron density of singe-point mPW1PW calculation with the MG3S[54] basis set. A nonlinear optimization procedure was used for the minimization.

The parameters for CM4 and CM4M for the 6-31G(d) basis set are given in Tables 3 and 4, respectively. The 6-31G(d) parameters in Table 4 differ from those previously reported[7] for lithium, silicon, and phosphorus. The Li−F parameter for the 6-31B basis sets were fixed at a value of 1.4. The corresponding mean unsigned errors broken down by functional group are given in Tables 5 and 6. A summary of the errors for CM4 and CM4M charges obtained from the M06-2X density functional and the 6-31G(d) basis set are given in Table 7, where they are compared to errors in dipole moments calculated from LPA charges or from the electron density itself. The CM4 and CM4M parameters and errors (as well as root-mean-square errors) for the remaining basis sets can be found in the Supporting Information.

2.5. Computational Methods. All calculations were run with the M06-2X density functional using a locally modified version of the *Gaussian 03* (*G03*) electronic structure program.[55] All CM*x* charges were calculated using the MN-GSM[56] module. Molecular geometries were optimized using the 6-31+G(d,p) basis set. Partial atomic charges using Löwdin population analysis and the CM2, CM3, CM4, and CM4M models were calculated at the optimized geometries using the 6-31G(d) basis set. The CM2 model is not parametrized for M06-2X; therefore, all reported CM2 charges were calculated using BPW91[57]/6-31G(d). To avoid confusion, dipole moments calculated from the quantum mechanical operator are referred to as density dipole moments. Second-order Møller-Plesset perturbation theory[58] (MP2) with the aug-cc-pVTZ triple-ζ basis set[59] was used to calculate density dipole moments.

Table 7. Mean-Signed (MSE), Mean-Unsigned (MUE), and Root-Mean Squared (RMS) Errors (in Debyes) for Dipole Moments Calculated Using Löwdin (LPA), General CM4, and Optimized CM4M Partial Charges for the M06 Series of Functionals Using the 6-31G(d) Basis Set

	M06-L			M06			M06-2X			M06-HF		
	MSE	MUE	RMS	MSE	MUE	RMS	MSE	MUE	RMS	MSE	MUE	RMS
LPA	0.35	0.62	1.06	0.32	0.63	1.08	0.35	0.65	1.10	0.30	0.66	1.12
CM4	−0.08	0.24	0.32	−0.01	0.21	0.29	0.12	0.24	0.32	0.25	0.36	0.44
CM4M	0.00	0.21	0.29	0.01	0.20	0.28	0.00	0.20	0.28	0.01	0.20	0.28
density	0.01	0.20	0.25	−0.03	0.17	0.22	−0.04	0.19	0.24	−0.14	0.23	0.30

Table 8. Charge (au) on Hydrogens in Ethylene and Benzene Calculated Using M06-2X/6-31G(d)//M06-2X/6-31+G(d,p)

	CM4M	CM4	CM3	CM2	Löwdin
ethylene	0.07	0.06	0.09	0.08	0.15
benzene	0.08	0.07	0.10	0.09	0.16

Table 9. Partial Atomic Charges (au) and Molecular Dipole Moments (in Debye) Calculated Using CM4M, CM4, CM3, CM2,[a] and NPA with M06-2X/6-31G(d)//M06-2X/6-31+G(d,p)

	CM4M	CM4	CM3	CM2[a]	Löwdin	NPA
			CH_4			
C	−0.31	−0.27	−0.40	−0.37	−0.66	−0.93
H	0.08	0.07	0.10	0.09	0.16	0.23
			CH_3Cl (1.93 D)[b]			
C	−0.13	−0.11	−0.22	−0.185	−0.485	−0.67
H	0.095	0.09	0.12	0.11	0.18	0.25
Cl	−0.15	−0.15	−0.13	−0.15	−0.055	−0.075
dipole moment	1.79	1.69	1.71	1.85	1.37	1.88
			CH_2Cl_2 (1.63 D)[b]			
C	−0.01	−0.01	−0.10	−0.05	−0.38	−0.50
H	0.11	0.10	0.13	0.13	0.195	0.27
Cl	−0.11	−0.10	−0.08	−0.10	−0.01	−0.22
dipole moment	1.67	1.55	1.575	1.75	1.21	2.21
			$CHCl_3$ (1.06 D)[b]			
C	0.08	0.07	−0.01	0.055	−0.30	−0.37
H	0.13	0.12	0.15	0.145	0.205	0.29
Cl	−0.07	−0.06	−0.045	−0.07	0.03	0.03
dipole moment	1.19	1.09	1.12	1.275	0.82	1.32
			CCl_4			
C	0.15	0.125	0.06	0.15	−0.24	−0.29
Cl	−0.04	−0.03	−0.015	−0.04	0.06	0.07

[a] CM2 charges are not defined for M06-2X. The CM2 charge listed was calculated using BPW91/6-31G(d)//M06-2X/6-31+G(d,p). [b] The value in parentheses is the density dipole moment calculated using MP2/aug-cc-pVTZ.

3. Polarization Effects

3.1. C−H Bond Polarity. As noted in section 2.4, the major difference between the CM3 and CM4 models is the treatment of the C−H bond polarity. Since the parameter describing the C−H bond (D_{HC}) was the first parameter that was optimized, and all other parameters are optimized given a fixed value of D_{HC}, the value of the parameter D_{HC} plays a critical role in how the model assigns partial atomic charges. Our general experience with the CM3 charge model had convinced us that the C−H bonds were somewhat too polar; therefore, we changed the strategy for obtaining D_{HC} in the CM4 model, as compared to CM3. The choice we made, optimizing gas-phase charges to the OPLS charges, is formally inconsistent because OPLS charges are designed for use in liquid-phase simulations and should be slightly more polar than gas-phase charges. However, this strategy produced partial charges less polar than those we used in CM2 and CM3, and it provided accurate solvation free energies in the SM6 implicit polarizable continuum solvation model, and the fitting strategy seems to be a good compromise between the considerations that led to the more polar C−H bonds of CM2 and CM3 and the practical experience that dictated less polar C−H bonds than CM3. As shown in Table 8, the CM3 model predicts the most polar C−H bond of any of the CM*x* models; however, all CM*x* models predict significantly less polar C−H bonds than Löwdin population analysis.

Polarization effects from substituting chlorine atoms for hydrogen atoms in methane are given in Table 9. The table shows that C−H is less polar in CM4 than in either CM2 or CM3. Furthermore, this table illustrates a basic intramolecular polarization effect in that the atoms in the C−H bond take on increasing positive charge as more chlorines are added, because the chlorines withdraw electron density. The majority of the charge comes from the carbon atom, which goes from having a negative partial atomic charge to a positive one along the series. A small amount of increase in the proton partial charge is also observed, consistent with the known hydrogen-bond donating capability of chloroform > dichloromethane > chloromethane > methane. The table also illustrates that the Löwdin population analysis does not yield qualitatively correct charges, especially for CCl_4; however, the trends in the Löwdin series are correct, which makes a systematic mapping from Löwdin charges (as employed in CM4) a sensible procedure.

The last column of Table 9 gives charges obtained by natural population analysis (NPA).[8] Comparing, for example, the charges in CH_2Cl_2, we see that $|q_H^{NPA}| > |q_{Cl}^{NPA}|$, whereas $|q_H^{CM4}| \approx |q_{Cl}^{CM4}|$; furthermore, $|q_C^{NPA}| < |q_{Cl}^{NPA}|$, whereas $|q_C^{CM4}| > |q_{Cl}^{CM4}|$, where the latter relation is expected based on electronegativity. Although one must be careful to use partial charges for the purposes for which they were intended, in solvation models it is essential that partial charges yield realistic physical observables like electrostatic potentials and multipole moments. In this context, it is interesting to compare the dipole moments for CH_2Cl_2 calculated from

Table 10. CM4M, CM4, CM3, and Löwdin Partial Atomic Charges (au) of Nitrobenzene Calculated Using M06-2X/6-31G(d)//M06-2X/6-31+G(d,p)[a]

	CM4M	CM4	CM3	Löwdin
C1	0.07	0.065	0.055	−0.00
C2, C6 (*ortho*)	−0.06	−0.05	−0.08	−0.14
C3, C5 (*meta*)	−0.07	−0.06	−0.09	−0.15
C4 (*para*)	−0.05	−0.045	−0.08	−0.14
H7, H11 (*ortho*)	0.11	0.10	0.135	0.20
H8, H10 (*meta*)	0.09	0.08	0.11	0.18
H9 (*para*)	0.09	0.08	0.11	0.17
N12	0.17	0.13	0.14	0.32
O13, O14	−0.21	−0.19	−0.19	−0.25
dipole moment (Debye)	4.39	4.155	4.22	4.58

[a] Refer to Figure 2 for atom labels.

Table 11. CM4M, CM4, CM3, and Löwdin Partial Atomic Charges (au) of Phenol Calculated Using M06-2X/6-31G(d)//M06-2X/6-31+G(d,p)[a]

	CM4M	CM4	CM3	Löwdin
C1	0.13	0.12	0.10	0.11
C2 (*ortho*)	−0.11	−0.10	−0.13	−0.19
C3, C5 (*meta*)	−0.07	−0.06	−0.09	−0.15
C4 (*para*)	−0.10	−0.09	−0.125	−0.19
C6 (*ortho*)	−0.14	−0.13	−0.16	−0.22
H7 (*ortho*)	0.09	0.08	0.11	0.175
H8, H10 (*meta*)	0.08	0.07	0.10	0.17
H9 (*para*)	0.08	0.07	0.10	0.16
H11 (*ortho*)	0.075	0.07	0.10	0.16
O12	−0.39	−0.37	−0.35	−0.40
H13	0.335	0.33	0.33	0.36
dipole moment (Debye)	1.125	1.12	1.10	1.24

[a] Refer to Figure 1 for atom labels.

partial charges to the density dipole (1.63 D, see Table 9) obtained using MP2/aug-cc-pVTZ; CM4 charges give 1.67 D while NPA charges give 2.21 D.

3.2. Aromatic Molecules. Tables 10 and 11 provide charges for nitrobenzene and phenol. The charges on the ring carbons at the *ipso*, *ortho*, and *para* positions are seen to vary by 0.05−0.08 when the substituent is changed from the electron-withdrawing nitro group to the electron-donating hydroxy group, but the charges at the *meta* position are changed by less than 0.01. The changes are such that in nitrobenzene the *ortho* and *para* CH groups become net positive (cf. benzene, where the CH groups are necessarily net uncharged; Table 8), while in phenol they become negative. Such behavior is in line with what would be expected from conventional resonance arguments in benzene rings substituted with electron-withdrawing and electron-donating groups, respectively. Note that while the hydrogens vary by 0.01−0.02 upon substitution, they are 0.02−0.03 less positive than in CM3, reflecting the more physical reduced polarity of CH bonds in the CM4 models.

3.3. Fluoromethanol. Fluoromethanol is a small molecule that was the subject of a number of early theoretical studies because of the influence of the anomeric effect on its internal rotational coordinate.[60,61] The anomeric effect,[62] also sometimes referred to as negative hyperconjugation or the Lemieux-Edwards effect, refers to the evident stabilization of conformers having gauche compared to anti dihedral angles associated with atomic linkages WXYZ, where W and Y are electronegative atoms with associated lone pairs, and X and Z may be any atoms but are most often H or Group 14 atoms. In fluoromethanol, W is F, X is C, Y is O, and Z is H, and the gauche conformer is indeed predicted to be substantially lower in energy than the anti conformer.[63]

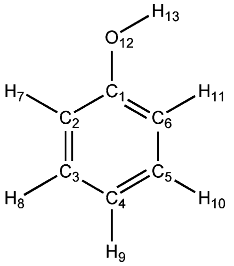


Figure 1. Atom labels in phenol.

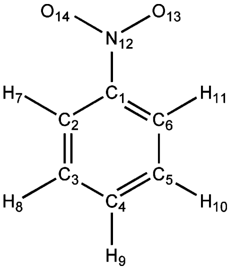


Figure 2. Atom labels in nitrobenzene.

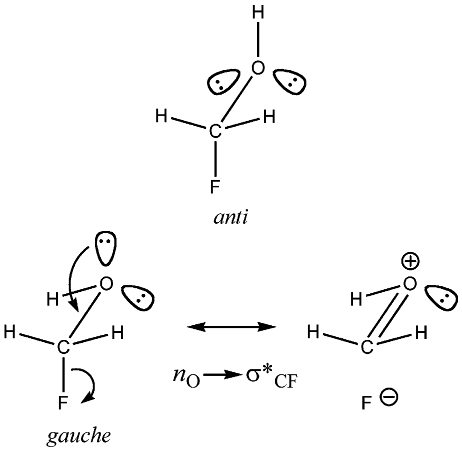


Figure 3. Anomeric delocalization in the gauche conformer of fluoromethanol compared to the anti.

The effect has been invoked in the conformational analysis of many different organic and inorganic systems[64] and is usually rationalized as deriving from stabilizing delocalization of lone-pair density on atom Y into the low-energy σ^* virtual orbital associated with atoms W and X. The overlap between the relevant orbitals is maximized for the gauche conformation, and in the limit of full negative hyperconjugation this delocalization has sometimes been called double-bond−no-bond resonance[65] (Figure 3). Given this electronic structure description, one might expect to see polarization in the gauche conformer associated with a transfer of negative charge from oxygen to fluorine. This effect has been analyzed in terms of partial atomic charges in other systems exhibiting

Table 12. Atomic and Group Partial Charges (au) and Dipole Moments (in Debye) in Anti and Gauche Conformers of Fluoromethanol from M06-2X/6-31G(d)//M06-2X/6-31+G(d,p) Analyses

charge model	atom/fragment				dipole moment[a]
	H(O)	O	CH_2	F	
Löwdin	0.36/0.36[f]	−0.48/−0.44	0.29/0.31	−0.18/−0.23	3.24/2.01
CM3	0.32/0.32	−0.44/−0.40	0.28/0.30	−0.16/−0.22	2.66/1.68
CM4	0.33/0.33	−0.46/−0.43	0.31/0.33	−0.18/−0.24	2.67/1.69
CM4M	0.33/0.33	−0.47/−0.44	0.34/0.37	−0.21/−0.26	2.87/1.80
ChelpG ESP[b]	0.40/0.43	−0.60/−0.62	0.42/0.47	−0.22/−0.28	2.98/1.80
MK ESP[c]	0.40/0.43	−0.59/−0.61	0.39/0.44	−0.19/−0.26	2.99/1.81
NPA[d]	0.49/0.49	−0.78/−0.76	0.67/0.68	−0.38/−0.41	5.25/3.23
μ[e]					2.99/1.78

[a] Computed from partial atomic charges. [b] Electrostatic potential fitting method of ref 10. [c] Electrostatic potential fitting method of ref 11. [d] Natural population analysis of ref 8. [e] Computed from the density as an expectation value. [f] Values before and after solidus refers to anti and gauche conformers, respectively.

anomeric delocalization,[66] and we here examine a variety of charge models for the particular case of fluoromethanol (Table 12).

Considering the various models, the first issue meriting discussion is the poor performance of the NPA charges for the prediction of the molecular dipole moment. The NPA procedure involves the assignment of all electrons to orbitals associated either with a single atom (lone pairs and core orbitals) or pairs of atoms (bonding and antibonding orbitals). Assigning lone pairs entirely to individual atoms may contribute to the greater magnitude of NPA charges and hence the larger charge-derived dipole moment compared to the other models.

Focusing now on changes in charges as a function of conformation, all of the seven charge models do predict that the fluorine partial atomic charge becomes more negative in the gauche conformer, and the absolute magnitudes of the charges are fairly consistent across all models other than NPA. All charge models except for the two ESP algorithms predict that one-half to two-thirds of the charge shift onto F comes from the oxygen atom, and the remainder from the CH_2 group, with the partial atomic charge of the H on oxygen being insensitive to conformation. The ESP charges, by contrast, predict that the O atom becomes more *negative* in the gauche conformation, forcing both the H atom on O and the CH_2 group to become more positive to preserve charge neutrality. This charge arrangement does not degrade the quality of the predicted molecular dipole moment, but there are an infinite number of combinations of monopoles at the nuclear positions that will give identical dipole moments. While it is not unreasonable to imagine the H on O becoming more acidic (more positive) in the gauche conformation, it seems counterintuitive that the O should become more negative.

4. Concluding Remarks

The partial charges calculated by Charge Models 4 and 4M (CM4 and CM4M) are stable and realistic and should be useful for parametrization of force fields or for direct use in molecular mechanics calculations where partial atomic charge parameters are lacking. CM4 and CM4M charges should also be useful for representing molecular charge distributions in solvation models, particularly because their simple algorithmic dependence on Hartree−Fock or Kohn−Sham density matrix elements, through population analysis, permits their straightforward inclusion into self-consistent reaction field models. Finally, the CM4 and CM4M models provide a balanced and chemically intuitive framework within which to discuss intramolecular charge polarization effects.

Acknowledgment. This work was supported in part by the office of Naval Research (award no. N00014-05-1-0538) and the National Science Foundation (grant nos. CHE-0610183 and CHE-0704974).

Supporting Information Available: CM4 and CM4M parameters for additional basis sets, mean unsigned errors and root-mean-square errors for all charge models, and geometries of all optimized molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) MacKerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584.

(2) Jorgensen, W. L.; Tirado-Rives, *J. Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665.

(3) Vizcarra, C. L.; Mayo, S. L. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622.

(4) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(5) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(6) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.

(7) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(8) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

(9) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(10) Baker, J. *Theor. Chim. Acta* **1985**, *68*, 221.

(11) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.

(12) Hirshfeld, F. *Theor. Chim. Acta* **1977**, *44*, 129.

(13) Cioslowski, J. *J. Am. Chem. Soc.* **1989**, *111*, 8333.

(14) Bader, R. F. W.; Matta, C. F. *J. Phys. Chem. A* **2004**, *108*, 8385.

(15) Chirlian, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *8*, 894.

(16) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(17) Arroyo, S. T.; Martin, J. A. S.; Garcia, A. H. *Chem. Phys. Lett.* **2002**, *357*, 279.

(18) Besler, B. H.; Merz, K. M.; Kollman, P. A., Jr. *J. Comput. Chem.* **1990**, *11*, 431.

(19) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

(20) Zhang, W.; Hou, T.; Qiao, X.; Xu, X. *J. Phys. Chem. B* **2003**, *107*, 9071. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157. Laio, A.; Gervasio, F. L.; VandeVondele, J.; Sulpizi, M.; Rothlosberger, U. *J. Phys. Chem. B* **2004**, *108*, 7963.

(21) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367.

(22) Green, D. F.; Tidor, B. *J. Phys. Chem. B* **2003**, *107*, 10261.

(23) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.

(24) Li, J.; Xing, J.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 885.

(25) Li, J.; Williams, B.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 724.

(26) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 10707.

(27) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(28) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.

(29) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

(30) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(31) Binning, R. C.; Curtiss, L. A., Jr. *J. Comput. Chem.* **1990**, *11*, 1206.

(32) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

(33) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(34) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(35) Li, J.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1998**, *99*, 192.

(36) Thompson, J. D.; Winget, P.; Truhlar, D. G. *Phys. Chem. Comm.* **2001**, *4*, 4116.

(37) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

(38) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(39) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654.

(40) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.

(41) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1.

(42) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(43) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* Published online; DOI: 10.1007/s00214-007-0310-x (accessed July 21, 2007).

(44) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 73.

(45) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 477.

(46) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.

(47) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Comm.* **2002**, *5*, 117.

(48) Mayer, I. *Chem. Phys. Lett.* **2004**, *393*, 209.

(49) Bruhn, G.; Davidson, E. R.; Mayer, I.; Clark, A. E. *Int. J. Quantum Chem.* **2006**, *106*, 2065.

(50) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(51) Perdew, J. P. In *Electronic Structrure of Solids '91*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

(52) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.

(53) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(54) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.

(55) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Munnucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatusjui, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gromperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Lui, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, A.; Peng, C. Y.; Nanyakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(56) Chamberlin, A. C.; Kelly, C. P.; Xidos, J. D.; Li, J.; Thompson, J. D.; Hawkins, G. D.; Winget, P. D.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J. *MN-GSM, version 6.2*; University of Minnesota: Minneapolis, MN, 2007.

(57) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. Perdew, J. P. In *Electronic Structrure of Solids '91*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11

(58) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(59) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(60) Radom, L.; Hehre, W. J.; Pople, J. A. *J. Am. Chem. Soc.* **1971**, *93*, 289.

(61) Wolfe, S.; Rauk, A.; Tel, L. M.; Czismadia, I. G. *J. Chem. Soc. B* **1971**, 136.

(62) Kirby, A. J. *The Anomeric Effect and Related Stereoelectronic Effects at Oxygen*; Springer-Verlag: Berlin, 1983.

(63) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons: Chichester, 2004; p 23.

(64) Cramer, C. J. *J. Mol. Struct. (THEOCHEM*) **1996**, *370*, 135.

(65) Hine, J. *J. Am. Chem. Soc.* **1963**, *85*, 3239.

(66) Cramer, C. J. *J. Org. Chem.* **1992**, *57*, 7034.

CT7001607

JCTC Journal of Chemical Theory and Computation

Polarization Effects in Aqueous and Nonaqueous Solutions

Aleksandr V. Marenich, Ryan M. Olson, Adam C. Chamberlin,
Christopher J. Cramer,* and Donald G. Truhlar*

*Department of Chemistry and Supercomputing Institute, University of Minnesota,
207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431*

Received June 26, 2007

Abstract: Polarization effects in aqueous and nonaqueous solutions were analyzed for nine neutral and three charged organic solutes by the SM8 universal implicit solvation model and class IV partial atomic charges based on Charge Model 4M (CM4M) with the M06-2X density functional. The CM4M partial atomic charges in neutral and ionic solutes and in the corresponding clustered solutes (supersolutes), which included one solute molecule and one or two solvent molecules, were modeled in three solvents (benzene, methylene chloride, and water) and compared to those in the gas phase. The use of the supersolute approach (microsolvation) allows one to account for charge transfer from the solute to the solvent, and we find charge transfers as large as 0.06 atomic units for neutral solutes (pyridine in water) and 0.32 atomic units for ions (methoxide anion in water). Relaxation of the electronic structure of the solute in the presence of solvent increases the polarization free energy of the neutral solutes studied here, on average, by 16% in benzene, 30% in methylene chloride, and 43% in water. The increase for the ions in water averaged 43%.

1. Introduction

The polarization of molecules as they pass from the gas phase into a condensed phase gives rise to a number of chemically interesting phenomena, e.g., changes in solute electrical multipole moments,[1−4] environmental effects on hydrogen bonds and other complexation, binding, and dissociation processes,[5−11] solvent-induced shifts in isomeric equilibria,[12−14] solvatochromic shifts,[15−22] solvent effects on circular dichroism,[23−30] and solvent effects on chemical reactivity.[14,31−42] The accurate prediction of these phenomena poses an interesting challenge to theory. One approach to this problem is to treat both the solute and a significant number of solvent molecules explicitly using suitably high levels of electronic structure theory. However, this approach is currently not practical owing to the large size of the system that is required to converge the solvent effect and the need to sample over many degrees of solvent freedom in a thermodynamically meaningful fashion. Another approach is to replace the explicit surrounding medium by a homogeneous continuum that is characterized by one or more bulk properties of the medium; for example, for the purpose of computing electrostatic phenomena, the continuum might be assigned the dielectric constant of the solvent.[12−14,42−44]

A large number of methods have been developed specifically for modeling liquid-phase polarization effects.[12−14,42−91] Many of these methods involve some sort of classical mechanical model and parametrization. One expects that a more fully quantum mechanical model based on density functional theory can be more broadly accurate,[42,43] and in the present work, we study polarization effects by density functional theory combined with the charge model CM4M, which is presented in a previous paper[92] in this issue of the journal. Additionally CM4M was developed to reproduce the gas-phase dipole moments of an extensive database of compounds using small to medium-sized basis sets. This allows the CM4M model to improve the accuracy of low-level quantum mechanical calculations without sacrificing the flexibility of quantum mechanical calculations and

* Corresponding author e-mail: cramer@chem.umn.edu (C.J.C), truhlar@umn.edu (D.G.T.).

10.1021/ct7001539 CCC: $37.00
Published on Web 10/16/2007

without introducing a significantly greater expense in computational cost.

By invoking the continuum approximation, the electronic structure problem is reduced to the size of the solute of interest. However, accurate computation of the solute polarization using continuum solvation models still poses several challenges. One important issue is that many continuum solvation models assume that the interactions of the solute and the surrounding solvent do not depend on the molecular structure of the solvent and that the dielectric response of the medium is uniform and linear at all positions outside the space that defines the solute. This assumption is particularly poor when strong, specific interactions between a solute and one or more first-shell solvent molecules are present, for example strong hydrogen bonding or $\pi-\pi$ stacking interactions. Continuum solvation models also are problematic in cases where there is significant charge transfer between the solute and the solvent; in such an instance, the solute itself is not characterized by an integral charge. This problem is particularly acute for solutes that are themselves charged or that contain a number of charged residues, as do proteins, for instance; models incorporating charge transfer have been only rarely studied.[93,94] Another concern is the applicability of continuum models to small charged species, e.g., metal ions, or to charged species with highly localized charges, e.g., oxyanions or transition-metal cations. For these cases, it is typically more appropriate to consider the first solvation shell of the ion as true ligands in a supermolecular complex,[95−97] and this suggests that supermolecular approaches incorporating explicit solvent molecules at sites having strong, specific interactions might be a general approach for improving the performance of continuum models (including more than a small number of explicit solvent molecules, however, tends to reintroduce the problem of sampling over the range of accessible conformational space).[42,98]

In addition to accounting for the effect of solvent molecularity and charge transfer on the polarization of the solute, there is the issue of how the solute charge distribution is represented. Modern quantum mechanical continuum models may work with the continuous charge distribution or with a truncated multipolar expansion of that distribution at one or many centers. For example, in the case of generalized Born continuum solvation models,[12−14,42,99] a truncated monopole expansion at the nuclear centers, i.e., atomic partial charges, is used. With such methods it is critical that the charges are physically accurate—one measure of such accuracy, since atomic charges themselves are not uniquely defined, is the degree to which the charges reproduce molecular electric moments.

Representation of a solute's charge distribution as a collection of atom-centered point charges has a very long history from a qualitative, conceptual standpoint. With respect to quantitative details, models for assigning partial atomic charges may be categorized into four broad classes. Class I models involve partial atomic charges that may be derived unambiguously from experimental data, e.g., charges assigned to reproduce the dipole moment of a diatomic molecule. Class II charge models are associated with necessarily arbitrary population analyses of a quantum mechanical wave function. Popular examples of such models include schemes based on Mulliken[100−102] and Löwdin[103−106] population analysis and natural population analysis.[107] Such charge models are sensitive to the choice of basis set, and the calculated charges are somewhat arbitrary except possibly in some cases for small, well balanced basis sets. Class III charge models assign partial atomic charges to fit a computed physical observable, e.g., electric multipole moments or the electrostatic potential at particular points around a molecule. (Such fitting problems are known to be ill conditioned, so that charges for buried atoms can be unreliable.[108,109]) Like class II charges, class III charges may show variations as a function of the molecular conformation or even the quality of the level of theory employed, although they tend to be less sensitive to basis set effects, particularly as more complete basis sets are employed. Class IV charge models are similar to class III ones in the sense that they assign partial charges in order to reproduce a physical observable, but in this case the observable is taken from experiment, not from the incomplete level of computation used for the application at hand, and the fitting involves a parametrized mapping starting from systematic class II or class III charges with mapping parameters optimized to maximize the accuracy of the charge model over a diverse training set.

The CM1,[110,111] CM2,[112,113] CM3,[114−116] and CM4[98] models are all class IV charge models, with CM4 being the most recent and robust generation of mappings. Here we apply our newest CM4 parametrization, CM4M,[92] developed for the Minnesota 2006 (M06) suite of density functionals,[117−119] to study the polarization and charge transfer in clustered and unclustered neutral and ionic solutes using the SM8 aqueous and organic continuum solvation models.[120]

We consider polarization effects in nine neutral solutes (acetic acid, benzaldehyde, chloroform, ethanol, methanethiol, methanol, nicotinamide, propionic acid, and pyridine) solvated by the three solvents, which are benzene, methylene chloride, and water. We also include three ionic solutes (acetate anion, methoxide anion, and pyridinium cation) in the set of aqueous solutes. The choice of these solutes is dictated by our intent to represent major classes of chemical compounds with various functionalities in this analysis. The set of solvents is chosen to span a range of dielectric constants[121] and other solvent descriptors such as Abraham's hydrogen bond acidity and basicity parameters[122−125] and indices[121] of refraction (see Table 1). Methylene chloride (dichloromethane) is a particularly interesting case for study because of its weak hydrogen bonding and coordinative properties.[126] To study the charge transfer between a solute and solvent molecules we replace 10 of the 12 solutes by solvent−solute clusters, also called supermolecules or supersolutes. The clusters include two solvent molecules in the case of water and only one solvent molecule in the case of benzene and methylene chloride (totally 24 clusters). This approach to microsolvation is reasonable with respect to the physical nature of intramolecular interactions between the selected solutes and solvents as well as practical with respect to computational time and choice of supersolute conformation. The gas-phase charge distribution in these solutes and

Table 1. Solvent Descriptors for the Three Solvents

descriptor	C_6H_6	CH_2Cl_2	H_2O
ϵ^a	2.27	8.93	78.36
α^b	0.00	0.10	0.82
β^c	0.14	0.05	0.38
n^d	1.5011	1.4242	1.3328
γ^e	40.62	39.15	104.71
ϕ^f	1.000	0.000	0.000
ψ^g	0.000	0.667	0.000

[a] Static dielectric constant[121] at 298 K. [b] Abraham's hydrogen bond acidity parameter[122−125] (which Abraham denotes as $\Sigma\alpha_2$). [c] Abraham's hydrogen bond basicity parameter[122−125] (which Abraham denotes as $\Sigma\beta_2$). [d] Index of refraction.[121] [e] $\gamma = \gamma_m/\gamma^o$, where γ_m is the macroscopic surface tension[121] at a liquid−air interface at 298 K expressed in cal mol^{-1}Å^{-2}, and γ^o is 1 cal mol^{-1} Å^{-2}. [f] Aromaticity: fraction of non-hydrogenic solvent atoms that are aromatic carbon atoms. [g] Electronegative halogenicity: fraction of non-hydrogenic solvent atoms that are halogens.

clusters (36 species total) was also calculated so that it may be compared to the charge distribution in solution. Thus we performed 36 gas-phase charge distribution calculations, 16 charge distribution calculations in benzene, 16 charge distribution calculations in methylene chloride, and 22 charge distribution calculations in aqueous solution.

2. Computational Methods

The geometries of selected solutes and the corresponding solvent−solute clusters are optimized with the M06-2X density functional[117,119] and the 6-31+G(d,p) basis set.[127,128] The M06-2X density functional was previously recommended for applications involving main-group thermochemistry and noncovalent interactions,[117−119] and it is especially appropriate for treating solvation in benzene because of its good ability to handle noncovalent interactions of π systems. The M06-2X/6-31+G(d,p) conformational analysis was carried out including calculation of harmonic frequencies to find the global minimum conformations in the gas phase. (Only gas-phase geometries are used in this article.) The molecular structures of the 10 solutes (out of 12) studied in this paper are shown in Figure 1. Figures 2−4 show the molecular structures of the gas-phase solute−solvent clusters for benzene, methylene chloride, and water, respectively. The Cartesian coordinates for all of the clustered solutes are given in the Supporting Information. All these calculations were carried out using *Gaussian 03*.[129]

The partial atomic charges of the solutes and supersolutes in the gas phase are calculated using Charge Model 4M (CM4M)[92] with a locally modified version[130] of the *Gaussian 03* electronic structure package.[129] CM4M, like its antecedent, CM4, is a class IV charge model that empirically maps class II charges to reproduce experimentally observable properties. As with CM4, the CM4M algorithm involves a mapping from Löwdin charges[103−106] when the basis set used to compute the electronic structure of the solute molecule does not include diffuse functions, and it uses redistributed Löwdin charges[131] when the basis set is diffuse. Also, as with its CM4 predecessor, CM4M maps the charges using an empirical scheme based upon the Mayer bond orders[132−134] between individual atoms in the solute. The primary difference between CM4 and CM4M is that while the former was designed to be generally applicable to any level of theory for a given basis set, CM4M is designed to be especially accurate for a small number of theoretical methods, i.e., the M06 methods. This approach is motivated by the observation that our chosen level of theory, the M06 suite of density functionals, has been shown to be significantly more accurate than any other density functional for a broad range of applications[117] allowing CM4M to be equally broadly applicable. A complete description of CM4M is provided in ref 92.

It is known that partial atomic charges obtained from population analysis are sensitive to basis set size, and, in particular, one can obtain unphysical charges when extended basis sets are used.[131] This is apparently a consequence of the fact that a large basis set on a given atom can mathematically describe electron density on neighboring atoms. For example, with a large, diffuse basis set, one can obtain a reasonably accurate electronic wave function for methane even with all of the basis functions centered only on carbon,[135] and either Mulliken or Löwdin analysis based on such a wave function would assign a partial charge of −4 to carbon and +1 to each of the hydrogen atoms. Since the CM4M model yielding class IV partial atomic charges uses class II charges from Löwdin (and, for diffuse basis sets, redistributed Löwdin) population analysis, and, since the population analysis is most meaningful for small basis sets, we employ the 6-31G(d)[127,128] basis set for calculation of partial atomic charges in all solutes and supersolutes and in both gas and liquid phases.

The liquid-phase partial atomic charges were calculated using the universal continuum solvation model SM8[120] with a locally modified version[130] of the *Gaussian 03* electronic structure package.[129] According to SM8, the free energy of solvation is written as

$$\Delta G_S^o = \Delta E_E + \Delta E_N + G_P + G_{CDS} + \Delta G_{conc}^o \qquad (1)$$

The first term in eq 1 refers to the energy of reorganization of the electronic structure of the solute (electronic relaxation) that is equal to the change in the solute's internal electronic (E) energy in moving from the gas phase to the liquid phase at the same geometry. The second term in eq 1 is the change in the solute's internal energy due to changes in the equilibrium nuclear (N) positions in the solute that accompany the solvation process (we call it geometry relaxation). The quantity of G_P is the free energy of polarization of solvent molecules by the solute. G_{CDS} is the portion of the free energy of solvation that is nominally associated with cavitation, dispersion, and solvent structure effects (CDS), and it is parametrized in terms of atomic surface tensions. The final term, ΔG_{conc}^o, of eq 1 is the free energy of liberation,[136] and it is zero in the present article because we use the same standard-state concentration for the gas phase as for solution. (It is conventional when this is done to say that the standard states are a 1 mol/L vapor and an ideal 1 mol/L solution, but actually the only issue that matters is that the concentration does not change.) Thus the SM8 model partitions the free energy of solvation into two contributions, one ($\Delta E_E + \Delta E_N + G_P$) arising from long-range bulk electrostatic effects and the other (G_{CDS}) from those electrostatic interactions between the solute and solvent molecules in the first solvation shell that are different from bulk electrostatic polarization and from other short-range effects

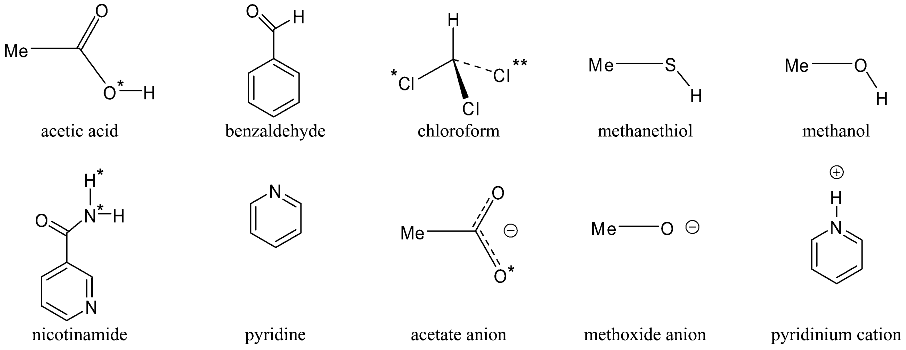


Figure 1. Molecular structures of solutes.

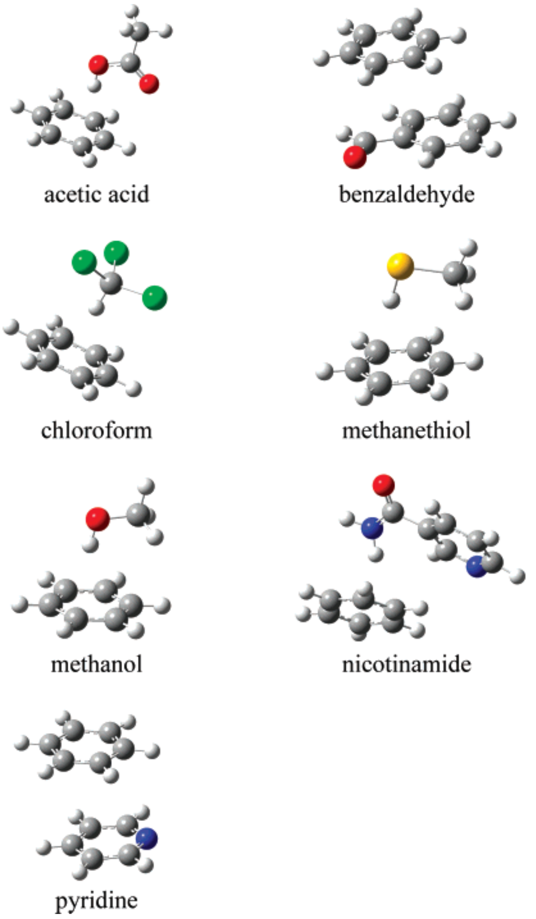


Figure 2. Solute−benzene clusters. The distances between the geometric center of the benzene ring and the geometric center of the solute ring in the benzaldehyde and pyridine clusters are 3.71 and 3.69 Å, respectively, for benzaldehyde and pyridine. The distances between the geometric center of the benzene ring and the closest hydrogen atom of the solute molecule in all other clusters are as follows (in Å): 2.25 (acetic acid), 2.19 (chloroform), 2.45 (methanethiol), 2.29 (methanol), and 2.52 (nicotinamide).

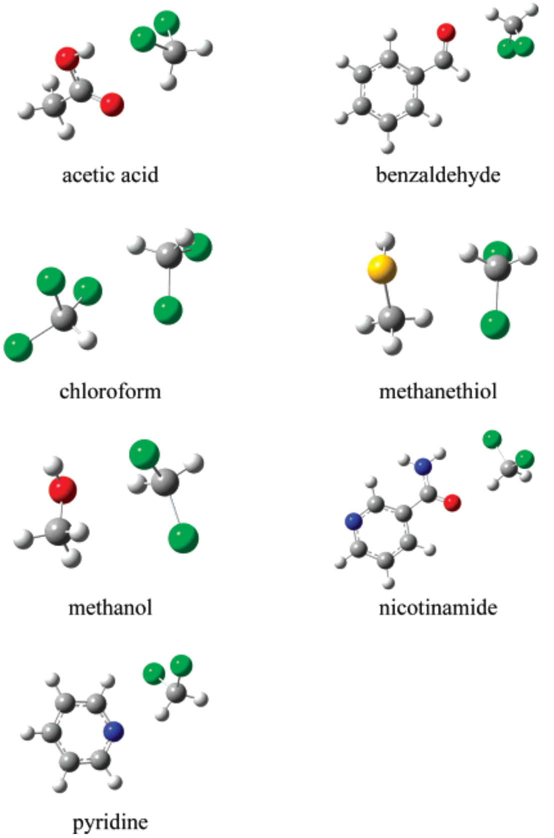


Figure 3. Solute−methylene chloride clusters. The lengths of the N···H, O···H, or S···H hydrogen bonds are as follows (in Å): 2.24 (acetic acid), 2.25 (benzaldehyde), 2.82 (methanethiol), 2.21 (methanol), 2.13 (nicotinamide), and 2.24 (pyridine). The shortest Cl···H distances for Cl in the methylene chloride and H in the solute are as follows (in Å): 2.46 (acetic acid), 3.06 (benzaldehyde), 2.83 (chloroform), 3.48 (methanethiol), 3.09 (methanol), 2.59 (nicotinamide), and 3.05 (pyridine).

beyond bulk electrostatics. Since in the present calculation we use the supermolecule (supersolute) approach in which one or two solvent molecules are treated as part of the solute, the supersolute already partially includes part of one explicit solvent shell. Therefore, the "first solvation shell" of the SM8 model actually includes part of the second solvation shell of the original solute and most of the first solvation shell.

Because solute electronic relaxation is included in the calculation of the polarization energy but not in the calculation of first-solvation-shell effects, the polarization of the solute depends on the partition of solvation effects into bulk electrostatics and first-solvation-shell effects. This partition is not unique because it depends on the choice of solute atomic radii used in the polarization calculation. These radii are part of the parametrization; however, they are not well determined by parametrizing to free energies of solvation of neutrals because the atomic surface tensions used to account for first-solvation-shell terms are very good at semiempirically making up for deficiencies in the electrostatic contributions. However the free energies of solvation of ions are very sensitive to these radii. We might hope that

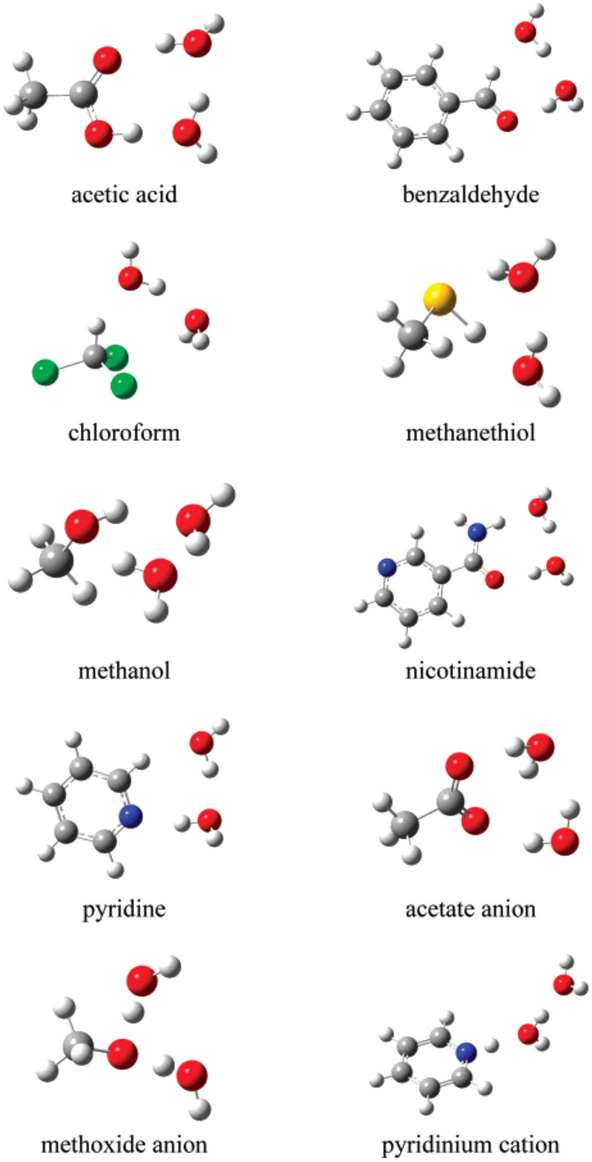


Figure 4. Solute−water clusters. The lengths of the $O\cdots H$ hydrogen bonds where H is the most polar hydrogen atom in the solute molecule are as follows (in Å): 1.63 (acetic acid), 2.21 (benzaldehyde), 2.02 (chloroform), 2.28 (methanethiol), 1.91 (methanol), 1.86 (nicotinamide), 2.24 (pyridine), and 1.56 (pyridinium cation). The shortest $H\cdots N$, $H\cdots O$, $H\cdots S$, or $H\cdots Cl$ distances where H is an aqueous hydrogen atom are as follows (in Å): 1.81 (acetic acid), 1.84 (benzaldehyde), 2.68 (chloroform), 2.38 (methanethiol), 1.88 (methanol), 1.78 (nicotinamide), 1.82 (pyridine), 1.83−1.93 (acetate anion), and 1.47−1.49 (methoxide anion). The O−O distances between the aqueous oxygen atoms in the clusters involving two water molecules are as follows (in Å): 2.68 (acetic acid), 2.78 (benzaldehyde), 2.79 (chloroform), 2.78 (methanethiol), 2.75 (methanol), 2.70 (nicotinamide), 2.77 (pyridine), 2.95 (acetate anion), 4.22 (methoxide anion), and 2.70 (pyridinium cation) (for comparison, the equilibrium O−O distance in the isolated water dimer $(H_2O)_2$ calculated at the same level of theory is 2.88 Å).

the SM8 solvation model gives a reasonable partition of solvation free energy into electrostatic and other contributions (and therefore gives a reasonable estimate of solute polarization) because SM8 is based on a very large number of ionic data in aqueous solution, and—unlike previously parametrized solvation models—it is also based on ionic solvation data in nonaqueous solvents.[120]

The bulk electrostatic contribution ($\Delta E_E + \Delta E_N + G_P$) to the total solvation free energy is calculated from a self-consistent molecular orbital calculation, where the generalized Born approximation[99,137−140] is used to compute the polarization term according to

$$G_P = \sum_k G_P(k) \qquad (2)$$

where

$$G_P(k) = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\left(q_k^{\,2}\gamma_{kk} + q_k \sum_{k'} q_{k'}\gamma_{kk'}\right) \qquad (3)$$

In the above equations, the summations go over atoms k in the solute. The quantity of ϵ is the dielectric constant of the solvent, q_k is the partial atomic charge of atom k, and $\gamma_{kk'}$ is a Coulomb integral involving atoms k and k'.

The self-consistently polarized partial atomic charges in solution differ from those obtained with the gas-phase electronic wave function even at the same geometry, and they depend on the nature of solvent. Polarization effects in solution can also be analyzed by comparison of atomic contributions to the polarization free energy (eqs 2 and 3) obtained in relaxed and unrelaxed calculations. The relaxed G_P terms are calculated using the liquid-phase electronic wave function optimized by solving the self-consistent reaction field equations. This is the G_P used in eq 1. The unrelaxed calculation uses charges obtained from the gas-phase electronic wave function. In other words, we neglect the electronic structure relaxation (polarization effect) upon solvation in the case called unrelaxed. Since the solute's geometry change upon solvation gives a much smaller contribution to the solvation free energy than the electronic structure relaxation does,[3] we use the same gas-phase geometries in both gas-phase and liquid-phase calculations, i.e., we neglect the nuclear relaxation, which means that we assume that the ΔE_N term in eq 1 is equal to 0.

3. Results

Figure 1 shows the molecular structures of the unclustered solutes studied in the present paper. The molecular structures of the solutes clustered in benzene, methylene chloride, and water are depicted in Figures 2−4, respectively. Table 2 shows the partial atomic charges and the partial charges on selected functional groups in seven neutral solutes in the gas phase and three solvents calculated using the CM4M charge model. Table 3 shows the CM4M partial atomic and group charges of acetate anion, methoxide anion, and pyridinium cation in the gas phase and water. Atomic contributions to the polarization energies are calculated by partitioning the cross terms equally between the two atoms, as is already done in eqs 2 and 3. Group contributions are calculated by summing atomic contributions for a given group. The atomic and group contributions of bare and clustered neutral solutes in the three media are listed in Tables 4 and 5, respectively. Atomic and group contributions to the polarization energies

Table 2. Partial Atomic and Group Charges of Neutral Solutes Calculated in the Gas Phase and Solution Using the CM4M Charge Model[a]

atom or group	unclustered solute				clustered solute		
	gas	C_6H_6	CH_2Cl_2	H_2O	C_6H_6	CH_2Cl_2	H_2O
			Acetic Acid				
H	0.34	0.34	0.34	0.36	0.34	0.33	0.35
O	−0.42	−0.44	−0.46	−0.49	−0.44	−0.45	−0.47
O*	−0.39	−0.39	−0.39	−0.39	−0.39	−0.39	−0.41
total charge	0.00	0.00	0.00	0.00	−0.02	0.00	−0.05
CH_3	0.08	0.09	0.10	0.11	0.08	0.10	0.10
CO	−0.03	−0.04	−0.05	−0.08	−0.05	−0.05	−0.08
COOH	−0.08	−0.09	−0.10	−0.11	−0.10	−0.10	−0.15
OH	−0.05	−0.05	−0.05	−0.03	−0.05	−0.06	−0.06
			Benzaldehyde				
H	0.04	0.05	0.06	0.07	0.05	0.06	0.09
O	−0.38	−0.40	−0.42	−0.47	−0.40	−0.40	−0.42
total charge	0.00	0.00	0.00	0.00	−0.01	0.02	0.04
C_6H_5	0.04	0.05	0.06	0.07	0.05	0.07	0.09
CO	−0.09	−0.10	−0.12	−0.14	−0.10	−0.10	−0.14
HCO	−0.04	−0.05	−0.06	−0.07	−0.06	−0.05	−0.05
			Chloroform				
Cl	−0.07	−0.07	−0.07	−0.08	−0.08	−0.08	−0.08
Cl*	−0.07	−0.07	−0.07	−0.08	−0.08	−0.07	−0.09
Cl**	−0.07	−0.07	−0.07	−0.08	−0.08	−0.07	−0.09
H	0.13	0.13	0.14	0.15	0.13	0.14	0.16
total charge	0.00	0.00	0.00	0.00	−0.03	0.00	−0.03
CH	0.20	0.21	0.22	0.23	0.21	0.21	0.23
			Methanethiol				
H	0.11	0.11	0.11	0.12	0.12	0.12	0.14
S	−0.21	−0.23	−0.24	−0.25	−0.24	−0.23	−0.25
total charge	0.00	0.00	0.00	0.00	−0.02	0.01	0.02
CH_3	0.10	0.12	0.13	0.13	0.10	0.12	0.13
SH	−0.10	−0.12	−0.13	−0.13	−0.12	−0.11	−0.11
			Methanol				
H	0.32	0.32	0.33	0.34	0.32	0.33	0.33
O	−0.48	−0.49	−0.50	−0.52	−0.49	−0.48	−0.51
total charge	0.00	0.00	0.00	0.00	−0.02	0.02	0.00
CH_3	0.16	0.17	0.17	0.18	0.15	0.17	0.18
OH	−0.16	−0.17	−0.17	−0.18	−0.17	−0.15	−0.18
			Nicotinamide				
H	0.33	0.34	0.35	0.35	0.33	0.35	0.35
H*	0.33	0.34	0.34	0.35	0.34	0.33	0.34
N	−0.42	−0.45	−0.46	−0.47	−0.45	−0.46	−0.47
N*	−0.65	−0.64	−0.64	−0.63	−0.64	−0.63	−0.62
O	−0.44	−0.47	−0.50	−0.57	−0.47	−0.48	−0.50
total charge	0.00	0.00	0.00	0.00	−0.02	0.02	0.01
C_5H_4N	0.01	0.00	0.00	0.02	0.00	0.02	0.02
CO	−0.02	−0.04	−0.06	−0.09	−0.05	−0.05	−0.08
NH_2	0.01	0.04	0.06	0.07	0.03	0.05	0.07
			Pyridine				
N	−0.42	−0.45	−0.47	−0.48	−0.45	−0.45	−0.45
total charge	0.00	0.00	0.00	0.00	0.00	0.03	0.06

[a] For atomic charges, only heteroatoms and polar hydrogen atoms are shown (see Figure 1). Group charges are summed over the atoms indicated. Total charge is the sum of the partial atomic charges of the solute.

of aqueous ions are presented in Table 6. The partial atomic and group charges and the polarization energy contributions in acetic acid and methanol are compared to those in their homologous analogs, in propionic acid and ethanol, respectively, in Tables 7 and 8. Only heteroatoms and polar hydrogen atoms of these solutes are listed in Tables 2−8, whereas the data on all atoms are given in the Supporting Information. Table 9 contains solute dipole moments. All charges are in atomic units, in which the charge on a bare proton is unity.

Table 3. Partial Atomic and Group Charges of Ions Calculated in the Gas Phase and Water Using the CM4M Charge Model[a]

atom or group	unclustered solute		clustered solute in H_2O
	gas	H_2O	
	Acetate Anion		
O	−0.60	−0.63[b]	−0.57
O*	−0.60	−0.64[b]	−0.57
total charge	−1.00	−1.00	−0.82
CH_3	−0.13	−0.05	0.00
COO	−0.87	−0.95	−0.82
	Methoxide Anion		
O	−0.81	−0.87	−0.70
total charge	−1.00	−1.00	−0.68
CH_3	−0.19	−0.13	0.02
	Pyridinium Cation		
H	0.37	0.39	0.37
N	−0.32	−0.31	−0.35
total charge	1.00	1.00	0.88

[a] See footnote *a* in Table 2. [b] The lower oxygen in Figure 4 has charge −0.64, and the higher oxygen has charge −0.63.

4. Discussion

First we analyze polarization effects in unclustered solutes. Comparison of the gas-phase partial atomic charges to those in the three solvents indicates that the charges on heteroatoms typically undergo a larger change upon a solute passing from the gas phase to solution (Tables 2 and 3) than do carbon and hydrogen atoms present in the hydrocarbon parts of these solutes. For instance, the charge on O in the unclustered benzaldehyde molecule varies from −0.38 (gas) to −0.40 in benzene (5% change), −0.42 in methylene chloride (11%), and −0.47 in water (24%) in accord with the increase of dielectric constant in the series $C_6H_6 \rightarrow CH_2Cl_2 \rightarrow H_2O$.

Analysis of the molecular structures of solute−solvent clusters (Figures 2−4) indicates a physically meaningful trend that the most polar hydrogen of one molecule is bound to the center with the most negative charge in another molecule. In the case of the benzene clusters (Figure 2), a polar hydrogen atom in the molecules of acetic acid, chloroform, methanethiol, and methanol is attracted to the nucleophilic aromatic ring. The clusters of benzaldehyde and nicotinamide with benzene are additionally stabilized by $\pi-\pi$ stacking interactions. The structure of the gas-phase water dimer is preserved in all aqueous clusters with two water molecules, except for methoxide anion (Figure 4) where the $H_3C-O^-\cdots H-OH$ bond is likely to be stronger than the $H_2O\cdots H-OH$ bond in the isolated water dimer. Addition of one or two explicit solvent molecules to the solute allows one to include (at least to some extent) the charge-transfer effect corresponding to the redistribution of the electronic density between the solute particle and the first solvation shell; this effect is not included by fully implicit solvent models (except perhaps in an average way by parametriza-

Table 4. Atomic and Group Contributions to Polarization Energy (kcal/mol) for Unclustered Neutral Solutes[a]

atom or group	C_6H_6		CH_2Cl_2		H_2O	
	unrelaxed	relaxed	unrelaxed	relaxed	unrelaxed	relaxed
			Acetic Acid			
H	−0.55	−0.54	−1.15	−1.15	−2.42	−2.80
O	−0.71	−0.80	−1.45	−1.80	−3.76	−5.31
O*	0.48	0.50	0.82	0.88	0.66	0.90
total G_P	−1.44	−1.61	−2.73	−3.29	−5.74	−7.59
CH_3	−0.44	−0.53	−0.68	−0.95	−0.70	−1.05
CO	−0.94	−1.04	−1.71	−2.07	−3.26	−4.64
COOH	−1.00	−1.08	−2.04	−2.34	−5.03	−6.54
OH	−0.06	−0.04	−0.33	−0.27	−1.77	−1.90
			Benzaldehyde			
H	−0.11	−0.15	−0.17	−0.30	−0.14	−0.32
O	−0.83	−1.01	−1.63	−2.27	−3.55	−5.95
total G_P	−1.70	−2.05	−2.89	−4.04	−4.30	−7.11
C_6H_5	−0.73	−0.89	−1.12	−1.58	−1.12	−1.69
CO	−0.87	−1.02	−1.60	−2.16	−3.04	−5.10
HCO	−0.97	−1.16	−1.76	−2.46	−3.18	−5.42
			Chloroform			
Cl	0.03	0.03	0.05	0.06	0.03	0.04
Cl*	0.03	0.03	0.05	0.06	0.03	0.04
Cl**	0.03	0.03	0.05	0.06	0.03	0.04
H	−0.31	−0.34	−0.57	−0.71	−0.84	−1.17
total G_P	−0.30	−0.33	−0.58	−0.70	−0.94	−1.28
CH	−0.39	−0.43	−0.72	−0.87	−1.04	−1.42
			Methanethiol			
H	−0.02	−0.01	−0.03	0.00	−0.04	0.00
S	−0.10	−0.13	−0.22	−0.29	−0.44	−0.62
total G_P	−0.48	−0.57	−0.79	−1.06	−1.01	−1.40
CH_3	−0.36	−0.43	−0.55	−0.77	−0.54	−0.78
SH	−0.12	−0.14	−0.24	−0.29	−0.47	−0.62
			Methanol			
H	−0.28	−0.28	−0.56	−0.58	−1.00	−1.12
O	−0.28	−0.30	−0.65	−0.72	−2.31	−2.69
total G_P	−0.84	−0.88	−1.62	−1.79	−3.58	−4.19
CH_3	−0.28	−0.30	−0.41	−0.49	−0.28	−0.38
OH	−0.56	−0.58	−1.21	−1.30	−3.30	−3.81
			Nicotinamide			
H	−2.83	−3.28	−4.26	−5.46	−3.80	−5.39
H*	−2.12	−2.39	−3.24	−3.96	−2.86	−3.68
N	−1.20	−1.44	−1.94	−2.55	−2.29	−2.65
N*	3.13	3.47	4.60	5.52	3.58	4.73
O	−0.77	−0.91	−1.69	−2.29	−4.38	−7.75
total G_P	−4.85	−5.78	−7.83	−10.53	−9.68	−14.88
C_5H_4N	−1.87	−2.24	−2.90	−3.89	−3.00	−3.95
CO	−1.15	−1.34	−2.03	−2.73	−3.59	−6.58
NH_2	−1.83	−2.20	−2.90	−3.90	−3.09	−4.35
			Pyridine			
N	−1.96	−2.44	−3.14	−4.50	−3.56	−5.19
total G_P	−2.30	−2.94	−3.58	−5.39	−3.69	−5.70

[a] For atomic contributions, only heteroatoms and polar hydrogen atoms are shown (see Figure 1). Group contributions are summed over the atoms indicated. G_P is the total polarization energy of the solute.

tion). The magnitude of the charge-transfer effect depends strongly on the nature of intermolecular (specific) interactions between the solvent and the solute. For instance, the presence of hydrogen-bonding enhances this effect. Indeed, Tables 2 and 3 show that the most significant charge transfer (up to 0.06 for neutrals and up to 0.32 in ions) is observed in aqueous clusters stabilized by hydrogen bonding (Figure 4). The magnitude of the charge transfer in ions is especially impressive so that it clearly indicates the desirability of the supersolute approach in modeling solvation effects involving ions. Although the total charge of the whole cluster is an integer and is equal to the charge of the unclustered solute, the total charge of the solute in the cluster calculated by summation over all solute's atoms need not be integral because of the charge transfer between solute and solvent. Concerning the neutral solutes, there is a slight trend in the charge transfer with respect to solute's hydrogen-bonding capability (Table 2). A stronger base (for instance, pyridine) acquires more positive charge, whereas a stronger acid (for instance, acetic acid) acquires more negative charge.

It is interesting to compare the magnitudes of the charge transfer and polarization effects. For example, the charge transfer of 0.06 and 0.04 for pyridine and benzaldehyde, respectively, in water has the same size as the largest changes in charge on any of the atoms of these solutes (except O in benzaldehyde) when polarization is considered without charge transfer. For another example, the charge transfers of 0.02−0.03 for acetic acid, chloroform, and methanol in benzene are greater than or equal to the largest pure polarization changes for any of the solute atomic charges.

The difference between the total polarization energies calculated for unclustered neutral solutes using the gas-phase electronic wave function (unrelaxed G_P) and those calculated using the liquid-phase electronic wave function (relaxed G_P) varies from 0.04 kcal/mol for methanol in benzene to 5.20 kcal/mol for nicotinamide in water (Table 4). Indeed, the latter value for nicotinamide in water comprises 37% of the magnitude of the corresponding solvation free energy (−13.95 kcal/mol) calculated by the SM8 model. For comparison, the solvation free energy of methanol in benzene calculated by the SM8 model is −2.25 kcal/mol. The solvation free energies of other solutes are listed in the Supporting Information, and they can also be compared to the corresponding G_P values. The comparison of relaxed and unrelaxed G_P values shows the importance of incorporating electronic relaxation into implicit modeling of solvation effects. Electronic relaxation is most significant for solutes in water (the most polarizable medium), where, on average, it increases the polarization free energy by 43%, and it is least important for solutes in benzene, where the average increase in polarization free energy is 16% (Table 4). Tables 4−6 show that the total polarization energy is heavily dominated by the atomic contributions from solute nitrogen and oxygen heteroatoms.

The supersolute approach leads to an apparent quenching of the polarization energy of the solute because G_P in eqs 1−3 only includes the polarization due to implicit solvent, and now some solvent is explicit. For instance, the total (relaxed) G_P energy of the unclustered pyridinium cation in water is −61.60 kcal/mol, whereas the total G_P energy of the pyridinium cation in the supersolute including two water molecules calculated by summation only over the solute's atoms is −43.40 kcal/mol. Part of the reason for this difference is the charge transfer in the clustered pyridinium

Table 5. Atomic and Group Contributions to Polarization Energy (kcal/mol) for Clustered Neutral Solutes[a]

	C_6H_6		CH_2Cl_2		H_2O	
atom or group	unrelaxed	relaxed	unrelaxed	relaxed	unrelaxed	relaxed
Acetic Acid						
H	0.16	0.19	0.03	0.09	0.93	1.24
O	−0.90	−1.00	−0.48	−0.58	−3.05	−3.89
O*	−0.14	−0.14	0.06	0.04	−1.66	−1.97
total G_P (solute)	−1.02	−1.17	−1.31	−1.60	−2.89	−3.68
total G_P (cluster)	−1.66	−1.88	−1.90	−2.31	−7.61	−8.96
CH_3	−0.35	−0.43	−0.68	−0.91	−0.53	−0.76
CO	−0.70	−0.79	−0.72	−0.83	−1.64	−2.19
COOH	−0.68	−0.74	−0.63	−0.70	−2.36	−2.92
OH	0.02	0.05	0.09	0.13	−0.73	−0.73
Benzaldehyde						
H	−0.06	−0.07	−0.09	−0.09	0.08	0.17
O	−0.80	−0.96	−0.49	−0.78	−2.00	−3.06
total G_P (solute)	−1.36	−1.61	−1.90	−2.46	−2.63	−3.67
total G_P (cluster)	−1.93	−2.26	−2.41	−3.04	−8.03	−9.49
C_6H_5	−0.61	−0.73	−1.17	−1.59	−1.11	−1.57
CO	−0.70	−0.82	−0.64	−0.79	−1.60	−2.27
HCO	−0.76	−0.88	−0.73	−0.87	−1.52	−2.09
Chloroform						
Cl	−0.03	−0.03	0.12	0.13	−0.12	−0.18
Cl*	−0.03	−0.03	0.07	0.07	0.08	0.09
Cl**	−0.03	−0.03	0.03	0.03	0.06	0.06
H	−0.07	−0.07	−0.35	−0.38	−0.03	0.00
total G_P (solute)	−0.16	−0.16	−0.25	−0.28	−0.04	−0.06
total G_P (cluster)	−0.77	−0.84	−1.03	−1.23	−5.06	−5.50
CH	−0.08	−0.08	−0.48	−0.51	−0.06	−0.03
Methanethiol						
H	0.11	0.12	−0.15	−0.14	0.48	0.58
S	−0.29	−0.33	0.19	0.20	−0.49	−0.62
total G_P (solute)	−0.39	−0.47	−0.31	−0.36	−0.10	−0.19
total G_P (cluster)	−1.12	−1.31	−1.02	−1.21	−5.37	−6.04
CH_3	−0.20	−0.26	−0.35	−0.42	−0.10	−0.14
SH	−0.19	−0.21	0.04	0.06	0.00	−0.05
Methanol						
H	0.38	0.40	−1.11	−1.15	0.95	1.21
O	−1.02	−1.08	1.00	1.04	−2.12	−2.54
total G_P (solute)	−0.66	−0.70	−0.70	−0.72	−1.37	−1.56
total G_P (cluster)	−1.44	−1.58	−1.27	−1.40	−6.19	−6.87
CH_3	−0.02	−0.03	−0.59	−0.62	−0.20	−0.23
OH	−0.64	−0.67	−0.11	−0.10	−1.17	−1.33
Nicotinamide						
H	−1.71	−1.89	−4.17	−5.12	−2.76	−3.50
H*	−1.69	−1.87	−1.62	−1.87	−0.27	−0.34
N	−1.02	−1.22	−1.46	−2.05	−2.19	−2.83
N*	2.27	2.44	3.89	4.48	1.58	1.91
O	−0.87	−1.00	−0.19	−0.26	−2.56	−3.69
total G_P (solute)	−3.79	−4.42	−6.00	−7.86	−6.57	−9.08
total G_P (cluster)	−4.19	−4.85	−6.33	−8.22	−10.57	−13.30
C_5H_4N	−1.59	−1.88	−2.96	−3.98	−3.06	−4.17
CO	−1.08	−1.22	−1.15	−1.37	−2.06	−2.99
NH_2	−1.13	−1.32	−1.90	−2.51	−1.45	−1.92
Pyridine						
N	−1.86	−2.27	−0.95	−1.42	−1.38	−2.17
total G_P (solute)	−1.99	−2.48	−2.17	−3.10	−2.14	−3.37
total G_P (cluster)	−2.66	−3.26	−2.76	−3.77	−7.96	−9.90

[a] For atomic contributions, only heteroatoms and polar hydrogen atoms are shown (see Figure 1). Group contributions are summed over the atoms indicated. G_P (solute) is the sum of the polarization energy contributions from each atom of the solute in the cluster. G_P (cluster) is the total polarization energy of the cluster (including solvent atoms of the supersolute).

Table 6. Atomic and Group Contributions to Polarization Energy (kcal/mol) for Aqueous Ions[a]

	unclustered solute		clustered solute	
atom or group	unrelaxed	relaxed	unrelaxed	relaxed
	Acetate Anion			
O	−45.75	−49.56	−32.50	−34.98
O*	−46.20	−49.74	−32.90	−34.59
total G_P (solute)	−75.68	−79.81	−49.37	−51.77
total G_P (cluster)			−60.55	−62.87
CH_3	−8.44	−3.84	−3.38	−0.90
COO	−67.25	−75.97	−45.99	−50.87
	Methoxide Anion			
O	−74.33	−81.61	−44.74	−47.50
total G_P (solute)	−87.97	−91.04	−44.91	−46.23
total G_P (cluster)			−68.31	−70.46
CH_3	−13.64	−9.43	−0.17	1.28
	Pyridinium Cation			
H	−24.95	−27.17	−17.75	−18.05
N	20.04	19.89	17.10	17.25
total G_P (solute)	−60.51	−61.60	−42.78	−43.40
total G_P (cluster)			−51.23	−51.71

[a] See footnote *a* in Table 5.

Table 7. Partial Atomic and Group Charges for Selected Homologous Analogs[a]

atom or group	gas	C_6H_6	CH_2Cl_2	H_2O
		Acetic Acid		
H	0.34	0.34	0.34	0.36
O	−0.42	−0.44	−0.46	−0.49
O*	−0.39	−0.39	−0.39	−0.39
total charge	0.00	0.00	0.00	0.00
CH_3	0.08	0.09	0.10	0.11
CO	−0.03	−0.04	−0.05	−0.08
COOH	−0.08	−0.09	−0.10	−0.11
OH	−0.05	−0.05	−0.05	−0.03
		Propionic Acid		
H	0.34	0.34	0.34	0.36
O	−0.42	−0.44	−0.46	−0.49
O*	−0.39	−0.39	−0.39	−0.39
total charge	0.00	0.00	0.00	0.00
C_2H_5	0.08	0.09	0.09	0.10
CH_3	0.03	0.04	0.04	0.04
CO	−0.03	−0.04	−0.05	−0.08
COOH	−0.08	−0.09	−0.09	−0.10
OH	−0.05	−0.04	−0.04	−0.03
		Methanol		
H	0.32	0.32	0.33	0.34
O	−0.48	−0.49	−0.50	−0.52
total charge	0.00	0.00	0.00	0.00
CH_3	0.16	0.17	0.17	0.18
OH	−0.16	−0.17	−0.17	−0.18
		Ethanol		
H	0.32	0.32	0.33	0.34
O	−0.48	−0.49	−0.49	−0.51
total charge	0.00	0.00	0.00	0.00
C_2H_5	0.16	0.16	0.17	0.17
CH_3	−0.01	−0.01	0.00	0.00
OH	−0.16	−0.16	−0.17	−0.17

[a] See footnote *a* in Table 2.

cation because the total polarization energy of the whole cluster is −51.71 kcal/mol. The other reason is that a significant amount of the polarization energy of the solute is included explicitly in the supersolute calculation.

Although Table 6 shows that polarization free energies for ions are much larger than those for neutrals in Tables 4 and 5, it also shows that the percentage increases due to electronic relaxation is much smaller, averaging only 4%. In absolute energy units, though, the effect is very large, averaging 2.8 kcal/mol, whereas for neutrals the effect averages to 1.9 kcal/mol in water, 1.0 kcal/mol in methylene chloride, and 0.3 kcal/mol in benzene.

One should keep in mind that the atomic G_P contributions calculated in the present study within the generalized Born approximation (eqs 2 and 3) are not physical observables and they can have positive values, whereas the total polarization energy (which is the sum of these contributions) should always be negative because of the spontaneous nature of polarization in solution (Tables 4−8). Since the Coulomb integral involving atoms k and k' in eq 3 is a function of the distance between k and k' and the summation in eq 3 runs over all atoms of the solute molecule, the G_P contribution from any individual atom is a function of both the molecular geometry and the partial charges of all of the atoms in the molecule. Thus it is understandable that atomic polarization energies of similar atoms in different solutes can be substantially different. For instance, the atomic polarization energy of the hydroxylic oxygen atom in acetic acid is substantially different from that of the same oxygen in methanol: cf. $G_P(O) = +0.90$ kcal/mol for acetic acid in water and $G_P(O) = -2.69$ kcal/mol for methanol in water. However, the hydroxylic oxygen in acetic acid is considerably less charged. This is due in part to the effects of the carbonyl oxygen geminal to it, while the charge on the hydroxylic hydrogen remains largely unchanged. This contributes to the apparently positive contribution to the polarization energy of the hydroxylic oxygen in acetic acid through the cross terms of eq 3. Additionally, the presence of the geminal carbonyl oxygen interferes with favorable polarization interactions of the hydroxylic oxygen with the surrounding solvent through simple dielectric descreening effects. As a consequence of these two physical phenomena it is unsurprising that the interactions between the hydroxylic oxygen in acetic acid and the surrounding solvent are less favorable than for the corresponding oxygen in methanol.

One might ask if the different polarization contributions discussed in the previous paragraph are unsystematic or if, in contrast, they are characteristics of functional groups. To examine this question, we carried out calculations on ethanol and propionic acid. We found that the trends in partial atomic charges and in atomic polarization energies obtained for methanol and acetic acid in different media are similar to those obtained for ethanol and propionic acid, respectively (Tables 7 and 8). Indeed, the charge on the most polar (hydroxylic) hydrogen atom remains unchanged within 0.04 atomic units in any of the four solutes placed in different media. The corresponding polarization contributions from the hydroxylic oxygen vary only a little within the same homologous series. We observe the same trends for polariza-

Table 8. Contributions to Polarization Energy (kcal/mol) for Selected Homologous Analogs[a]

atom or group	C_6H_6		CH_2Cl_2		H_2O	
	unrelaxed	relaxed	unrelaxed	relaxed	unrelaxed	relaxed
			Acetic Acid			
H	−0.55	−0.54	−1.15	−1.15	−2.42	−2.80
O	−0.71	−0.80	−1.45	−1.80	−3.76	−5.31
O*	0.48	0.50	0.82	0.88	0.66	0.90
total G_P	−1.44	−1.61	−2.73	−3.29	−5.74	−7.59
CH_3	−0.44	−0.53	−0.68	−0.95	−0.70	−1.05
CO	−0.94	−1.04	−1.71	−2.07	−3.26	−4.64
COOH	−1.00	−1.08	−2.04	−2.34	−5.03	−6.54
OH	−0.06	−0.04	−0.33	−0.27	−1.77	−1.90
			Propionic Acid			
H	−0.50	−0.49	−1.07	−1.08	−2.33	−2.73
O	−0.75	−0.85	−1.51	−1.86	−3.80	−5.33
O*	0.41	0.42	0.70	0.73	0.54	0.77
total G_P	−1.22	−1.34	−2.37	−2.81	−5.34	−7.02
C_2H_5	−0.26	−0.31	−0.42	−0.57	−0.48	−0.68
CH_3	−0.08	−0.10	−0.14	−0.19	−0.18	−0.26
CO	−0.86	−0.95	−1.58	−1.90	−3.08	−4.39
COOH	−0.96	−1.03	−1.95	−2.24	−4.87	−6.35
OH	−0.09	−0.08	−0.37	−0.34	−1.79	−1.96
			Methanol			
H	−0.28	−0.28	−0.56	−0.58	−1.00	−1.12
O	−0.28	−0.30	−0.65	−0.72	−2.31	−2.69
total G_P	−0.84	−0.88	−1.62	−1.79	−3.58	−4.19
CH_3	−0.28	−0.30	−0.41	−0.49	−0.28	−0.38
OH	−0.56	−0.58	−1.21	−1.30	−3.30	−3.81
			Ethanol			
H	−0.20	−0.20	−0.43	−0.43	−0.79	−0.87
O	−0.32	−0.35	−0.69	−0.79	−2.24	−2.62
total G_P	−0.66	−0.70	−1.31	−1.45	−3.10	−3.62
C_2H_5	−0.13	−0.15	−0.19	−0.23	−0.07	−0.12
CH_3	−0.01	−0.02	−0.03	−0.07	−0.11	−0.17
OH	−0.53	−0.55	−1.12	−1.22	−3.03	−3.50

[a] See footnote *a* in Table 4.

Table 9. Dipole Moments (debye) of Unclustered Neutral Solutes in the Gas Phase and Solution[a]

solute	gas (exp)	gas	C_6H_6	CH_2Cl_2	H_2O
acetic acid	1.70 ± 0.03	1.94	2.13	2.29	2.52
benzaldehyde	3.0	3.10	3.50	3.87	4.38
chloroform	1.04 ± 0.02	1.19	1.25	1.31	1.39
ethanol	1.69 ± 0.03	1.56	1.65	1.71	1.80
methanethiol	1.52 ± 0.08	1.42	1.57	1.67	1.72
methanol	1.70 ± 0.02	1.57	1.63	1.68	1.76
nicotinamide		1.94	2.12	2.30	2.75
propionic acid	1.75 ± 0.09	2.02	2.19	2.34	2.59
pyridine	2.215 ± 0.010	2.12	2.57	2.95	3.02

[a] Dipole moments are calculated using the CM4M partial atomic charges in the gas phase and in solution. The corresponding experimental gas-phase values are taken from ref 121.

tion energy contributions from individual functional groups present in these compounds.

To conclude this section, we will discuss the values of dipole moments calculated for neutral solutes in three media and in the gas phase, as presented in Table 9. We note that experimental dipole moments are readily available for gas-phase molecules but not for molecules in solution, where the dipole moment is not even uniquely defined. The water molecule is apparently the only molecule for which the effective dipole moment in the liquid phase is known from experiment.[141] Nevertheless dipole moments in solution can be calculated using various theoretical approaches, at least for unclustered solutes or other approaches that assume no charge transfer, and analysis of the theoretical values seems to be useful for better understanding polarization effects in solution.

The increase of the magnitudes of dipole moments in solution in comparison with those in the gas phase for all of the solutes is similar to the results of previous work.[3] For instance, the dipole moment of pyridine in water is 42% larger than that in the gas phase. This indicates that the change of the electronic structure of the solute upon passing from the gas phase to solution (called electronic relaxation) is significant and cannot be neglected in modeling solvation effects in various systems. Solutes in water are more polar than the same solutes in methylene chloride, whereas they are more polar in methylene chloride than in benzene.

Although the present article has focused on small solutes in liquid solvents, similar polarization effects also occur for more complex situations. For example, very large polarization effects have been observed (computationally) for substrates in enzymes[142,143] and for proteins in water.[144]

5. Conclusions

The role of polarization effects in liquid-phase solution was studied by employing the new SM8 universal solvation model. Using SM8, the bulk electrostatic contribution to the total solvation free energy (this term contains the change in the internal energy of the solute upon solvation, the free energy of polarization of solvent molecules by the solute, and the free energy cost of polarizing the solvent) is calculated from a self-consistent molecular orbital calculation, where the generalized Born approximation is used to compute the polarization term using class IV partial atomic charges self-consistently polarized in solution.

We consider polarization effects in nine neutral solutes (acetic acid, benzaldehyde, chloroform, ethanol, methanethiol, methanol, nicotinamide, propionic acid, and pyridine) and in three solvents, namely benzene, methylene chloride, and water. We also include three ionic solutes (acetate anion, methoxide anion, and pyridinium cation) in the set of aqueous solutes. These solutes are chosen to include major organic chemical functionalities in this analysis. The results indicate the importance of electronic relaxation in solvation effects. Electronic relaxation is most significant for solutes in water (the most polarizable medium), and it is least important in benzene.

To study the charge transfer between a solute and solvent molecules we replace 10 of the 12 solutes by solvent−solute clusters, also called supermolecules or supersolutes. The clusters include two solvent molecules in the case of water and only one solvent molecule in the case of benzene and methylene chloride (totally 24 clusters). The magnitude of the charge transfer in ions is especially large, and it indicates the importance of including charge transfer in modeling solvation effects involving ions. Although there is currently considerable interest in including implicit polarization, the present study shows that in some cases the explicit treatment of charge transfer is equally or more important. The most significant charge transfer (up to 0.06 for neutrals and up to 0.32 for ions) is observed in aqueous clusters stabilized by hydrogen bonding.

Acknowledgment. This work was supported by the National Institutes of Health training grant for Neurophysical-computational Sciences (traineeship to A.C.C.), by the National Science Foundation (grant CHE06-10183 and grant CHE07-04974), by the Office of Naval Research under grant No. 00014-05-01-0538, and by the U.S. Army Research Office under the Multidisciplinary Research Program of the University Research Initiative through grant No. DAAD19-02-1-0176.

Supporting Information Available: Complete sets of partial atomic charges and atomic contributions to polarization energy for clustered and unclustered neutral solutes in benzene, methylene chloride, and water and for clustered and unclustered aqueous ions; solvation free energy components of the SM8 model; and the Cartesian coordinates of the solutes optimized at the M06-2X/6-31+G(d,p) level of theory. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Alagona, G.; Ghio, C.; Igual, J.; Tomasi, J. *J. Am. Chem. Soc.* **1989**, *111*, 3417.

(2) Alagona, G.; Ghio, C.; Igual, J.; Tomasi, J. *J. Mol. Struct.: THEOCHEM* **1990**, *204*, 253.

(3) Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1992**, *198*, 74.

(4) Luque, F. J.; Orozco, M.; Bhadane, P. K.; Gadre, S. R. *J. Chem. Phys.* **1994**, *100*, 6718.

(5) Cieplak, P.; Kollman, P. A. *J. Phys. Chem.* **1988**, *110*, 3734.

(6) Lim, C.; Bashford, D.; Karplus, M. *J. Am. Chem. Soc.* **1991**, *95*, 5610.

(7) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.

(8) Florián, J.; Šponer, J.; Warshel, A. *J. Phys. Chem. B* **1999**, *103*, 884.

(9) Meot-Ner, M.; Elmore, D. E.; Scheiner, S. *J. Am. Chem. Soc.* **1999**, *121*, 7625.

(10) Aquino, A. J. A.; Tunega, D.; Haberhauer, G.; Gerzabek, M. H.; Lischka, H. *J. Phys. Chem. A* **2002**, *106*, 1862.

(11) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558.

(12) Cramer, C. J.; Truhlar, D. G. In *Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P., Murray, J. S., Eds.; Elsevier: Amsterdam, 1994; p 9.

(13) Cramer, C. J.; Truhlar, D. G. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 6, p 1.

(14) Cramer, C. J.; Truhlar, D. G. In *Solvent Effects and Chemical Reactivity*; Tapia, O., Bertrán, J., Eds.; Kluwer: Boston, 1996; p 1.

(15) Mikkelsen, K. V.; Jørgensen, P.; Jensen, H. J. A. *J. Chem. Phys.* **1994**, *100*, 6597.

(16) Mikkelsen, K. V.; Sylvester-Hvid, K. O. *J. Phys. Chem.* **1996**, *100*, 9116.

(17) Cammi, R.; Mennucci, B. *J. Chem. Phys.* **1999**, *110*, 9877.

(18) Christiansen, O.; Mikkelsen, K. V. *J. Chem. Phys.* **1999**, *110*, 8348.

(19) Li, J.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **2000**, *77*, 264.

(20) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708.

(21) Cammi, R.; Frediani, L.; Mennucci, B.; Ruud, K. *J. Chem. Phys.* **2003**, *119*, 5818.

(22) Caricato, M.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2004**, *108*, 6248.

(23) Ruiz-Lopez, M. F.; Rinaldi, D. *Chem. Phys.* **1984**, *86*, 367.

(24) Ruiz-Lopez, M. F.; Rinaldi, D.; Rivail, J. L. *Chem. Phys.* **1986**, *110*, 403.

(25) Furche, F.; Ahlrichs, R.; Wachsmann, C.; Weber, E.; Sobanski, A.; Vögtle, F.; Grimme, S. *J. Am. Chem. Soc.* **2000**, *122*, 1717.

(26) Autschbach, J.; Ziegler, T.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 6930.

(27) Diedrich, C.; Grimme, S. *J. Phys. Chem. A* **2003**, *107*, 2524.

(28) Kongsted, J.; Hansen, A. E.; Pedersen, T. B.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *Chem. Phys. Lett.* **2004**, *391*, 259.

(29) Kongsted, J.; Pedersen, T. B.; Osted, A.; Hansen, A. E.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2004**, *108*, 3632.

(30) Pecul, M.; Marchesan, D.; Ruud, K.; Coriani, S. *J. Chem. Phys.* **2005**, *122*, 024106.

(31) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1994**, *116*, 3892.

(32) Barrows, S. E.; Cramer, C. J.; Truhlar, D. G.; Elovitz, M. S.; Weber, E. J. *Environ. Sci. Technol.* **1996**, *30*, 3028.

(33) Moliner, V.; Castillo, R.; Safont, V. S.; Oliva, M.; Bohn, S.; Tuñón, I.; Andrés, J. *J. Am. Chem. Soc.* **1997**, *119*, 1941.

(34) Truong, T. N. *Int. Rev. Phys. Chem.* **1998**, *17*, 525.

(35) Chuang, Y.-Y.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **1998**, *70*, 887.

(36) Chuang, Y.-Y.; Radhakrishnan, M. L.; Fast, P. L.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 4893.

(37) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187.

(38) Patterson, E. V.; Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **2001**, *123*, 2025.

(39) Mo, S. J.; Vreven, T.; Mennucci, B.; Morokuma, K.; Tomasi, J. *Theor. Chem. Acc.* **2004**, *111*, 154.

(40) Moreau, Y.; Loos, P.-F.; Assfeld, X. *Theor. Chem. Acc.* **2004**, *112*, 228.

(41) Tondo, D. W.; Pliego, J. R., Jr. *J. Phys. Chem. A* **2005**, *109*, 507.

(42) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(43) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(44) Rivail, J. L.; Rinaldi, D. In *Computational Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific: Singapore, 1996; Vol. 1, p 139.

(45) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Comm.* **2005**, *172*, 69.

(46) Wu, Y.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*, 024503/1.

(47) Drude, P. *The Theory of Optics*; Longmans, Green, and Co.: New York, 1902.

(48) Dias, L. G.; Shimizu, K.; Farah, J. P. S.; Chaimovich, H. *Chem. Phys.* **2002**, *282*, 237.

(49) Born, M.; Huang, K. *Dynamical Theory of Crystal Lattices*; Oxford University Press: U.K., 1954.

(50) Dick, B. G., Jr.; Overhauser, A. W. *Phys. Rev.* **1958**, *112*, 90.

(51) Cao, J.; Berne, B. J. *J. Chem. Phys.* **1993**, *99*, 6998.

(52) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.

(53) Rick, S. W.; Stuart, S. J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 2002; Vol. 18, p 89.

(54) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27.

(55) Jordan, P. C.; van Maaren, P. J.; Mavri, J.; van der Spoel, D.; Berendsen, H. J. C. *J. Chem. Phys.* **1995**, *103*, 2272.

(56) Stuart, S. J.; Berne, B. J. *J. Phys. Chem.* **1996**, *100*, 11934.

(57) de Leeuw, N. H.; Parker, S. C. *Phys. Rev. B* **1998**, *58*, 13901.

(58) van Maaren, P. J.; van der Spoel, D. *J. Phys. Chem. B* **2001**, *105*, 2618.

(59) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.

(60) Yu, H.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221.

(61) Straatsma, T. P.; McCammon, J. A. *Mol. Simul.* **1990**, *5*, 181.

(62) Yu, H.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *121*, 9549.

(63) Yu, H.; Geerke, D. P.; Liu, H.; van Gunsteren, W. F. *J. Comput. Chem.* **2006**, *27*, 1494.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(65) Teleman, O.; Jönsson, B.; Engström, S. *Mol. Phys.* **1987**, *60*, 193.

(66) Tironi, I. G.; Brunne, R. M.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1996**, *250*, 19.

(67) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *115*, 10758.

(68) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.

(69) Liu, Y.-P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739.

(70) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741.

(71) Chelli, R.; Ciabatti, S.; Cardini, G.; Righini, R.; Procacci, P. *J. Chem. Phys.* **1999**, *111*, 4218.

(72) Chelli, R.; Procacci, P.; Righini, R.; Califano, S. *J. Chem. Phys.* **1999**, *111*, 8569.

(73) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730.

(74) Chen, B.; Xing, J.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 2391.

(75) Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2001**, *105*, 11470.

(76) Rick, S. W. *J. Chem. Phys.* **2001**, *114*, 2276.

(77) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237.

(78) Tabacchi, G.; Mundy, C. J.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **2002**, *117*, 1416.

(79) Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7829.

(80) Donchev, A. G.; Galkin, N. G.; Illarionov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8613.

(81) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233.

(82) Price, S. L.; Stone, A. J.; Alderton, M. *Mol. Phys.* **1984**, *52*, 987.

(83) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065.

(84) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047.

(85) Ángyán, J. G.; Chipot, C. *Int. J. Quantum Chem.* **1994**, *52*, 17.

(86) Price, S. L. *Faraday Trans.* **1996**, *92*, 2997.

(87) Dang, L. X.; Chang, T.-M. *J. Chem. Phys.* **1997**, *106*, 8149.

(88) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

(89) Tsiper, E. V. *Phys. Rev. Lett.* **2005**, *94*, 013204/1.

(90) Dyer, P. J.; Cummings, P. T. *J. Chem. Phys.* **2006**, *125*, 144519/1.

(91) Tan, Y.-H.; Luo, R. *J. Chem. Phys.* **2007**, *126*, 094103.

(92) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2046−2054.

(93) van der Vaart, A.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1999**, *121*, 9182.

(94) Gogonea, V.; Merz, K. M., Jr. *J. Phys. Chem. B* **2000**, *104*, 2117.

(95) Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L. *Inorg. Chem.* **1996**, *35*, 4694.

(96) Uudsemaa, M.; Tamm, T. *Chem. Phys. Lett.* **2004**, *400*, 54.

(97) Jaque, P.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 5783.

(98) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(99) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

(100) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 564.

(101) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(102) Mulliken, R. S. *J. Chem. Phys.* **1962**, *36*, 3428.

(103) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365.

(104) Golebiewski, A.; Rzeszowska, E. *Acta Phys. Pol., A* **1974**, *45*, 563.

(105) Baker, J. *Theor. Chim. Acta* **1985**, *68*, 221.

(106) Kar, T.; Sannigrahi, A. B.; Mukherjee, D. C. *J. Mol. Struct.: THEOCHEM* **1987**, *153*, 93.

(107) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.

(108) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367.

(109) Francl, M. M.; Chirlian, L. E. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley: New York, 2000; Vol. 14, p 1.

(110) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

(111) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chim. Phys.* **1997**, *94*, 1448.

(112) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 1820.

(113) Li, J.; Williams, B.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 724.

(114) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 10707.

(115) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291.

(116) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2005**, *113*, 133.

(117) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* In press.

(118) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1.

(119) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(120) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011−2033.

(121) *CRC Handbook of Chemistry and Physics*; Lide, D. R., Ed.; Taylor and Francis: Boca Raton, FL, 2007; Vol. 87. (Internet Version 2007, http://www.hbcpnetbase.com).

(122) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. *J. Chem. Soc., Perkin Trans. II* **1989**, 699.

(123) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.

(124) Abraham, M. H. *J. Phys. Org. Chem.* **1993**, *6*, 660.

(125) Abraham, M. H. In *Quantitative Treatment of Solute/Solvent Interactions*; Theoretical and Computational Chemistry Series Vol. 1; Politzer, P., Murray, J. S., Eds.; Elsevier: Amsterdam, 1994; p 83.

(126) Visentin, T.; Kochanski, E.; Moszynski, R.; Dedieu, A. *J. Phys. Chem. A* **2001**, *105*, 2031.

(127) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(128) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(129) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revisions C.01, C.02, and D.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(130) Olson, R. M.; Marenich, A. V.; Chamberlin, A. C.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Hawkins, G. D.; Winget, P.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J. *MN-GSM, version 2007-beta*; University of Minnesota: Minneapolis, MN 55455-0431, 2007.

(131) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Comm.* **2002**, *5*, 117.

(132) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.

(133) Mayer, I. *Chem. Phys. Lett.* **1985**, *117*, 396.

(134) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 73.

(135) Hoyland, J. R. *J. Chem. Phys.* **1967**, *47*, 3556.

(136) Ben-Naim, A. *Solvation Thermodynamics*; Plenum: New York, 1987.

(137) Daudel, R. *Quantum Theory of Chemical Reactivity*; Reidel: Dordrecht, 1973.

(138) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(139) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(140) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.

(141) Badyal, Y. S.; Saboungi, M.-L.; Price, D. L.; Shastri, S. D.; Haeffner, D. R.; Soper, A. K. *J. Chem. Phys.* **2000**, *112*, 9206.

(142) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *J. Mol. Biol.* **2003**, *327*, 549.

(143) Hensen, C.; Hermann, J. C.; Nam, K.; Ma, S.; Gao, J.; Höltje, H.-D. *J. Med. Chem.* **2004**, *47*, 6673.

(144) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III *J. Comput. Chem.* **2004**, *25*, 1504.

CT7001539

JCTC Journal of Chemical Theory and Computation

Theoretical Study of Aqueous Solvation of K^+ Comparing ab Initio, Polarizable, and Fixed-Charge Models

Troy W. Whitfield,[†] Sameer Varma,[§] Edward Harder,[||] Guillaume Lamoureux,[‡]
Susan B. Rempe,[*,§] and Benoit Roux[*,||]

Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6323, Computational Bioscience Department, Sandia National Laboratories, Albuquerque, New Mexico 87185, and Department of Biochemistry, University of Chicago, Chicago, Illinois 60615

Received July 11, 2007

Abstract: The hydration of K^+ is studied using a hierarchy of theoretical approaches, including ab initio Born−Oppenheimer molecular dynamics and Car−Parrinello molecular dynamics, a polarizable force field model based on classical Drude oscillators, and a nonpolarizable fixed-charge potential based on the TIP3P water model. While models based more directly on quantum mechanics offer the possibility to account for complex electronic effects, polarizable and fixed-charges force fields allow for simulations of large systems and the calculation of thermodynamic observables with relatively modest computational expense. A particular emphasis is placed on investigating the sensitivity of the polarizable model to reproduce key aspects of aqueous K^+, such as the coordination structure, the bulk hydration free energy, and the self-diffusion of K^+. It is generally found that, while the simple functional form of the polarizable Drude model imposes some restrictions on the range of properties that can simultaneously be fitted, the resulting hydration structure for aqueous K^+ agrees well with experiment and with more sophisticated computational models. All the computational models yield a similar hydration structure, with a first peak in the radial distribution function near 2.7 Å, though the distribution functions obtained from the two ab initio simulations are less sharply peaked. A counterintuitive result, seen in Car−Parrinello molecular dynamics and in simulations with the Drude polarizable force field, is that the average induced molecular dipole of the water molecules within the first hydration shell around K^+ is slightly smaller than the corresponding value in the bulk. In final analysis, the perspective of K^+ hydration emerging from the various computational models is broadly consistent with experimental data, though at a finer level there remain a number of issues that should be resolved to further our ability in modeling ion hydration accurately.

I. Introduction

Small ions such as K^+ and Na^+ play a ubiquitous role in biology. For this reason, understanding how they are solvated by water molecules remains an issue of great relevance. A powerful approach to investigate ion solvation is to rely on computer simulations of atomic models based on potential functions.[1−5] For meaningful simulation studies it is important to use models that represent the microscopic interactions as accurately as possible. In the past few decades, a number of fixed-charge nonpolarizable force fields have been parametrized to model ion solvation[6−9] and are now used on a regular basis to investigate diverse problems. Induced

* Corresponding author e-mail: roux@uchicago.edu (B.R.), slrempe@sandia.gov (S.B.R.).

[†] Argonne National Laboratory.

[‡] University of Pennsylvania.

[§] Sandia National Laboratories.

[||] University of Chicago.

10.1021/ct700172b CCC: $37.00
Published on Web 10/27/2007

electronic polarization, which is generally neglected in standard molecular dynamics simulations of biomolecular systems, remains of particular concern in the case of ionic systems where nonadditive many-body effects could be important. In principle, accurate computational models can be developed, validated, and refined by comparing with experimental data (gas and bulk phase) as well as high level ab initio computations. In practice, however, this presents a difficult challenge for a number of reasons.

The individual microscopic interactions that are involved in ion hydration are most directly probed by single-ion thermochemical gas-phase experimental data on small water clusters.[10−12] Nonetheless, how this information must be extrapolated to the bulk phase is uncertain, because the properties of small clusters can be both similar and different from their bulk counterparts. Interpretation of experimental data about ions in the bulk phase is also not without any difficulties. Analysis of the neutron scattering data used to measure the coordination structure of Na^+ and K^+ in liquid water must rely on simulation models to determine the partial radial distribution functions.[13] These problems are reflected in the lack of consensus concerning the structural properties of hydrated ions, especially their hydration numbers.[14] An additional piece of information in developing meaningful ion solvation models is the experimentally measured hydration free energies. Experimental determination of the hydration free energies of charged species is a challenging problem that has been revisited numerous times over the years.[4,5,12,15−20] Single ion solvation properties in the infinite dilution limit must be extracted from electrochemical data using extra-thermodynamic assumptions, which are uncertain.[5] These difficulties are further compounded by the fact that, in a real physical system, the total reversible work to take an ion from the gas phase and transfer it into a bulk liquid phase includes a contribution from the electrostatic potential associated with the vacuum/liquid interface. The currently available experimental data are, by themselves, insufficient to establish a definitive picture of the solvation of simple ions such as K^+ and Na^+ in water.

Computations can be used to extend the information extracted from experiments. Because they can account for a wide range of complex electronic effects, simulations based on quantum mechanical ab initio methods offer an important source of information to deepen and extend our knowledge of ion solvation. However, bulk phase ab initio simulations are computationally intensive and can be burdened by finite size effects, short sampling time, and any approximations inherent to the treatment of electron correlation. In the particular case of density functional theory (DFT), approximations in available exchange-correlation functionals and the neglect of van der Waals dispersive attraction must also be kept in mind.[21−23] Alternatively, simulations based on physically realistic classical potential functions offer a path for estimating statistically converged thermodynamic averages, in terms of size and configurational sampling, although the validity of the simplifying assumptions upon which these potential functions are constructed must be assessed. In spite of these difficulties, it is our hope that a well-defined (if not definitive) perspective on the aqueous solvation of small ions can emerge by critically examining and contrasting data from simulations and experiments.

In the present effort, aqueous solvation of K^+ is investigated using a hierarchy of computational approaches. This includes two quantum mechanical ab initio simulation methods, Born−Oppenheimer molecular dynamics (BOMD) and Car−Parrinello molecular dynamics (CPMD), as well as two classical force field methods, TIP3P, a widely used nonpolarizable effective fixed charge model,[8,24] and SWM4-NDP, a polarizable model based on classical Drude oscillators.[25] The polarizable model of ion solvation presented here is based upon the classical Drude oscillator.[26−31] In this model, electronic induction is represented by the displacement of a charge-carrying auxiliary particle harmonically bound to a polarizable atom under the influence of the local electric field. The familiar self-consistent field (SCF) regime of induced polarization is reproduced in molecular dynamics simulations if the classical Drude oscillators are kept near their local energy minima for a given configuration of the atoms in the system.[31]

In the following, the ability of the models to reproduce the single-ion thermochemical gas-phase data in small clusters is examined. In addition, a particular emphasis is placed on examining the sensitivity in the Drude model of key aspects of aqueous solvation of K^+, such as the coordination structure, the hydration free energy, and the coefficient of self-diffusion. It is found that, while the simple functional form of potential functions imposes some restrictions on the range of properties that can simultaneously be fitted, the resulting hydration structure for aqueous K^+ is in broad accord with experiment and with ab initio simulations. In conclusion, MD studies based on properly parametrized models can yield meaningful results, although there remain a number of small discrepancies that shall be critically examined.

II. Methods

The hydration of K^+ was studied using four distinct computational models, the details of which are outlined below. The four computational models are as follows: (i) a fixed charge model based upon the TIP3P[24] water model, (ii) a Drude polarizable model based upon the SWM4-NDP water model,[25] (iii) a density functional theory (DFT) model based upon the gradient-corrected PW91 approximate density functional,[32,33] and (iv) a second DFT model using the gradient-corrected BLYP approximate density functional.[34,35] In all periodic simulations with a net charge, a uniform canceling background charge is assumed.

A. Fixed Charge Model. The fixed charge model of aqueous K^+ is based on the Lennard-Jones parameters that were previously optimized[8] to give reasonable monohydrate energy and hydration free energies for K^+ when used in conjunction with the TIP3P water model;[24] the parameters for K^+ are $E_{\min} = 0.0870$ kcal/mol, and $\sigma = 2.142645$ Å assuming a Lorentz−Berthelot combination rule with the TIP3P parameters. A system consisting of a box of 500 TIP3P water molecules and a single K^+ ion was simulated with periodic boundary conditions. Long-range electrostatic interactions were computed using Ewald summation.[36] The canonical ensemble was simulated using Nosé-Hoover thermostats[37] and a 1 fs time-step. The internal geometry of the TIP3P water molecule was fixed using the SHAKE[38] algorithm. After an initial equilibration of 100 ps, equilibrium properties were averaged over a 1 ns molecular dynamics simulation.

B. Drude Polarizable Model. The model for K^+ is consistent with the recently developed SWM4-NDP polarizable water model with a negatively charged Drude oscillator bound to its oxygen site.[25] The SWM4-NDP potential reproduces most properties of bulk water under ambient conditions (density, vaporization enthalpy, radial distribution function, dielectric constant, self-diffusion constant, shear viscosity, and free energy of hydration). In particular, the SWM4-NDP model yields a correct static dielectric constant, which makes it appropriate to study systems dominated by water-mediated electrostatic interactions. Accordingly, polarization of the cation is represented with a negatively charged particle bound to its nucleus. All atomic dispersion and electronic overlap effects are represented in a pairwise additive way using the Lennard-Jones potential.

The interaction energy of a single ion of charge q_{ion} with N water molecules is

$$U_{\text{iw}}(\mathbf{r}_{is},\mathbf{r},\mathbf{r}_{\text{D}}) = \frac{1}{2}k_{\text{D}}|\mathbf{r}-\mathbf{r}_{\text{D}}|^2 + \sum_{i=1}^{N}\sum_{s=1}^{4}\left[\frac{(q_{\text{ion}}-q_{\text{D}})q_s}{|\mathbf{r}-\mathbf{r}_{is}|} + \frac{q_{\text{D}}q_s}{|\mathbf{r}_{\text{D}}-\mathbf{r}_{is}|}\right] + \sum_{i=1}^{N}4\epsilon_{\text{ion}-\text{O}}\left[\left(\frac{\sigma_{\text{ion}-\text{O}}}{|\mathbf{r}-\mathbf{r}_{i\text{O}}|}\right)^{12} - \left(\frac{\sigma_{\text{ion}-\text{O}}}{|\mathbf{r}-\mathbf{r}_{i\text{O}}|}\right)^{6}\right] \quad (1)$$

where the vectors $\mathbf{r}$ and $\mathbf{r}_{\text{D}}$ are the positions of the ionic core and the ionic Drude particle, respectively. The ionic core has a charge ($q_{\text{ion}} - q_{\text{D}}$) and the Drude particle has a charge q_{D}. The spring constant k_{D} is set to 1000 kcal/mol/Å^2 for all Drude oscillators in the system. This value dictates the magnitude of the charge the Drude particle should carry to produce an ionic polarizability α, i.e., $q_{\text{D}} = -\sqrt{\alpha k_{\text{D}}}$.[25] In eq 1, the vector $\mathbf{r}_{is}$ is the position of the interaction site s of water molecule i. The SWM4-NDP water model comprises five sites: the oxygen atom "O'' (charge $= -q_{\text{D}}$), the hydrogen atoms "H_1'' and "H_2'' (charged), a massless site "M'' (charged), and a Drude particle "D'' attached to the oxygen atom (negatively charged). The Lennard-Jones parameters for the ion−water oxygen interaction are determined via the Lorentz−Berthelot combination rule,[39] $\epsilon_{\text{ion}-\text{O}} = \sqrt{\epsilon_{\text{ion}}\epsilon_{\text{O}}}$ and $\sigma_{\text{ion}-\text{O}} = (\sigma_{\text{ion}} + \sigma_{\text{O}})/2$. The parameters for K^+ were chosen to give agreement with experimental monohydrate properties[10] and a hydration free energy for the cation that was consistent with published values.[5,40,41]

The simulation protocol for studying the bulk hydration structure of the polarizable Drude model is identical to that of the fixed charge model, except that a dual thermostat scheme was used to keep the Drude particles at a low temperature (1 Kelvin) and therefore close to the (self-consistent field) ground state.[31]

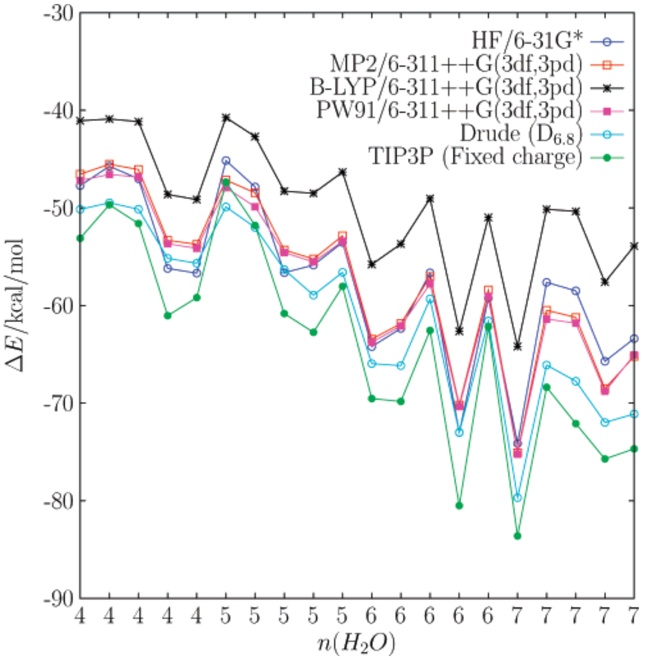


Figure 1. Interaction energies for a series of K^+ $(H_2O)_n$ clusters at various levels of ab initio theory and for a fixed charge and Drude polarizable model. Each cluster was extracted from MD simulation of aqueous K^+. The *x*-axis indexes the number of water molecules, *n*, coordinating the cation.

The adjustable parameters for monatomic ions within the classical Drude scheme to build a polarizable biomolecular force field are the Lennard-Jones parameters of the ion, σ_{ion} and ϵ_{ion}. Rather than try to determine these parameters by scanning in the space of $\{\sigma_{\text{ion}},\epsilon_{\text{ion}}\}$, it has proved more convenient to explore the space of monohydrate interaction energies and minimum-energy ion-oxygen distances $\{U_{\text{min}},d_{\text{min}}\}$.[5] Furthermore, quadratic response functions are fitted to the data from explicit computations, defined by coordinates in $\{U_{\text{min}},d_{\text{min}}\}$, to interpolate predicted properties between simulated models.[5,25] A set of polarizable models for K^+ were thus constructed by determining the Lennard-Jones parameters spanning a regular grid in the $\{U_{\text{min}},d_{\text{min}}\}$ coordinates. For each model on the grid, MD simulations were then carried out to compute the aqueous bulk hydration number, $n(r_c)$, and the bulk hydration free energy, ΔG_{hydr}. These properties were then fitted to a polynomial response function with a quadratic dependence on $\{U_{\text{min}},d_{\text{min}}\}$. The results of these computations are summarized in Figures 2 and 3. To monitor consistency between K^+ and Na^+ models, the hydration free energy of Na^+ is also reported in Figure 4 for a set of Na^+ polarizable Drude models.

The hydration free energy of the ions was decomposed into three contributions[42]

$$\Delta G_{\text{hydr}} = \Delta G_{\text{hydr}}^{\text{rep}} + \Delta G_{\text{hydr}}^{\text{disp}} + \Delta G_{\text{hydr}}^{\text{elec}} \quad (2)$$

where $\Delta G_{\text{hydr}}^{\text{rep}}$ and $\Delta G_{\text{hydr}}^{\text{disp}}$ are the repulsive and attractive (dispersive) components, respectively, of the Lennard-Jones interaction in eq 1. The electrostatic component of the hydration free energy is $\Delta G_{\text{hydr}}^{\text{elec}}$. Each component of the total hydration free energy was computed from independent

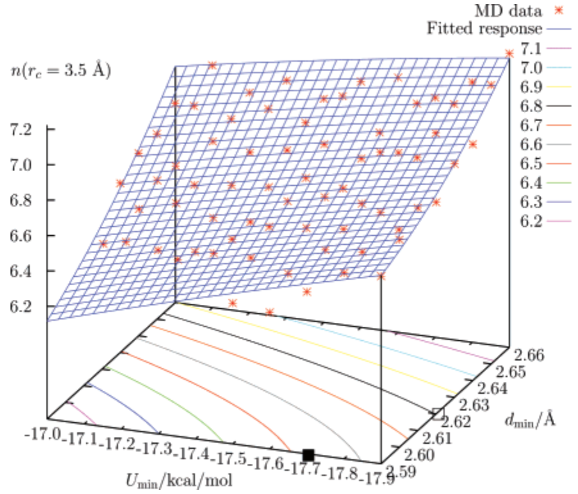


Figure 2. Coordination number, $n(r_c = 3.5$ Å$)$, for a family of putative K^+ ions as a function of monohydrate properties for the polarizable model. The open square (□) indicates the location of the $D_{6.8}$ model, while the filled square (■) indicates that of the $D_{6.5}$ model.

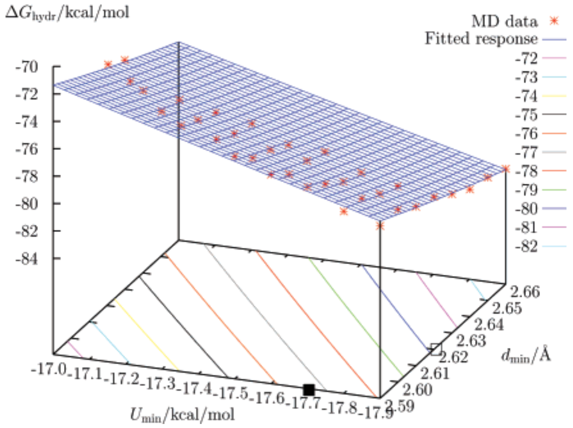


Figure 3. Computed hydration free energy for a family of putative K^+ ions as a function of monohydrate properties for the polarizable model. The open square (□) indicates the location of the $D_{6.8}$ model, while the filled square (■) indicates that of the $D_{6.5}$ model.

simulations in which an ion was placed at the center of a droplet of 200 explicit SWM4-NDP water molecules, contained by the reactive spherical solvent boundary potential (SSBP).[8] The repulsive contribution, ΔG^{rep}_{hydr}, was computed using a soft-core scheme as described elsewhere[42] and was unbiased using the weighted histogram analysis method (WHAM),[43] while ΔG^{disp}_{hydr} and ΔG^{elec}_{hydr} were computed using thermodynamic integration (TI). In discussions of the hydration free energies of ionic species, one may consider the *real* physical value, which includes the contribution of the phase potential arising from crossing the physical air/water interface, and the *intrinsic* bulk-phase value, which is independent of any interfacial potential.[5,18] Because the interfacial potential in SSBP is nearly identical to the one from a simulation of a vacuum-liquid interface,[5] the charging free energy computed with SSBP effectively includes the interfacial potential contribution that an ion gains by crossing the physical interface from the gas phase to the bulk water.

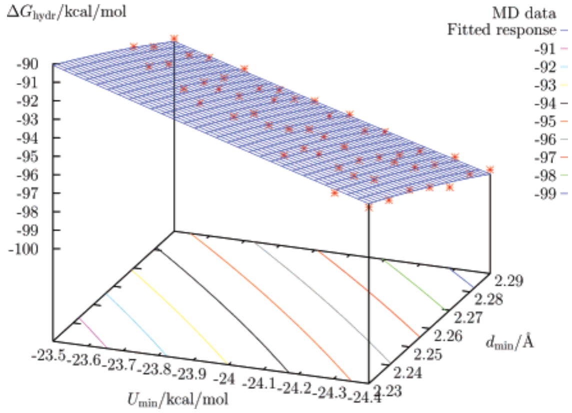


Figure 4. Computed hydration free energy of Na^+ as a function of monohydrate properties for the polarizable model.

It follows that the results from SSBP computations can readily be interpreted as *real* hydration free energies. Unless specified otherwise, *real* hydration free energies are discussed in the rest of the paper.

For convenience, the upper bound on the radial integral used throughout to define the hydration number was set to $r_c = 3.5$ Å. While this choice for r_c may not always coincide with the conventional definition that r_c is the position of the first minimum in the radial distribution function for the $O-K^+$ contact in all of the models of aqueous K^+ studied here, it is necessary when comparing so many different models. As it turns out, $r_c = 3.5$ Å is a good approximation for the position of the first minimum in $g_{OK^+}(r)$ for all of the Drude models, the fixed charge model and the PW91/pw representation of the system. Since the only radial distribution function examined here is for the $O-K^+$ contact, the definitions $g_{OK^+}(r) \equiv g(r)$ and $n_{OK^+}(r) \equiv n(r)$ are employed in the remainder of this paper.

C. Ab Initio Models. The fixed charge and Drude polarizable models of aqueous K^+ are compared with two different ab initio density functional theory (DFT) models of the same system, each using a different gradient-corrected approximate density functional: BLYP[34,35] and PW91.[32,33] Although both ab initio simulations were performed at the Γ-point, there are many methodological differences between the two computations. Simulations with the PW91 exchange-correlation functional were performed within a Born−Oppenheimer molecular dynamics (BOMD) scheme using the VASP software package,[44,45] while simulations using the BLYP functional were performed within the Car−Parrinello molecular dynamics scheme[46] using the PINY_MD software package.[47,48] Some results from this BOMD simulation have previously been published elsewhere.[14,49] The simulation details are given below.

In the BOMD simulation of aqueous K^+, core-valence interactions are described using the projector augmented-wave (PAW) method.[50,51] Convergence was accepted for the electronic structure calculation when the energy difference between successive self-consistent iterations is less than 10^{-6} eV and the valence orbitals are expanded in plane waves with a kinetic energy cutoff of 36.75 Ry (500 eV). This model of the aqueous K^+ system is henceforth referred to as PW91/pw.

The system consisted of 64 water molecules and one K^+ ion in a cubic box of length 12.4171 Å with periodic boundary conditions. The fixed volume was chosen such that the water density matches the experimental density of liquid water at standard conditions. Initial conditions come from a well-equilibrated classical MD run of pure liquid water at standard conditions using SPC/E[52] water for 20 ps, followed by a 10 ps BOMD simulation of pure water. In the BOMD simulation of pure water, a Nosé-Hoover thermostat was applied to constrain the temperature to 375 K, after which a K^+ ion was inserted into the box of pure water and all hydrogens were deuterated. The 3p semicore electrons were explicitly included in the valence orbitals for K^+ ion. During the equilibration phase, constant temperature was maintained at $T = 330$ K with velocity scaling and the equations of motion were integrated using a 1 fs time-step for 14.5 ps. The equilibrated system was then simulated in the microcanonical ensemble with a 0.5 fs time-step for 40 ps of production. During the course of the BOMD simulation, the temperature was 313 ± 21 Kelvin.

The CPMD simulations of aqueous K^+ used the gradient-corrected BLYP approximate density functional[34,35] and a plane-wave basis set. Calculations were performed with a 70 Ry energy cutoff and norm-conserving pseudopotentials.[53] Following the prescription of the initial fully ab initio simulations carried out on this system,[54] the semicore 3 *s* and 3 *p* states of potassium have been included with the valence electrons. A baseline fictitious electronic mass of 475 au was used with mass preconditioning.[55] The canonical ensemble was sampled using Nosé-Hoover chain thermostats[37,56−59] and a 0.125 fs time-step. In order to ensure adiabaticity, the hydrogen masses were substituted with oxygen masses. The temperature over the course of the CPMD simulation was 296 ± 15 Kelvin. This model of the aqueous K^+ system is henceforth referred to as BLYP/pw.

The BLYP/pw system consisted of the same equilibrated BOMD simulation box as above, containing 64 water molecules with a single potassium cation and with periodic boundary conditions. The system was further equilibrated for 5 ps of CP molecular dynamics. Results were collected during a subsequent 50 ps CPMD simulation. An error analysis and finite system size study for this small system and the relatively short simulation times of the ab initio systems presented in the Appendix indicate that the simulations are statistically accurate and representative of the properties of a system with a large number of water molecules (no significant finite size effect on the ion−water radial distribution function).

III. Results and Discussion

A. Monohydrate and Cluster Energies. The interaction energy of the K^+ monohydrate was computed with various methods. The geometry of the monohydrate was optimized for the fixed charge and polarizable Drude models using the CHARMM[60] software package and also quantum mechanically at the Hartree−Fock level with the 6-31G* basis set. In each case the K^+ ion was coplanar with the plane of the water molecule, coordinated with the oxygen atom (that is, had C_{2v} symmetry). As a further comparison, interaction energies have also been computed for DFT optimized geometries (see Tables 1 and 2). The resulting geometries are summarized in Table 2. The monohydrate interaction energies for these geometries, at various levels of theory,[61] are presented in Table 1, with and without the Boys−Bernardi counterpoise correction to basis set superposition error.[62] The experimental gas-phase enthalpy for this system has been measured to be −17.9 kcal/mol[10] (see Table 3 and footnote to Table 1).

Table 1. Interaction Energy, ΔE, for $K^+\cdots OH_2$ Binding[c]

geometry	basis	method	ΔE	ΔE (CPC)
HF/6-31G*	6-31G*	HF	−20.29	−18.76
HF/6-31G*	6-311^{++}G(3df,3pd)	MP2	−17.47	−17.18
HF/6-31G*	6-311^{++}G(3df,3pd)	CCSD	−17.31	−17.03
HF/6-31G*	6-311^{++}G(3df,3pd)	BLYP	−16.57	−16.44
BLYP/6-311^{++}G(3df,3pd)	6-311^{++}G(3df,3pd)	BLYP	−16.68	−16.56
HF/6-31G*	pw (70 Ry)	BLYP	−16.50	N/A
HF/6-31G*	pw (140 Ry)	BLYP	−16.62	N/A
HF/6-31G*	pw (280 Ry)	BLYP	−16.62	N/A
HF/6-31G*	6-311^{++}G(3df,3pd)	PW91	−17.69	−17.55
PW91/6-311^{++}G(3df,3pd)	6-311^{++}G(3df,3pd)	PW91	−17.82	−17.68
HF/6-31G*	pw (70 Ry)	PW91	−17.25	N/A
HF/6-31G*	pw (140 Ry)	PW91	−17.37	N/A
HF/6-31G*	pw (280 Ry)	PW91	−17.37	N/A
fixed charge		fixed charge	−18.9	
$D_{6.5}$		Drude model	−17.7	
$D_{6.8}$		Drude model	−17.9	
expt.[a]		Drude model	−18.3 [b]	

[a] Reference 10. [b] Interaction energy estimated from an experimentally measured enthalpy, −17.9 kcal/mol plus −0.4 kcal/mol taken from Drude model computations of the monohydrate enthalpy (see Table 3). [c] In kcal/mol. Interaction energies are compared from quantum chemical basis set computations, classical force fields, and experiment. Unless otherwise noted, the quantum chemical interaction energies are presented for geometries optimized at the HF/6-31G* level. Data are presented both with and without counterpoise corrections (CPC) to the basis set superposition error (BSSE).

Table 2. Optimized Geometries for $K^+\cdots OH_2$[a]

theory level	r_{OK^+}	r_{OH}	θ_{HOH}
HF/6-31G*	2.6481	0.9511	105.03
BLYP/6-311^{++}G(3df,3pd)	2.6395	0.9736	104.16
PW91/6-311^{++}G(3df,3pd)	2.6116	0.9709	104.06
fixed charge	2.6243	0.9572	104.52
Drude ($D_{6.8}$)	2.6196	0.9572	104.52

[a] Distances are reported in Å; energies are reported in kcal/mol.

The interaction energies presented in Table 1 demonstrate the variability and accuracy of the various quantum chemical approaches that have subsequently been applied to larger aqueous K^+ clusters. The interaction energies in Table 1 are all roughly in accord with the experimental estimate of −18.3 kcal/mol (see footnote to Table 1), though there are small differences that deserve to be noted. Nearly all of the quantum chemical interaction energies appear to be slightly less negative than the experimental estimate. The Hartree−Fock calculation, which overestimates the binding by as much as ∼2 kcal/mol, is the lone exception to this rule. Previous analysis showed that the larger binding energy is directly related to the overestimated dipole of the water

Table 3. Hydration Enthalpy, ΔH, for Gas-Phase K$^+$ $(H_2O)_n$ Clusters[c]

n	fixed charge	Drude $D_{6.8}$	exp.[a]	exp.[b]
1	-18.5 ± 0.02	-17.5 ± 0.04	-17.9	-18.1
2	-35.6 ± 0.03	-33.0 ± 0.04	-34.0	-34.2
3	-51.2 ± 0.05	-46.2 ± 0.05	-47.2	-47.4
4	-64.3 ± 0.06	-57.7 ± 0.05	-59.0	-59.2
5	-74.6 ± 0.07	-67.0 ± 0.06	-69.7	-69.9
6	-84.4 ± 0.08	-76.0 ± 0.08	-79.7	-79.9

[a] Reference 10. [b] Reference 12. [c] In kcal/mol.

molecule, due to neglect of electron correlation.[3] The fixed charge K$^+$ monohydrate, which was originally parametrized to give a reasonable bulk hydration free energy in TIP3P,[8] also overestimates the monohydrate binding energy. As expected, the counterpoise corrections become smaller for larger basis sets.

In order to assess both the magnitude of the fluctuations in the potential energy within the first hydration shell of K$^+$ and the level of consistency with which these are represented by various computational models, a series of K$^+$ $(H_2O)_n$ clusters was examined and compared. First, the enthalpy of hydration, ΔH, is reported in Table 3 for a series of K$^+$ $(H_2O)_n$ clusters with $1 \le n \le 6$ using simulations based on the fixed charge TIP3P and Drude polarizable force fields (model $D_{6.8}$). The enthalpy of the small clusters of one K$^+$ ion and n water molecules were calculated as, $\Delta H = (\langle U_n \rangle - nk_BT)$, where $\langle U_n \rangle$ is the average potential energy of the cluster estimated from a 1 ns trajectory at a temperature of 300 K. Examination of Table 3 indicates that the experimentally observed trend is reproduced by both models (more accurately by the polarizable model), although neither model reproduces the experimental gas-phase data exactly.

In addition, the energy of instantaneous snapshots of water molecules surrounding K$^+$ extracted from a simulation generated using the polarizable force field with model $D_{6.8}$ was calculated and compared for the various models. Configurations with $4 \le n \le 7$ were extracted. For each configuration, all the O−K$^+$ distances were within a 3.5 Å radius from the K$^+$, serving here as the standard definition of the first hydration shell of the ion (see above). The ranking of cluster interaction energies for the instantaneous configurations, shown in Figure 1, follows that for the K$^+$ monohydrates, with a few variations. Interestingly, both the polarizable ($D_{6.8}$) and the fixed charge model closely follow the trends of the quantum chemical interaction energies. While both models yield similar bulk hydration free energies, the polarizable $D_{6.8}$ model is in closer agreement with the MP2 and PW91 (with an atom-centered basis set) interaction energies for this set of configurations. Despite the difference in magnitude, the energies of the instantaneous snapshots are highly correlated. A normalized correlation coefficient can be defined as

$$C_{ij} = \frac{\langle \Delta E_i \Delta E_j \rangle}{\sqrt{\langle \Delta E_i^{\,2} \rangle \langle \Delta E_j^{\,2} \rangle}} \tag{3}$$

where $\Delta E_i = E_i - \langle E_i \rangle$. The C_{ij} vary between 0.94 (e.g., Drude with HF/6-31G*) to 0.99 (e.g., Drude with MP2, or Drude with PW91) for all the models. The high degree of correlation suggests that, while the magnitude of the energies are different, the structure of the potential energy surface is similar in all the models.

B. Hydration Free Energy. The hydration free energy provides an important reference to assess the validity of various models. The Lennard-Jones parameters of K$^+$ were explored to ascertain the sensitivity of the polarizable potential energy function. Lennard-Jones parameters could not be found to generate polarizable models of K$^+$ which had both very small O−K$^+$ monohydrate distances and lower interaction energies. As is evident in Figure 2, it was nevertheless possible to find polarizable models for K$^+$ that had hydration numbers of ∼6.5. Looking at both Figures 2 and 3, it is observed that polarizable models for K$^+$ that have a hydration number of ∼6.5 also have hydration free energies of about −77 kcal/mol. The hydration free energy of a set of Na$^+$ models was also calculated to assess the consistency, or lack thereof, with the putative K$^+$ models. The absolute Na$^+$ hydration free energies are shown in Figure 4. This consistency is important because, while there are inherent uncertainties concerning the absolute scale of single-ion hydration free energy, the relative hydration free energy between monovalent cations is known from experiment very accurately.[12] The relative hydration free energy between K$^+$ and Na$^+$ from experiments is 17.2 kcal/mol.[12]

From an exploration of the $\{U_{\min}, d_{\min}\}$ space for aqueous hydration of K$^+$, two models were selected for further study: one which accurately captures the monohydrate geometry and interaction energy $(U_{\min},d_{\min}) = (-17.9$ kcal/mol, 2.62 Å), and another which sacrifices some of this accuracy in order to yield a hydration number that is in closer accord with that predicted by a recent analysis of neutron scattering experiments, $(U_{\min},d_{\min}) = (-17.7$ kcal/mol, 2.59 Å). The first model has a hydration number of 6.8, while the second model has a coordination number of 6.5 (integrating the radial distribution functions up to a distance of 3.5 Å). These two polarizable models are referred to as $D_{6.8}$ and $D_{6.5}$, respectively. They are indicated in Figures 2 and 3. As an example of how these Drude models must work in conjunction with other Drude polarizable ions, consider a Drude model of Na$^+$ that can be matched with the $D_{6.8}$ model of K$^+$: $D_{6.8}$ has a hydration free energy of $\Delta G_{\mathrm{hydr}} = -80.15$ kcal/mol. Any Drude model for Na$^+$ that has a hydration free energy of $\Delta G_{\mathrm{hydr}} = -97.35$ kcal/mol (representable as a contour in Figure 4) might be suitable. The best choice, however, would also accurately reproduce the monohydrate properties.[5] In this case, a Drude Na$^+$ model with $(U_{\min},d_{\min}) = (-24.0$ kcal/mol, 2.288 Å) is an optimal choice. For the $D_{6.5}$ model of K$^+$, with a hydration free energy of $\Delta G_{\mathrm{hydr}} = -76.60$ kcal/mol, a Drude model for Na$^+$ would be found along the $\Delta G_{\mathrm{hydr}} = -93.80$ kcal/mol contour of Figure 4. While such a model can be found for Na$^+$, it lies close to the boundary of physically realizable models in the $(U_{\min},d_{\min})$ variables (see Figure 4): it is not possible in general to find a reasonable Drude model of Na$^+$ that is consistent with an arbitrarily chosen K$^+$ model.

The hydration free energy for the PW91 and BLYP models are estimated to be −74.3 and −66.1 kcal/mol using a

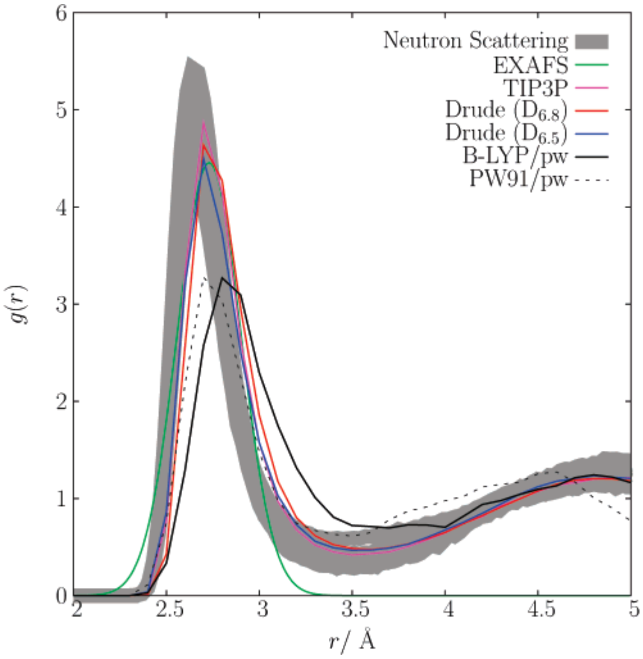


Figure 5. Radial distribution function extracted from the analysis of neutron scattering experimental data[13] and different simulations based on the fixed charge model, two polarizable models, and the BLYP/pw and PW91/pw models. The position of the main peak is as follows: neutron data, 2.65; TIP3P, 2.71; $D_{6.8}$, 2.71; $D_{6.5}$, 2.71; BLYP, 2.83; PW91, 2.73 (in Å). The Gaussian radial distribution function extracted from EXAFS (mean at 2.73 Å and width 0.1712 Å), normalized to a coordination number of 6, is also shown.[63]

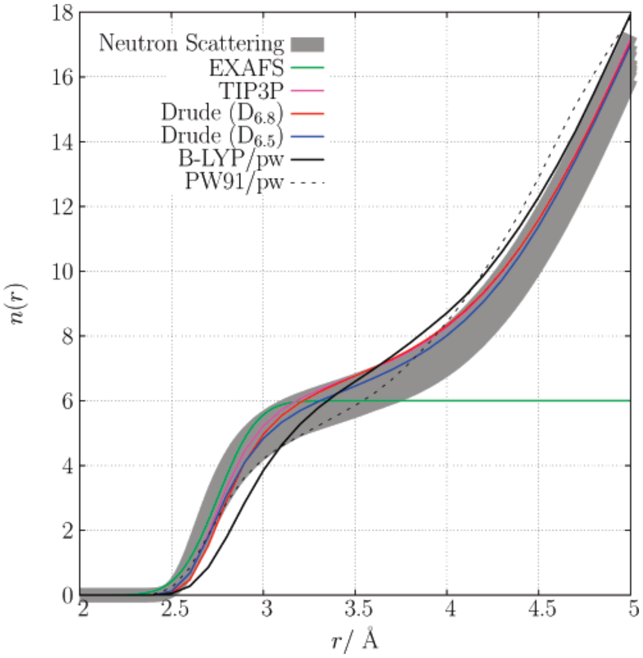


Figure 6. Hydration number, *n*(*r*), of K^+ contact from several different models: the fixed charge model, two polarizable models, and the BLYP/pw and PW91/pw models. The coordination number is as follows: neutron data, 5.5−6.4; TIP3P, 6.77; $D_{6.8}$, 6.8; $D_{6.5}$, 6.5; BLYP, 6.6; PW91, 5.86 (all integrated up to a distance of 3.5 Å). The hydration number *n*(*r*) estimated from EXAFS is also shown.[63]

computational scheme based on the quasi-chemical theory.[20] In this computational scheme, the K^+ ion and the 4 nearest water molecules (within ~3.0 Å, see Figure 6) are modeled explicitly with the exchange-correlation density functional, while the influence of the remaining liquid is incorporated via a far-field treatment. Dispersion interactions and packing effects have been neglected in these particular estimates. These effects are expected to contribute with opposite signs and yield an overall slightly less favorable hydration free energy. It is also worth noting that the quasi-chemical estimates for the BLYP and PW91 density functionals are based on ab initio computations including all electrons, differing slightly with the models of the BOMD and CPMD simulations, which represent the core electrons using a pseudopotential. Superficially, these estimates appear to differ from the molecular dynamics based free energy perturbation (FEP/MD) calculations based on the potential functions by as much as 14 kcal/mol, but this is deceptively incorrect. The FEP/MD calculations with SSBP include the phase potential arising from the vacuum-liquid interface (e.g., they are *real* hydration free energies),[5] whereas the calculations carried out according to the quasi-chemical theory report the *intrinsic* hydration free energy. In calculations based on potential functions, the phase potential is on the order of −500 mV in the liquid, thus contributing favorably to the solvation of a cation by about 12 kcal/mol.[5] Adding this contribution from potential functions to the estimated *intrinsic* hydration free energy based on the quasi-chemical theory yields a *real* hydration free energy on the order of about −86 kcal/mol for the PW91 approximate exchange-correlation functional or a little less if packing and dispersion effects are incorporated. While further work would be required to ascertain the validity of this comparison, the present analysis suggests that the hydration free energy from the $D_{6.8}$ polarizable model is consistent with the value obtained from the quasi-chemical treatment.

C. Hydration Structure in the Bulk Liquid. The radial distribution functions, *g*(*r*), for the O−K^+ contact, for each of the computational models studied here, are presented in Figure 5, along with recently reported radial distributions extracted from neutron scattering experimental data.[13] In order to gauge the spread in the experimental data, they are presented as a set of overlapping distributions, each one deduced from neutron scattering measurements on K^+ solutions made with different salts (KF, KCl, KBr, and KI) and of different concentrations (data for a total of 12 different solutions are shown). The radial distribution of the fixed charge model, based upon TIP3P water, agrees closely with that of the $D_{6.8}$ model. The $D_{6.5}$ model, adjusted to yield a slightly lower coordination number, remains within the range of the experimentally refined distributions. All the radial distribution functions are peaked around 2.7−2.8 Å, though the distributions from the two ab initio simulations (BLYP and PW91) are clearly more diffuse and less sharply peaked than those from classical simulations (TIP3P and Drude polarizable models). On average, the position of the peak in *g*(*r*) is shifted outward by about 0.10 Å relative to the energy minimum ion−water oxygen distance in the monohydrate (Table 2), except for BLYP/pw where it is shifted by almost 0.2 Å. It is worth noting that the distribution functions

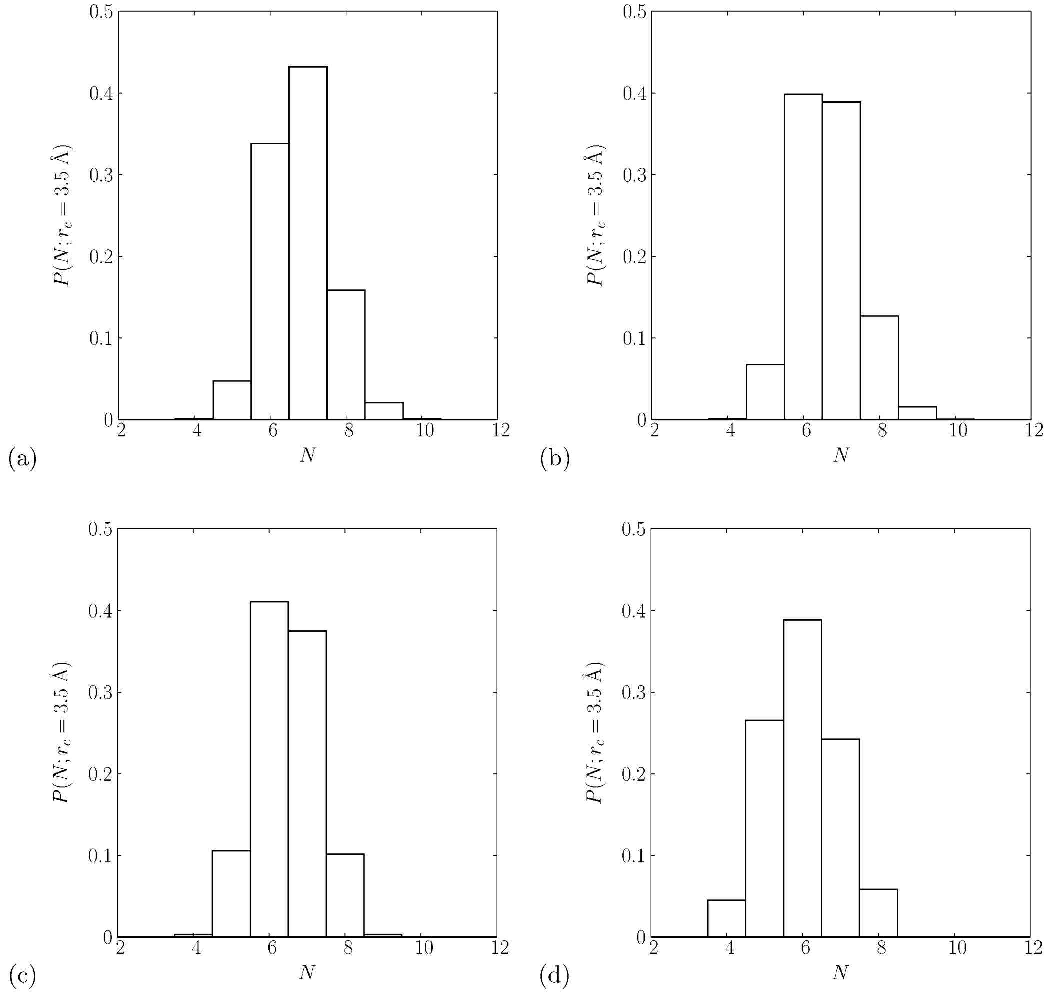


Figure 7. The probability distributions, $P(N;r_c = 3.5$ Å$)$, for the hydration number of aqueous K^+ in the (a) fixed charge, (b) Drude polarizable ($D_{6.8}$), and (c) BLYP/pw and (d) PW91/pw descriptions of the system.

extracted from neutron scattering were also obtained from classical simulations, which were constrained to fit the experimental data.[13] An estimate of the first peak (represented as a Gaussian) based on an analysis of the anomalous diffraction of K^+ by X-ray absorption fine structure (EXAFS) spectra is also shown.[63]

The hydration numbers for K^+ in each of the computational models as well as those deduced from experiments are presented in Figure 6. A 3.5 Å radial cutoff, which is near the minimum between the first and second peaks in $g(r)$, is used throughout to define a unique standard for comparing the calculated coordination number of aqueous K^+ (see earlier discussion). Recently reported hydration numbers deduced from neutron diffraction experiments[13] range from $5.5 \leq n(r_c = 3.5$ Å$) \leq 6.4$. The estimated hydration number from EXAFS is 6 ± 1.[63] Earlier experiments had estimated this number anywhere from 4 to 8 water molecules in the first shell.[14] Density functional models estimate the hydration number to be slightly below (PW91) or above 6 (BLYP). The coordination numbers are 6.77 and 6.8, for TIP3P and $D_{6.8}$, respectively. Although the computational models studied here all differ in their details, the calculated number of water molecules in the first hydration shell consistently lies within the range of what can currently be estimated from experiment.

In Figure 7, the probability distribution, $P(N;r_c = 3.5$ Å$)$, of finding N water molecules that have their oxygen atoms within 3.5 Å from the ion is presented for the different models. In BLYP/pw and PW91/pw ab initio simulations, the number of water molecules found with the highest probability within the first hydration shell is 6. For the polarizable model $D_{6.8}$, the probability distribution has a maximum at 6, while it is 7 for the fixed charge model. The maximum of $P(N;r_c = 3.5$ Å$)$ is 6 for the BLYP/pw simulation, partly due to the use of the $r_c = 3.5$ Å cutoff. As can be seen from Figure 5, a $r_c = 3.75$ Å cutoff would be closer to the minimum, and, indeed this larger cutoff was previously determined by Ramaniah et al.[54] in their simulation using the same BLYP approximate exchange-correlation functional and semicore K^+ pseudopotential that has been employed here. If a cutoff of $r_c = 3.75$ Å is used, the maximum in $P(N)$ becomes 7 for the BLYP/pw simulation.

The fluctuations about the mean hydration number offer a measure of the dynamics within the coordination shell of solvent surrounding K^+. Remarkably, all of the distributions is Figure 7 are well described by Gaussian distributions with similar variances. The standard deviation for the fixed charge model is $\sigma_N = 0.86$, for the $D_{6.8}$ model it is $\sigma_N = 0.86$, while for the ab initio models it is $\sigma_N = 0.84$ and $\sigma_N = 0.96$ for BLYP/pw and PW91/pw, respectively. Thus, while the mean

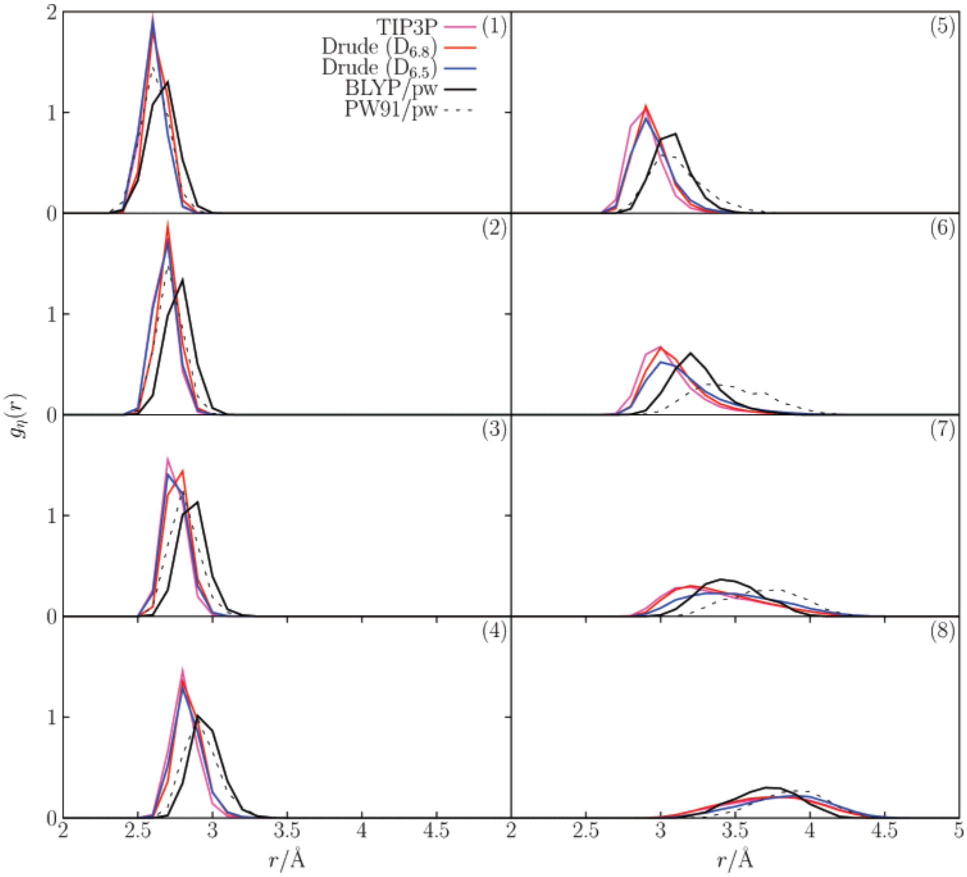


Figure 8. Partial radial distribution functions of the O−K$^+$ contact for the fixed charge, Drude ($D_{6.8}$) polarizable, and BLYP/pw and PW91/pw descriptions of the system. The panels contain partial radial distribution functions for the (1) nearest contact, (2) next nearest, (3) third nearest, (4) fourth, (5) fifth, (6) sixth, (7) seventh, and (8) eighth nearest contact.

coordination number varies slightly among the different models, coordination states within ±1 water molecules about the mean occur approximately 70% of the time in all the models. The significant fluctuations in coordination suggests that the hydration structure around K$^+$ is quite dynamic. This is expected, as the density at the minimum between the first and second hydration shell ($r = 3.5$ Å) is about 50−60% of the bulk solvent density.

A useful way to characterize the hydration structure of an ion is to examine "partial'' radial distribution functions. For example, the radial distribution function of whichever oxygen atom is closer to the K$^+$ ion than is any of the other oxygen atoms in the system, or whichever oxygen is the second closest, and so on. For each such partial radial distribution function, the radial integral converges to 1 at some finite distance, by construction. In Figure 8, the partial radial distribution functions for each of the first 8 nearest oxygen atoms are presented. The differences between the various descriptions of aqueous K$^+$ that were apparent in the $g(r)$ are also seen in the partial radial distribution functions. On average the two ab initio simulations display a looser hydration structure than the classical models, with the third- to sixth-nearest contacts shifted to larger separations, though they also display some differences with one another. This is especially noticeable for the O−K$^+$ distances of the first- to fourth-nearest contacts of BLYP/pw, which are further on average than for those of PW91/pw. The fixed charge partial radial distributions are closely matched with those of the $D_{6.8}$ and $D_{6.5}$ polarizable models. The partial radial distribution functions from the BLYP/pw and PW91/pw simulations are similar to one another for the 5 nearest water molecules around K$^+$. For the sixth-, seventh-, and eighth-nearest O−K$^+$ contacts, the PW91/pw coordination structure is looser compared with that of BLYP/pw—that is, the oxygen atoms of these three partial radial distributions are further from the K$^+$ ion in the PW91/pw representation of this system than they are in the BLYP/pw representation.

In addition, Figure 9 displays the cumulative partial hydration numbers for each of the models studied here. From Figure 9, it can easily be seen which of the nearby water molecules is contributing significant density to the radial distribution function features within $r_c = 3.5$ Å. For example, with the fixed charge model, there are essentially 6 oxygen atoms entirely within the 3.5 Å cutoff; the remainder of the $n(r_c) = 6.77$ coordination number is contributed by both the seventh- and eighth-nearest water molecules. In the PW91/pw simulation, 5 oxygens lie within the 3.5 Å cutoff, while the sixth- and seventh-nearest water molecules also contribute to the density within the first hydration shell.

D. Self-Diffusion of K$^+$. The diffusion constant of K$^+$ has been computed for the $D_{6.5}$ and $D_{6.8}$ polarizable Drude models from the mean-square displacement. Because there is only a single ion in the system, relatively long simulations are required to obtain well converged estimates. Accordingly, 5 independent simulations of 1 ns length were averaged together for each polarizable model. The diffusion constant

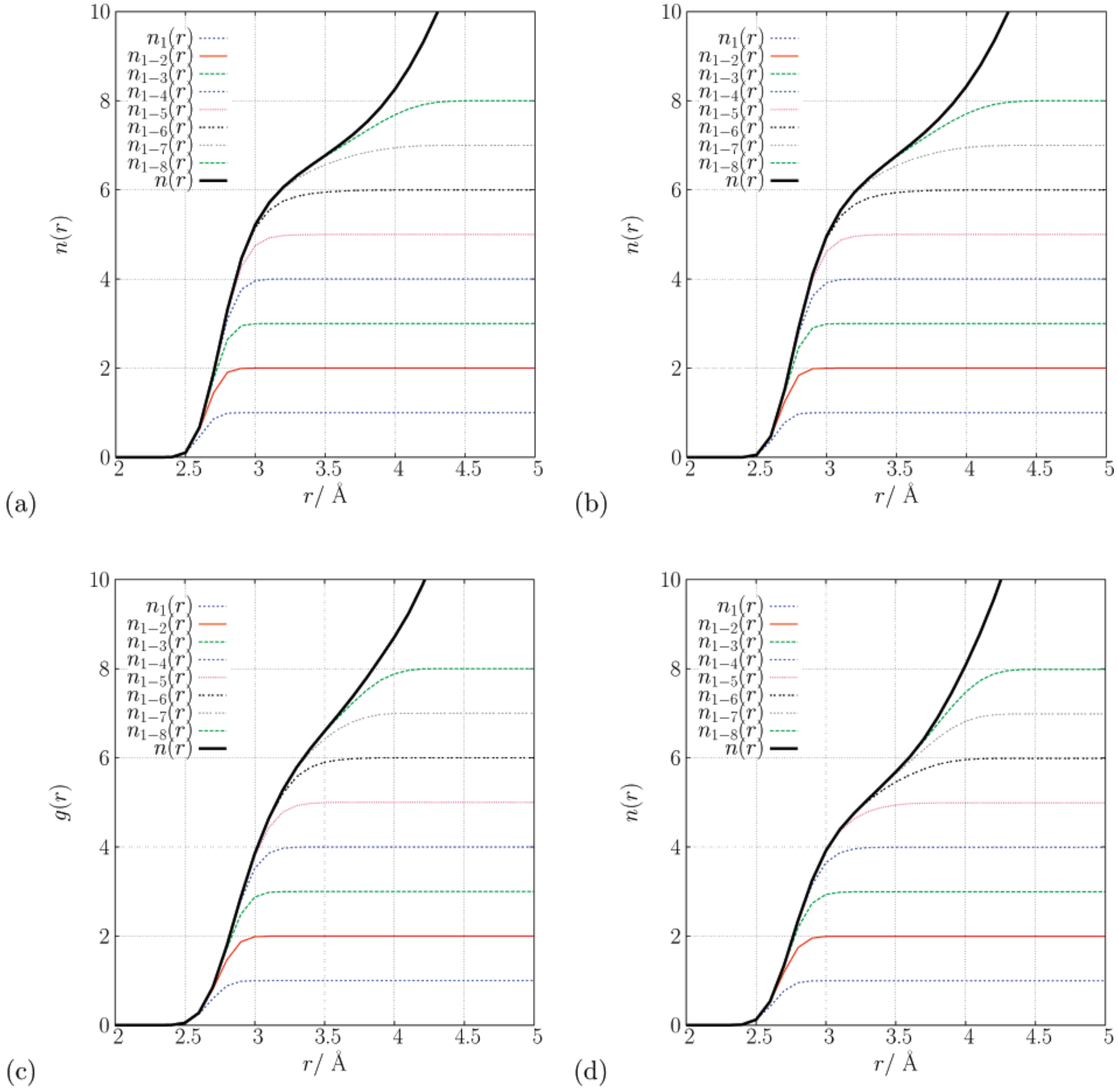


Figure 9. Cumulative partial hydration numbers, $n_{1-\alpha}(r)$, of aqueous K$^+$ in the (a) fixed charge, (b) Drude polarizable ($D_{6.8}$), and (c) BLYP/pw and (d) PW91/pw descriptions of the system.

of the $D_{6.5}$ model was 1.71 ± 0.2 × 10^{-5} cm^2/s, and for the $D_{6.8}$ model it was 1.83 ± 0.2 × 10^{-5} cm^2/s. Both of these values are in excellent agreement with the experimental value of 1.96 × 10^{-5} cm^2/s.[64] One may note that, in this particular case, the model with the lower hydration number actually diffuses slightly more slowly (though the difference is very small). However, a systematic analysis of a family of models shows that the diffusion coefficient does tend to decrease when the hydration number increases (by about −0.092 × 10^{-5} cm^2/s), in accord with the expected hydrodynamic trend).

E. Electronic Polarization near and far from K$^+$. The induction effects of the K$^+$ ion on its first hydration shell were compared between the polarizable model and the ab initio models by computing the respective distributions of molecular dipole magnitudes. For models of neutral molecules based on point charges, calculating the molecular dipole amounts to a straightforward sum over molecular charges. The situation is more ambiguous for ab initio simulations of condensed-phase systems, where the electronic charge density is continuously distributed. One approach that has been used in the past[65−68] is to transform from the Kohn−Sham orbitals to the basis of maximally localized Wannier functions.[69−71] In the localized basis, the Wannier function centers (WFCs) allow for an assignment of molecular dipoles. In the present study, analysis of the WFCs allows for comparison between the water dipole distributions in the bulk and in the nearest solvation shell as well as between computational models for K$^+$ hydration.

The effects of polarization within the first hydration shell of K$^+$ were studied by computing the distribution of molecular dipole magnitudes for water molecules within the first hydration shell and for those outside. The molecular dipoles in the CPMD simulation were assigned using the WFCs, and the distributions are shown in Figure 10. In total, WFCs were computed for 94 different configurations of the equilibrated BLYP/pw system. These configurations were taken from the final 47 ps of the production run, and each

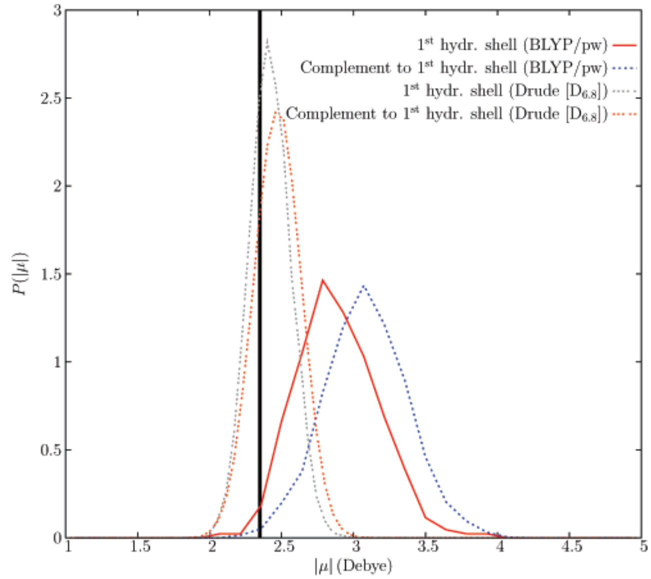


Figure 10. Probability distributions of molecular dipole magnitudes, $P(|\mu|)$, for water molecules in the aqueous K$^+$ system. Distributions are shown for water molecules in the first hydration shell, defined by a 3.5 Å O−K$^+$ distance, and for water molecules outside of the first hydration shell, for the Drude polarizable, and BLPY/pw descriptions, respectively. For reference, the vertical line at $|\mu| = 2.35$ Debye indicates the magnitude of the TIP3P molecular dipole.

was 500 fs apart from the next. The average dipole magnitude for water molecules outside of the first hydration shell is consistent with previously reported pure liquid water values for both the SWM4-NDP[25] and BLYP/pw.[66] Of particular interest, there is a small downward shift in the average dipole magnitude for water molecules within the first hydration shell for both the BLYP/pw and Drude polarizable models, The shift is 0.2 Debye and 0.05 Debye for the BLYP/pw and the $D_{6.8}$ polarizable models, respectively. Relative to the value of the average molecular dipole magnitude in the bulk, $\delta\langle|\mu|\rangle/\langle|\mu|\rangle$, the shifts are 6.5% in the BLYP/pw simulation and 2% in the polarizable force field simulation. A qualitatively similar shift has been observed by comparing, for a polarizable force field model, the distribution of molecular dipole magnitudes in K$^+$ $(H_2O)_n$ clusters with that in pure bulk water.[72]

The observation that the molecular dipole of water within the first hydration shell of K$^+$ has a slightly smaller average value than that in bulk water is rather counterintuitive. A water molecule in the first hydration shell would be expected to be polarized by the electric field from the ion. This is certainly observed for a K$^+$ monohydrate, but the situation is more complex in the bulk phase. The surprising electrostatic properties revealed by Figure 10 result from a balance of competing factors. There is a net benefit to align the water molecules and induce dipoles within the first hydration shell. There is also an unfavorable energy cost arising from the interaction between those dipoles pointing toward a central point. Furthermore, the molecular dipole of water increases from the vapor to the liquid phases due to the hydrogen-bonding network structure of liquid water.[73−75] As this network is disrupted in the neighborhood of K$^+$, the average magnitude of the molecular dipole decreases.[67] Finally, it is worth noting that, because the shift in the $\langle|\mu|\rangle$ is small, fixed charged models like TIP3P closely approximate the hydration structure of the polarizable models near K$^+$. This may partly explain the surprising ability of nonpolarizable models to represent bulk hydration of ions.

F. On Differences and Similarities. The present study shows that our current knowledge of K$^+$ hydration is satisfactory, with different models being in broad agreement with the available experimental data. The interaction energy of the monohydrate is about −18 kcal/mol, near the experimental gas-phase estimate. The hydration structure in the bulk is consistent with a coordination number on the order of 6−7 and with a first peak around 2.7 Å, as indicated by the analysis of neutron scattering from solutions. The total solvation free energy is about −80 kcal/mol, consistent with a variety of thermodynamic estimates from experiments[12,15,16,40,41] or computations.[4−9] In comparison, the AMOEBA model of Grossfield, Ren, and Ponder[4] yields a *real* hydration free energy for K$^+$ that is roughly 4−5 kcal/mol larger than the present estimate and a coordination number of 7.0. Such differences appear to be within acceptable bounds.

Nevertheless, at a finer level, there remain some discrepancies that should be better understood to further refine our models of ion hydration. For example, there are notable differences between the position and the shape of the main peak extracted from the neutron scattering data and the results from the two ab initio simulations (see Figure 5). The average radial distribution function extracted from neutron scattering for 12 solutions is sharply peaked at 2.65 Å, whereas the peak from the two ab initio simulations are more diffuse. In the case of the simulation based on BLYP, the peak is also slightly shifted toward larger distances. What is puzzling is the fact that the two classical models (including the non-polarizable force field) are in closer agreement with the results from neutron scattering experiments than the two ab initio simulations. Normally, the average coordination structure obtained from ab initio simulations is quite reliable. However, it is important to keep in mind that the radial distribution functions are extracted from the neutron scattering data using a refinement procedure, which relies on a set of simulations biased to fit the experiments.[13] Those simulations are not not exempt from assumptions. For example, the K$^+$-oxygen minimum distance is set to 2.6 Å (Alan Soper, personal communication), based on an earlier estimate from Herdman and Neilson.[76] Furthermore, the ion−water repulsion is modeled after a Lennard-Jones potential, which is generally steeper than the core repulsion calculated from ab initio. In spite of these caveats, the $g(r)$ extracted from the neutron scattering data shown in Figure 5 is in reasonable accord with a variety of experimental X-ray and neutron scattering data indicating that the peak in the K$^+$-oxygen distribution function should be somewhere between 2.60 and 2.80 Å (though some older estimates were as high as 2.92 Å).[77] Furthermore, the coordination number extracted from the neutron scattering data via the refinement procedure, ranging from 5.5 to 6.4, appears to be nearly reproduced by all the models (see Figure 6). In excellent accord with the

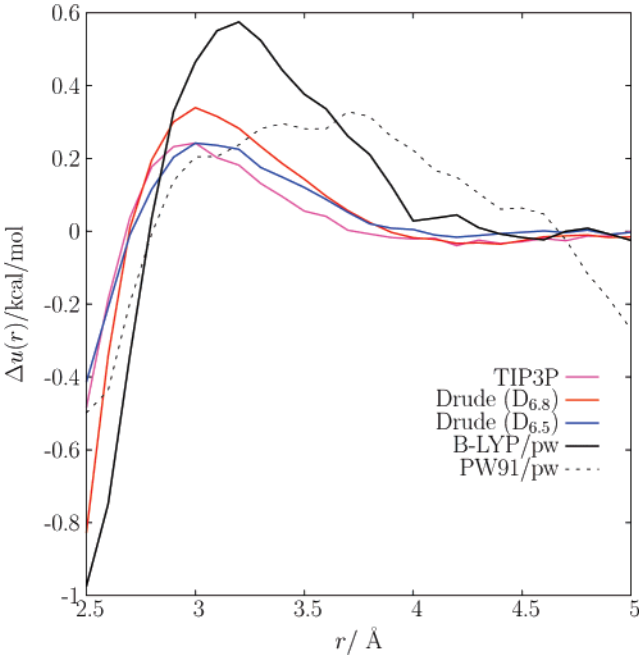


Figure 11. Perturbative analysis of the K$^+$-water oxygen interaction using the average radial distribution function extracted from the neutron scattering data as a reference.

current results, a recent estimate based on an analysis of the anomalous diffraction of K$^+$ by X-ray absorption fine structure (EXAFS) spectra estimates the average distance between the K$^+$ and the water oxygen in the first shell at 2.730 ± 0.05 Å and the coordination number at 6 ± 1.[63]

While an assessment of the sensitivity of the results extracted from neutron scattering data to all input assumptions would be required to ascertain the accuracy of the different computational models, an important question remains whether the observed differences in the radial distribution functions signal some fundamental underlying problems in our understanding of K$^+$ hydration. At the simplest level, differences in the radial distribution of K$^+$-water oxygen observed in Figure 5 could be caused simply by differences in the direct ion−water interaction. Such small differences, on the order of ~0.5 kcal/mol, can already be noted in Table 1. In fact, nearly all the ab initio calculations yield a K$^+$-water binding energy that is slightly weaker than the experimental estimate (the exception being the HF/6-31G* calculation). By a low order perturbative treatment, one can express the small differences observed between the various radial distribution functions from the various models in terms of a putative difference in the direct ion−water interaction. Taking the average radial distribution function extracted from the neutron scattering data as a reference $g_{\mathrm{ref}}(r)$, we define the potential $\Delta u_i(r)$

$$\Delta u_i(r) = -k_{\mathrm{B}}T \ln\left[\frac{g_{\mathrm{ref}}(r)}{g_i(r)}\right] \tag{4}$$

To lowest order, $\Delta u_i(r)$ is the potential that needs to be added to the ion−water interaction of a model i in order to recover $g_{\mathrm{ref}}(r)$. Of course, such analysis is valid only if the perturbation is small. At higher order, the ability of a liquid to coordinate an ion is also related to the amount of cohesion that exists in the pure liquid, e.g., hydration of an ion would be reduced in a water model that attributes more internal cohesion to the liquid, and it should be increased in a model that attributes less cohesion to the liquid. Nonetheless, an analysis based on eq 4 is informative. The results for $\Delta u_i(r)$ are plotted in Figure 11. According to this perturbative analysis, it appears that all the models (except the ab initio simulations from BLYP), would require a fairly small perturbation in the ion−water interaction to yield $g_{\mathrm{ref}}(r)$. At near-contact ($r \approx 2.6-2.7$ Å), the perturbation amounts to a fraction of kcal/mol, which is consistent with the magnitude of the variations observed in the binding energy of the monohydrates given in Table 1. From this perspective, it is possible that the differences observed between the various models might reflect the relatively small differences in the direct ion−water interaction.

IV. Conclusion

A hierarchy of computational models have been used to study the properties of aqueous K$^+$, including two ab initio models, a fixed charge model, and a polarizable model based on classical Drude oscillators. The O−K$^+$ radial distribution functions of the models have been compared with those derived from neutron scattering experiments.[13] Among the different computational representations of the system, the polarizable model and fixed charge model appear to agree more closely with the shape of the radial distribution functions deduced from experiments, while those from the two ab initio simulations seems to be not as sharply peaked. All the computational models yield hydration number between 5.86 (PW91/pw) and 6.8 ($D_{6.8}$), in good accord with the experimental estimates (see Figure 6), and yield a reasonable monohydrate binding energy as well as hydration free energy.

A somewhat counterintuitive observation made on the basis of the $D_{6.8}$ and CPMD simulations concerns the induced dipolar of water molecules nearest to the K$^+$. The electronic polarization effects of the K$^+$ ion on the water molecules in the first hydration shell have been examined using a BLYP/pw ab initio simulation and a polarizable force field simulation of aqueous K$^+$. In both cases, a slight shift to lower average dipole magnitudes for molecules in the first hydration shell, compared to that in the bulk liquid, has been observed. This observation contradicts the intuitive notion that water molecules in direct contact with a cation must be overpolarized compared to the bulk value. In fact, it appears that in the case of K$^+$ they are, if anything, slightly less polarized than the water molecules in the bulk. This is, perhaps, one reason for the relative success of simple fixed charged models.[6−9] It may be that K$^+$ has a size that renders it similar to water in its “polarizing strength’’, suggesting that only smaller ions require a treatment of induced polarization. In view of this result, one might be tempted to suggest that a polarizable force field is not really needed for K$^+$. In the context of a homogeneous bulk liquid phase, this is partly true. However, one must be careful in overextending this conclusion to inhomogeneous environments such as interfaces or the interior of narrow pores. In those systems, the limitations of nonpolarizable force fields in the case of K^+ have been clearly documented.[78]

Although a fairly consistent perspective of K^+ hydration emerges from the current study, resolving a number of issues could further our ability in modeling ion hydration accurately. In particular, a sensitivity analysis of the hydration structure properties extracted from scattering experimental data would be very useful. Contrasting the results from different computational models also helps delineate the limits of present knowledge about K^+ hydration. Simulations based on quantum mechanical ab initio methods can account for a wide range of complex electronic effects. But the complete information about the thermodynamic properties in the bulk phase of those ab initio models is not easily accessible to ascertain the implications of the results. The properties in the bulk phase can be fully explored for computationally simpler models based on a potential function, such as the polarizable force field based on Drude oscillators. Such models use parametrized mathematical functional forms to represent complex microscopic interactions. While those parameters can be freely adjusted to reproduce various properties for any cation, the structure of the potential function places internal constraints on the range of possible models that can be constructed. This idea is illustrated in Figures 2−4. In the present study, the coordination numbers of K^+ and Na^+ are strongly correlated with the monohydrate binding energies and thus with the bulk hydration free energies. This correlation was illustrated here by considering two different polarizable models for K^+. One model, referred to as $D_{6.8}$, was fitted to agree with the K^+ monohydrate properties. The other model, referred to as $D_{6.5}$, was adjusted to interact less strongly with water, in order to yield a lower hydration number in closer accord with the ab initio simulations. However, the hydration free energy of the $D_{6.5}$ model of K^+ is decreased, and it becomes challenging to parametrize a model of Na^+ with a relative hydration free energy that is consistent with the experiment.[12,40,41] Thus, in the context of the polarizable potential function based on classical Drude oscillators, the relative hydration free energy of K^+ and Na^+ (or any other ion) limits the range of accessible coordination numbers. Such internal constraints deduced from simulations based on a given functional form of force field are model-specific. Nonetheless, qualitatively similar observations are made from the AMOEBA model of K^+, where a slightly larger coordination number is correlated with a slightly larger hydration free energy.[4] This correlation suggests that such internal constraints qualitatively reflect inherent trends (e.g., one cannot arbitrarily shift the first peak in $g(r)$ to larger distances and expect to decrease the hydration number while reproducing the monohydrate properties and achieving a reasonable hydration free energy), though particular results could change quantitatively if a different functional form was used. Thus, development of a microscopic perspective on K^+ hydration, integrating the information provided by experiments and computational models, remains partly subjective at this point.

Acknowledgment. We are very grateful to Alan K. Soper for sharing his recently published[13] data and for insightful comments. T.W.W. thanks Mark E. Tuckerman for sharing the potassium pseudopotential and G.J. Martyna for helpful discussions. T.W.W. is supported by a strategic LDRD grant from Argonne National Laboratory. This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, award number PHS 2 PN2 EY016570B, and the grant GM072558 (B.R.). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Dept. of Energy. This work was also supported by the LDRD program under Contract DE-AC04-94A185000.

Appendix: Analysis of Statistical Error and Finite Size Effects

The statistical error in the radial distribution functions calculated from the BLYP/pw simulation was determined by dividing the 50 ps trajectory into 25 2 ps parts and calculating the error in the mean of each histogram window, r, to generate $\Delta(r) = \sigma(r)/\sqrt{25}$, the r-dependent error in $g(r)$. This estimate compares well with a more accurate one obtained by performing 50 simulations of length 40 ps (the same length as the BOMD simulation) using the Drude force field and computing $\Delta(r) = \sum_k \sigma^{(k)}(r)/50$. In Figure 12, the radial distribution function computed from each of the 50 independent 40 ps simulations is plotted along with the average $g(r)$. The spread in the distributions in Figure 12 gives an excellent estimate of the statistical uncertainty from a short simulation (also shown are the error bars resulting from the above analysis of the Drude model trajectories).

In order to assess the significance of finite size effects in the relatively small system containing 64 water molecules, the radial distribution function for the O−K^+ contact is compared, in Figure 13, with that generated from a much

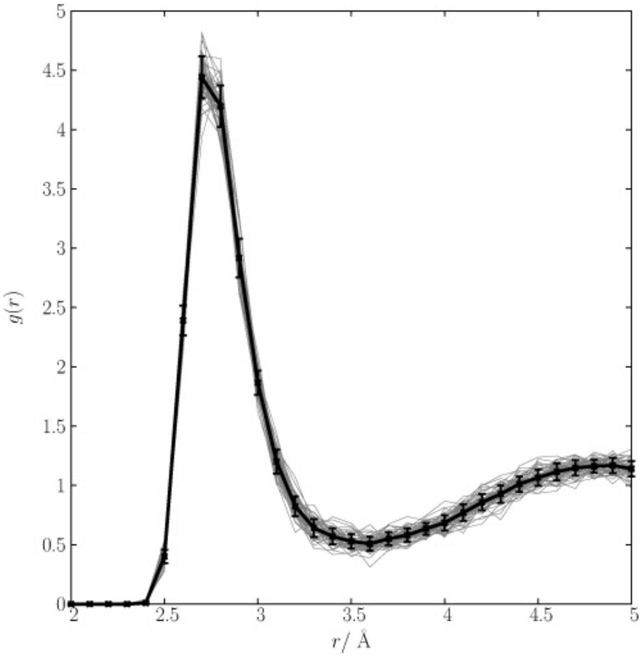


Figure 12. Statistical spread in $g(r)$ of the O−K^+ contact taken from 40 ps of molecular dynamics. The Drude polarizable model was used simulate the system. The black line is the average $g(r)$.

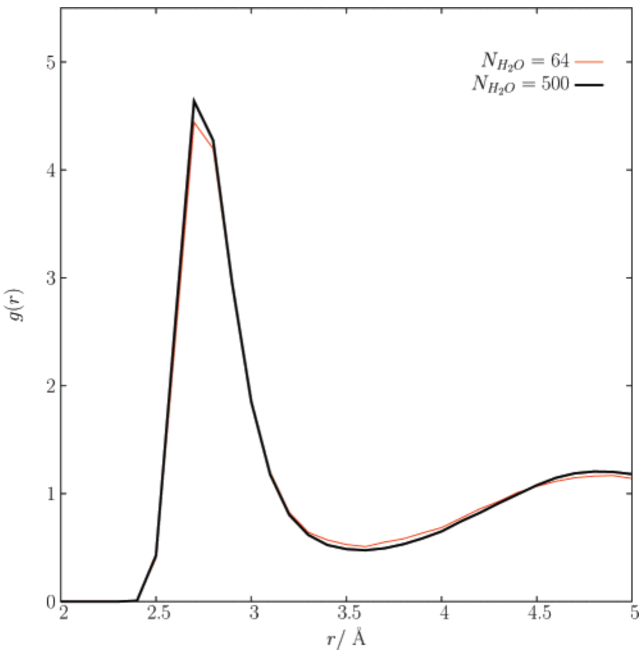


Figure 13. Radial distribution function of the O−K^+ contact taken from two different system sizes: a smaller system with 64 water molecules and a larger system containing 500 water molecules. Both system sizes were modeled using the Drude $D_{6.8}$ polarizable force field.

larger system containing 500 water molecules. In both cases, it is a polarizable model system that is being simulated. It is evident that finite size effects are not significant for this property of aqueous K^+.

References

(1) Lybrand, T. P.; Kollman, P. A. *J. Chem. Phys.* **1985**, *83*, 2923−2933.

(2) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 2481−2486.

(3) Roux, B.; Karplus, M. *J. Comput. Chem.* **1995**, *16*, 690−704.

(4) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671−15682.

(5) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308−3322.

(6) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(7) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021−8024.

(8) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050−9063.

(9) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*(6), 1499−1509.

(10) Džidić, I.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1466−1474.

(11) Klassen, J. S.; Anderson, S. G.; Blades, A. T.; Kebarle, P. *J. Phys. Chem.* **1996**, *100*, 14218−14227.

(12) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R., Jr. *J. Phys. Chem. A* **1998**, *102*, 7787−7794.

(13) Soper, A. K.; Weckström, K. *Biophys. Chem.* **2006**, *124*, 180−191.

(14) Varma, S.; Rempe, S. B. *Biophys. Chem.* **2006**, *124*, 192−199.

(15) Gomer, R.; Tryson, G. *J. Chem. Phys.* **1977**, *66*, 4413−4424.

(16) Klots, C. E. *J. Phys. Chem.* **1981**, *85*, 3585−3588.

(17) Zhan, C.-G.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 11534−11540.

(18) Asthagiri, D.; Pratt, L. R.; Ashbaugh, H. S. *J. Chem. Phys.* **2003**, *119*, 2702−2708.

(19) Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem And Models Of Molecular Solutions;* Cambridge University Press: Cambridge, MA, 2006.

(20) Pratt, L. R.; Rempe, S. B. In *Simulation and Theory of Electrostatic Interactions in Solution: Computational Chemistry, Biophysics, and Aqueous Solutions*; number 492 in AIP Conference Proceedings, Pratt, L. R., Hummer, G., Eds.; American Institute of Physics: New York, 1999; pp 172−201.

(21) Pérez-Jordá, J. M.; San-Fabián, E.; Pérez-Jiménez, A. J. *J. Chem. Phys.* **1999**, *110*, 1916−1920.

(22) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Phys. Chem.* **2004**, *120*, 2693−2699.

(23) Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2005**, *122*, 114102.

(24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(25) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *Chem. Phys. Lett.* **2006**, *418*, 245−249.

(26) Drude, P. *Lehrbuch der Optik;* S. Hirzel: Leipzig, Germany, 1900.

(27) Sangster, M. J. L.; Dixon, M. *Adv. Phys.* **1976**, *25*, 247.

(28) Stillinger, F. H.; David, C. W. *J. Chem. Phys.* **1978**, *69*, 1473.

(29) Pratt, L. R. *Mol. Phys.* **1980**, *40*, 347.

(30) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1989**, *89*, 7556.

(31) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

(32) Wang, Y.; Perdew, J. P. *Phys. Rev. B* **1991**, *44*(24), 13298−13307.

(33) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*(11), 6671−6687.

(34) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(35) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(36) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(37) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635−2643.

(38) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(39) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Oxford University Press: Oxford, 1987.

(40) Noyes, R. M. *J. Am. Chem. Soc.* **1962**, *84*, 513−522.

(41) Marcus, Y. *Biophys. Chem.* **1994**, *51*, 111−127.

(42) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567−16576.

(43) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(44) Kresse, G.; Hafner, J. *Phys. Rev. B* **1993**, *47*(1), 558−561.

(45) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*(16), 11169−11186.

(46) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(47) Samuelson, S.; Martyna, G. *J. Chem. Phys.* **1998**, *109*, 11061−11073.

(48) Tuckerman, M. E.; Yarne, D.; Samuelson, S. O.; Hughes, A. L.; Martyna, G. J. *Comput. Phys. Commun.* **2000**, *128*, 333−376.

(49) Varma, S.; Rempe, S. B. *Biophys. J.* **2007**, *93*.

(50) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*(24), 17953−17979.

(51) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*(3), 1758−1775.

(52) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269−6271.

(53) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993−2006.

(54) Ramaniah, L. M.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 1587−1591.

(55) Tassone, F.; Mauri, F.; Car, R. *Phys. Rev. B* **1994**, *50*(15), 10561−10573.

(56) Tuckerman, M. E.; Parrinello, M. *J. Chem. Phys.* **1994**, *101*, 1302−1315.

(57) Tuckerman, M. E.; Parrinello, M. *J. Chem. Phys.* **1994**, *101*, 1316−1329.

(58) Hutter, J.; Tuckerman, M. E.; Parrinello, M. *J. Chem. Phys.* **1995**, *102*, 859−871.

(59) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117−1157.

(60) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision 02*; Gaussian, Inc.: Pittsburgh, PA.

(62) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.

(63) Glezakou, V.-A.; Chen, Y.; Fulton, J. L.; Schenter, G. K.; Dang, L. X. *Theor. Chem. Acc.* **2006**, *115*, 86−99.

(64) Lide, D. R., Ed. *CRC Handbook of Chemistry and Physics*, 87th ed.; Taylor and Francis: Boca Raton, FL, 2007.

(65) Silvistrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, 82, 3308−3311.

(66) Silvistrelli, P. L.; Parrinello, M. *J. Chem. Phys.* **1999**, *111*, 3572−3580.

(67) Boero, M.; Terakura, K.; Ikeshoji, T.; Liew, C. C.; Parrinello, M. *Phys. Rev. Lett.* **2000**, *85*(15), 3245−3248.

(68) Whitfield, T. W.; Crain, J.; Martyna, G. J. *J. Chem. Phys.* **2006**, *124*, 094503.

(69) Wannier, G. H. *Phys. Rev.* **1937**, *52*, 191−197.

(70) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300−302.

(71) Resta, R.; Sorella, S. *Phys. Rev. Lett.* **1999**, *82*, 370−373.

(72) Carrillo-Tripp, M.; Saint-Martin, H.; Ortega-Blake, I. *J. Chem. Phys.* **2003**, *118*, 7062−7073.

(73) Bernal, J. D.; Fowler, R. H. *J. Chem. Phys.* **1933**, *1*, 515.

(74) Stillinger, F. H. *Science* **1980**, *25*, 451−547.

(75) Chen, B.; Ivanov, I.; Klein, M. L.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *91*, 215503.

(76) Herdman, G. J.; Neilson, G. W. *J. Mol. Liq.* **1990**, *46*, 165−179.

(77) Marcus, Y. *Chem. Rev.* **1988**, *88*, 1475−1498.

(78) Roux, B.; Berneche, S. *Biophys. J.* **2002**, *82*, 1681−1684.

CT700172B

JCTC Journal of Chemical Theory and Computation

Polarizable Atomic Multipole Solutes in a Generalized Kirkwood Continuum

Michael J. Schnieders[†] and Jay W. Ponder[*,‡]

Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, and Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110

Received June 2, 2007

Abstract: The generalized Born (GB) model of continuum electrostatics is an analytic approximation to the Poisson equation useful for predicting the electrostatic component of the solvation free energy for solutes ranging in size from small organic molecules to large macromolecular complexes. This work presents a new continuum electrostatics model based on Kirkwood's analytic result for the electrostatic component of the solvation free energy for a solute with arbitrary charge distribution. Unlike GB, which is limited to monopoles, our generalized Kirkwood (GK) model can treat solute electrostatics represented by any combination of permanent and induced atomic multipole moments of arbitrary degree. Here we apply the GK model to the newly developed Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field, which includes permanent atomic multipoles through the quadrupole and treats polarization via induced dipoles. A derivation of the GK gradient is presented, which enables energy minimization or molecular dynamics of an AMOEBA solute within a GK continuum. For a series of 55 proteins, GK electrostatic solvation free energies are compared to the Polarizable Multipole Poisson−Boltzmann (PMPB) model and yield a mean unsigned relative difference of 0.9%. Additionally, the reaction field of GK compares well to that of the PMPB model, as shown by a mean unsigned relative difference of 2.7% in predicting the total solvated dipole moment for each protein in this test set. The CPU time needed for GK relative to vacuum AMOEBA calculations is approximately a factor of 3, making it suitable for applications that require significant sampling of configuration space.

1. Introduction

The solvent environment influences the structure and behavior of solutes within it. For example, the scaling of the radius of gyration of a polymer with chain length in dilute aqueous solution can be predicted by considering whether solvent molecules prefer interactions among themselves to those with the polymer.[3] This scaling law serves to emphasize that rigorous results can be obtained without treating the solvent in explicit atomic detail. Here we present an analytic model of the electrostatic interactions between a solute represented by polarizable atomic multipoles and a continuum environment characterized by its permittivity, dispensing with the expense of representing explicit solvent molecules.

Our approach can be traced to work presented by Born in 1920 to describe the electrostatic solvation energy of a charged, spherical ion in terms of macroscopic continuum theory.[4] In 1934, Kirkwood extended this approach to a spherical particle with arbitrary electrostatic multipole moments with application to the study of zwitterions, which have a large dipole moment.[1] More recently, Kong and Ponder revisited Kirkwood's theory to allow analytic treatment of off-center point multipoles.[5] For a single spherical

[*] Corresponding author phone: (314)362-4195; fax: (314)362-7183; e-mail: ponder@dasher.wustl.edu.

[†] Washington University in St. Louis.

[‡] Washington University School of Medicine.

10.1021/ct7001336 CCC: $37.00
Published on Web 10/18/2007

particle in isolation, therefore, the theoretical foundations to enable use of macroscopic continuum theory have already been established.

However, a general analytic solution to the Poisson equation for an arbitrarily spaced collection of spherical dielectric particles embedded in solvent is tenable only via approximations. For example, the generalization of Born's method to a collection of monopoles began to be considered in the 1990s by a number of groups including Schaefer et al.,[6,7] Hawkins et al.,[8,9] Still et al.,[10,11] Feig et al.,[12,13] and Onufriev et al.[14−18] This generalized Born (GB) approach is intended to approximate the numerical solution of the Poisson equation for realistic molecular geometries and monopole charge distributions.[19−22] Given highly accurate self-energies, GB has been shown to be remarkably quantitative.[13,14,16,23,24] The goal of the present work is to extend the ideas underlying GB to more accurate charge distributions, specifically to the treatment of polarizable atomic multipoles, which might be termed generalized Kirkwood (GK) by analogy.

In order to further motivate the present work, we recall the electrostatic solvation energy is a key component of an implicit solvent model, which typically also includes apolar contributions due to cavitation and dispersion.[11,25,26] Given a solute potential and implicit solvent, a broad range of physical properties can be predicted, including conformational preferences such as radius of gyration, binding energies, and pK_as.[24] Recent work by a number of groups to explicitly include higher order permanent moments and polarization within the functional form of empirical force field electrostatics may improve the quality of theoretical predictions based on implicit solvent approaches.[27−32] However, this step forward can only be realized if the improved detail of the molecular mechanics electrostatic model is propagated through to the reaction potential.

For an excellent introduction to the fundamentals of GB theory, including treatment of salt effects, we recommend the review by Bashford and Case.[33] Feig and Brooks present a review of recent improvements in GB methodology as well as novel applications.[12] Assuming this level of familiarity, we immediately outline the key components of GB that need to be further generalized in order to incorporate polarizable atomic multipoles.

1.1. Effective Radii and the Self-Energy. Definition of the "perfect" effective radius a_i for site i under the GB approximation[16] guarantees an exact self-energy. It is based on the following equality

$$a_i = \frac{1}{2}\left(\frac{1}{\epsilon_\text{s}} - \frac{1}{\epsilon_\text{h}}\right)\frac{q_i^{\ 2}}{\Delta W_{\text{self},i}^{\text{Poisson}}} \tag{1}$$

where the factor of $^1/_2$ accounts for the cost of polarizing the continuum, q_i is a partial charge, ϵ_h is the permittivity of a homogeneous reference state, and ϵ_s is the permittivity of the solvent. The self-energy $\Delta W_{\text{self},i}^{\text{Poisson}}$ can be determined to high precision numerically. In this manner, the self-energy for each fixed partial charge of a solute is mapped onto the Born equation.[4] Alternatively, an analytic solution for the self-energy in terms of an energy density is possible after making the Coulomb field approximation

$$\Delta W_{\text{self},i}^{\text{GB}} = \frac{1}{2}\left(\frac{1}{\epsilon_\text{s}} - \frac{1}{\epsilon_\text{h}}\right)\frac{q_i^{\ 2}}{4\pi}\int_{\text{solvent}}\frac{1}{r^4}dV \tag{2}$$

although other methods will be elaborated below. Substituting for $\Delta W_{\text{self},i}^{\text{Poisson}}$ in eq 1 with $\Delta W_{\text{self},i}^{\text{GB}}$ from eq 2 and changing the limits of integration for convenience it can be shown that each effective Born radius is[33]

$$a_i = \left(\frac{1}{r_i} - \frac{1}{4\pi}\int_{\text{solute},r>r_i}\frac{1}{r^4}dV\right)^{-1} \tag{3}$$

where the integration over the solute does not include the region within the atomic radius r_i. A number of analytic methods have been developed for determining this integral, notably the pairwise descreening method of Hawkins, Cramer, and Truhlar that we will refer to as HCT,[8,9] a method by Qiu et al. that assumes constant energy density within each descreening atom,[11] and more recently a parameter free approach by Gallicchio et al.[34] Although effective radii determine the reaction potential at atomic centers, we note that the electrostatic solvation energy of a polarizable atomic multipole also depends on its higher order gradients.

After computing effective radii, the total self-energy of a solute within GB is

$$\Delta W_{\text{self}}^{\text{GB}} = \frac{1}{2}\left(\frac{1}{\epsilon_\text{s}} - \frac{1}{\epsilon_\text{h}}\right)\sum_i \frac{q_i^{\ 2}}{a_i} \tag{4}$$

For permanent multipoles, the self-energy of higher order components must be considered. Furthermore, if the solute is polarizable, self-consistent induced moments elicit a reaction potential that leads to an additional contribution to the electrostatic solvation free energy. We will avoid decomposing the polarization energy into self-energy and cross-term contributions, since it is inherently many-body and therefore any partitioning is somewhat artificial.

1.2. Cross-Term Energy. An analytic continuum electrostatics model designed to match results from the Poisson equation must also include an estimate of the pairwise cross-term energy between all multipole pairs. Given effective radii, the GB cross-term energy for fixed partial charges is given by

$$\Delta W_{\text{cross}}^{\text{GB}} = \frac{1}{2}\left(\frac{1}{\epsilon_\text{s}} - \frac{1}{\epsilon_\text{h}}\right)\sum_i\sum_{j\neq i}\frac{q_i q_j}{f} \tag{5}$$

where the empirical generalizing function f usually takes the form[10]

$$f = \sqrt{r_{ij}^{\ 2} + a_i a_j e^{-r_{ij}^2/c_f a_i a_j}} \tag{6}$$

where r_{ij} is the distance between sites i and j and the tuning parameter c_f is chosen in the range 2−8. As r_{ij} goes to zero, the Born formula is recovered, such that the self-energy is simply a special case of the cross-term energy. Derivation of a general form for the pairwise cross-term energy between two multipole components will be presented, which is similar in spirit to GB in that the limiting cases of superimposition and wide separation for a pair of solvated multipoles are reproduced. The accuracy of the proposed interpolation at intermediate separations will be investigated via a series of tests ranging from simple systems consisting of only two sites up to the electrostatic solvation energy and dipole moment of proteins.

Our tests of GK rely on the Polarizable Multipole Poisson−Boltzmann (PMPB) model[2] as a standard of accuracy, which has been implemented for solutes described by the Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field.[35−37] In our previous work, excellent agreement was seen in the electrostatic response of proteins solvated by the PMPB continuum when compared to ensemble average explicit water simulations, indicating that at the length scale of proteins treatment of solvent as a continuum is valid. As an alternative to numerical PMPB electrostatics, the analytic GK formulation for the AMOEBA force field is orders of magnitude more efficient.

The description of GK will be subdivided into four sections. First, determination of the self-energy for a permanent multipole will be considered. Second, we will propose a functional form for the cross-term energy between arbitrary degree multipole moments. Third, given the underlying GK theory, we continue on to the derivation of the electrostatic solvation energy and gradient in the specific case of solutes described by the AMOEBA force field. Finally, we apply the GK continuum model to a series of proteins and compare their electrostatic solvation free energy and total dipole moment to analogous calculations with the PMPB continuum.

2. Multipole Self-Energy

We begin by reiterating that the self-energy of a multipole depends not only on the reaction potential at atomic centers but also on the reaction field, the reaction field gradient, and so on. Unlike GB, the perfect effective radius is not enough information to guarantee the higher order features of the reaction potential are correct, unless the multipole site happens to be at the center of a spherical cavity. Two methods have been investigated to describe the self-energy of a permanent atomic multipole. The first method reduces to the Coulomb-field approximation (CFA) for a monopole and requires knowledge of the analytic solution for the field in solvent based on a multipole at the center of a spherical dielectric cavity.[1] We term this the solvent field approximation (SFA), as it is consistent with the CFA but requires more information. A second approach makes use of Grycuk's method for determining effective radii based on the reaction potential of an off-center charge within a spherical solute.[38] We refer to this approach as the reaction potential approximation (RPA).

Before detailing the SFA and RPA methods, a brief introduction to the electrostatic energy of a dielectric media will be given. The work required to assemble a localized fixed charge distribution in a linearly polarizable medium[33,39,40] can be formulated by a volume integral of the product of the charge density $\rho(r)$ with the potential $\phi(r)$ or by the scalar product of the electric field $\mathbf{E}$ with the electric displacement $\mathbf{D}$

$$W = \frac{1}{2}\int_V \rho(r)\phi(r)dV$$
$$= \frac{1}{8\pi}\int_V \mathbf{E}\cdot\mathbf{D}dV \tag{7}$$

where the displacement is proportional to the electric field in regions of constant permittivity ϵ

$$\mathbf{D} = \epsilon\mathbf{E} \tag{8}$$

For our purposes, the system of interest is composed of a solute with a different permittivity than the solvent. The electrostatic free energy of this system relative to a homogeneous reference state is[33,39]

$$\Delta G = \frac{1}{8\pi}\int_V (\mathbf{E}\cdot\mathbf{D} - \mathbf{E}_\mathrm{h}\cdot\mathbf{D}_\mathrm{h})\mathrm{d}V \tag{9}$$

where in the homogeneous case the field is Coulombic and can be defined relative to the vacuum field as $\mathbf{E}_\mathrm{h} = \mathbf{E}_\mathrm{vac}/\epsilon_\mathrm{h}$ using the homogeneous permittivity ϵ_h. The homogeneous displacement is simply $\mathbf{D}_\mathrm{h} = \mathbf{E}_\mathrm{vac}$. A less intuitive but equivalent definition of the electrostatic free energy given in eq 9 is[33,39]

$$\Delta G = \frac{1}{8\pi}\int_V (\mathbf{E}\cdot\mathbf{D}_\mathrm{h} - \mathbf{D}\cdot\mathbf{E}_\mathrm{h})dV \tag{10}$$

This expression can be subdivided into integrals over the solute and solvent volumes as

$$\Delta G = \frac{1}{8\pi}\int_\mathrm{solute} (\mathbf{E}\cdot\mathbf{D}_\mathrm{h} - \mathbf{D}\cdot\mathbf{E}_\mathrm{h})dV + \frac{1}{8\pi}\int_\mathrm{solvent} (\mathbf{E}\cdot\mathbf{D}_\mathrm{h} - \mathbf{D}\cdot\mathbf{E}_\mathrm{h})dV \tag{11}$$

In both the homogeneous and mixed permittivity states the solute retains the homogeneous permittivity. By using the relationships for the homogeneous field and displacement described above it can be seen that the integral over the solute vanishes

$$\Delta G = \frac{1}{8\pi}\int_\mathrm{solute} (\mathbf{E}\cdot\mathbf{E}_\mathrm{vac} - \epsilon_\mathrm{h}\mathbf{E}\cdot\mathbf{E}_\mathrm{vac}/\epsilon_\mathrm{h})dV + \frac{1}{8\pi}\int_\mathrm{solvent} (\mathbf{E}_\mathrm{s}\cdot\mathbf{E}_\mathrm{vac} - \epsilon_\mathrm{s}\mathbf{E}_\mathrm{s}\cdot\mathbf{E}_\mathrm{vac}/\epsilon_\mathrm{h})dV \tag{12}$$

to leave only the integral over the solvent

$$\Delta G = \frac{1}{8\pi}\left(1 - \frac{\epsilon_\mathrm{s}}{\epsilon_\mathrm{h}}\right)\int_\mathrm{solvent} (\mathbf{E}_\mathrm{s}\cdot\mathbf{E}_\mathrm{vac})dV \tag{13}$$

Having made no assumptions to this point, the remaining challenge can be simplified to defining the field within the solvent $\mathbf{E}_\mathrm{s}$ for the mixed permittivity case. This is the starting point for the SFA. In general, the solvent field does not have an exact analytic form for a union of spheres. However, many molecular systems of interest are globular, and therefore an approximation based on the assumption of a spherical solute is not only qualitatively reasonable but in many cases quantitative.

2.1. Solvent Field Approximation. The SFA is similar to the CFA but is based on evaluating eq 13 using Kirkwood's solution for the field outside a spherical solute with a central multipole moment[1,41]

$$\mathbf{E}_{\mathrm{s}} = \sum_{l=0}^{\infty} \frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}} \mathbf{E}_{\mathrm{vac}}^{(l)} \tag{14}$$

where $\mathbf{E}_{\mathrm{vac}}^{(l)}$ is the vacuum field due to all multipole moments of degree l, defined using either irregular spherical harmonics or Cartesian tensors. Throughout the current work we neglect salt effects, although their addition to a future GK formulation is straightforward. This definition of the self-energy is equivalent to the CFA for a monopole and becomes approximate for off-center multipole sites or for nonspherical solute geometries.

Under the SFA, the self-energy of a permanent multipole site i is given by

$$\Delta G_i^{\mathrm{SFA}} = \frac{1}{8\pi}\left(1 - \frac{\epsilon_{\mathrm{s}}}{\epsilon_{\mathrm{h}}}\right)\int_{\mathrm{solvent}} \mathbf{E}_{\mathrm{vac},i}\cdot\sum_{l=0}^{\infty}\left(\frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}}\mathbf{E}_{\mathrm{vac},i}^{(l)}\right)dV \tag{15}$$

It is possible to invert the integration domain by adding and subtracting an integral over the solute region outside the radius R_i of atom i to eq 15 giving

$$\Delta G_i^{\mathrm{SFA}} = \frac{1}{8\pi}\left(1 - \frac{\epsilon_{\mathrm{s}}}{\epsilon_{\mathrm{h}}}\right)\int_{r>R_i} \mathbf{E}_{\mathrm{vac},i}\cdot\sum_{l=0}^{\infty}\left(\frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}}\mathbf{E}_{\mathrm{vac},i}^{(l)}\right)dV - \frac{1}{8\pi}\left(1 - \frac{\epsilon_{\mathrm{s}}}{\epsilon_{\mathrm{h}}}\right)\int_{\mathrm{solute},r>R_i} \mathbf{E}_{\mathrm{vac},i}\cdot\sum_{l=0}^{\infty}\left(\frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}}\mathbf{E}_{\mathrm{vac},i}^{(l)}\right)dV \tag{16}$$

The first integral is the solvation energy of a lone multipole ΔG_i^{M} and the second represents the effect of descreening sites. Substituting ΔG_i^{M} into eq 16 gives

$$\Delta G_i^{\mathrm{SFA}} = \Delta G_i^{\mathrm{M}} - \frac{1}{8\pi}\left(1 - \frac{\epsilon_{\mathrm{s}}}{\epsilon_{\mathrm{h}}}\right)\int_{\mathrm{solute},r>R_i} \mathbf{E}_{\mathrm{vac},i}\cdot\sum_{l=0}^{\infty}\left(\frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}}\mathbf{E}_{\mathrm{vac},i}^{(l)}\right)dV \tag{17}$$

where

$$\Delta G_i^{\mathrm{M}} = \frac{1}{2}\left[c_0 \frac{q_i^{\,2}}{a_i} + c_1 \frac{\mu_{i,\alpha}^{\;2}}{a_i^{\,3}} + c_2 \frac{2}{3}\frac{\Theta_{i,\alpha\beta}^{\;2}}{a_i^{\,5}}\right] \tag{18}$$

and

$$c_l = \frac{1}{\epsilon_{\mathrm{h}}}\frac{(l+1)(\epsilon_{\mathrm{h}} - \epsilon_{\mathrm{s}})}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}} \tag{19}$$

In eq 18 we have assumed the Einstein convention for summation over Greek subscripts α and β, which can take the value x, y, or z. The descreening integral in eq 17, which we will refer to as I_i, can be decomposed into a sum of pairwise integrals I_{ij}[8,9]

$$I_i(r_{ij},R_i,R_j) = \sum_{j\neq i}\int\int_0^{\xi_{ij}}\int_0^{2\pi} \mathbf{E}_{\mathrm{vac},i}\cdot\sum_{l=0}^{\infty}\left(\frac{(2l+1)\epsilon_{\mathrm{h}}}{(l+1)\epsilon_{\mathrm{s}} + l\epsilon_{\mathrm{h}}}\mathbf{E}_{\mathrm{vac},i}^{(l)}\right)r^2 \sin\theta\, d\phi\, d\theta\, dr = \sum_{j\neq i} I_{ij}(r_{ij},R_i,R_j) \tag{20}$$

where ξ_{ij} is the angle formed between the pairwise axis and any ray that begins at the center of atom i and passes through the circle of intersection between the integration shell and atom j

$$\xi_{ij} = \cos^{-1}\left(\frac{r_{ij}^{\;2} - R_j^{\;2} + r^2}{2r_{ij}r}\right) \tag{21}$$

where r_{ij} is the distance between atoms i and j, R_j is the radius of atom j, and r is the radial integration variable. The integration limits for the radial coordinate depend on what extent atoms i and j intersect, and therefore the solution to eq 20 is presented as an indefinite integral that is to be evaluated at limits described below. Typically the radius of the descreening atom is scaled down to prevent over counting due to atomic overlap, although parameter free approaches are being explored.[34] Specifically, R_j is replaced with sR_j where the HCT scale factor s is a parameter between 0 and 1 fit to reproduce PMPB results (see section 2.3 below).

Unlike the field due to a partial charge, the field due to a multipole of arbitrary order has an angular dependence. Our approach has been to represent the field using a spherical harmonic basis, rather than Cartesian tensors, to determine the analytic solution to eq 20 through quadrupole order. Additionally, it is assumed that the positive z-axis of the multipole frame is directed toward the center of the descreening atom. This imposes symmetry that greatly reduces the number of nonvanishing terms in the solution but requires rotation of multipole moments for each pairwise descreening interaction.

A complex definition of spherical harmonics is commonly used in the formulation of quantum mechanics; however, this work uses the following real form

$$Y_l^{(m)}(\theta,\phi) = \begin{cases} (-1)^m\sqrt{2}\sqrt{\dfrac{(l-m)!}{(l+m)!}}\,P_l^{(m)}(\cos\theta)\cos m\phi & m > 0 \\ \sqrt{\dfrac{(l-m)!}{(l+m)!}}\,P_l^{(m)}(\cos\theta) & m = 0 \\ (-1)^{|m|}\sqrt{2}\sqrt{\dfrac{(l-|m|)!}{(l+|m|)!}}\,P_l^{(|m|)}(\cos\theta)\sin|m|\phi & m < 0 \end{cases} \tag{22}$$

where $Y_l^{(m)}(\theta,\phi)$ is of degree $l \geq 0$ and order $|m| \leq l$, $P_l^{(m)}$ are the associated Legendre polynomials, the polar angle ranges from $0 \leq \theta \leq \pi$, and the azimuth ranges from $0 \leq \varphi \leq 2\pi$. We chose to use the Racah normalization, which has the property that $Y_l^{(0)}(0,0) = 1$. In combination with our choice of phase factors, this ensures formulas for the conversion between Cartesian multipole moments, and those

Table 1. Indefinite Integrals for Pairwise Descreening of Multipoles Through Quadrupole

$(l,m)_1$	$(l,m)_2$	$D_{(l,m)_i,(l,m)_j}\,(r_{ij},R_j)$
(0,0)	(0,0)	$-(2\ln(r)r^2 + 4r_{ij}r - r_{ij}^2 + R_j^2)/16r_{ij}r^2$
	(1,0)	$-(4r^4\ln(r) + 4r_{ij}^2r^2 + 4r^2R_j^2 - r_{ij}^4 + 2r_{ij}^2R_j^2 - R_j^4)/64r^4r_{ij}^2$
	(2,0)	$-(12r^6\ln(r) + 6r_{ij}^2r^4 + 18r^4R_j^2 + 3r^2r_{ij}^4 + 6r^2r_{ij}^2R_j^2 - 9r^2R_j^4 - 2r_{ij}^6 + 6r_{ij}^4R_j^2 - 6r_{ij}^2R_j^4 + 2R_j^6)/256r_{ij}^3r^6$
(1,0)	(1,0)	$-(12r^6\ln(r) - 42r_{ij}^2r^4 + 18r^4R_j^2 + 64r^3r_{ij}^3 - 21r^2r_{ij}^4 + 30r^2r_{ij}^2R_j^2 - 9r^2R_j^4 - 2r_{ij}^6 + 6r_{ij}^4R_j^2 - 6r_{ij}^2R_j^4 + 2R_j^6)/384r^6r_{ij}^3$
	(2,0)	$-(24r^8\ln(r) - 48r^6r_{ij}^2 + 48r^6R_j^2 + 60r_{ij}^4r^4 + 72r_{ij}^2r^4R_j^2 - 36r^4R_j^4 - 16r^2r_{ij}^6 + 48r^2r_{ij}^4R_j^2 - 48r^2r_{ij}^2R_j^4 + 16r^2R_j^6 - 3r_{ij}^8 + 12r_{ij}^6R_j^2 - 18r_{ij}^4R_j^4 + 12r_{ij}^2R_j^6 - 3R_j^8)/1024r^8r_{ij}^4$
(1,1) (1,−1)	(1,1) (1,−1)	$(12r^6\ln(r) + 102r_{ij}^2r^4 + 18r^4R_j^2 - 128r^3r_{ij}^3 + 51r^2r_{ij}^4 - 42r^2r_{ij}^2R_j^2 - 9r^2R_j^4 - 2r_{ij}^6 + 6r_{ij}^4R_j^2 - 6r_{ij}^2R_j^4 + 2R_j^6)/768r^6r_{ij}^3$
	(2,1) (2,−1)	$\sqrt{3}\,(24r^8\ln(r) + 96r^6r_{ij}^2 + 48r^6R_j^2 - 84r_{ij}^4r^4 - 72r_{ij}^2r^4R_j^2 - 36r^4R_j^4 + 32r^2r_{ij}^6 - 48r^2r_{ij}^4R_j^2 + 16r^2R_j^6 - 3r_{ij}^8 + 12r_{ij}^6R_j^2 - 18r_{ij}^4R_j^4 + 12r_{ij}^2R_j^6 - 3R_j^8)/3072r^8r_{ij}^4$
(2,0)	(2,0)	$-3(120r^{10}\ln(r) - 140r^8r_{ij}^2 + 300r^8R_j^2 - 540r^6r_{ij}^4 + 360r^6r_{ij}^2R_j^2 - 300r^6R_j^4 + 1024r^5r_{ij}^5 - 360r^4r_{ij}^6 + 600r^4r_{ij}^4R_j^2 - 440r^4r_{ij}^2R_j^4 + 200r^4R_j^6 - 35r^2r_{ij}^8 + 180r^2r_{ij}^6R_j^2 - 330r^2r_{ij}^4R_j^4 + 260r^2r_{ij}^2R_j^6 - 75r^2R_j^8 - 12r_{ij}^{10} + 60r_{ij}^8R_j^2 - 120r_{ij}^6R_j^4 + 120r_{ij}^4R_j^6 - 60r_{ij}^2R_j^8 + 12R_j^{10})/20480r^{10}r_{ij}^5$
(2,1) (2,−1)	(2,1) (2,−1)	$(120r^{10}\ln(r) + 180r^8r_{ij}^2 + 300r^8R_j^2 + 900r^6r_{ij}^4 - 120r^6r_{ij}^2R_j^2 - 300r^6R_j^4 - 1536r^5r_{ij}^5 + 600r^4r_{ij}^6 - 680r^4r_{ij}^4R_j^2 - 120r^4r_{ij}^2R_j^4 + 200r^4R_j^6 + 45r^2r_{ij}^8 - 60r^2r_{ij}^6R_j^2 - 90r^2r_{ij}^4R_j^4 + 180r^2r_{ij}^2R_j^6 - 75r^2R_j^8 - 12r_{ij}^{10} + 60r_{ij}^8R_j^2 - 120r_{ij}^6R_j^4 + 120r_{ij}^4R_j^6 - 60r_{ij}^2R_j^8 + 12R_j^{10})/10240r^{10}r_{ij}^5$
(2,2) (2,−2)	(2,2) (2,−2)	$-(120r^{10}\ln(r) + 1140r^8r_{ij}^2 + 300r^8R_j^2 - 4380r^6r_{ij}^4 - 1560r^6r_{ij}^2R_j^2 - 300r^6R_j^4 + 6144r^5r_{ij}^5 - 2920r^4r_{ij}^6 + 1880r^4r_{ij}^4R_j^2 + 840r^4r_{ij}^2R_j^4 + 200r^4R_j^6 + 285r^2r_{ij}^8 - 780r^2r_{ij}^6R_j^2 + 630r^2r_{ij}^4R_j^4 - 60r^2r_{ij}^2R_j^6 - 75r^2R_j^8 - 12r_{ij}^{10} + 60r_{ij}^8R_j^2 - 120r_{ij}^6R_j^4 + 120r_{ij}^4R_j^6 - 60r_{ij}^2R_j^8 + 12R_j^{10})/40960r^{10}r_{ij}^5$

consistent with this definition of real spherical harmonics are identical to the conversions commonly used for complex spherical harmonics. The conversion formulas through quadrupole degree are given in Table A-1 of the Supporting Information.[42]

The potential due to a unit magnitude multipole moment $\Phi_l^{(m)}(r,\theta,\phi)$ is obtained by multiplication of the real spherical harmonics by a radial factor of $1/r^{l+1}$ to give

$$\Phi_l^{(m)}(r,\theta,\phi) = \frac{Y_l^{(m)}(\theta,\phi)}{r^{l+1}} \tag{23}$$

and are listed in Table A-2 (Supporting Information) through quadrupole order. The unit field can then be calculated as the negative gradient of the unit potential

$$\mathbf{E}_l^{(m)} = -\nabla\Phi_l^{(m)}(r,\theta,\phi)$$

$$= -\frac{\partial\Phi_l^{(m)}(r,\theta,\phi)}{\partial r}\hat{\mathbf{r}} - \frac{1}{r}\frac{\partial\Phi_l^{(m)}(r,\theta,\phi)}{\partial\theta}\hat{\boldsymbol{\theta}} - \frac{1}{r\sin\theta}\frac{\partial\Phi_l^{(m)}(r,\theta,\phi)}{\partial\phi}\hat{\phi} \tag{24}$$

The field for 9 multipole components through degree 2, which are listed in Table A-3 of the Supporting Information, lead to 36 scalar products that must be integrated via eq 20 to determine the descreening energy due to atom j. However, due to the symmetry of the integration domain only 14 scalar products lead to nonzero integrals, and these are listed in Table A-4 (Supporting Information). The integration results are given in Table 1, showing 10 unique terms and 4 duplicates. Schaeffer et al. originally presented the same result for a monopole,[6,7] and the higher order formulas are presented here for the first time. If the descreening angle ξ_{ij} is π as a result of atom j completely engulfing atom i, then the indefinite integrals simplify to those given in Table 2. This situation can occur for hydrogen atoms bonded to a heavy atom, for example, or in more artificial structures where one still wishes to have a continuous potential.

We note that after performing the integration no angular dependence remains. Therefore, although the derivation is based on spherical harmonics, our solution is equally useful for Cartesian tensors by using the conversion formulas in Table A-1 (Supporting Information). We can now define the pairwise descreening integral for a permanent atomic multipole at site i being descreened by site j under the SFA as

$$I_{ij}(r_{ij},R_i,R_j) = \sum_{l_i=0}^{n}\frac{(2l_i+1)\epsilon_\mathrm{h}}{(l_i+1)\epsilon_\mathrm{s} + l_i\epsilon_\mathrm{h}}\sum_{m_i=-l_i}^{l_i}Q_{l_i}^{(m_i)} \times \sum_{l_j=0}^{n}\sum_{m_j=-l_j}^{l_j}Q_{l_j}^{(m_j)}D_{(l,m)_i,(l,m)_j}(r_{ij},R_i,R_j) \tag{25}$$

where $Q_{l_i}^{(m_i)}$ is the magnitude of a spherical harmonic of site i, $Q_{l_j}^{(m_j)}$ is the magnitude of a spherical harmonic of site j, and $D_{(l,m)_i,(l,m)_j}(r_{ij},R_i,R_j)$ is given by

$$D_{(l,m)_i,(l,m)_j}(r_{ij},R_i,R_j) = \begin{cases} \delta_{(l_1,l_2)}\delta_{(m_1,m_2)}D_{l_i}\big|_{r=R_i}^{r=R_j-r_{ij}} + & \\ D_{(l,m)_i,(l,m)_j}(r_{ij},R_j)\big|_{r=R_j-r_{ij}}^{r=r_{ij}+R_j} & R_j - r_{ij} > R_i \\ \text{Case 1: Engulfment by the descreener} & \\ D_{(l,m)_i,(l,m)_j}(r_{ij},R_j)\big|_{r=R_i}^{r=r_{ij}+R_j} & R_j - r_{ij} <= R_i \\ & r_{ij} < R_i + R_j \\ \text{Case 2: Partial overlap} & \\ D_{(l,m)_i,(l,m)_j}(r_{ij},R_j)\big|_{r=r_{ij}-R_j}^{r=r_{ij}+R_j} & r_{ij} > R_i + R_j \\ \text{Case 3: No overlap} & \end{cases} \tag{26}$$

Radial limits are detailed for three cases including engulfment by the descreener, partial overlap, and no overlap. These limits are applied in conjunction with the indefinite integrals $D_{(l,m)_i,(l,m)_j}(r_{ij},R_j)$ and D_{l_i} listed in Tables 1 and 2, respectively. We note that the Kronecker delta functions δ specify that the engulfment integrals between orthogonal spherical harmonics vanish. In our implementation of eq 25, the magnitudes of the spherical harmonic moments are found via conversion from AMOEBA traceless Cartesian multipoles.

2.2. Reaction Potential Approximation. An alternative to the CFA for determining effective radii based on the analytic solution for the reaction potential of an off-center charge within a spherical dielectric cavity[1,43] has been proposed by Grycuk.[38] We briefly outline this RPA method and its application to the self-energy of a permanent multipole.

The reaction potential at $\boldsymbol{r}$ due to an off-center charge at $\boldsymbol{r_0}$ inside a spherical dielectric cavity of permittivity ϵ_h surrounded by solvent with permittivity ϵ_s is given by

$$\Phi(\mathbf{r}) = \frac{q}{a\epsilon_h}\sum_{l=0}^{\infty}\frac{(l+1)(\epsilon_h - \epsilon_s)}{(l+1)\epsilon_s + l\epsilon_h}\left(\frac{rr_0}{a^2}\right)^l P_l(\cos\theta) \tag{27}$$

where a is the cavity radius, q is the magnitude of the charge, and P_l is the Legendre polynomial of degree l whose argument is the cosine of the angle θ between r and r_0.[1,43] The self-energy of a charge based on eq 27 is

$$W(d) = \frac{1}{2}\frac{q^2}{a\epsilon_h}\sum_{l=0}^{\infty}\frac{(l+1)(\epsilon_h - \epsilon_s)}{(l+1)\epsilon_s + l\epsilon_h}\left(\frac{d^2}{a^2}\right)^l P_l(1) \tag{28}$$

where d is used to specify the distance between the multipole site and the center of the sphere. For $d = 0$, all asymmetric self-interactions vanish, for example the charge with a dipole component, but for off-center multipole sites these interactions are generally nonzero.[5] Noting that $P_l(1) = 1$ for all l, the summation in eq 28 can be reduced to a closed form if the factor $(l + 1)$ can be canceled by setting $l\epsilon_h$ in the denominator to $(l + 1)\epsilon_h$ or to 0, giving quantities that are more positive $W_+(d)$ or more negative $W_-(d)$ than the true self-energy, respectively

$$\begin{aligned} W_+(d) &= \frac{1}{2}\frac{(\epsilon_h - \epsilon_s)}{\epsilon_s\epsilon_h + \epsilon_h{}^2}\frac{q^2}{a}\sum_{l=0}^{\infty}\left(\frac{d^2}{a^2}\right)^l \\ &= \frac{1}{2}\frac{(\epsilon_h - \epsilon_s)}{(\epsilon_s\epsilon_h + \epsilon_h{}^2)}q^2\frac{a}{(a^2 - d^2)} \\ W_-(d) &= \frac{1}{2}\left(\frac{1}{\epsilon_s} - \frac{1}{\epsilon_h}\right)\frac{q^2}{a}\sum_{l=0}^{\infty}\left(\frac{d^2}{a^2}\right)^l \\ &= \frac{1}{2}\left(\frac{1}{\epsilon_s} - \frac{1}{\epsilon_h}\right)q^2\frac{a}{(a^2 - d^2)} \end{aligned} \tag{29}$$

Both the upper and lower bound approach the true self-energy if $\epsilon_s \gg \epsilon_h$ allowing the simpler form to be used as an approximation

$$W(d) \approx W_-(d) \tag{30}$$

As shown by Grycuk, it is possible to calculate the factor $a_r = a/(a^2 - d^2)$, which is equivalent to the inverse of an effective radius, as

$$a_r = \left(\frac{3}{4\pi}\int_{ex}\frac{1}{r'^6}dV\right)^{1/3} \tag{31}$$

This expression can be motivated by the analytic solution for a spherical geometry

$$\begin{aligned} \int_{ex}\frac{1}{r'^6}dV &= 2\pi\int_a^{\infty}\int_0^{\pi}\frac{r^2\sin\theta}{(r^2 + d^2 - 2dr\cos\theta)^3}d\theta dr \\ &= \frac{4\pi}{3}\frac{a^3}{(a^2 - d^2)^3} \end{aligned} \tag{32}$$

As d approaches zero, the multipole site approaches the center of the dielectric sphere such that a_r equals the radius of the sphere a. In practice this integral is evaluated using the pairwise descreening approach described in the previous section for the SFA and elsewhere.[8,9,38] After determining effective radii, the self-energy for each permanent atomic multipole under the RPA is evaluated via eq 18.

2.3. Self-Energy Accuracy. We now demonstrate that for a series of proteins the RPA is superior to the SFA, which is consistent with findings for fixed partial charge models.[38,44] The perfect self-energy and perfect effective radii for all permanent atomic multipole sites in five protein structures retrieved from the Protein Databank,[45] including 1CRN,[46] 1ENH,[47] 1FSV,[48] 1PGB,[49] and 1VII,[50] were determined using the PMPB model.[2] The grid size for all calculations was 257 × 257 × 257 using a grid spacing of 0.31 Å to give approximately 10 Å of continuum solvent between the low dielectric boundary and the grid boundary. The Bondi radii set (H 1.2, C 1.7, N 1.55, O 1.52, S 1.8) was used to define a step-function solute−solvent boundary with the solute dielectric set to unity and that of the solvent to 78.3.[51] Multiple Debye-Hückel boundary conditions were used to complete the definition of the Dirichlet problem. We also tried larger grids, up to 353 × 353 × 353, and therefore

Table 2. Indefinite Integrals for Pairwise Descreening of Multipoles through Quadrupole When $\xi_{ij} = \pi$

l_i	D_{l_i}
0	$-1/2r$
1	$-1/3r^3$
2	$-3/10r^5$

Table 3. Shown Is a Comparison of the Performance of the SFA and RPA in Determining the Perfect Self-Energy (kcal/mol) for a Series of Five Folded Proteins[a]

	self-energy			signed % difference		unsigned % difference	
	PMPB	SFA	RPA	SFA	RPA	SFA	RPA
CRN	−8141	−8191	−8196	−0.6	−0.7	0.6	0.7
ENH	−11919	−11852	−11878	0.6	0.3	0.6	0.3
FSV	−6254	−6341	−6287	−1.4	−0.5	1.4	0.5
PGB	−11794	−11743	−11803	0.4	−0.1	0.4	0.1
VII	−7206	−7132	−7133	1.0	1.0	1.0	1.0
mean				0.0	0.0	0.8	0.5

[a] Optimization of a single HCT scale factor for each method removes systematic error as shown by the mean signed percent differences. However, the mean RPA unsigned percent difference of 0.5 is smaller than that of the SFA.

smaller grid spacing, which leads to the PMPB electrostatic solvation energy increasing by less than 2%. We opted for efficiency, since the important conclusion of this section, that the RPA is superior to the SFA, is not altered.

The SFA was fit using nonlinear optimization to determine one HCT scale factor per atomic number that minimized the rms percent error in the permanent atomic multipole self-energies against numerical PMPB results for 3032 data points. As discussed previously, these HCT parameters scale down the radius of the descreening atom to prevent over counting due to atomic overlap. This lead to a mean unsigned relative difference (MURD) between the perfect self-energy for each multipole site and the SFA self-energy of 5.5%. However, using only a single scale factor (0.568), rather than one per atomic number, increased the MURD by just 0.4 to 5.9%.

Similarly, the RPA was fit using nonlinear optimization to determine a second set of scale factors to minimize the rms percent difference between analytic effective radii and perfect effective radii. The achieved MURD in the effective radii was 1.1%. Alternatively, using a single scale factor (0.690) increased the MUPD by only 0.2% to 1.3%. Therefore, given the negligible improvements of using one HCT parameter per atomic number, we prefer implementations of the SFA and RPA that are each based on a single parameter.

The total analytic self-energy for each protein is compared to the total computed by summing the numerical permanent multipole self-energies as shown in Table 3. Fitting of a single HCT parameter for each method as described above eliminated the systematic error for both the SFA and RPA. However, the mean unsigned percent difference of the RPA (0.5) is smaller than that of the SFA (0.8). Considering that the RPA is more efficient and more accurate than the SFA, it is our preferred method to compute effective radii and permanent multipole self-energies.

3. Multipole Cross-Term Energy

There are two concepts needed to extend the GB cross-term to the interaction between two arbitrary multipole components. First, we describe the simplest possible definition for the reaction potential of any multipole component in the presence of a second multipole site, where an effective radius characterizes each site. Second, using this auxiliary definition of the reaction potential for each site, we formulate the cross-term energy in a consistent fashion. The electrostatic solvation free energy for the interaction between multipole components will be reproduced in the limiting cases of superimposition and wide separation.

3.1. Generalized Kirkwood Auxiliary Reaction Potential. The generalized Kirkwood auxiliary reaction potential is a building block for defining the interaction energy and its gradients for any pair of multipole components. It is motivated by noting that the only difference between the analytic solution for the reaction potential inside and outside of a spherical solute with central multipole is exchange of the solute radius a in the former case with separation distance r_{ij} in the latter, where $\mathbf{r}_{ij} = (x_j - x_i, y_j - y_i, z_j - z_i)$.[1,41] For example, substitution for f in eq 33 below by a or r_{ij} gives the analytic formulas for the reaction potential inside and outside of the dielectric boundary, respectively.

Rather than using radial factors of $1/r_{ij}^{l+1}$ as was done earlier in defining the unit vacuum potential in terms of real spherical harmonics, the factor $r_{ij}^{l}/f^{\,2l+1}$ is used to define the unit GK auxiliary reaction potential $A_l^{(m)}$ for a multipole component of degree l and order m

$$A_l^{(m)}(\mathbf{r}_{ij},a_i,a_j,\theta,\phi) = c_l \frac{r_{ij}^{l}}{f^{\,2l+1}} Y_l^{(m)}(\theta,\phi) \tag{33}$$

where f is the generalizing function defined in eq 6, and c_l is a function of the permittivity inside and outside the solute defined in eq 19. We note that for $r_{ij}^2 \gg a_i a_j$, $r_{ij}^{l}/f^{\,2l+1}$ approaches $1/r_{ij}^{l+1}$ to give the reaction potential in solvent. When $r_{ij} = 0$ and therefore $a_i = a_j = a$, then $r_{ij}^{l}/f^{\,2l+1}$ simplifies to r_{ij}^{l}/a^{2l+1} to give the reaction potential at the center of the two concentric atoms. In this case the reaction potential is nonzero only for the monopole.

A definition in terms of Cartesian tensors is possible by first taking successive gradients of $1/r_{ij}$ and then substituting for factors of r_{ij} in the denominator with factors of f. For example, neglecting the ij subscript, the vacuum tensors are[42]

$$T = \frac{1}{r}$$

$$T_\alpha = \nabla_\alpha \frac{1}{r} = -\frac{r_\alpha}{r^3}$$

$$T_{\alpha\beta} = \nabla_\alpha \nabla_\beta \frac{1}{r} = \frac{3 r_\alpha r_\beta}{r^5} - \frac{\delta_{\alpha\beta}}{r^3}$$

$$T_{\alpha\beta\gamma} = \nabla_\alpha \nabla_\beta \nabla_\gamma \frac{1}{r} = -\frac{15 r_\alpha r_\beta r_\gamma}{r^7} + \frac{3(r_\alpha \delta_{\beta\gamma} + r_\beta \delta_{\alpha\gamma} + r_\gamma \delta_{\alpha\beta})}{r^5}$$

$$T_{\alpha\beta\gamma\delta} = \nabla_\alpha \nabla_\beta \nabla_\gamma \nabla_\delta \frac{1}{r} = \frac{105 r_\alpha r_\beta r_\gamma r_\delta}{r^9} - \frac{15(r_\alpha r_\beta \delta_{\gamma\delta} + r_\alpha r_\gamma \delta_{\beta\delta} + r_\alpha r_\delta \delta_{\beta\gamma} + r_\beta r_\gamma \delta_{\alpha\delta} + r_\beta r_\delta \delta_{\alpha\gamma} + r_\gamma r_\delta \delta_{\alpha\beta})}{r^7} + \frac{3(\delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma})}{r^5} \tag{34}$$

where α, β, γ, and δ can take the values x, y, or z, and the Kronecker delta function is unity if its subscripts are equal, but zero otherwise. Applying the substitution gives

$$A = c_0 \frac{1}{f}$$

$$A_\alpha = -c_1 \frac{r_\alpha}{f^3}$$

$$A_{\alpha\beta} = c_2 \frac{3r_\alpha r_\beta}{f^5}$$

$$A_{\alpha\beta\gamma} = -c_3 \frac{15 r_\alpha r_\beta r_\gamma}{f^7}$$

$$A_{\alpha\beta\gamma\delta} = c_4 \frac{105 r_\alpha r_\beta r_\gamma r_\delta}{f^9} \tag{35}$$

which represents the GK auxiliary reaction potential tensors. We have removed terms that require summing over a trace by requiring use of traceless multipoles. Unlike the vacuum case, the GK auxiliary reaction potential tensor of degree *l* is not simply a gradient of a degree *l*-1 tensor.

The total auxiliary reaction potential due to multipole *i*, up to quadrupole order, at site *j* is

$$\phi^{(i)}(\mathbf{r}_{ij},a_i,a_j) = q_i A - \mu_{i,\alpha} A_\alpha + \frac{1}{3}\Theta_{i,\alpha\beta} A_{\alpha\beta} \tag{36}$$

where the Einstein convention for repeated summation over Greek subscripts is implied. The total auxiliary potential due to multipole *j*, up to quadrupole order, at site *i* is given by

$$\phi^{(j)}(\mathbf{r}_{ji},a_i,a_j) = q_j A - \mu_{j,\alpha} A_\alpha + \frac{1}{3}\Theta_{j,\alpha\beta} A_{\alpha\beta} \tag{37}$$

where $\boldsymbol{r}_{ji}$ is defined from site *j* to site *i*.

3.2. Generalized Kirkwood Cross-Term. Given the auxiliary reaction potentials, we define the auxiliary cross-term energy using eq 36 to be

$$U^{(i)}(\mathbf{r}_{ij},a_i,a_j) = \frac{1}{2}\left(q_j\phi^{(i)} + \mu_{j,\gamma}\nabla_\gamma\phi^{(i)} + \frac{1}{3}\Theta_{j,\gamma\delta}\nabla_\gamma\nabla_\delta\phi^{(i)}\right) \tag{38}$$

such that substituting for $\phi^{(i)}$ gives

$$\begin{aligned} U^{(i)}(\mathbf{r}_{ij},a_i,a_j) = \frac{1}{2}\Big[& q_j\Big(q_i A - \mu_{i,\alpha}A_\alpha + \frac{1}{3}\Theta_{i,\alpha\beta}A_{\alpha\beta}\Big) + \\ & \mu_{j,\gamma}\nabla_\gamma\Big(q_i A - \mu_{i,\alpha}A_\alpha + \frac{1}{3}\Theta_{i,\alpha\beta}A_{\alpha\beta}\Big) + \\ & \frac{1}{3}\Theta_{j,\gamma\delta}\nabla_\gamma\nabla_\delta\Big(q_i A - \mu_{i,\alpha}A_\alpha + \frac{1}{3}\Theta_{i,\alpha\beta}A_{\alpha\beta}\Big)\Big] \end{aligned} \tag{39}$$

while the auxiliary cross-term energy using eq 37 is

$$U^{(j)}(\mathbf{r}_{ji},a_i,a_j) = \frac{1}{2}\left(q_i\phi^{(j)} + \mu_{i,\gamma}\nabla_\gamma\phi^{(j)} + \frac{1}{3}\Theta_{i,\gamma\delta}\nabla_\gamma\nabla_\delta\phi^{(j)}\right) \tag{40}$$

such that substituting for $\phi^{(j)}$ gives

$$\begin{aligned} U^{(j)}(\mathbf{r}_{ji},a_i,\mathrm{a}_j) = \frac{1}{2}\Big[& q_i\Big(q_j A - \mu_{j,\alpha}A_\alpha + \frac{1}{3}\Theta_{j,\alpha\beta}A_{\alpha\beta}\Big) + \\ & \mu_{i,\gamma}\nabla_\gamma\Big(q_j A - \mu_{j,\alpha}A_\alpha + \frac{1}{3}\Theta_{j,\alpha\beta}A_{\alpha\beta}\Big) + \\ & \frac{1}{3}\Theta_{i,\gamma\delta}\nabla_\gamma\nabla_\delta\Big(q_j A - \mu_{j,\alpha}A_\alpha + \frac{1}{3}\Theta_{j,\alpha\beta}A_{\alpha\beta}\Big)\Big] \end{aligned} \tag{41}$$

In the case of superimposition, either $U^{(i)}$ or $U^{(j)}$ exactly reproduces the correct self-energies. In the case of wide separation, both $\phi^{(i)}$ and $\phi^{(j)}$ neglect the bending of field lines near the spherical dielectric cavity surrounding site *j* and site *i*, respectively. The density of field lines in the case of wide separation is not an issue for a fixed partial charge interaction, although neglect of this effect introduces an error of less than 1% for dipole interactions in the case of a solute with unit permittivity in water.

Gradients of the auxiliary reaction potential can easily be obtained, although it is important to note that

$$\nabla_\alpha A \neq A_\alpha \tag{42}$$

Namely, $\nabla_\alpha A$ includes a factor of $(1 - e^{-r_{ij}^2/c_f a_i a_j}/c_f)$ relative to A_α such that equality is only achieved for r_{ij} equal to zero or infinity. This subtle point implies, not surprisingly, the auxiliary reaction potential is too simple for intermediate r_{ij}. An important consequence is that $U^{(i)} \neq U^{(j)}$. A consistent model requires that the α-component of the potential gradient at site *j* of a unit charge at site *i* should equal the potential at site *i* of the dipole's unit magnitude α-component at site *j*. This reciprocity condition is a well-known property of linear dielectric continuums.[40] We note that in practice $\nabla_\alpha A \approx A_\alpha$, and, therefore, we simply take the average of the energies to obtain a consistent interaction model.

$$\Delta G_{ij} = \frac{1}{2}(U^{(i)} + U^{(j)}) \tag{43}$$

The qualitative behavior of the GK cross-term formulation for multipole permutations through quadrupole degree is seen in Figure 1. The system is composed of two spheres, each with a radius of 3.0 Å and unit permittivity, in a solvent with permittivity 78.3. The total electrostatic solvation energy was evaluated using the PMPB and GK models. In the case of superimposition, the GK value is exact. When the two spheres are widely separated, GK asymptotes to the PMPB results for all permutations. For intermediate separations, the behavior is promising but not exact.

4. Amoeba Solutes under Generalized Kirkwood

4.1. Electrostatic Solvation Free Energy. Derivation of the electrostatic solvation free energy for an AMOEBA solute[35−37] within the GK continuum resembles the derivation of the PMPB electrostatic solvation free energy.[2] Each permanent atomic multipole site can be considered as a vector of coefficients including charge, dipole, and quadrupole components

$$\mathbf{M}_i = [q_i, d_{i,x}, d_{i,y}, d_{i,z}, \Theta_{i,xx}, \Theta_{i,xy}, \Theta_{i,xz}, \ldots, \Theta_{i,zz}]^t \tag{44}$$

where the superscript *t* denotes the transpose. The interaction

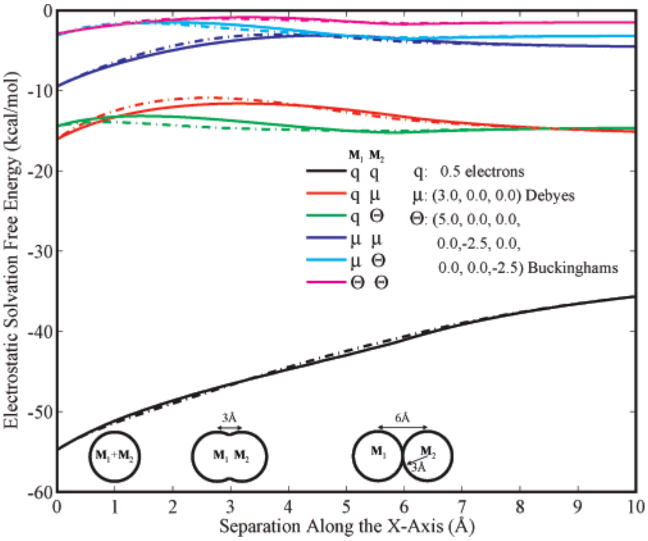


Figure 1. The solvation energy for a system composed two spheres, each with a radius of 3 Å and a permittivity of 1, and a variety of multipole combinations are computed as a function of separation along the *x*-axis using numerical Poisson solutions (solid lines) and generalized Kirkwood (dashed lines). The solvent permittivity was 78.3. The limiting cases of wide separation and superimposition are reproduced for all combinations, while intermediate separations are seen to be a reasonable approximation.

potential energy between two sites *i* and *j* separated by the distance r_{ij} in a homogeneous permittivity ϵ_h can then be represented in tensor notation as

$$U(\mathbf{r}_{ij}) = \mathbf{M}_i^t \mathbf{T}_{ij} \mathbf{M}_j$$

$$= \begin{bmatrix} q_i \\ d_{i,x} \\ d_{i,y} \\ d_{i,z} \\ \Theta_{i,xx} \\ \vdots \end{bmatrix}^t \begin{bmatrix} 1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \cdots \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \cdots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \cdots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \frac{1}{\epsilon_h r_{ij}} \begin{bmatrix} q_j \\ d_{j,x} \\ d_{j,y} \\ d_{j,z} \\ \Theta_{j,xx} \\ \vdots \end{bmatrix} \tag{45}$$

Similarly, the GK energy for two multipoles (self or cross-term) is given by

$$\Delta G_{ij}(\mathbf{r}_{ij}, a_i, a_j) = \frac{1}{2} \mathbf{M}_i^t \mathbf{K}_{ij} \mathbf{M}_j \tag{46}$$

where the factor of $^1/_2$ accounts for the cost of charging the continuum, and the GK interaction matrix $\mathbf{K}_{ij}$ depends on the coordinates of all atoms via the effective radii a_i and a_j. As introduced above, GK requires averaging of the auxiliary reaction potentials and their respective gradients to obtain a consistent interaction matrix

$$\mathbf{K}_{ij} = \frac{1}{2}[\mathbf{K}^{(i)} + \mathbf{K}^{(j)}]$$

$$\mathbf{K}^{(i)}(\mathbf{r}_{ij}, a_i, a_j) = \begin{bmatrix} A & \frac{\partial A}{\partial x} & \frac{\partial A}{\partial y} & \frac{\partial A}{\partial z} & \cdots \\ A_x & \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial z} & \cdots \\ A_y & \frac{\partial A_y}{\partial x} & \frac{\partial A_y}{\partial y} & \frac{\partial A_y}{\partial z} & \cdots \\ A_z & \frac{\partial A_z}{\partial x} & \frac{\partial A_z}{\partial y} & \frac{\partial A_z}{\partial z} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\mathbf{K}^{(j)}(\mathbf{r}_{ji}, a_i, a_j) = (\mathbf{K}^{(i)}(\mathbf{r}_{ji}, a_i, a_j))^t \tag{47}$$

Each site may also be polarizable, such that an induced dipole is formed in vacuum μ_i^v proportional to the strength of the local field

$$\boldsymbol{\mu}_i^v = \alpha_i \mathbf{E}_i^v$$

$$= \alpha_i\Big(\sum_{j \neq i} \mathbf{T}_{d,ij}^{(1)} \mathbf{M}_j + \sum_{k \neq i} \mathbf{T}_{ik}^{(11)} \mu_k\Big) \tag{48}$$

Here α_i is an isotropic atomic polarizability, and $\mathbf{E}_i^v$ is the total vacuum field, which can be decomposed into contributions from permanent multipole sites and induced dipoles, and the summations run over all multipole sites. The interaction tensors $\mathbf{T}_{d,ij}^{(1)}$ and $\mathbf{T}_{ik}^{(11)}$ are, respectively

$$\mathbf{T}_{d,ij}{}^{(1)} = \begin{bmatrix} \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \cdots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \cdots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \cdots \end{bmatrix} \frac{1}{\epsilon_h r_{ij}} \tag{49}$$

and

$$\mathbf{T}_{ik}^{(11)} = \begin{bmatrix} \frac{\partial^2}{\partial x_i \partial x_k} & \frac{\partial^2}{\partial x_i \partial y_k} & \frac{\partial^2}{\partial x_i \partial z_k} \\ \frac{\partial^2}{\partial y_i \partial x_k} & \frac{\partial^2}{\partial y_i \partial y_k} & \frac{\partial^2}{\partial y_i \partial z_k} \\ \frac{\partial^2}{\partial z_i \partial x_k} & \frac{\partial^2}{\partial z_i \partial y_k} & \frac{\partial^2}{\partial z_i \partial z_k} \end{bmatrix} \frac{1}{\epsilon_h r_{ik}} \tag{50}$$

where the *d* in $\mathbf{T}_{d,ij}^{(1)}$ denotes that masking rules for the AMOEBA group-based polarization model are applied. Upon adding the GK reaction field due to the permanent multipoles and induced dipoles, the self-consistent induced dipoles are proportional to the self-consistent reaction field

$$\boldsymbol{\mu}_i = \alpha_i \mathbf{E}_i$$

$$= \alpha_i\Big[\sum_j [(1 - \delta_{ij})\mathbf{T}_{d,ij}^{(1)} + \mathbf{K}_{ij}^{(1)}]\mathbf{M}_j + \sum_k [(1 - \delta_{ik})\mathbf{T}_{ik}^{(11)} + \mathbf{K}_{ik}^{(11)}]\boldsymbol{\mu}_k\Big] \tag{51}$$

where the sums now include self-contributions to the reaction field but exclude Coulomb self-interactions via Kronecker delta functions. The GK interaction matrices $\mathbf{K}_{ij}^{(1)}$ and $\mathbf{K}_{ik}^{(11)}$ are, respectively

$$\mathbf{K}_{ij}^{(1)} = \frac{1}{2}\left(\mathbf{K}_{ij}^{(1,i)}(\mathbf{r}_{ij},a_i,a_j) + \mathbf{K}_{ij}^{(1,j)}(\mathbf{r}_{ji},a_i,a_j)\right) \tag{52}$$

where

$$\mathbf{K}_{ij}^{(1,i)}(\mathbf{r}_{ij},a_i,a_j) = \begin{bmatrix} A_x & \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial z} & \ldots \\ A_y & \frac{\partial A_y}{\partial x} & \frac{\partial A_y}{\partial y} & \frac{\partial A_y}{\partial z} & \ldots \\ A_z & \frac{\partial A_z}{\partial x} & \frac{\partial A_z}{\partial y} & \frac{\partial A_z}{\partial z} & \ldots \end{bmatrix} \tag{53}$$

$$\mathbf{K}_{ij}^{(1,j)}(\boldsymbol{r}_{ji},a_i,a_j) = \begin{bmatrix} \frac{\partial A}{\partial x} & \frac{\partial A_x}{\partial x} & \frac{\partial A_y}{\partial x} & \frac{\partial A_z}{\partial x} & \ldots \\ \frac{\partial A}{\partial y} & \frac{\partial A_x}{\partial y} & \frac{\partial A_y}{\partial y} & \frac{\partial A_z}{\partial z} & \ldots \\ \frac{\partial A}{\partial z} & \frac{\partial A_x}{\partial z} & \frac{\partial A_y}{\partial z} & \frac{\partial A_z}{\partial z} & \ldots \end{bmatrix} \tag{54}$$

and

$$\mathbf{K}_{ik}^{(11)} = \begin{bmatrix} \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial z} \\ \frac{\partial A_y}{\partial x} & \frac{\partial A_y}{\partial y} & \frac{\partial A_y}{\partial z} \\ \frac{\partial A_z}{\partial x} & \frac{\partial A_z}{\partial y} & \frac{\partial A_z}{\partial z} \end{bmatrix} \tag{55}$$

where averaging cancels for the matrix $\mathbf{K}_{ik}^{(11)}$ that produces the field at site *i* due to the induced dipole at site *k* as a result of symmetry.

The linear system of equations, both for the vacuum and solvated systems, can be solved via a number of approaches, including direct matrix inversion or iterative schemes such as successive over-relaxation (SOR). The total vacuum electrostatic energy $\mathrm{U}_{\mathrm{elec}}^{\mathrm{v}}$ includes pairwise permanent multipole interactions and many-body polarization

$$\mathrm{U}_{\mathrm{elec}}^{\mathrm{v}} = \frac{1}{2}\left[\mathbf{M}^t\mathbf{T} - (\boldsymbol{\mu}^{\mathrm{v}})^t\mathbf{T}_{\mathrm{p}}^{(1)}\right]\mathbf{M} \tag{56}$$

where the factor of $^1/_2$ avoids double-counting of permanent multipole interactions in the first term and accounts for the cost of polarizing the system in the second term. Furthermore, **M** is a column vector of 13N multipole components

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \vdots \\ \mathbf{M}_{\mathrm{N}} \end{bmatrix} \tag{57}$$

T is a N × N supermatrix with $\mathbf{T}_{ij}$ off-diagonal elements

$$\mathbf{T} = \begin{bmatrix} 0 & \mathbf{T}_{12} & \mathbf{T}_{13} & \ldots \\ \mathbf{T}_{21} & 0 & \mathbf{T}_{23} & \ldots \\ \mathbf{T}_{31} & \mathbf{T}_{32} & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{58}$$

$\boldsymbol{\mu}^{\mathrm{v}}$ is a 3N column vector of converged induced dipole components in vacuum

$$\boldsymbol{\mu}^{\mathrm{v}} = \begin{bmatrix} \mu_{1,x}^{v} \\ \mu_{1,y}^{v} \\ \mu_{1,z}^{v} \\ \vdots \\ \mu_{N,z}^{v} \end{bmatrix} \tag{59}$$

and $\mathbf{T}_{\mathrm{p}}^{(1)}$ is a 3N × 13N supermatrix with $\mathbf{T}_{\mathrm{p},ij}^{(1)}$ as off-diagonal elements

$$\mathbf{T}_{\mathrm{p}}^{(1)} = \begin{bmatrix} 0 & \mathbf{T}_{\mathrm{p},12}^{(1)} & \mathbf{T}_{\mathrm{p},13}^{(1)} & \ldots \\ \mathbf{T}_{\mathrm{p},21}^{(1)} & 0 & \mathbf{T}_{\mathrm{p},23}^{(1)} & \ldots \\ \mathbf{T}_{\mathrm{p},31}^{(1)} & \mathbf{T}_{\mathrm{p},32}^{(1)} & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{60}$$

The subscript p denotes a tensor matrix that operates on the permanent multipoles to produce the electric field in which the polarization energy is evaluated, while the subscript d is used to specify an analogous tensor matrix that produces the field that induces dipoles. The differences between the two are masking rules that scale the 1−2, 1−3, and 1−4 interactions in the former case and use the AMOEBA group based polarization scheme for the later.[35]

For the solvated system, the total electrostatic energy is similar to the vacuum case

$$\mathrm{U}_{\mathrm{elec}} = \frac{1}{2}\left[\mathbf{M}^t(\mathbf{T} + \mathbf{K}) - \boldsymbol{\mu}^t(\mathbf{T}_{\mathrm{p}}^{(1)} + \mathbf{K}^{(1)})\right]\mathbf{M} \tag{61}$$

where the GK matrices are

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \mathbf{K}_{13} & \ldots \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \mathbf{K}_{23} & \ldots \\ \mathbf{K}_{31} & \mathbf{K}_{32} & \mathbf{K}_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{62}$$

and

$$\mathbf{K}^{(1)} = \begin{bmatrix} \mathbf{K}_{11}^{(1)} & \mathbf{K}_{12}^{(1)} & \mathbf{K}_{13}^{(1)} & \ldots \\ \mathbf{K}_{21}^{(1)} & \mathbf{K}_{22}^{(1)} & \mathbf{K}_{23}^{(1)} & \ldots \\ \mathbf{K}_{31}^{(1)} & \mathbf{K}_{32}^{(1)} & \mathbf{K}_{33}^{(1)} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{63}$$

The total electrostatic solvation free energy is determined as the difference between the vacuum electrostatic energy and total electrostatic energy in solvent as

$$\mathrm{U}_{\mathrm{solv}} = \frac{1}{2}\left(\mathbf{M}^t\mathbf{K} - \boldsymbol{\mu}^{\Delta}\mathbf{T}_{\mathrm{p}}^{(1)} - \boldsymbol{\mu}\mathbf{K}^{(1)}\right)\mathbf{M} \tag{64}$$

where $\boldsymbol{\mu}^{\Delta}$ represents the change in the induced dipoles upon solvation

$$\boldsymbol{\mu}^{\Delta} = \boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{v}} \tag{65}$$

4.2. Permanent Multipole Energy Gradient. The permanent multipole electrostatic solvation energy gradient between sites *i* and *j* only depends on the gradient of the GK interaction tensor

$$\frac{\partial \mathbf{K}_{ij}}{\partial r_{i,\sigma}} = \frac{1}{2}\left[\left(\frac{\partial \mathbf{K}^{(i)}}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{K}^{(j)}}{\partial r_{i,\sigma}}\right)_{a_i,a_j} + \left(\frac{\partial \mathbf{K}^{(i)}}{\partial a_i} + \frac{\partial \mathbf{K}^{(j)}}{\partial a_i}\right)\frac{\partial a_i}{\partial r_{i,\sigma}} + \left(\frac{\partial \mathbf{K}^{(i)}}{\partial a_j} + \frac{\partial \mathbf{K}^{(j)}}{\partial a_j}\right)\frac{\partial a_j}{\partial r_{i,\sigma}}\right] \quad (66)$$

and subscript a_i and a_j denote keeping the effective radii fixed in this case. Generation of the GK interaction tensors that make up $\partial\mathbf{K}^{(i)}/\partial r_{i,\sigma}$, $\partial\mathbf{K}^{(i)}/\partial a_i$, $\partial\mathbf{K}^{(i)}/\partial a_j$, $\partial\mathbf{K}^{(j)}/\partial r_{i,\sigma}$, $\partial\mathbf{K}^{(j)}/\partial a_i$, and $\partial\mathbf{K}^{(j)}/\partial a_j$ are described in Appendix B of the Supporting Information. The derivatives of the effective radii with respect to an atomic displacement follow from the pairwise descreening implementation of the RPA and will not be discussed here.[8,9,38] We also point out that there is a torque on the permanent dipoles due to the permanent reaction field and also on the permanent quadrupoles due to the permanent reaction field gradient. All torques, including contributions from the polarization energy gradient discussed below, are converted to forces on adjacent atoms that define the local coordinate frame of the multipole.

4.3. Polarization Energy Gradient. The polarization energy gradient when using either the "direct" or "mutual" polarization models within the GK continuum will now be derived. The definition of the starting point for the iterative convergence of the self-consistent reaction field (SCRF) is the total "direct" field $\mathbf{E}_{\text{direct}}$ at each polarizable site. This field is the sum of the permanent atomic multipoles (PAM) intramolecular field

$$\mathbf{E}_{\text{d}} = \mathbf{T}_{\text{d}}^{(1)}\mathbf{M} \quad (67)$$

where $\mathbf{T}_{\text{d}}^{(1)}$ is analogous to the tensor matrix defined in deriving the AMOEBA vacuum energy in eq 56 and the PAM GK reaction field

$$\mathbf{E}_{\text{RF}} = \mathbf{K}^{(1)}\mathbf{M} \quad (68)$$

The product of the direct field $\mathbf{E}_{\text{direct}}$ with a vector of atomic polarizabilities determines the initial induced dipoles $\boldsymbol{\mu}_{\text{direct}}$

$$\begin{aligned}\boldsymbol{\mu}_{\text{direct}} &= \boldsymbol{\alpha}\mathbf{E}_{\text{direct}}\\ &= \boldsymbol{\alpha}(\mathbf{T}_{\text{d}}^{(1)} + \mathbf{K}^{(1)})\mathbf{M}\end{aligned} \quad (69)$$

At this point the induced dipoles do not act upon each other nor do they elicit a reaction field. This is defined as the direct model of polarization.

In contrast to the direct polarization model, the total SCRF $\mathbf{E}$ has two additional contributions due to the induced dipoles and their reaction field

$$\mathbf{E} = (\mathbf{T}_{\text{d}}^{(1)} + \mathbf{K}^{(1)})\mathbf{M} + (\mathbf{T}^{(11)} + \mathbf{K}^{(11)})\boldsymbol{\mu} \quad (70)$$

for a sum of 4 contributions. The induced dipoles

$$\boldsymbol{\mu} = \boldsymbol{\alpha}[(\mathbf{T}_{\text{d}}^{(1)} + \mathbf{K}^{(1)})\mathbf{M} + (\mathbf{T}^{(11)} + \mathbf{K}^{(11)})\boldsymbol{\mu}] \quad (71)$$

can be solved for in an iterative fashion using successive over-relaxation (SOR) to accelerate convergence.[52] Alternatively, the induced dipoles can be solved for directly as a mechanism for deriving the polarization energy gradient with respect to an atomic displacement. Moving all terms containing the induced dipoles to the LHS allows their isolation

$$(\boldsymbol{\alpha}^{-1} - \mathbf{T}^{(11)} - \mathbf{K}^{(11)})\boldsymbol{\mu} = (\mathbf{T}_{\text{d}}^{(1)} + \mathbf{K}^{(1)})\mathbf{M} \quad (72)$$

For convenience, a matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = \boldsymbol{\alpha}^{-1} - \mathbf{T}^{(11)} - \mathbf{K}^{(11)} \quad (73)$$

which is substituted into eq 72 above to show the induced dipoles are a linear function of the PAM $\mathbf{M}$, directly via the intramolecular interaction tensor $\mathbf{T}_{\text{d}}^{(1)}$ that implicitly contains the AMOEBA group based polarization scheme, and also through their reaction field

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{C}^{-1}(\mathbf{T}_{\text{d}}^{(1)} + \mathbf{K}^{(1)})\mathbf{M}\\ &= \mathbf{C}^{-1}(\mathbf{E}_{\text{d}} + \mathbf{E}_{\text{RF}})\end{aligned} \quad (74)$$

The polarization energy can now be described in terms of the permanent reaction field and solute field $\mathbf{E}_{\text{p}}$

$$\text{U}_{\mu} = -\frac{1}{2}(\mathbf{E}_{\text{p}} + \mathbf{E}_{\text{RF}})^{t}\boldsymbol{\mu} \quad (75)$$

To find the polarization energy gradient, we wish to avoid terms that rely on the change in induced dipoles with respect to an atomic displacement. Therefore, the induced dipoles in eq 75 are substituted for using eq 74 to yield

$$\text{U}_{\mu} = -\frac{1}{2}(\mathbf{E}_{\text{p}} + \mathbf{E}_{\text{RF}})^{t}\mathbf{C}^{-1}(\mathbf{E}_{\text{d}} + \mathbf{E}_{\text{RF}}) \quad (76)$$

By the chain rule, the polarization energy gradient is

$$\frac{\partial \text{U}_{\mu}}{\partial r_{\text{i},\sigma}} = -\frac{1}{2}\left[\left(\frac{\partial \mathbf{E}_{\text{p}}}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}}\right)^{t}\boldsymbol{C}^{-1}(\mathbf{E}_{\text{d}} + \mathbf{E}_{\text{RF}}) + (\mathbf{E}_{\text{p}} + \mathbf{E}_{\text{RF}})^{t}\frac{\partial \mathbf{C}^{-1}}{\partial r_{i,\sigma}}(\mathbf{E}_{\text{d}} + \mathbf{E}_{\text{RF}}) + (\mathbf{E}_{\text{p}} + \mathbf{E}_{\text{RF}})^{t}\mathbf{C}^{-1}\left(\frac{\partial \mathbf{E}_{\text{d}}}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}}\right)\right] \quad (77)$$

For convenience a mathematical quantity $\boldsymbol{\nu}$ is defined, which is similar to $\boldsymbol{\mu}$, as

$$\boldsymbol{\nu} = (\mathbf{E}_{\text{p}} + \mathbf{E}_{\text{RF}})\mathbf{C}^{-1} \quad (78)$$

We can now greatly simplify eq 77 above using eqs 74 and 78 along with the identity $\partial\mathbf{C}^{-1}/\partial r_{i,\sigma} = -\mathbf{C}^{-1}\ \partial\mathbf{C}/\partial r_{i,\sigma}\ \mathbf{C}^{-1}$ to give

$$\frac{\partial \text{U}_{\mu}}{\partial r_{\text{i},\sigma}} = -\frac{1}{2}\left[\left(\frac{\partial \mathbf{E}_{\text{p}}}{\partial r_{i,\sigma}}\right)^{t}\boldsymbol{\mu} + \boldsymbol{\nu}^{t}\frac{\partial \mathbf{E}_{\text{d}}}{\partial r_{i,\sigma}} + \left(\frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}}\right)^{t}\boldsymbol{\mu} + \boldsymbol{\nu}^{t}\frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}} - \boldsymbol{\nu}^{t}\frac{\partial \mathbf{C}}{\partial r_{i,\sigma}}\boldsymbol{\mu}\right] \quad (79)$$

Under the direct polarization model, $\mathbf{C}$ is an identity matrix whose derivative is zero, and therefore eq 79 simplifies to

$$\frac{\partial \text{U}_{\mu_{\text{direct}}}}{\partial r_{i,\sigma}} = -\frac{1}{2}\left[\left(\frac{\partial \mathbf{E}_{\text{p}}}{\partial r_{i,\sigma}}\right)^{t}\boldsymbol{\mu} + \boldsymbol{\nu}^{t}\frac{\partial \mathbf{E}_{\text{d}}}{\partial r_{i,\sigma}} + \left(\frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}}\right)^{t}\boldsymbol{\mu} + \boldsymbol{\nu}^{t}\frac{\partial \mathbf{E}_{\text{RF}}}{\partial r_{i,\sigma}}\right] \quad (80)$$

The first two terms on the RHS appear in the polarization energy gradient even in the absence of a continuum reaction field and are described elsewhere.[35] The third and fourth terms are specific to GK and can be combined. We require the derivative of the GK reaction field due to permanent multipoles with respect to movement of any atom

$$\frac{\partial \mathbf{E}_{\mathrm{RF}}}{\partial r_{i,\sigma}} = \frac{\partial \mathbf{K}^{(1)}}{\partial r_{i,\sigma}} \mathbf{M} \tag{81}$$

It is therefore sufficient to describe the gradient of any $\mathbf{K}_{ij}^{(1)}$ submatrix of $\mathbf{K}^{(1)}$ as

$$\frac{\partial \mathbf{K}_{ij}^{(1)}}{\partial r_{i,\sigma}} = \frac{1}{2}\left[\left(\frac{\partial \mathbf{K}_{ij}^{(1,i)}}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{K}_{ij}^{(1,j)}}{\partial r_{i,\sigma}}\right)_{a_i,a_j} + \left(\frac{\partial \mathbf{K}_{ij}^{(1,j)}}{\partial a_i} + \frac{\partial \mathbf{K}_{ij}^{(1,j)}}{\partial a_i}\right)\frac{\partial a_i}{\partial r_{i,\sigma}} + \left(\frac{\partial \mathbf{K}_{ij}^{(1,i)}}{\partial a_j} + \frac{\partial \mathbf{K}_{ij}^{(1,j)}}{\partial a_j}\right)\frac{\partial a_j}{\partial r_{i,\sigma}}\right] \tag{82}$$

The tensors that make up $\partial \mathbf{K}_{ij}^{(1,i)}/\partial r_{i,\sigma}$, $\partial \mathbf{K}_{ij}^{(1,i)}/\partial a_i$, $\partial \mathbf{K}_{ij}^{(1,i)}/\partial a_j$, $\partial \mathbf{K}_{ij}^{(1,j)}/\partial r_{i,\sigma}$, $\partial \mathbf{K}_{ij}^{(1,j)}/\partial a_i$, and $\partial \mathbf{K}_{ij}^{(1,j)}/\partial a_j$ are described in Appendix B. In this case there is a torque on the permanent dipoles and quadrupoles due to the reaction field and reaction field gradient of $(\mu + \nu)/2$, respectively.

The full mutual polarization gradient has an additional term compared to the direct polarization gradient, in addition to the implicit difference due to the induced dipoles being converged self-consistently. Specifically, the derivative of the matrix **C** leads to two terms

$$\boldsymbol{\nu}^t \frac{\partial \mathbf{C}}{\partial r_{i,\sigma}} \boldsymbol{\mu} = -\boldsymbol{\nu}^t\left(\frac{\partial \mathbf{T}^{(11)}}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{K}^{(11)}}{\partial r_{i,\sigma}}\right)\boldsymbol{\mu} \tag{83}$$

The first term on the RHS occurs in vacuum and is described elsewhere;[35] however, the final term is specific to GK. The gradient of one submatrix of the $\partial \mathbf{K}^{(11)}/\partial r_{i,\sigma}$ supermatrix is

$$\frac{\partial \mathbf{K}_{ij}^{(11)}}{\partial r_{i,\sigma}} = \left(\frac{\partial \mathbf{K}_{ij}^{(11)}}{\partial r_{i,\sigma}}\right)_{a_i,a_j} + \frac{\partial \mathbf{K}_{ij}^{(11)}}{\partial a_i}\frac{\partial a_i}{\partial r_{i,\sigma}} + \frac{\partial \mathbf{K}_{ij}^{(11)}}{\partial a_j}\frac{\partial a_j}{\partial r_{i,\sigma}} \tag{84}$$

The expression for the gradient of $\mathbf{K}_{ij}^{(11)}$ is simpler than those for the other GK interaction matrices because it is symmetric.

The veracity of the AMOEBA/GK energy gradients was checked using finite-differences of the energy, optimization of proteins to an rms convergence criterion of 10^{-4} kcal/mol/Å, and constant energy molecular dynamics. For example, at a mean temperature of 300 K the protein 1ETL showed a mean total energy of −361.20 kcal/mol with a standard deviation of just 0.25 kcal/mol over 1 ns.

5. Validation and Application

GK is an approximation to the Poisson solution that extends GB to arbitrary order polarizable atomic multipoles. Here we test GK by comparing to numerical PMPB solutions in the limit of using a van der Waals definition of the solute−solvent interface parametrized using the Bondi radii set.[51] Specifically, the electrostatic solvation free energy and total solvated dipole moment for a series of 55 proteins was compared using the PMPB and GK continuums. This test set based on PDB entries[45] was recently proposed by Tjong and Zhou for studying the accuracy of analytic solvation models and is characterized by structures with less than 10% sequence identity, resolution better than 1.0 Å, and less than 250 residues.[44] Amino acids with missing side chains were changed to alanine if the C_β carbon was present and to glycine if it was not. The TINKER[53] pdbxyz program added missing hydrogen atoms. Histidine residues were made neutral with the δ-nitrogren protonated. All structures were optimized in vacuum to an rms gradient of 5.0 kcal/mol/Å, with the goal being to remove bad contacts. The average heavy atom rms distance from the crystal structure was 0.07 Å after optimization.

5.1. Electrostatic Solvation Free Energy of Proteins. Previous studies have shown that given accurate effective radii, GB predicts the electrostatic solvation energy of proteins to a mean unsigned relative difference of approximately 1% relative to numerical Poisson calculations.[16] In this section we investigate whether it is reasonable to expect similar performance from GK by comparing the electrostatic solvation free energy for a series of folded proteins to values computed using the PMPB model.

The PMPB calculations used a grid spacing of 0.31 Å and at least 10 Å between the edge of the solute−solvent boundary and the grid boundary. A finer grid spacing of 0.23 Å was also tried, which lowered the PMPB energy by approximately 2% but did not change the quality of the agreement between the two models. The interior of the protein was assigned a permittivity of 1.0, while the solvent was set to 78.3. The induced dipoles were deemed to have converged at a tolerance of 0.01 rms Debye. Converging to a tighter tolerance of 10^{-6} rms Debye only changed the electrostatic solvation free energy by 0.1% relative to the looser criteria and was therefore deemed unnecessary. The constant in the generalizing function, c_f, was optimized by hand to eliminate systematic error, which was found to occur at a value of 2.455.

The results are shown in Table 4. The mean signed relative difference is 0.0% a result of tuning the cross-term parameter. The mean unsigned relative difference is 0.9%, which is comparable to the most accurate GB methods.[13,15,17,34,44,54] We anticipate using a different cross-term parameter when optimizing GK to reproduce PMPB calculations based on a molecular surface definition.

5.2. Dipole Moment of Solvated Proteins. The change in dipole moment as a function of environment for a polarizable solute is a relevant observable in terms of validating GK because it indicates whether or not the reaction field strength is consistent. The PMPB calculations are exactly equivalent to those described in the previous section. Furthermore, the same constant was used in the GK cross-term. In Table 5 it is observed that the total dipole moment of proteins within the GK continuum achieve a mean signed relative difference of −2.7% and a mean unsigned percent difference of 2.7%. This indicates a small, but systematic underestimation of the reaction field. In all cases, for both PMPB and GK models, the reaction field factor was greater than one, except for 1P9G. In this case, the vacuum dipole

Table 4. Electrostatic Solvation Free Energy (kcal/mol) for 55 Proteins within the PMPB and GK Continuum Models[a]

			energy		% difference	
	N_{atoms}	Q	PMPB	GK	signed	unsigned
1A6M	2435	2	−2831	−2765	2.3	2.3
1AHO	936	0	−1161	−1158	0.3	0.3
1BYI	3383	−4	−3861	−3873	−0.3	0.3
1C75	985	−6	−1733	−1742	−0.5	0.5
1C7K	1927	−5	−2523	−2481	1.7	1.7
1CEX	2867	1	−3161	−3212	−1.6	1.6
1EB6	2566	−15	−5044	−5042	0.1	0.1
1EJG	642	0	−580	−614	−6.0	6.0
1ETL	140	−1	−246	−247	−0.5	0.5
1EXR	2240	−25	−8656	−8620	0.4	0.4
1F94	967	2	−1240	−1226	1.1	1.1
1F9Y	2535	−5	−2964	−2968	−0.2	0.2
1G4I	1842	−1	−2356	−2345	0.5	0.5
1G66	2794	−2	−2826	−2824	0.1	0.1
1GQV	2135	7	−2708	−2723	−0.6	0.6
1HJE	175	1	−264	−269	−2.0	2.0
1IQZ	1171	−17	−4663	−4729	−1.4	1.4
1IUA	1207	−1	−1400	−1419	−1.4	1.4
1J0P	1597	8	−2975	−2934	1.4	1.4
1K4I	3253	−6	−4085	−4099	−0.3	0.3
1KTH	885	0	−1469	−1448	1.4	1.4
1L9L	1226	11	−3182	−3150	1.0	1.0
1M1Q	1236	−4	−2084	−2077	0.3	0.3
1NLS	3564	−7	−4743	−4756	−0.3	0.3
1NWZ	1912	−6	−2768	−2760	0.3	0.3
1OD3	1893	−3	−2105	−2104	0.0	0.0
1OK0	1076	−5	−1578	−1571	0.5	0.5
1P9G	519	4	−814	−817	−0.4	0.4
1PQ7	3065	4	−2946	−2942	0.1	0.1
1R6J	1230	0	−1486	−1477	0.6	0.6
1SSX	2755	8	−3000	−2980	0.7	0.7
1TG0	1029	−12	−3017	−3014	0.1	0.1
1TQG	1660	−7	−2920	−2900	0.7	0.7
1TT8	2676	1	−2762	−2758	0.1	0.1
1U2H	1495	2	−2038	−2002	1.8	1.8
1UCS	997	0	−1027	−1042	−1.4	1.4
1UFY	1911	0	−2130	−2145	−0.7	0.7
1UNQ	1947	−1	−3217	−3155	1.9	1.9
1VB0	913	3	−1246	−1232	1.1	1.1
1VBW	1056	8	−1931	−1927	0.2	0.2
1W0N	1756	−5	−2380	−2356	1.0	1.0
1WY3	560	1	−750	−747	0.3	0.3
1X6Z	1720	−1	−2170	−2198	−1.3	1.3
1X8Q	2815	−1	−3739	−3714	0.7	0.7
1XMK	1268	1	−1723	−1724	0.0	0.0
1YK4	774	−8	−1893	−1920	−1.4	1.4
1ZZK	1243	1	−1730	−1699	1.8	1.8
2A6Z	3430	−3	−4203	−4186	0.4	0.4
2BF9	560	−2	−933	−940	−0.8	0.8
2CHH	1624	−3	−2128	−2131	−0.1	0.1
2CWS	3400	−3	−3651	−3616	1.0	1.0
2ERL	567	−6	−1178	−1179	0.0	0.0
2FDN	731	−8	−1746	−1796	−2.9	2.9
2FWH	1830	−6	−2495	−2502	−0.3	0.3
3LZT	1960	8	−2754	−2723	1.1	1.1
mean	1692	−1.9	−2458	−2454	0.0	0.9

[a] The number of atoms and total charge of each protein is listed along with the signed and unsigned relative difference of the GK model to PMPB.

Table 5. Total Dipole Moment (Debye) for 55 Proteins in Vacuum and within the PMPB and GK Continuum Models[a]

	dipole moment			% difference		reaction field factor	
	vacuum	PMPB	GK	signed	unsigned	PMPB	GK
1A6M	191.5	252.1	242.6	−3.7	3.7	1.32	1.27
1AHO	119.3	143.6	142.6	−0.7	0.7	1.20	1.20
1BYI	295.8	357.4	343.1	−4.0	4.0	1.21	1.16
1C75	125.0	167.2	165.7	−0.9	0.9	1.34	1.33
1C7K	229.3	310.3	302.7	−2.4	2.4	1.35	1.32
1CEX	451.0	599.7	574.3	−4.2	4.2	1.33	1.27
1EB6	217.9	281.0	274.6	−2.3	2.3	1.29	1.26
1EJG	37.4	49.0	48.9	−0.3	0.3	1.31	1.31
1ETL	29.3	42.9	41.2	−3.8	3.8	1.46	1.41
1EXR	352.5	395.6	384.2	−2.9	2.9	1.12	1.09
1F94	90.7	116.7	113.0	−3.2	3.2	1.29	1.25
1F9Y	138.4	166.0	161.9	−2.5	2.5	1.20	1.17
1G4I	87.9	102.1	97.9	−4.1	4.1	1.16	1.11
1G66	226.5	279.9	273.5	−2.3	2.3	1.24	1.21
1GQV	314.6	394.5	385.2	−2.4	2.4	1.25	1.22
1HJE	48.3	61.2	60.4	−1.4	1.4	1.27	1.25
1IQZ	86.1	110.7	107.2	−3.1	3.1	1.29	1.25
1IUA	107.5	146.1	141.5	−3.2	3.2	1.36	1.32
1J0P	105.2	148.7	142.3	−4.3	4.3	1.41	1.35
1K4I	130.1	163.0	159.6	−2.1	2.1	1.25	1.23
1KTH	117.1	152.1	148.9	−2.1	2.1	1.30	1.27
1L9L	422.8	525.9	517.0	−1.7	1.7	1.24	1.22
1M1Q	261.7	318.1	311.2	−2.2	2.2	1.22	1.19
1NLS	244.9	331.8	313.0	−5.7	5.7	1.35	1.28
1NWZ	83.2	130.2	126.9	−2.5	2.5	1.56	1.53
1OD3	115.2	165.9	160.9	−3.0	3.0	1.44	1.40
1OK0	149.4	193.7	189.1	−2.4	2.4	1.30	1.27
1P9G	17.7	14.6	13.0	−10.7	10.7	0.82	0.74
1PQ7	46.4	49.6	49.1	−1.1	1.1	1.07	1.06
1R6J	86.8	108.8	106.7	−1.9	1.9	1.25	1.23
1SSX	66.0	93.8	89.9	−4.2	4.2	1.42	1.36
1TG0	236.9	316.8	311.1	−1.8	1.8	1.34	1.31
1TQG	355.4	489.5	477.3	−2.5	2.5	1.38	1.34
1TT8	339.6	450.3	434.3	−3.6	3.6	1.33	1.28
1U2H	157.1	206.0	200.6	−2.6	2.6	1.31	1.28
1UCS	111.1	133.0	132.9	0.0	0.0	1.20	1.20
1UFY	94.0	105.9	102.3	−3.4	3.4	1.13	1.09
1UNQ	601.1	735.2	718.8	−2.2	2.2	1.22	1.20
1VB0	132.2	158.2	155.0	−2.0	2.0	1.20	1.17
1VBW	94.4	117.0	114.0	−2.6	2.6	1.24	1.21
1W0N	114.9	155.4	150.0	−3.5	3.5	1.35	1.31
1WY3	63.7	96.4	93.6	−3.0	3.0	1.51	1.47
1X6Z	294.2	366.7	355.9	−2.9	2.9	1.25	1.21
1X8Q	183.8	244.2	237.6	−2.7	2.7	1.33	1.29
1XMK	272.8	356.1	347.0	−2.6	2.6	1.31	1.27
1YK4	66.1	83.7	83.6	−0.2	0.2	1.27	1.26
1ZZK	195.2	246.5	241.6	−2.0	2.0	1.26	1.24
2A6Z	84.1	105.0	101.4	−3.4	3.4	1.25	1.21
2BF9	255.7	290.6	288.4	−0.7	0.7	1.14	1.13
2CHH	267.4	335.7	329.2	−1.9	1.9	1.26	1.23
2CWS	168.6	220.5	211.0	−4.3	4.3	1.31	1.25
2ERL	81.2	108.1	105.6	−2.3	2.3	1.33	1.30
2FDN	78.3	93.2	93.4	0.3	0.3	1.19	1.19
2FWH	104.9	146.3	142.9	−2.3	2.3	1.39	1.36
3LZT	178.5	214.6	209.8	−2.3	2.3	1.20	1.18
mean	173.2	220.9	215.0	−2.7	2.7	1.28	1.24

[a] The signed and unsigned relative difference of the GK model to PMPB is given along with their reaction field factors, μ/μ_v.

moment decreased from 18 to 15 and 13 Debye in the PMPB and GK models, respectively. Overall, the mean reaction field factor for the 55 proteins was 1.28 in the PMPB model and 1.24 in GK.

6. Conclusions

Over the course of the past several years GB has been shown to be capable of capturing the electrostatic response of the solvent environment to solutes. It has been successfully applied to molecular dynamics simulations, scoring protein conformations, and the prediction of binding affinities.[24] However, GB models are generally limited to use with fixed atomic partial charge electrostatic representations. Applications of recent interest, including high-resolution homology modeling, design of protein−protein interactions, and design of proteins with enzymatic activity may require improved accuracy in force field electrostatics.[28,29,55] We suggest that the AMOEBA force field coupled with the GK continuum model is a promising improvement.

There are two main differences between GB and GK. First, the GK self-energy of a permanent multipole site depends on Kirkwood's solution for the electrostatic solvation energy of a spherical particle with arbitrary charge distribution, which is reduced to Born's formula in the case of a monopole. Second, the GK cross-term is formulated by averaging a simple auxiliary potential for each multipole site, which reduces to the GB cross-term for monopole interactions.

We have implemented GK for the AMOEBA force field, including energy gradients, within the TINKER package.[53] The model was tested against numerical PMPB calculations of the electrostatic solvation free energy for a series of 55 diverse proteins and showed a mean unsigned relative difference of 0.9%. The fidelity of the reaction field of GK relative to PMPB can be inferred from the total solvated dipole moment of each protein, which showed GK to have a mean unsigned relative difference of 2.7%.

The next step in the implementation of GK for AMOEBA solutes will be parametrization of a complete implicit solvent model by addition of an apolar term.[26] The overall model will be parametrized against small molecule solvation free energies, which has been a successful approach in the past.[9,10,25,56,57] Alternatively, the electrostatic, dispersion, and cavitation components of solvation can be matched to explicit solvent free energy perturbation results.[58−61]

GK may be useful for developing new continuum models based on electron densities derived from electronic structure calculations. For example, Cramer and Truhlar have successfully employed GB in their SMX series of solvation models.[8,9,62,63] GK would also offer an analytic alternative to the numerical distributed multipole solvation model of Rinaldi et al. [64,65]

Further improvements in both the PMPB and GK continuum electrostatics models may depend on reconciling deficiencies that emerge in treating local, specific molecular interactions. For example, both the Clausius-Mossotti[40] and Onsager[66] theories for the permittivity of a liquid break down for substances that "associate" such as water. Here association is defined as short range ordering that leads to correlations in the orientations and positions of neighboring groups, such as hydrogen-bonding pairs. Formalisms introduced by Kirkwood[67] and Fröhlich[68] include a correction factor to explicitly account for this deviation from continuum behavior. More recently, Rick and Berne showed that no parametrization of the dielectric boundary for a water molecule in liquid water could simultaneously fit the electrostatic free energy and reaction potential to within 20%, mainly due to nonlinear electrostriction.[69−73] This effect, inherent in both numerical and analytic continuum electrostatic models, may be a limiting factor in the accuracy of current implicit solvation models.

Acknowledgment. We would like to thank Dr. Pengyu Ren, Dr. Nathan Baker, Dr. Chuanjie Wu, Dr. Sergio Urahata, and Justin Xiang for helpful discussions. M.J.S. was partially supported by a Computational Biology Training Grant from the NIH and a Grace Norman Scholarship. J.W.P. was supported by NSF grants CHE-0535675 and MCB-0344670 and NIH grant GM069553.

Supporting Information Available: Intermediate terms in the derivation of the solvent field approximation (Appendix A and Tables A-1−A-4) and a procedure for factoring the auxiliary tensors needed to compute the gradient of the GK potential (Appendix B and Tables B-1−B-10). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351−361.

(2) Schnieders, M. J.; Baker, N. A.; Ren, P. Y.; Ponder, J. W. *J. Chem. Phys.* **2007**, *126*.

(3) Flory, P. J. *Statistical Mechanics Of Chain Molecules*; Butterworth-Heinemann Ltd.: 1969.

(4) Born, M. *Z. Phys.* **1920**, *1*, 45−48.

(5) Kong, Y.; Ponder, J. W. *J. Chem. Phys.* **1997**, *107*, 481−492.

(6) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(7) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045−1066.

(8) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(9) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(10) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(11) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

(12) Feig, M.; Im, W.; Brooks, C. L. *J. Chem. Phys.* **2004**, *120*, 903−911.

(13) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265−284.

(14) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712−3720.

(15) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.

(16) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297−1304.

(17) Sigalov, G.; Fenley, A.; Onufriev, A. *J. Chem. Phys.* **2006**, *124*.

(18) Sigalov, G.; Scheffel, P.; Onufriev, A. *J. Chem. Phys.* **2005**, *122*.

(19) Baker, N. A. *Methods Enzymol.* **2004**, *383*, 94−118.

(20) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137−143.

(21) Holst, M.; Saied, F. *J. Comput. Chem.* **1993**, *14*, 105−113.

(22) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435−445.

(23) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606−10614.

(24) Feig, M.; Brooks, C. L., III *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−24.

(25) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517−529.

(26) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1−20.

(27) Maple, J. R.; Cao, Y. X.; Damm, W. G.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694−715.

(28) Jaramillo, A.; Wodak, S. J. *Biophys. J.* **2005**, *88*, 156−71.

(29) Marshall, S. A.; Vizcarra, C. L.; Mayo, S. L. *Protein Sci.* **2005**, *14*, 1293−1304.

(30) Piquemal, J. P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*.

(31) Friesner, R. A.; Baldwin, R. L. In *Advances in Protein Chemistry*; Academic Press: 2005; Vol. 72, pp 79−104.

(32) Ponder, J. W.; Case, D. A. In *Advances in Protein Chemistry*; Academic Press: 2003; Vol. 66, pp 27−85.

(33) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.

(34) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.

(35) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497−1506.

(36) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(37) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2004**, *108*, 13427−13437.

(38) Grycuk, T. *J. Chem. Phys.* **2003**, *119*, 4817−4826.

(39) Jackson, J. D. *Classical Electrodynamics*, 3rd ed.; John Wiley & Sons, Inc.: New York, 1998.

(40) Böttcher, C. J. F. *Dielectrics in Static Fields*, 2nd ed.; Elsevier Pub. Co.: Amsterdam, 1993; Vol. 1.

(41) Böttcher, C. J. F. *Dielectrics in Static Fields*, 1st ed.; Elsevier Pub. Co.: Amsterdam, 1952; Vol. 1.

(42) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon Press: Oxford, 1996; Vol. 32.

(43) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.

(44) Tjong, H.; Zhou, H. X. *J. Phys. Chem. B* **2007**, *111*, 3055−3061.

(45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(46) Teeter, M. M. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 6014−6018.

(47) Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O. *Protein Sci.* **1994**, *3*, 1779−1787.

(48) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82−87.

(49) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721−4729.

(50) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol.* **1997**, *4*, 180−184.

(51) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441−451.

(52) Young, D. M. *Iterative Solutions of Large Linear Systems*; Academic Press: New York, 1971.

(53) Ponder, J. W. *TINKER: Software Tools for Molecular Design*, 4.2; Saint Louis, MO, 2004.

(54) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348−1356.

(55) Vizcarra, C. L.; Mayo, S. L. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622−6.

(56) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264−16270.

(57) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210−3221.

(58) Banavali, N. K.; Im, W.; Roux, B. *J. Chem. Phys.* **2002**, *117*, 7381−7388.

(59) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.

(60) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239−5248.

(61) Nina, M.; Im, W.; Roux, B. *Biophys. Chem.* **1999**, *78*, 89−96.

(62) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257−3271.

(63) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133−1152.

(64) Rinaldi, D.; Bouchy, A.; Rivail, J. L. *Theor. Chem. Acc.* **2006**, *116*, 664−669.

(65) Rinaldi, D.; Bouchy, A.; Rivail, J. L.; Dillet, V. *J. Chem. Phys.* **2004**, *120*, 2343−2350.

(66) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486−1493.

(67) Kirkwood, J. G. *J. Chem. Phys.* **1939**, *7*, 911−919.

(68) Fröhlich, H. *Theory of Dielectrics*, 2nd ed.; Oxford University Press: London, 1958.

(69) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1994**, *116*, 3949−3954.

(70) Challacombe, M.; Schwegler, E.; Almlof, J. *Chem. Phys. Lett.* **1995**, *241*, 67−72.

(71) McMurchie, L. E.; Davidson, E. R. *J. Comput. Phys.* **1978**, *26*, 218−231.

(72) Applequist, J. *J. Phys. A: Math. Gen.* **1989**, *22*, 4303−4330.

(73) Applequist, J. *Theor. Chem. Acc.* **2002**, *107*, 103−115.

CT7001336

JCTC Journal of Chemical Theory and Computation

Induced-Polarization Energy Map: A Helpful Tool for Predicting Geometric Features of Anion-π Complexes

Daniel Escudero,[†] Antonio Frontera,[*,†] David Quiñonero,[†] Antoni Costa,[†]
Pablo Ballester,[‡] and Pere M. Deyà[*,†]

Department of Chemistry, Universitat de les Illes Balears, Crta. de Valldemossa km 7.5, E-07122 Palma de Mallorca, Spain, and ICREA and Institute of Chemical Research of Catalonia (ICIQ), Avinguda Països Catalans 16, 43007 Tarragona, Spain

Received May 21, 2007

Abstract: In this manuscript we propose the use of a new tool that we have found useful to predict the geometries of ion-π complexes. This tool is entitled the Induced-Polarization Energy map (IPE map). The novelty of this representation is that in the map only the contribution of the ion-induced polarization term to the total interaction energy for a given noncovalent interaction is contoured in a 2D region. The IPE map has been found useful to predict and explain geometries of several complexes of a tetrahedral 2 anion (BF_4^-) with perfluoropyrazine, perfluoropyridazine, perfluoropyrimidine, the three isomers of perfluorotriazine, and the three isomers of perfluorotetrazine.

Introduction

In modern chemistry, noncovalent interactions are decisive. This is especially true in the field of supramolecular chemistry and molecular recognition.[1] In particular, interactions involving aromatic rings[2] are key processes in both chemical and biological recognition since aromatic rings are omnipresent in biological systems. A classical example is the interaction of cations with aromatic systems, namely cation-π interactions,[3] which are supposed to be decisive in the ion selectivity in potassium channels.[4] Such interactions are also important for the binding of acetylcholine to the active site of the enzyme acetylcholine esterase.[5] Recently, the importance of cation-π interactions in neurotransmitter receptors has been demonstrated,[6] and they play an important role in transport of nitrogen through the membrane by the ammonia transport protein.[7] Anion-π interactions[8] are also important noncovalent forces that have attracted considerable attention in the last 3 years. They have been observed experimentally, supporting the theoretical predictions and the promising proposal for the use of anion receptors based on anion-π interactions in molecular recognition.[9] In addition, π-acidic oligonaphthalendiimide rods have been recently proposed as transmembrane anion-π slides.[10] A recent review of P. Gamez et al. deals with anion-binding involving π-acidic heteroaromatic rings.[11]

The cation-π and anion-π interactions are mainly dominated by electrostatic and ion-induced polarization terms.[12] The nature of the electrostatic term can be rationalized by means of the permanent quadrupole moment of the arene. The face-to-face interaction of the benzene−hexafluorobenzene complex is favorable due to the large and opposite permanent quadrupole moments of the two molecules.[13] The π−π interaction in the benzene dimer is governed by dispersion effects.[14] We have explained the dual binding mode of some molecules to form stable complexes with both cations and anions arguing polarization effects.[12c,15] Two examples are the triazine and trifluorobenzene rings, and the dual behavior is explained by means of the small quadrupole moment of these molecules. The interaction is thus dominated by polarization effects, and the electrostatic contribution to the interaction energy is negligible. As a consequence the binding energies of the complexes of these compounds with ions is small compared with benzene (cation-π complexes) or hexafluorobenzene (anion-π complexes), but recent reports have described the complexes formed between nitrate and the π face of triazine in solid phase.[16] Recently our group has published a theoretical MP2 study where the energetic and geometric characteristics of several π-complexes involv-

* Corresponding author: fax: +34 971 173426; e-mail: toni.frontera@uib.es (A.F.), pere.deya@uib.es (P.M.D.)

[†] Universitat de les Illes Balears.

[‡] ICIQ.

10.1021/ct700122y CCC: $37.00
Published on Web 10/11/2007

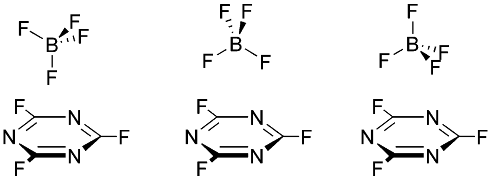


Figure 1. The three orientations of the complexes of trifluoro-*s*-triazine with BF_4^-.

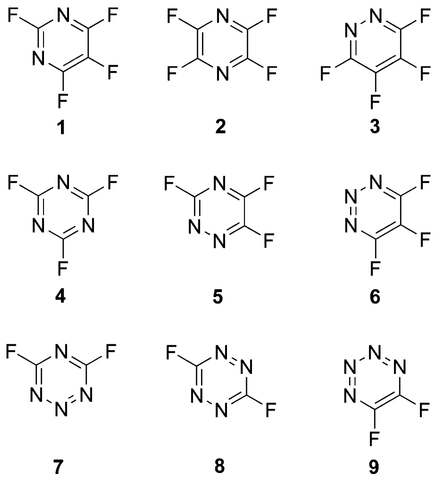


Figure 2. Perfluoropyrimidine (**1**), perfluoropyrazine (**2**), perfluoropyridazine (**3**), isomers of perflurotriazine (**4**−**6**), and isomers of perfluorotetrazine (**7**−**9**).

ing tetrahedral and octahedral anions have been analyzed and compared to X-ray structures retrieved from the Cambridge Structural Database.[17] In these complexes several orientations for the anion can be operative. For instance for the interaction of BF_4^- with perfluoro-*s*-triazine, the anion can interact with the aromatic ring using one, two, or three fluorine atoms (see Figure 1). In the present study, we propose the use of the Induced-Polarization Energy map (IPE map) as a useful tool to predict and explain the orientation of polyatomic anions in anion-π complexes, although it is expected that its use can be easily generalized to other systems such cation-π complexes.

In this manuscript we report a MP2 study, where we analyze complexes of the BF_4^- anion with perfluoropyrimidine (**1**), perfluoropyrazine (**2**), perfluoropyridazine (**3**), the three isomers of perfluorotriazine (**4**−**6**), and the three isomers of perfluorotetrazine (**7**−**9**) (see Figure 2). In addition we have computed the IPE maps for compounds **1**−**9** using the Molecular Interaction Potential with polarization (MIPp) partition scheme developed by Orozco and Luque.[18] The MIPp is a convenient tool for predicting binding properties. It has been successfully used for rationalizing molecular interactions such us hydrogen bonding and ion-π interactions and for predicting molecular reactivity.[19] The MIPp partition scheme is an improved generalization of the MEP where three terms contribute to the interaction energy: (i) an electrostatic term identical to the MEP,[20] (ii) a classical dispersion-repulsion term,[21] and (iii) a polarization term derived from perturbational theory.[22] The latter term has been used to construct the IPE maps, which has been compared to the MIPp maps. We have found that the geometric characteristics of the complexes of BF_4^- anion with compounds **1**−**9** are in agreement with the IPE maps, which are able to predict the orientation of the anion. This agreement confirms the importance of polarization effects in anion-π interactions not only energetically but also geometrically as well.

II. Theoretical Methods

The geometries of all compounds studied in this work were fully optimized using the MP2/6-31++G** level of theory within the Gaussian 03 package.[23] The minimum nature of the complexes was evaluated performing frequency analyses at the same level. In three complexes (**2**, **3**, and **7**) one small imaginary frequency has been found that corresponds to a rotational movement of the anion. The binding energies were calculated with correction for the basis set superposition error (BSSE) by using the Boys−Bernardi counterpoise technique.[24] The optimization of the complexes has been performed without imposing symmetry constraints unless otherwise noted. Calculation of the MIPp maps of **1**−**9** interacting with F^- was performed using the HF/6-31++G**//MP2/6-31++G** wave function by means of the MOPETE-98 program.[25] The ionic van de Waals parameters for F^- were taken from the literature.[26] Some basic concepts of MIPp follow (see refs 18 and 21 for a more comprehensive treatment). The MEP can be understood as the interaction energy between the molecular charge distribution and a classical point charge. The formalism used to derive MEP remains valid for any classical charge; therefore, it can be generalized using eq 1 where Q_B is the classical point charge at R_B. Q_B can adopt any value, but it has a chemical meaning only when $Q_B = 1$ (proton); ϕ stands for the set of basis functions used for the quantum mechanical molecule A; and $c_{\mu i}$ is the coefficient of atomic orbital μ in the molecular orbital i.

$$\mathrm{MEP} = \sum_A \frac{Z_A Q_B}{|R_B - R_A|} - \sum_i^{\mathrm{occ}} \sum_\mu \sum_\nu c_{\mu i} c_{\nu i} < \phi_\mu \left| \frac{Q_B}{|R_B - r|} \right| \phi_\nu > \quad (1)$$

The MEP formalism permits the rigorous computation of the electrostatic interaction between any classical particle and the molecule. Nevertheless, nuclear repulsion and dispersion effects are omitted. This can be resolved by the addition of a classical dispersion-repulsion term, which leads to the definition of MIP[23] (eq 2), where C and D are empirical van der Waals parameters.

$$\mathrm{MIP} = \mathrm{MEP} + \sum_{A'B'} \left(\frac{C_{A'B'}}{|R_{B'} - R_{A'}|^{12}} - \frac{D_{A'B'}}{|R_{B'} - R_{A'}|^{6}} \right) \quad (2)$$

The definition of MIPp is given by eq 3, where polarization effects are included at the second-order perturbation level;[24] ϵ stands for the energy of virtual (j) and occupied (i) molecular orbitals. It is worth noting that eq 3 includes three important contributions: first, the rigorous calculation of electrostatic interactions between quantum mechanical and classical particles; second, the introduction of an empirical

Table 1. Binding Energies without and with the Basis Set Superposition Error Correction (E and E_{BSSE}, kcal/mol, Respectively) and Equilibrium Distances (R_e, Å, from the Boron Atom to the Ring Centroid) at the MP2/6-31++G** Level of Theory Computed for the Complexes of BF_4^- with Compounds **1**−**9**[a]

complex	E	E_{BSSE}	NImag	R_e	Q_{zz} (B)	α_z (au)
1+BF_4^-	−17.9	−13.2	0	3.37	8.9	33.0
2+BF_4^-	−16.5	−12.8	1	3.55	8.4	32.2
3+BF_4^-	−18.4	−13.6	1	3.41	8.4	32.6
4+BF_4^-	−17.1	−13.0	0	3.25	8.2	30.3
5+BF_4^-	−18.5	−13.9	0	3.45	8.8	30.5
6+BF_4^-	−19.4	−14.3	0	3.41	9.4	30.9
7+BF_4^-	−17.5	−13.6	1	3.51	8.7	27.7
8+BF_4^-	−18.3	−14.2	0	3.47	8.5	29.4
9+BF_4^-	−18.9	−14.4	0	3.56	9.5	28.6

[a] Several properties of compounds **1**−**9** are also included.

dispersion-repulsion term; and third, the perturbative treatment of the polarization term.

$$\mathrm{MIPp} = \mathrm{MIP} + \sum_{j}^{\mathrm{vir}} \sum_{i}^{\mathrm{occ}} \frac{1}{\epsilon_i - \epsilon_j} \left\{ \sum_{\mu} \sum_{\nu} c_{\mu i} c_{\nu i} < \phi_\mu \left| \frac{Q_B}{|R_B - r|} \right| \phi_\nu > \right\}^2 \quad (3)$$

The "atoms-in-molecules" analysis[27] has been performed by means of the AIM2000 version 2.0 program[28] using the MP2/6-31++G** wave functions. The quadrupole moment of compounds **1**−**9** was computed using the CADPAC program[29] at the MP2/6-31G* level since previous studies[30] have demonstrated that quantitative results are obtained at this level of theory.

III. Results and Discussion

A. Energetic and Geometrical Results. In Table 1 we summarize the binding energies and equilibrium distances of the complexes of BF_4^- with compounds **1**−**9**. From the inspection of the results an interesting point arises. The interaction energy of perfluoroheteroaromatic compounds **1**−**9** with BF_4^- is large and negative, and its value is almost independent upon the number of the nitrogen atoms of the ring. This fact can be interpreted as a compensating effect, that is, the electron-withdrawing influence of a fluorine atom bonded to a carbon atom of the ring equals the atomic substitution of this carbon atom of the ring by one nitrogen atom, which is more electronegative than carbon. As a consequence, the π-acidity of the ring is essentially maintained. This fact can be corroborated by inspecting the values of quadrupole moments (Q_{zz}) and molecular polarizabilities (α_z), which have also been included in Table 1 for compounds **1**−**9**. As stated in the introduction, the physical nature of the anion-π interaction is mainly explained by the participation of two forces that contribute to the interaction: the electrostatic term and the ion-induced polarization. We have demonstrated that the former[12b] depends on the quadrupole moment of the aromatic compound (since the μ_z is negligible), and the latter depends on the molecular polarizability. The values of Q_{zz} and α_z present in Table 1 are comparable for all aromatic compounds and, thus, their interaction energies with BF_4^- are also comparable.

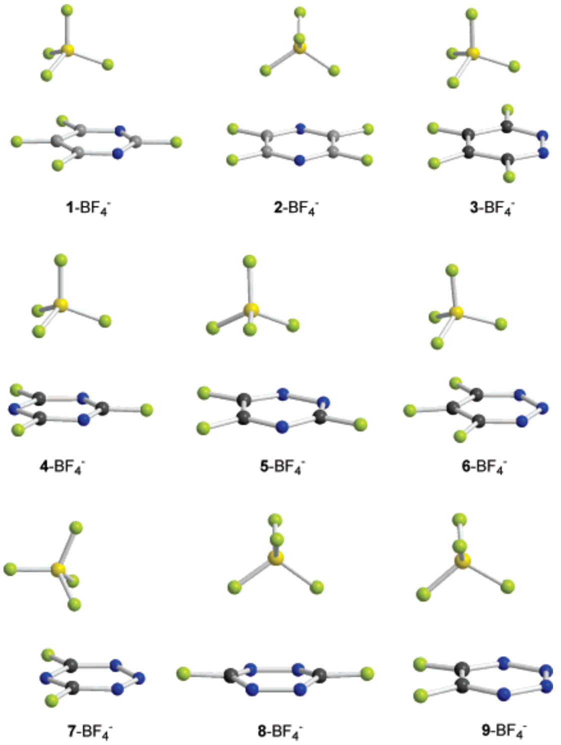


Figure 3. MP2/6-31++G** optimized structures of the π-complexes of BF_4^- with compounds **1**−**9**.

In Figure 3 we represent the MP2/6-31++G** optimized geometries of the anion-π complexes of compounds **1**−**9** interacting with BF_4^-. A variety of geometries is observed. In all cases the anion interacts with the π-cloud of the aromatic rings, but only in the more symmetric compounds the anion is located over the center of the ring (**2**, **4**, and **8**). In the other cases the anion is to some extent displaced from the ring centroid. In general the anion has the tendency to move over the region of the aromatic ring where a major number of carbon atoms is present (see Figure 3). The equilibrium distances observed for all complexes are similar, apart from complex **4**-BF_4^-. In this complex, the equilibrium distance (measured from boron atom to ring centroid) is shorter than the rest of complexes due to two factors. First, the boron atom is located over the center of the ring along the main symmetry axis. Second, in this complex three fluorine atoms of the anion are pointing at the ring. This geometry minimizes the distance from the boron atom to the ring centroid.

B. MIPp Analysis. We have performed the MIPp partition scheme calculation of compounds **1**−**9** interacting with F^- using the HF/6-31++G**//MP2/6-31++G** wave function. In the calculation F^- was considered as a classical nonpolarizable particle. The total MIPp energy is the sum of three terms, electrostatic, polarization, and van der Waals (dispersion-repulsion), as explained in the computational methods. We have computed bidimensional (2D) MIPp energy maps of compounds **1**−**9** interacting with F^- calculated at 2.6 Å above the molecular plane in order to explore their binding ability. In addition, using only the polarization contribution to the total MIPp energy, we have computed the 2D-IPE maps of compounds **1**−**9** interacting with F^- calculated at the same distance than the MIPp above the molecular plane. We have computed these maps to learn if this representation

Table 2. Interaction Energies (kcal/mol) Computed at the MIPp and IPE Minima Observed for Compounds **1**−**9** Interacting with F^{-a}

compound	MIPp	IPE	%IPE (%)
1	−22.2	−12.0	54
2	−22.1	−11.7	52
3	−21.5	−12.1	56
4	−23.0	−10.3	44
5	−22.8	−11.1	49
6	−21.5	−11.7	54
7	−23.2	−10.0	43
8	−22.2	−10.2	46
9	−22.4	−10.7	48

[a] The percentage of the polarization term to the total interaction energy is also shown.

can be useful to predict the geometric characteristics of anion-π complexes, especially when the anion is polyatomic. Bearing in mind that the anion-π interaction is mainly characterized by electrostatic and ion-induced polarization forces and that electrostatic forces are not directional, a mapping of the polarization contribution can give valuable information about the position of the anion over the aromatic ring. For all compounds the energetic value of the MIPp measured at the minimum ranges from −23.2 to −21.5 kcal/mol and the one for the IPE ranges from −12.0 to −10.0 kcal/mol (see Table 2), in agreement with the values of Q_{zz} and α_z of **1**−**9** (see Table 1). This result confirms that the polarization term is important, and its contribution accounts for approximately 50% of the total interaction energy. Previous studies[8a,9e,12b,c,15,17] demonstrate that the MIPp energies are in agreement with the interaction energies measured optimizing the complexes at the MP2 level of theory, which give reliability to the MIPp partition scheme. In the complexes studied in this work, a direct comparison of the MIPp energy values and the MP2/6-31++G** interaction energies is not possible, since the MIPp maps are computed using F$^-$ as the interacting particle instead of BF_4^-. The interaction energies present in Table 1 are lower than the MIPp values present in Table 2, because the formal charge assigned to the interacting particle is "−1" and the formal charge of the fluorine atoms of the BF_4^- anion is smaller. We have computed the 2D-IPE and 2D-MIPp maps 2.6 Å above the molecular plane because this is the location of the MIPp minimum when a F$^-$ approaches the center of the ring of compounds **1**−**9** following a perpendicular trajectory.

1. Heteroaromatic Rings with Two Nitrogen Atoms. The 2D-IPE(F$^-$) maps of perfluoropyrimidine (**1**), perfluoropyrazine (**2**), and perfluoropyridazine (**3**) are represented in Figures 4−6. In addition the 2D-MIPp(F$^-$) maps are represented in the figures for comparison purposes. The corresponding MP2/6-31++G** optimized complexes are also included in the figures in order to illustrate the agreement of the 2D-IPE/MIPp maps with the geometric features of the complexes. For compounds **1**−**3** the 2D-MIPp maps predict a minimum on the potential energy located approximately over the center of the ring. In contrast the 2D-IPE maps show a totally different distribution. For instance, the 2D-IPE map of compound **1** (see Figure 4) predicts a minimum which is located over the region of C5. The optimized geometry of the **1**-BF_4^- complex is in total agreement with the location of the IPE minimum. In addition, the geometry of the complex is not in disagreement with the 2D-MIPp map since one fluorine atom of the BF_4^- anion is located near the MIPp minimum.

The plots of 2D-IPE(F$^-$) and 2D-MIPp(F$^-$) maps computed for **2** are represented in Figure 5. The position of the anion is in accord with the MIPp map, since it is located

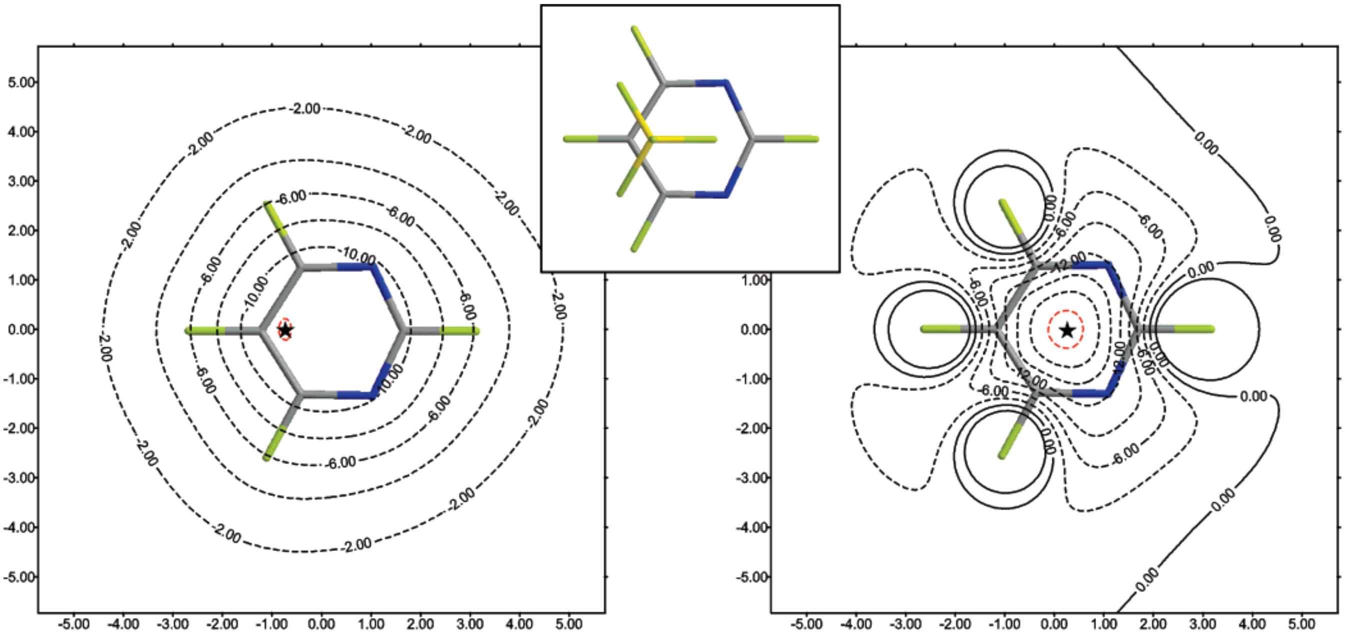


Figure 4. Left: 2D-IPE(F$^-$) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol. The red isocontour corresponds to −12 kcal/mol. The minimum is represented by a star. Right: The 2D-MIPp(F$^-$) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **1**-BF_4^- complex is represented (MP2/6-31++G**).

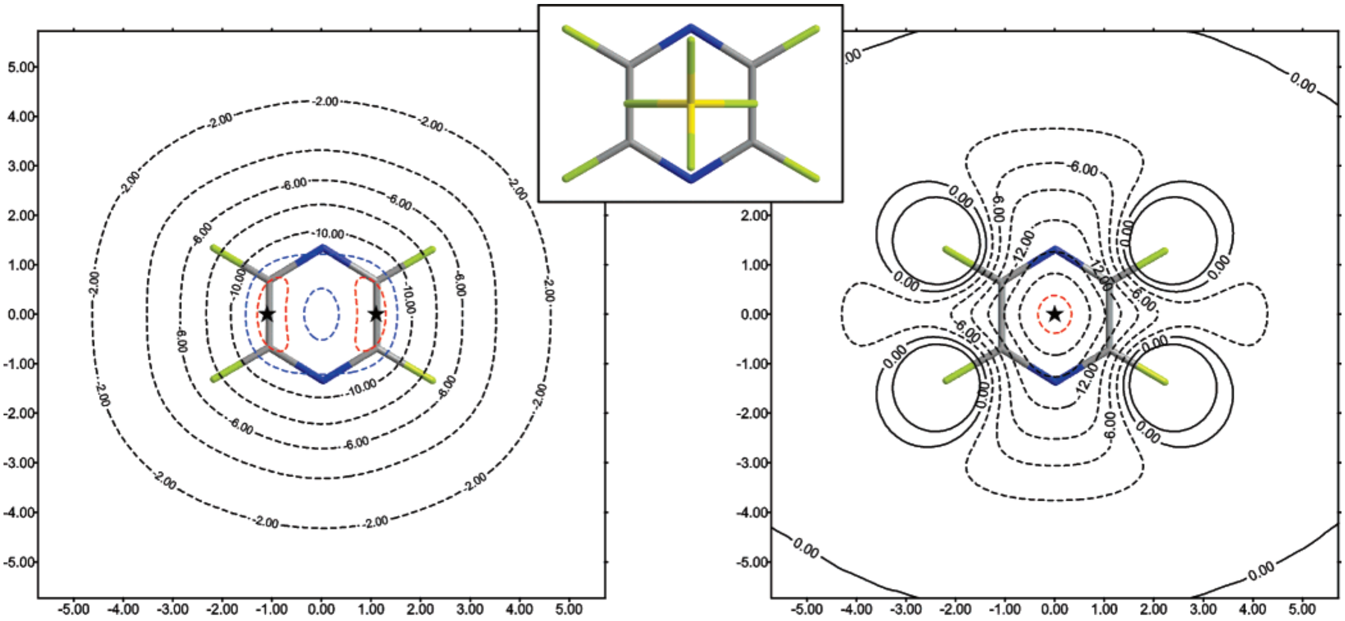


Figure 5. Left: 2D-IPE(F$^-$) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol. The blue isocontour corresponds to −11 kcal/mol, and the red isocontour corresponds to −11.5 kcal/mol. The minima are represented by stars. Right: The 2D-MIPp(F$^-$) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **2**-BF_4^- complex is represented (MP2/6-31++G**).

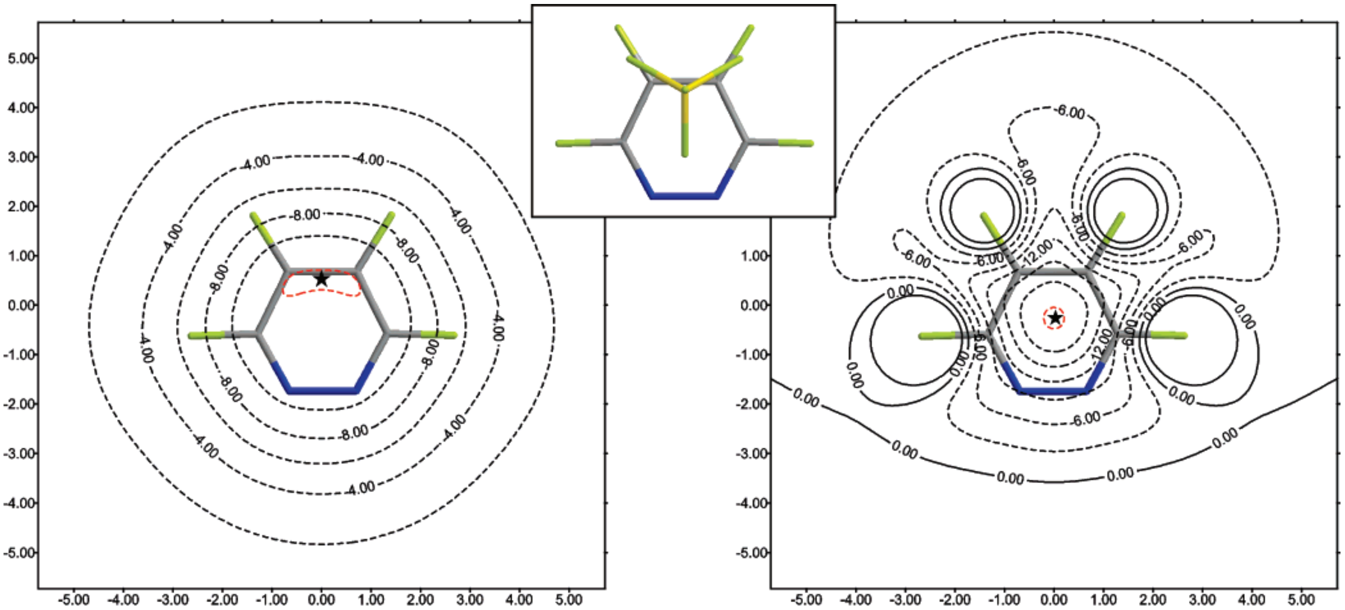


Figure 6. Left: 2D-IPE(F$^-$) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol. The red isocontour corresponds to −12 kcal/mol. The minimum is represented by a star. Right: The 2D-MIPp(F$^-$) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **3**-BF_4^- complex is represented (MP2/6-31++G**).

over the center of the ring. As expected, there is a good agreement between the location of two fluorine atoms of the anion and the position of the two minima found in the 2D-IPE(F$^-$) map. In the optimized complex **2**-BF_4^- two fluorine atoms of the BF_4^- are pointing to the middle of two C−C bonds, precisely where the IPE minima are found.

The plots of 2D-IPE(F$^-$) and 2D-MIPp(F$^-$) maps computed for **3** are represented in Figure 6. The IPE minimum is located over the middle of the C4−C5 bond, and the MIPp minimum is located approximately over the center of the ring, to a minor extent displaced toward the C4−C5 region. The MP2/6-31++G** optimized structure of the **3**-BF_4^- complex is in agreement with both maps. The anion is approximately located where the IPE map predicts with one fluorine atom located at the MIPp minimum. For compounds **2** and **3** (Figures 5 and 6) the 2D-IPE and 2D-MIPp maps

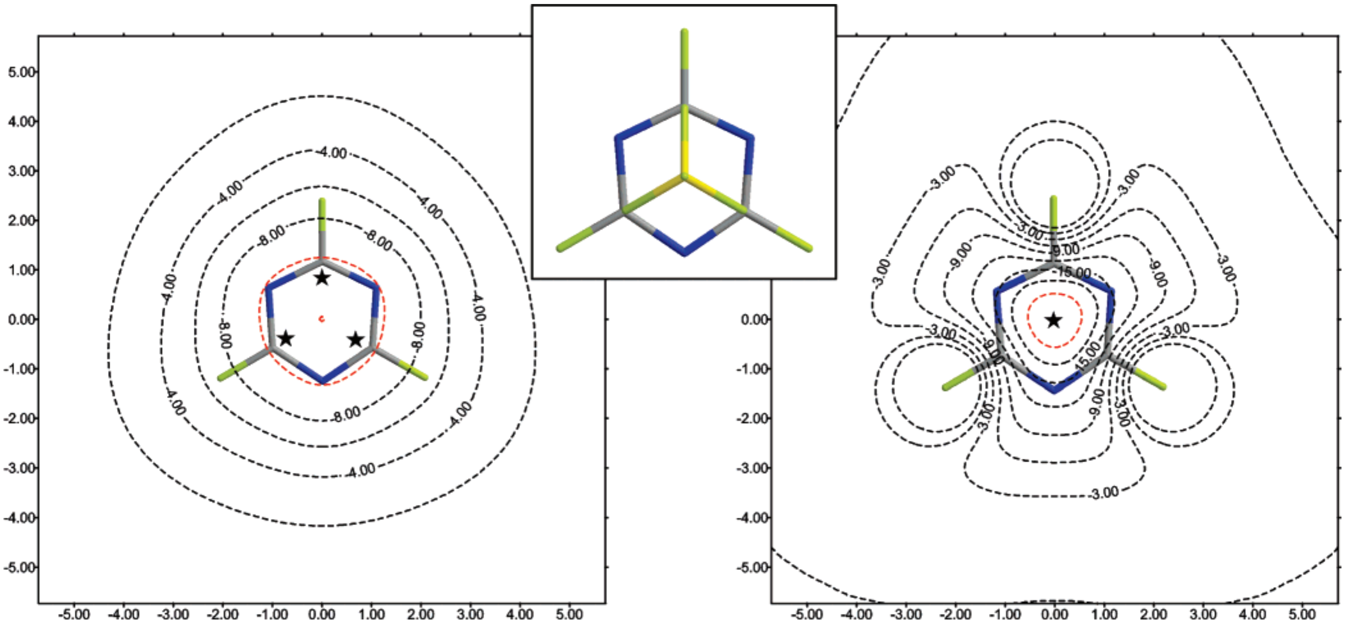


Figure 7. Left: IPE(F^-) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol. The red isocontour corresponds to −10 kcal/mol. The minima are represented by stars. Right: The 2D-MIPp(F^-) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **4**-BF_4^- complex is represented (MP2/6-31++G**).

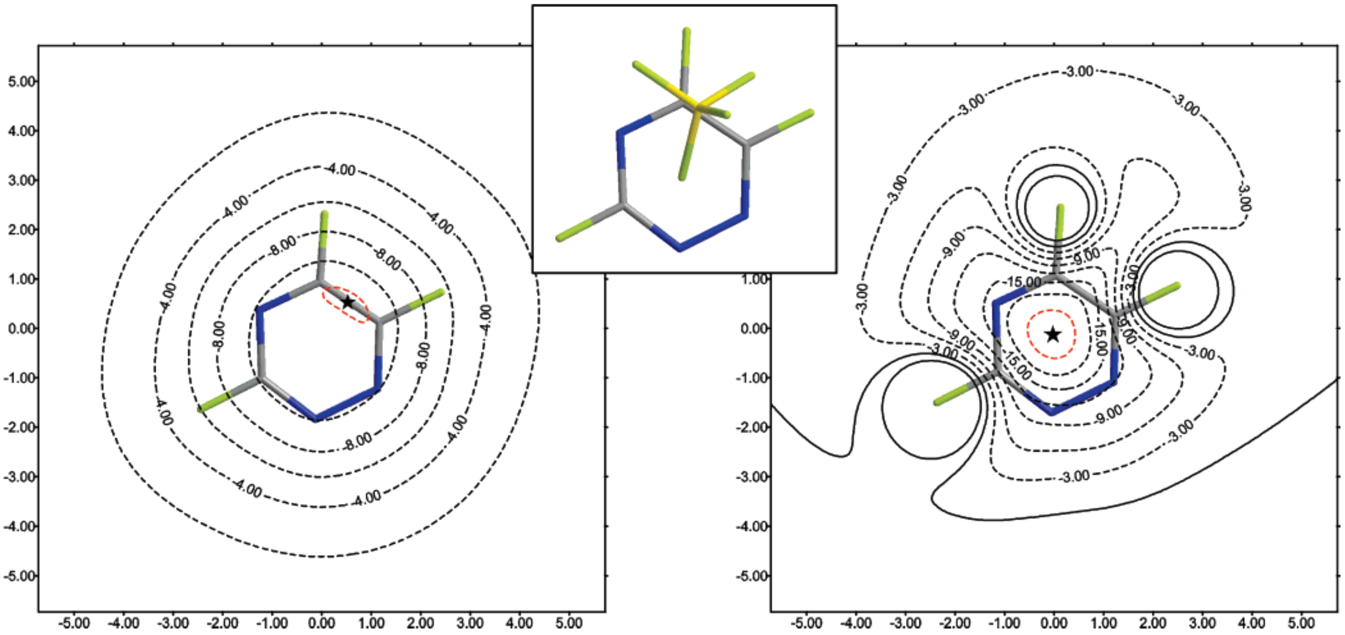


Figure 8. Left: IPE(F^-) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol from −2.0 to 10.0 kcal/mol. The red isocontour corresponds to −11 kcal/mol. The minimum is represented by a star. Right: The 2D-MIPp(F^-) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **5**-BF_4^- complex is represented (MP2/6-31++G**).

nicely complement each other and are useful tools to explain the observed geometric features of the optimized complexes.

2. Heteroaromatic Compounds with Three Nitrogen Atoms. The 2D-IPE(F^-) and 2D-MIPp maps of the three isomers of perfluorotriazine (**4**−**6**) are represented in Figures 7−9. In addition, the optimized complexes are also represented in the figures for comparison purposes. For trifluoro-*s*-triazine **4** the IPE and MIPp maps are depicted in Figure 7. The location of the minima at the 2D-IPE map and the geometry of the optimized complex are in good agreement. There are three IPE minima, and their position coincides with the location of three fluorine atoms of the anion. In this case both maps are useful, since the location of the BF_4^- anion agrees with the 2D-MIPp map (over the center of the ring)

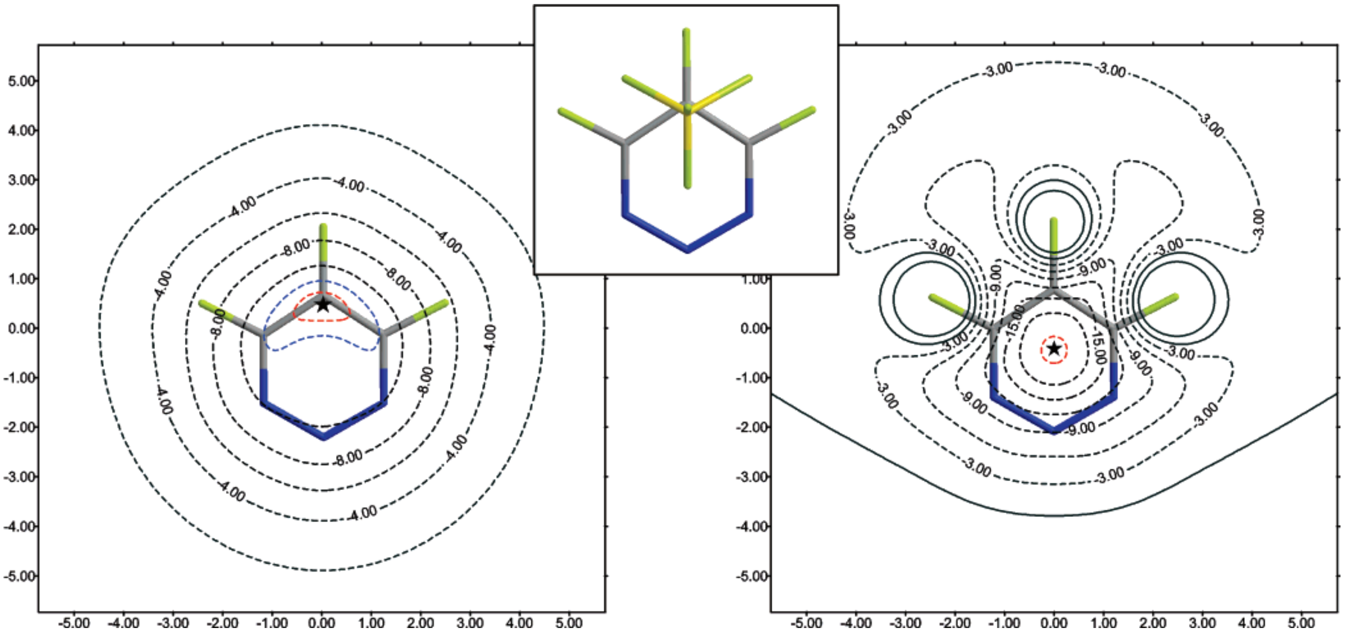


Figure 9. Left: IPE(F^-) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol from −2.0 to −10.0 kcal/mol. The blue isocontour corresponds to −11.0 kcal/mol, and the red isocontour corresponds to −11.5 kcal/mol. The minimum is represented by a star. Right: The 2D-MIPp(F^-) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **6**-BF_4^- complex is represented (MP2/6-31++G**).

and the orientation of the anion agrees with the 2D-IPE map (three F atoms pointing to the ring).

The 2D-IPE and 2D-MIPp maps computed for 1,2,4-trifluorotriazine (**5**) are represented in Figure 8. The position of the anion in the **5**-BF_4^- complex agrees with the location of the IPE minimum. In addition the position of one fluorine atom of the anion agrees with the MIPp minimum. A parallel finding has been found for 1,2,3-trifluorotriazine **6** (see Figure 9); the position of the anion in the **6**-BF_4^- complex agrees with the IPE minimum, and the position of one fluorine atom of the anion agrees with the MIPp minimum. These results confirm the utility of the IPE maps to predict the geometries of anion-π complexes of asymmetric aromatic compounds and that they usefully complement the MIPp maps.

3. Heteroaromatic Compounds with Four Nitrogen Atoms. The 2D-IPE(F^-) and 2D-MIPp maps of the three isomers of perfluorotetrazine (**7**−**9**) and the geometry of the optimized complexes of compounds **7**−**9** with BF_4^- are represented in Figures 10−12. The 2D-IPE(F^-) map of 1,2,3,5-tetrazine **7** is in sharp agreement with the optimized geometry of the complex, as can be observed in Figure 10. Two fluorine atoms of the anion are located at the IPE minima. In this case the position of the anion is not in total agreement with the MIPp minimum, nevertheless the spatial location of the BF_4^- is approximately on the region of the red contour.

The 2D-IPE(F^-) and 2D-MIPp(F^-) maps of 1,2,4,5-perfluorotetrazine **8** are shown in Figure 11. As expected, the agreement between both maps and the geometry of the complex is good. The location of two fluorine atoms of the anion agrees with the IPE minima, and the global location of the BF_4^- anion agrees with the MIPp minimum. This behavior has also been observed for the other symmetric complexes **2**-BF_4^- and **4**-BF_4^- (see Figures 5 and 7, respectively), for which the IPE map accurately predicts and explains the orientation of the anion and the MIPp explains the position of the anion.

The 2D-IPE(F^-) and 2D-MIPp(F^-) maps computed for 1,2,3,4-perfluorotetrazine (**9**) are represented in Figure 12. The position of the anion in the **9**-BF_4^- complex agrees with the location of the IPE minimum, and the position of one fluorine atom of the anion agrees with the MIPp minimum. A similar behavior has been observed for the complexes of low symmetry (C_s or C_1), i.e. complexes of compounds **1**, **3**, **5**, and **6**.

C. AIM Analysis. We have used the Bader's theory of "atoms-in-molecules" (AIM), which has been widely used to characterize a great variety of interactions,[31] to analyze the anion-π interaction of the complexes and to study if there is a relationship between the location of critical points and the IPE and MIPp minima. It has been demonstrated that the value of the electron charge density at the (3,+3) critical point (CP) that it is generated in anion-π complexes can be used as a measure of the bond order.[12b,c] In addition, the presence of bond critical points between the anion and the atoms of the ring is a clear indication of bonding.[27] In Figure 13 we show the distribution of (3,-1) and (3,+3) CPs in complexes **2**-BF_4^-, **4**-BF_4^-, and **8**-BF_4^-. We have chosen the symmetric complexes to illustrate the distribution of CPs for the sake of clarity. Moreover the ring CPs are not shown for the same reason. The other complexes exhibit a more complicated distribution of CPs, and they have been included in the Supporting Information (Figures S1−S3). For complex **2**-BF_4^-, the exploration of the CPs revealed the presence of two (3, −1) and one (3, +3) CPs. The bond CPs connect two fluorine atoms of the anion with the middle of two C−C

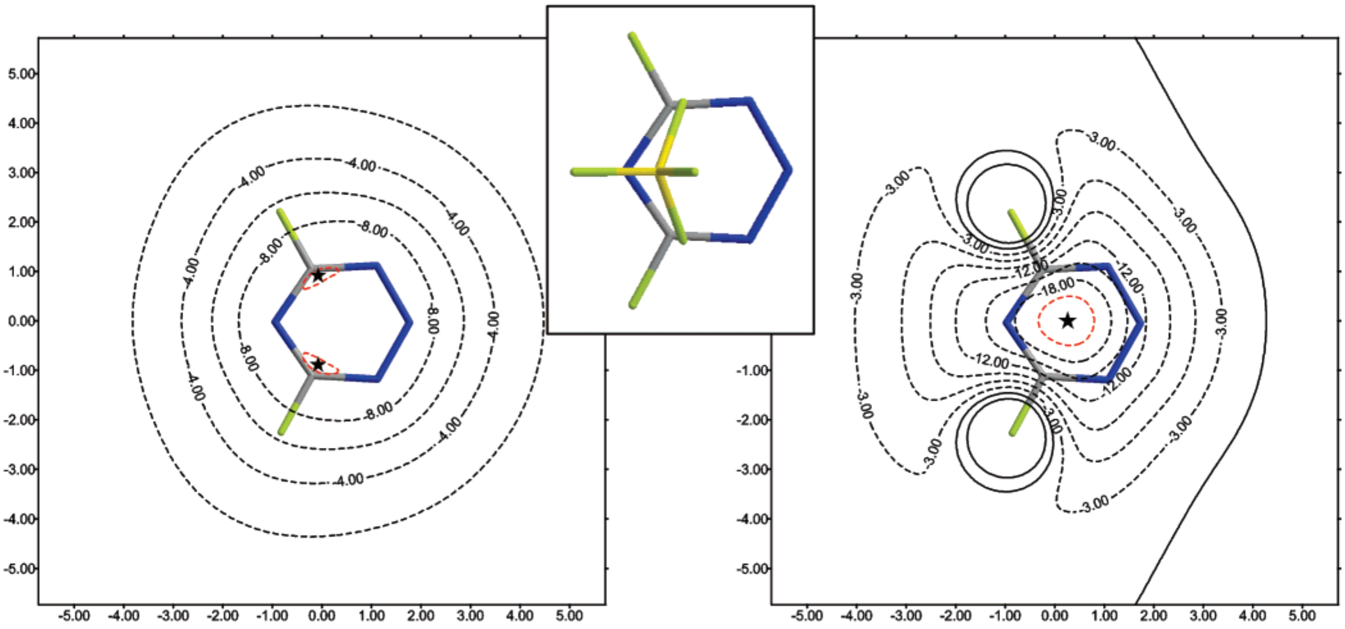


Figure 10. Left: IPE(F⁻) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol from −2.0 to −10.0 kcal/mol. The red isocontour corresponds to −10 kcal/mol. The minima are represented by stars. Right: The 2D-MIPp(F⁻) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **7**-BF_4^- complex is represented (MP2/6-31++G**).

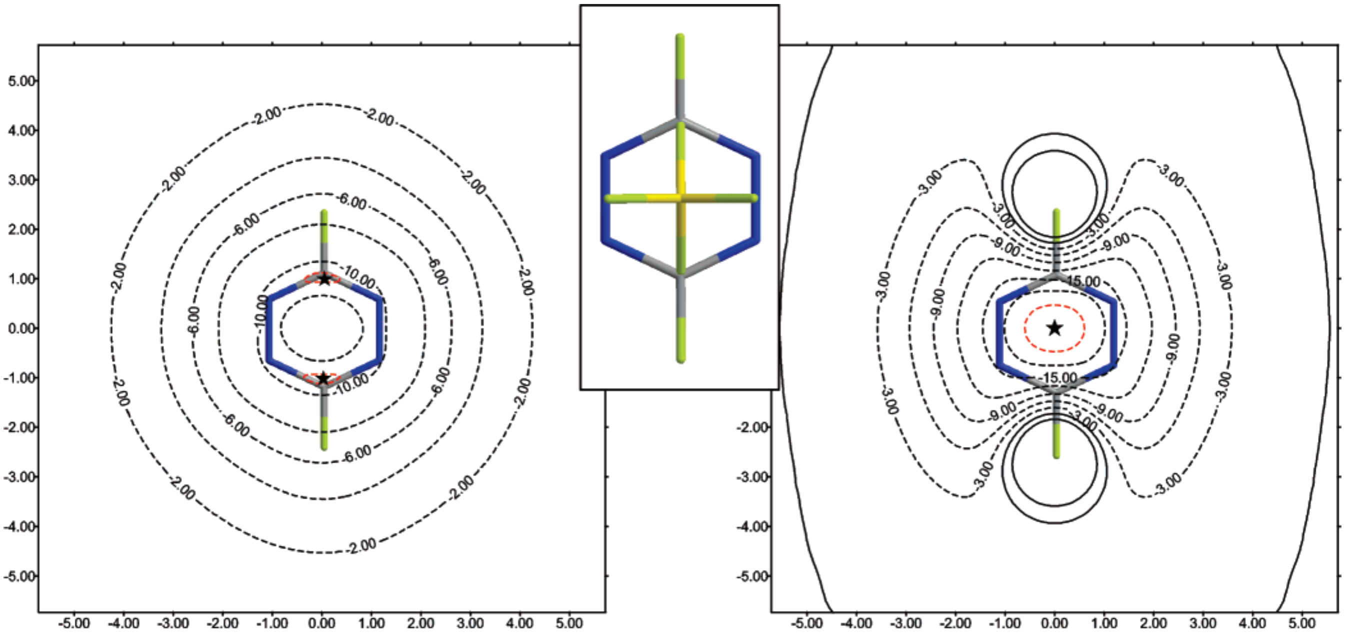


Figure 11. Left: IPE(F⁻) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol from −2.0 to −10.0 kcal/mol. The red isocontour corresponds to −10.2 kcal/mol. The minima are represented by stars. Right: The 2D-MIPp(F⁻) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **8**-BF_4^- complex is represented (MP2/6-31++G**).

bonds, and the cage CP connects the boron atom of the anion with the center of the aromatic ring. For complex **4**-BF_4^-, the exploration of the CPs revealed the presence of three (3, −1) and one (3, +3) CPs. The bond CPs connect three fluorine atoms of the anion with the carbon atoms of the ring, and the cage CP connects the boron atom of the anion with the ring centroid. For complex **8**-BF_4^-, the exploration of the CPs revealed the presence of two (3, −1) and one (3, +3) CPs. The bond CPs connect two fluorine atoms of the anion with the carbon atoms of the ring, and the cage CP connects the boron atom of the anion with the ring centroid. This distribution of CPs observed for the **2**-BF_4^-, **4**-BF_4^-,

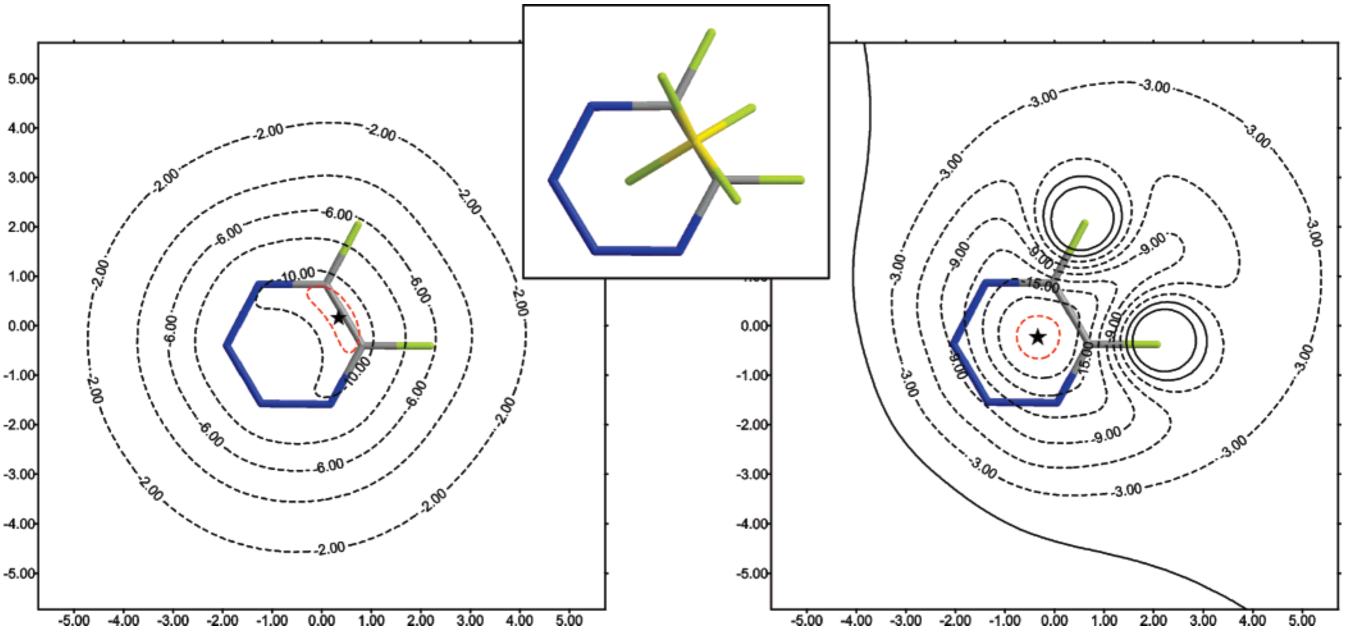


Figure 12. Left: IPE(F$^-$) map computed at 2.6 Å. Isocontour lines are drawn every 2 kcal/mol from −2.0 to −10.0 kcal/mol. The red isocontour corresponds to −10.5 kcal/mol. The minimum is represented by a star. Right: The 2D-MIPp(F$^-$) map is computed at 2.6 Å above the molecular plane. Isocontour lines are drawn every 3 kcal/mol, solid lines correspond to positive values of energy, and dashed lines correspond to negative values. The red isocontour line corresponds to −21 kcal/mol. The minimum is represented by a star. Middle: A zenithal view of the optimized **9**-BF_4^- complex is represented (MP2/6-31++G**).

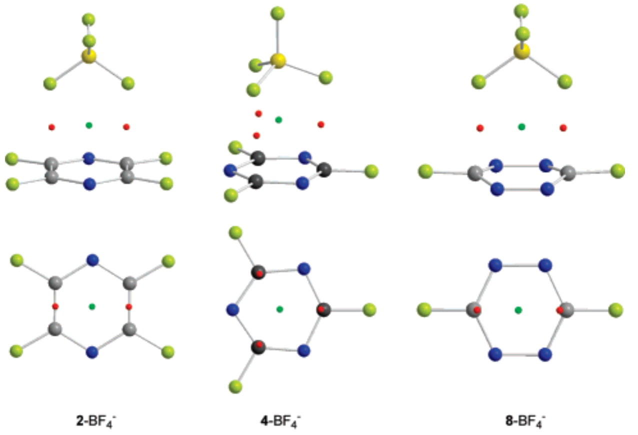


Figure 13. Schematic representation of the location of the (3,-1) bond CPs (red circles) and (3,+3) CP (green circle) originated upon complexation of the anion with compounds **2** (left), **4** (middle), and **8** (right). On the bottom the complexes are viewed perpendicular to the aromatic ring; for clarity, the anion has been omitted.

and **8**-BF_4^- is in good agreement with the 2D-IPE and 2D-MIPp maps and the location of the minima. The position of the bond CPs is related to the location of the IPE minima, and the position of the cage CPs corresponds to the location of the MIPp minima. For the rest of the complexes an acceptable agreement between the distribution of bond and cage CPs and the location of the IPE and MIPp minima is also found (see the Supporting Information), which is better for the C_s complexes **1**-BF_4^-, **3**-BF_4^-, **7**-BF_4^-, and **9**-BF_4^- than for complex **5**-BF_4^-. As aforementioned, the distribution of CPs is complicated, and, in general, the concentration of bond and cage CPs are mainly placed in the regions of the IPE and MIPp minima (see Figures S1−S3).

IV. Conclusion

The results derived from this study reveal that the 2D-IPE map is a good tool to predict and explain the geometric features of anion-π complexes. The polarization contribution to the total interaction energy is similar in magnitude to the electrostatic term. Moreover, it is more directional and is decisive to control the orientation of the tetrahedral BF_4^- anion.

The agreement between the MP2/6-31++G** optimized complexes and the 2D-IPE maps gives reliability to the MIPp partition scheme and supports the importance of polarization effects in both energetic and geometric characteristics of anion-π interactions. For symmetric complexes (C_{3v} and C_{2v}) the IPE map predicts the spatial disposition of the fluorine atoms, and the MIPp map predicts the position of the BF_4^-. For the rest of complexes the IPE map predicts the position of the BF_4^- anion, and the MIPp minimum coincides with the position of one fluorine atom of the anion. Finally, the AIM analysis is consistent with the IPE and MIPp maps.

Acknowledgment. We thank the DGICYT of Spain (projects CTQ2005-08989-01 and CTQ2005-08989-02) for financial support. We thank the CESCA for computational facilities. D.E. thanks the MEC for an undergraduate fellowship. D.Q. thanks the MEC for a “Juan de la Cierva” contract.

Supporting Information Available: Cartesian coordinates of MP2/6-31++G** optimized complexes and figures of AIM analysis of complexes of BF_4^- with compounds **1**, **3**, **5**−**7**, and **9** (Figures S1−S3). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.

(2) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed*. **2003**, *42*, 1210.

(3) (a) Ma, J. C.; Dougherty, D. A. *Chem. Rev*. **1997**, *97*, 1303. (b) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459. (c) Gokel, G. W.; Wall, S. L. D.; Meadows, E. S. *Eur. J. Org. Chem*. **2000**, 2967. (d) Gokel, G. W.; Barbour, L. J.; Wall, S. L. D.; Meadows, E. S. *Coord. Chem. Rev*. **2001**, *222*, 127. (e) Gokel, G. W.; Barbour, L. J.; Ferdani, R.; Hu, J. *Acc. Chem. Res*. **2002**, *35*, 878. (f) Hunter, C. A.; Singh, J.; Thorton, J. M. *J. Mol. Biol*. **1991**, *218*, 837.

(4) (a) Kumpf, R. A.; Dougherty, D. A. *Science* **1993**, *261*, 1708. (b) Heginbotham, L.; Lu, Z.; Abramson, T.; Mackinnon, R. *Biophys. J.* **1994**, *66*, 1061.

(5) Dougherty, D. A. *Science* **1996**, *271*, 163.

(6) Lummis, S. C. R.; Beene, D. L.; Harrison, N. J.; Lester, H. A.; Dougherty, D. A. *Chem. Biol*. **2005**, *12*, 993.

(7) Ishikita, H.; Knapp, E.-W. *J. Am. Chem. Soc*. **2007**, *129*, 1210.

(8) (a) Quiñonero, D.; Garau, C.; Rotger, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Angew. Chem., Int. Ed*. **2002**, *41*, 3389. (b) Alkorta, I.; Rozas, I.; Elguero, J. *J. Am. Chem. Soc*. **2002**, *124*, 8593. (c) Mascal, M.; Armstrong, A.; Bartberger, M. *J. Am. Chem. Soc*. **2002**, *124*, 6274.

(9) (a) Demeshko, S.; Dechert, S.; Meyer, F. *J. Am. Chem. Soc*. **2004**, *126*, 4508. (b) Schottel, B. L.; Bacsa, J.; Dunbar, K. R. *Chem. Commun*. **2005**, 46. (c) Rosokha, Y. S.; Lindeman, S. V.; Rosokha, S. V.; Kochi, J. K. *Angew. Chem., Int. Ed*. **2004**, *43*, 4650. (d) de Hoog, P.; Gamez, P.; Mutikainen, I.; Turpeinen, U.; Reedijk, J. *Angew. Chem., Int. Ed*. **2004**, *43*, 5815. (e) Frontera, A.; Saczewski, F.; Gdaniec, M.; Dziemidowicz-Borys, E.; Kurland, A.; Deyà, P. M.; Quiñonero, D.; Garau, C. *Chem. Eur. J*. **2005**, *11*, 6560. (f) Gil-Ramirez, G.; Benet-Buchholz, J.; Escudero-Adan, E. C.: Ballester, P. *J. Am. Chem. Soc*. **2007**, *129*, 3820. (g) Mascal, M. *Angew. Chem., Int. Ed*. **2006**, *45*, 2890.

(10) Gorteau, V.; Bollot, G.; Mareda, J.; Perez-Velasco, A.; Matile, S. *J. Am. Chem. Soc*. **2006**, *128*, 14788.

(11) Gamez, P.; Mooibroek, T. J.; Teat, S. J.; Reedijk, J. *Acc. Chem. Res*. **2007**, *40*, 435.

(12) (a) Cubero, E.; Luque, F. J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5976. (b) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *ChemPhysChem* **2003**, *4*, 1344. (c) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2004**, *108*, 9423.

(13) (a) Williams, J. H.; Cockcroft, J. K.; Fitch, A. N. *Angew. Chem., Int. Ed. Engl*. **1992**, *31*, 1655. (b) Williams, J. H. *Acc. Chem. Res*. **1993**, *26*, 593. (c) Adams, H.; Carver, F. J.; Hunter, C. A.; Morales, J. C.; Seward, E. M. *Angew. Chem., Int. Ed. Engl*. **1996**, *35*, 1542.

(14) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200.

(15) (a) Garau, C.; Quiñonero, D.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Org. Lett*. **2003**, *5*, 2227. (b) Quiñonero, D.; Garau, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Chem. Phys. Lett*. **2002**, *359*, 486.

(16) (a) Maheswari, P. U.; Modec, B.; Pevec, A.; Kozlevcar, B.; Massera, C.; Gamez, P.; Reedijk, J. *Inorg. Chem*. **2006**, *45*, 6637. (b) Mooibroek, T. J.; Gamez, P. *Inorg. Chim. Acta* **2007**, *360*, 381.

(17) (a) Garau, C.; Quiñonero, D.; Frontera, A.; Escudero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *Chem. Phys. Lett*. **2007**, *438*, 104. (b) Quiñonero, D.; Frontera, A.; Escudero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *ChemPhysChem* **2007**, *8*, 1182.

(18) Luque, F. J.; Orozco, M. *J. Comput. Chem*. **1998**, *19*, 866.

(19) (a) Hernández, B.; Orozco, M.; Luque, F. J. *J. Comput.-Aided Mol. Des*. **1997**, *11*, 153. (b) Luque, F. J.; Orozco, M. *J. Chem. Soc., Perkin Trans. 2* **1993**, 683. (c) Quiñonero, D.; Frontera, A.; Suner, G. A.; Morey, J.; Costa, A.; Ballester, P.; Deyà, P. M.; *Chem. Phys. Lett*. **2000**, *326*, 247. (d) Quiñonero, D.; Frontera, A.; Garau, C.; Ballester, P.; Costa, A.; Deyà P. M. *ChemPhysChem*. **2006**, *7*, 2487.

(20) Scrocco, E.; Tomasi, J. *Top. Curr. Chem*. **1973**, *42*, 95.

(21) Orozco, M.; Luque, F. J. *J. Comput. Chem*. **1993**, *14*, 587.

(22) Francl, M. M. *J. Phys. Chem*. **1985**, *89*, 428.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(24) Boys, S. B.; Bernardi, F. *Mol. Phys*. **1970**, *19*, 553.

(25) Luque, F. J.; Orozco, M. *MOPETE-98 computer program*; Universitat de Barcelona: Barcelona, 1998.

(26) Garau, C.; Quiñonero, D.; Frontera, A.; Costa, A.; Ballester, P.; Deyà, P. M. *Org. Lett*. **2003**, *5*, 2227.

(27) (a) Bader, R. F. W. *Chem. Rev*. **1991**, *91*, 893. (b) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon: Oxford, 1990.

(28) http://www.AIM2000.de (accessed Sep 7, 2007).

(29) Amos, R. D.; Alberts, I. L.; Andrews, J. S.; Colwell, S. M.; Handy, N. C.; Jayatilaka, D.; Knowles, P. J.; Kobayashi, R.; Laidig, K. E.; Laming, A. G.; Lee, M.; Maslen, P. E.; Murray, C. W.; Rice, J. E.; Simandiras, E. D.; Stone, A. J.; Su, M.-D.; Tozer. D. J. *CADPAC*: *The Cambridge Analytic Derivatives Package Issue 6*; Cambridge, U.K., 1995.

(30) Hernandez-Trujillo, J.; Vela, A. *J. Phys. Chem*. **1996**, *100*, 6524.

(31) (a) Cheeseman, J. R.; Carrol, M. T.; Bader, R. F. W. *Chem. Phys. Lett*. **1998**, *143*, 450. (b) Koch, U.; Popelier, P. L. A. *J. Phys. Chem*. **1995**, *99*, 9794. (c) Cubero, E.; Orozco, M.; Luque, F. J. *J. Phys. Chem. A* **1999**, *103*, 315.

CT700122Y

JCTC Journal of Chemical Theory and Computation

Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects

Kai Zhu, Michael R. Shirts, and Richard A. Friesner*

Department of Chemistry, Columbia University, New York, New York 10027

Received July 3, 2007

Abstract: This paper presents significant improvements in both accuracy and computational efficiency of protein side chain and loop predictions using the Protein Local Optimization Program (PLOP). We introduce a novel energy model in which the internal dielectric constant of the protein is allowed to vary as a function of the interacting residues and present a physical rationale for this model. Using this model, we achieve qualitative improvements in the accuracy of side chain predictions with respect to experimental crystal structure and substantially reduce the RMSDs for loop predictions, particularly those predictions involving charged side chains. For the single side chain prediction of lysine, 40% of the errors are eliminated, and the accuracy increases from 62.6% to 76.8%. The errors in glutamate and aspartate predictions are reduced by 19% and 24%, respectively. When applied to a set of 240 loop predictions with 6, 8, 10, and 13 residue of loop length, this new model yields unprecedented accuracies with average backbone root-mean-square deviations of 0.39 Å, 0.68 Å, 0.80 Å, and 1.00 Å for 6, 8, 10, and 13 residue loops, respectively. We also describe a series of technical improvements in the PLOP simulation algorithms, which lead to a speedup of a factor of 2−4 in loop predictions.

I. Introduction

In several previous publications, we have described a novel approach to high-resolution protein structure prediction, implemented in the protein local optimization program (PLOP).[1−4] This program combines sophisticated conformational sampling algorithms with a molecular mechanics force field[5,6] and a continuum solvation model based on the generalized Born approach,[7,8] in contrast to typical programs for loop and side chain modeling which rely on either simplified physical chemistry based scoring functions or knowledge based potentials inspired by bioinformatics approaches.[9−21] While such approximate methods have performed respectably for low-resolution protein modeling, it appears as though achievement of a truly accurate atomic level description of protein structure—as is required for many practical applications, for example structure based drug design—necessitates the use of more accurate energy functions and correspondingly efficient and precise sampling algorithms. Using this more physical approach, we have achieved significant reductions in root-mean-square-deviation (RMSD) from crystal structures in repredictions of loops and side chains within the context of the native structure of the protein. Particularly large advances[4] are apparent for long loops (up to ∼11−13 residues), which place severe demands upon the accuracy of the scoring function and the efficiency of the sampling algorithms.

Despite these successes, the previously published methodology still displays systematic errors for subsets of test cases such as the prediction of lysine side chain structures. It also suffers from substantial computational requirements, which becomes a significant problem when applying the methods to problems such as homology modeling where the “context” of the local region to be refined is imperfect, and a large number of calculations per loop region are presumably

* Corresponding author e-mail: rich@chem.columbia.edu. Corresponding author address: Department of Chemistry, Columbia University, 3000 Broadway, Mail Code 3110, New York, NY 10027.

10.1021/ct700166f CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/16/2007

required to achieve convergence of the energy function. These deficiencies have motivated further development of both the computational algorithms in PLOP and the energy model used to rank order structures. Our expectation is that such improvements will be ongoing for a number of years, and our experience to date has been that both quantitative and qualitative advances in the technology continue to be generated by these efforts.

This paper presents significant improvements in both accuracy and computational efficiency. Qualitative improvements which both increase the accuracy of side chain prediction and substantially reduce the RMSDs for loop prediction are achieved by a model in which the internal dielectric constant of the protein is allowed to vary as a function of the interacting residues. A novel physical rationale for this model, based on an analysis of the treatment of polarization in contemporary fixed charge force fields, is presented, and the model is shown to have a particularly large effect on predictions for charged side chains. Second, we describe a series of technical improvements in the PLOP simulation algorithms, which lead to a speedup of a factor of 2−4 in loop prediction. Further acceleration of the calculations is clearly possible but is left for another publication.

The paper is organized as follows. In section II, a brief review of the PLOP methodology for side chain and loop prediction is presented. Section III introduces the variable internal dielectric model and provides the physical interpretation of this model as well as detailing its particular implementation in the current version of PLOP. It also presents the algorithmic improvements in the current version of PLOP responsible for increased speed. Section IV discusses the data sets that we use to benchmark the methodology, for both side chain and loop prediction. Section V presents accuracy and timing results, comparing earlier versions of PLOP as well as some literature data with the current version. Finally, in the Conclusion, we summarize the results and discuss future directions.

II. Side Chain and Loop Prediction via PLOP

We have described our side chain and loop prediction algorithms in previous publications.[1−4] Here we give only a brief review. Initially, we use *single* side chain prediction (i.e., keeping the remainder of the protein fixed at the native configuration) to parametrize and validate the variable dielectric model. We originally developed this strategy to develop and test the new torsional parameters for the OPLS-AA force fields[3] and for a novel protonation state assignment algorithm.[22] We use a hierarchical approach to single side chain prediction. Initially, side chain conformations are sampled using a highly detailed rotamer library developed by Xiang and Honig.[15] This library contains, for example, 2086 rotamers for lysine. The use of such a detailed library ensures adequate sampling. The associated computational expense is reduced by prescreening the rotamers using only hard sphere overlap as a criterion, which can be made very rapid with the use of a cell list. Many rotamers can in this manner be excluded before performing energy evaluations. Then a rapid, reduced energy calculation is performed for each remaining rotamer. The reduced energy uses a short cutoff for nonbonded interactions and includes only the torsional energy among the bonded terms. Next we perform a clustering procedure on these rotamers. We start at the lowest energy structure and then find all neighbors in torsional space, working outward until the energy no longer goes up. This is the first energy basin (or cluster). Then we find the lowest energy structure among the remaining rotamers and continue to do this for all energy basins until we run out of rotamers. The representative of each cluster is chosen as the lowest energy structure of the energy basin. The entropic contribution is calculated by taking the configurational integral in the torsional space over each basin. Then the representative is completely energy minimized using a fast minimization algorithm previously developed.[23] The sum of the minimized energy and the entropic contribution is used to rank the structures and give the final prediction.

Loop predictions in PLOP also feature a hierarchical approach. The generation of loop conformations is accomplished via a dihedral angle buildup procedure which, at the limit of highest resolution, exhaustively searches the phase space of possible loop geometries connecting the two loop stems. The energy evaluation achieves both efficiency and high accuracy via deployment of a hierarchy of scoring functions; rapid screening functions are used to eliminate large numbers of high-energy loops, ultimately yielding a relatively small number of candidates which are then clustered. Representative members of each cluster are then evaluated via minimization of an all atom molecular mechanics energy function and continuum solvation model (in this study OPLS-AA force field[5,6] plus SGB/NP solvent model).[7,8] Furthermore, we have developed a powerful sampling algorithm for the long loop predictions, which involves multiple stage loop predictions and refinements, and achieved very high accuracy when combined with a hydrophobic term we have developed to fix a major flaw in the generalized Born model.[4] The crystal environment is explicitly included in our loop and side prediction algorithms by using dimensions and the space group reported in PDB files.[1,2] PLOP executables can be obtained from Matthew Jacobson at UCSF, free of charge for academic users, as per instructions on his Web site (http://francisco.compbio.ucsf.edu/∼jacobson/). An implementation of PLOP, with a graphical user interface, is also available to both academic and commercial users in the Prime program, distributed by Schrodinger, Inc.

III. Methods

A. Variable Internal Dielectric Model. The question of what value to use for the internal dielectric constant in Poisson−Boltzmann (PB) and generalized Born (GB) calculations has been the subject of a large number of papers over the past 20 years. Much of the early work was focused on PB methods, at a time when analytical gradients for PB calculations did not exist, and there was therefore no convenient way to carry out accurate conformational searches, geometry optimizations, or molecular dynamics simulations of PB based models. In this situation, movement of protein groups was often invoked to justify the use of a “high” internal dielectric constant, typically, in the 4−20 range. For example, such values were used extensively in PB based pK_a calculations.[24−28] However, our current methodology involves a more extensive exploration of conformational phase space, so this component of the protein dielectric should not be contributory.

A second alternative is to assign an internal dielectric to the protein based on the optical dielectric constant, which is based on electronic response, not nuclear motion, of a “typical” organic compound. This value is on the order of 2.[29,30] At first glance, this appears to be a satisfactory approach, as there is no question that reorganization of the electrons in the protein will occur in response to an applied field. However, when considering what internal dielectric to employ, one has to take into account how the molecular mechanics (MM) force fields used in PB or GB calculations were developed. Even hydrocarbons in the OPLS-AA force field, for example, have small, but noticeable, point charges associated with them. The critical point is that in general the charges used in any MM force field are *already* enhanced from gas-phase values, to take into account the “average” polarization in the condensed phase. For example, the OPLS-AA force field was parametrized to fit experimental thermodynamic data such as density and heat of vaporization of pure liquids. An alternative is to use quantum chemically derived charges that are “scaled up”, either by using a relatively small basis set that implicitly yields higher charges or by explicit use of a scale factor. Thus, one could argue that the optical polarization has been implicitly included in the force field, and using a dielectric of 2 results in double counting of this effect. One reasonable alternative is to employ specially derived charges that are fit to experiment to complement the model with a dielectric of 2, as in the PARSE model of Honig and co-workers.[31] However, this involves a complete redesign of the force field and discards the substantial amount of information obtained from fitting molecular dynamics simulations of pure liquids to thermodynamic data.

Our philosophy to date within PLOP has been to use a dielectric constant of 1, on the theory that optical polarization of the protein is incorporated into the force field as discussed above. However, there is one situation where this argument fails, and that is when a protein group is interacting with a charged species. Neutral groups, such as the hydroxyl group in serine, are parametrized to fit pure liquid simulations of methanol; hence, the assumed environment of the group is neutral hydrogen bonds to other hydroxyls. If a serine is instead hydrogen bonded to a lysine, the polarization of the group is presumably greater—but this is not reflected in the uniform internal dielectric of 1 that is used in the GB model in PLOP. Similarly, charged groups are parametrized to agree with solvation free energies of the charged species in water, again placing the group in question in a neutral environment. When a salt bridge is formed, the internal dielectric of 1 is then, again, presumably inappropriate.

The clearest solution to this problem is to employ a polarizable force field in high-resolution protein simulations. This is a promising future path that is being pursued by several research groups[32−42] but still requires significant additional effort along a number of directions to be fully practical, for example, the design of a continuum solvation model that is compatible with the polarizable force field.[43,44] If a polarizable force field is used to represent the protein, then the internal dielectric clearly should be unity, as all internal polarization effects are now being modeled explicitly.

In the context of a fixed charge force field, possible solutions of the internal dielectric problem must involve heuristic approximations. We describe one such approximation below and then demonstrate that significant improvement in both side chain and loop prediction is obtained with the use of a few adjustable parameters, which assume physically reasonable values. Furthermore, simpler approaches, such as increasing the internal dielectric constant from one to a higher value, lead to results that are qualitatively inferior.

The basic idea is to vary the value of the internal dielectric constant as a function of the interacting atoms. In the GB model, the total electrostatic free energy is expressed as the sum of the Coulomb interaction and the generalized Born solvation term

$$G_{\mathrm{es}} = \frac{1}{2}\sum_{i<j}\frac{q_i q_j}{r_{ij}\epsilon_{\mathrm{in}}} - \frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{sol}}}\right)\sum_{ij}\frac{q_i q_j}{f_{\mathrm{GB}}} \tag{1}$$

where

$$f_{\mathrm{GB}} = \sqrt{{r_{ij}}^2 + {\alpha_{ij}}^2 e^{-D}}$$

and

$$\alpha_{ij} = \sqrt{\alpha_i \alpha_j}, \quad D = \frac{{r_{ij}}^2}{(2\alpha_{ij})^2}$$

The α_i’s are generalized Born radii. In our variable dielectric model, the internal dielectric ϵ_{in} depends upon the pair of atoms that are involved in the specified electrostatic interaction. We write this explicitly:

$$G_{\mathrm{es}} = \frac{1}{2}\sum_{i<j}\frac{q_i q_j}{r_{ij}\epsilon_{\mathrm{in}(ij)}} - \frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}(ij)}} - \frac{1}{\epsilon_{\mathrm{sol}}}\right)\sum_{ij}\frac{q_i q_j}{f_{\mathrm{GB}}} \tag{2}$$

Note that this newly defined $\epsilon_{\mathrm{in}(ij)}$ enters both the solvation terms and the Coulomb interaction terms. We use a residue-based parametrization with the variation confined to side chain atoms of charged residues, i.e., we assign different dielectric constants for charged side chains, while all the backbone atoms and neutral side chains still use the dielectric 1. There are several reasons for not changing dielectric constants of backbone atoms: (1) We want to treat the backbones consistently with both charged and neutral residues, as they have the same parameters (charge, Lennard-Jones, etc.) independent of the residue type in the OPLS-AA force field. (2) Although the individual atoms in the backbone carbonyl and the amine group are significantly charged, the backbone as a whole is neutral, hence the argument that an appropriate polarization has already been incorporated via optimization of the charges in liquid-state simulations applies. (3) Experiments with structure predic-

Table 1. Internal Dielectric Constants Used in a Variable Dielectric Model[a]

residue	Lys	Glu	Asp	Arg	His	other
dielectric	4	3	2	2	2	1

[a] The new values other than 1 are only assigned to the side chain atoms. For a specific interacting pair, the internal dielectric uses the larger one of the two values associated with the two atoms.

tions prove it is a better choice to use unity dielectric for backbone atoms than varying it with residue type (data not shown).

For the inter-residue pair interaction, in the present paper we employ a simple rule in which the higher of the two “residues-based” dielectric constants are used. If both interacting atoms are neutral, then according to the arguments described above, the internal dielectric constant is set to 1. On the other hand, if one or both of the atoms belong to the side chain of a charged residue, then the higher internal dielectric associated with the two atoms is employed. The adjustment of the residue-based dielectrics is accomplished through the improvement of the single side chain predictions. We first use a uniform dielectric of 1 to 6, 8, and 20 for the entire protein and determine which value yields the most accurate predictions (measured by the percentage accuracy, see Results) for each residue and choose this value to be each residue’s dielectric. We did not try a finer parametrization because the structure prediction is not very sensitive to the slight changes in the dielectric constant any smaller than 1. Then, in the variable dielectric model where the combining rule is applied for inter-residue interactions, we further adjust these dielectric values (by a maximum of ±1) to maximize the overall accuracy of the single side chain predictions of all 11 polar or charged residues. The optimized set of values is presented in Table 1; the results obtained using these values are given in section V. Considering there are 2178 single side chain prediction test cases and only 5 adjustable parameters, it seems unlikely that the results are due to gross overfitting. The significant improvement of loop predictions by the variable dielectric model (see section V) also provides an independent validation test. A more sophisticated optimization could be employed and might yield better results; some possibilities along these lines are considered in the Discussion section.

Before examining detailed numerical results, it is useful to obtain an intuitive physical feeling for the results of the variable dielectric model outlined above and to see whether such results will move side chain predictions in the qualitatively correct direction. For charged residues a well-known, fundamental problem of dielectric continuum models is a tendency to form salt bridges considerably more frequently than is observed experimentally; this was demonstrated, and discussed, in our previous work.[45] In contrast, hydrogen bonds between neutral residues do not display an extreme bias one way or another. Thus, the hope would be that the variable dielectric model can reduce the frequency of salt bridges, while having minimal impact on neutral−neutral hydrogen bonding.

A simple physical argument suggests that this will indeed be the case. Take the case of the interaction of a lysine residue with the surrounding atoms of the protein. The internal dielectric refers to all of the atoms surrounding the lysine. The polarization of these atoms creates a reaction field around the charged atoms of the lysine group—not as large as the reaction field from water but larger than would be observed if the internal dielectric constant were unity. This reaction field then has an unfavorable interaction with the hydrogen-bonding partner of the lysine, e.g., a carboxylate group. Similarly, the carboxylate group has a reaction field around it that has an unfavorable interaction with the lysine. It is these reaction fields that reduce the magnitude of the effective interactions between the two groups, as in the case of charged groups in water.

This effect is most easily seen from eq 1 in the limit where the Born alpha values of the interacting atoms become very large, as would be the case for a significantly buried salt bridge. In the large alpha limit, the second term in eq 1 becomes negligible, and one is left with only the first term for the interaction energy; this term divides the Coulomb interaction by the internal dielectric constant. In this way, salt bridges become properly energetically disfavored using the variable dielectric model, due to the increased reaction field of the protein surrounding each component of the salt bridge. This diminishes the number of unphysical salt bridges as is demonstrated in more detail below.

For other, more complicated cases, in which alpha is not assumed to be much larger than the separation distance between the interacting groups, the analysis of increasing the internal dielectric becomes more complicated, but the basic physics (creation of a reaction field due to the protein in response to the electric fields from the interacting groups) is unchanged, and the adequacy of the quantitative treatment of this effect by our simple, variable dielectric approximation must be judged by the quality of the results for loop and side chain predictions as presented below.

Now consider the case of two serines interacting with each other, for which, in computing the electrostatic free energy via eq 1, we use an internal dielectric of unity. There is no enhancement of the reaction field surrounding the −OH groups, because the remainder of the protein force field was derived based on a neutral, hydrogen-bonding environment. In contrast, the charged groups will produce a reaction field in the protein in excess of what is incorporated into the force field, because the field exerted on the neighboring protein atoms is in excess of what was used in the parametrization of the model. The empirically tuned variable dielectrics represent a crude, but apparently quite useful approximation to the magnitude of this excess. The variation in the reaction field differential with Born alpha and other geometrical parameters of a given structure is implicit in eq 1; this apparently corresponds well enough to physical reality that substantial improvements in both side chain and loop predictions are obtained. It is worth noting that alternative empirical “fixes” such as changing the dielectric radius of various charged atoms did not yield significant improvements in structural predictions in our experiments (data not shown).

Implementation of the variable, residue-based dielectric model as described above is relatively trivial; the constant internal dielectric used previously is replaced by a variable determined by looking up the appropriate value in Table 1.

The extensive tests, carried out and described in section V, are considerably more time-consuming but necessary to evaluate the performance of the model, which, given its approximate and heuristic character, cannot be rigorously inferred from the theoretical arguments made in this section.

B. Increasing the Speed of PLOP. A number of software optimizations have been included in the current version of PLOP after extensive profiling of the previous versions. Some improvements involve simple steps to avoid unnecessarily expensive copying of large arrays and string compares. More substantially, all function calls and most conditionals have been removed from inner loops of the gradient and energy code. Instead of conditionals for the various solvation types and corrections being placed in the inner loops, separate inner loops for different solvation conditions are generated automatically with pseudocode. Any atom pairs requiring special additional correction terms (such as the hydrophobic pair term introduced previously[4]) are now placed into separate neighbor lists, removing the need for conditionals within the inner loops.

After the inner loops were optimized, the generation of the SGB surface, the integration over this surface, and the determination of Born alphas consumed the most time in both minimizations and loop predictions. A number of improvements to the surface generation and integration code have been performed to eliminate unnecessary checks and duplicated calculations that occurred when only some parts of the surface changed. Additionally, in the intermediate steps of side chain optimization, Born alphas are only updated within twice the solvent radius (1.4 Å for implicit water) of any moving atoms. This reduces the time for calculations done for intermediate steps where only approximate energy evaluations are necessary.

A previously implemented correction to the Generalized Born energy due to Ghosh et al.[8] with the aim of improving the consistency between GB and PB results for protein structures was determined to almost double the time required to determine the surface integral, taking 20−30% of the total time of a loop prediction. A set of side chain and loop prediction tests determined that this correction only negligibly affected the prediction results (average RMSDs differ less than 0.05 Å) and has therefore been removed from the code.

One new algorithmic improvement in the current implementation of PLOP is the replacement of residue-based cutoffs with dipole based cutoffs. In the previous versions of PLOP, potential energy cutoffs were residue based. If any two atoms of a residue were less distant than the specified cutoff, all atoms on those two residues were treated as interacting. Neutral−neutral residues had a first cutoff, neutral-charged residue interactions had a second cutoff, and charged−charged residue interactions had a third cutoff. This leads to an undesirable situation where extremely small changes in structure can result in a significantly larger change in energy for short cutoffs as residues move in and out of the cutoff distance. This residue-based approach can therefore cause instabilities with the multiscale minimization algorithm used in PLOP. In this scheme, inner loops of the minimization use only a short-range cutoff, and the long-range gradient is approximated as a constant. When the neighbor list is updated, however, entire residues may have moved into or out of the cutoff, changing the direction of the minimization significantly. Relatively long short-range cutoffs were therefore required; with cutoffs less than 8 Å, minimizations would frequently become numerically unstable. A default short-range cutoff of 10 Å cutoffs for neutral−neutral residues and neutral-charge residues and 15 Å cutoffs for charge−charge residues was determined to be safe for adequate convergence.

In the PLOP implementation presented in this paper, distance-based interaction cutoffs are still present but are dipole based, instead of residue based. Atomic charges are decomposed into formal charges and dipoles. As an example, we examine the case of a hypothetical neutral methane. If we assign hydrogens a partial charge of 0.1, the central carbon must have a partial charge of −0.4 (all charges are for illustrative purposes and not meant to reflect the actual force field used for predictions). We now represent the partial charges, instead of being atom based, being bond based, with each bond having positive and negative charges equal in magnitude at each end. This methane molecule then consists of four C−H dipoles, each of magnitude 0.1, with the negative poles toward the carbon and positive poles toward the hydrogen. For a hypothetical ammonium ion, with a total charge of 1.0, partial charges of 0.2 on the hydrogens, and a partial charge of 0.2 on the nitrogen, it would be represented as a formal charge of magnitude 1.0 centered on the nitrogen and four dipoles of magnitude 0.2, with negative poles toward the nitrogen. For a neutral molecule, there exists a unique decomposition with the exception of ring systems; a choice is made in the ring systems to minimize the magnitudes of the resulting dipoles.

The atoms of two dipoles interact only if all four atoms are within the cutoff distance of each other. For example, imagine two such methane molecules interacting. Suppose that all pairs of atoms between the two methanes are within the cutoff, with the exception of one hydrogen on methane A and one hydrogen on molecule B whose distance is slightly greater than the cutoff. These two hydrogens do not interact, and because atoms of these two dipoles do not interact, neither do the other atoms in the dipole. The partial charge of both carbons is now the sum of only the three dipoles, so it becomes 0.3. This partial charge is only with respect to the pair, meaning that the effective pairwise partial charge must be determined between all pairs of atoms. If there are two hydrogens on molecule A that are further than the cutoff from one hydrogen on molecule B, then the carbon on methane A would have partial charge 0.2, and the carbon on molecule B would have partial charge 0.3. This method of determining the electrostatics cutoffs has a useful property that the total sum of the product of all charge pairs in the system q_iq_j is independent of the cutoff, at least for those charges that are the result of sums of dipoles, not formal charges. This results in significantly smoother changes in the energy with respect to changes in structure than an abrupt atomic cutoff or even a residue-based cutoff. The basic algorithm, as described up to this point, was implemented in the Macromodel modeling package but has not previously been published. The additional modifications described below, however, are novel.

Determining the dipole interactions as a function of the distance between pairwise atoms can be somewhat time-consuming and must be redone every time the structure changes. To significantly improve the performance of the method, we group sets of dipoles together. If an atom has only one neighbor, then we call it a leaf. When it has more than one neighbor (or, for practical purposes, it has a formal charge), we call it a trunk. When two trunks interact, then all of their leaves also interact. Therefore, the product of the partial charges for the interaction of any two leaves is zero if their trunks are not within the cutoff. This is equivalent to treating the length of the trunk/leaf dipole as zero for the purpose of determining which dipoles interact (though *not* for the purposes of calculating the energy itself). For example, if we consider two methanol molecules under this system, the C−H dipoles will always interact if the C−C distance is sufficiently close, as will the O−H dipoles, if the O−O distance is sufficiently close. However, the intramolecular C−O dipole will only contribute to the total partial charges involved in the C−C interaction if all four carbons/oxygens intermolecular distances are less than the cutoff. In many cases, the interactions of this form of dipole based cutoffs may die off more quickly as a function of distance that in the original version, as multiple dipoles contributing to a single "trunk" will tend to orient in opposite directions, leading to a smaller average dipole.

In place of neutral−neutral, neutral−charged, and charged−charged residue cutoffs, all dipole−dipole interactions and all Lennard-Jones interactions are truncated with a single dipole−dipole cutoff, with dipole−formal charge and formal charge−formal charge cutoffs treated separately with longer cutoffs. Previously, Lennard-Jones interactions between charged−charged residues or charged−neutral residues as well as the electrostatics of nonpolar moieties were calculated at a larger distance than Lennard-Jones interactions in neutral−neutral sides. Although these deviations from a uniform treatment of the Lennard-Jones terms are relatively small, they can cause inconsistencies when, for example, the protonation state of a residue is changed.[22]

For multiscale minimization, PLOP uses two separate neighbor lists: a long-range list that is used to calculate the full gradient and energy and a short-range list that is used in calculating the short-range gradient and Hessian for the inner loops and preconditioner of the minimizer. The dipole-based cutoff scheme is applied to both sets of cutoffs.

With the new dipole-based cutoff system, the number of atoms included in the short-range energy and gradient evaluation can be reduced significantly, with a relatively small computational cost for the additional bookkeeping. Well converged results can be obtained with cutoffs as short as 6 Å, and the use of even shorter cutoffs is only ruled out because of bookkeeping difficulties with the 1−4 interactions, which must be treated separately. As the short-range cutoffs get smaller, however, the approximation used in the multiscale minimization of a long-range contribution to the gradient which is constant constant during the minimization inner loop becomes less accurate, increasing the number of iterations required to converge minimizations. Short-range cutoffs of 8 Å dipole−dipole interactions, 10 Å for dipole−formal charge interactions, and 16 Å for formal charge−formal charge interactions provide near optimal speed in most instances. Long-range cutoffs of 15 Å, 20 Å, and 30 Å were used. These long-range cutoffs are the same as used for the previous residue-based cutoffs and thus involve somewhat fewer atoms but yield similar relative energies between structures with either cutoff scheme.

IV. Test Set

For the single side chain prediction test, we use the test suite of 30 protein structures from the previously publication.[22] These proteins are selected such that all of them have resolution <2 Å and do not have any serious heavy atom steric clashes or nonpeptide ligands. The pairwise sequence identity is less than 30%. In this work, we add a new screening criterion based on the B-factor to eliminate the noise in the results due to the experimental uncertainty. If any side chain atom has a B-factor greater than a threshold of 40, then that residue will be excluded from our list. We focus on the predictions of 11 polar (and charged) residues since the hydrophobic residues are generally buried and trivial to predict and thus less affected by the solvent model. This yields a total of 2178 single residue side chain structure repredictions.

For the loop prediction targets, we use the combined filtered list in ref 3 for 6, 8, and 10 residue loops. The 13 residue loops are from ref 4. These loop targets were filtered by pH value, B-factor, steric clash, and other criteria to ensure the selection of high-quality structures. In total, there are 99, 65, 41, and 35 targets for 6, 8, 10, and 13 residue loops. The loop target 9-14 of 1xso from the 6 residue list and the target 606-613 of 1gof from the 8 residue list are removed because of serious steric clashes in these structures.

V. Results

A. Single Side Chain Predictions. Single side chain predictions represent one of the simplest tests that can be applied to evaluate the quality of a protein energy model. We focus on comparisons to crystallographic structural data, as opposed to NMR data, as it is not clear whether NMR data for side chains is precise enough to enable robust comparisons at this point in time. To compare realistically with X-ray crystallographic data from the PDB, the calculations must be carried out in the appropriate crystalline environment; many side chains form nonbonded contacts (e.g., salt bridges) with neighboring protein molecules in the crystal. In the calculation, all atoms other than those of the side chain in question are held fixed, the conformational phase space of the side chain is sampled as thoroughly as possible, and the energy model (molecular mechanics potential energy plus free energy due to the continuum solvation model plus some entropic term) is used to select the final prediction. Because the number of degrees of freedom in a single side chain is small as compared to a long loop, it is generally possible to converge the side chain sampling to the global free energy minimum, although in a small number of cases in the data sets, the presence of a

Table 2. Single Side Chain Prediction Accuracy of 11 Polar Residues on a Data Set of 30 Proteins and 2178 Side Chain Targets[a]

residue type	total no.	uniform dielectric 1 (%)	uniform dielectric 2 (%)	uniform dielectric 4 (%)	variable dielectric (%)	variable dielectric with ICDA assignment (%)
ASN	237	71.7	72.6	70.5	75.5	85.7
GLN	161	65.8	65.8	58.4	65.2	85.7
HIS	132	54.5	60.6	58.3	59.8	86.4
ASP	254	83.9	86.2	83.9	87.8	91.7
GLU	193	67.4	74.1	75.6	73.6	79.3
SER	297	77.4	63.0	42.8	80.1	79.1
THR	302	90.1	90.7	88.1	91.7	92.4
LYS	198	62.6	70.7	77.8	76.8	76.8
ARG	171	78.4	77.2	70.8	74.9	77.8
CYS	49	93.9	89.8	89.8	93.9	93.9
TYR	184	88.0	89.7	92.4	91.3	89.7
SUM	2178	76.2	76.3	72.5	79.8	85.0

[a] The accurate prediction is defined as having side chain heavy atom RMSD less than 1.5 Å. The variable dielectric model is compared with the uniform dielectric models assuming different dielectric constants. The last column shows the single side chain prediction results with the ICDA assignment structures using variable dielectric model.

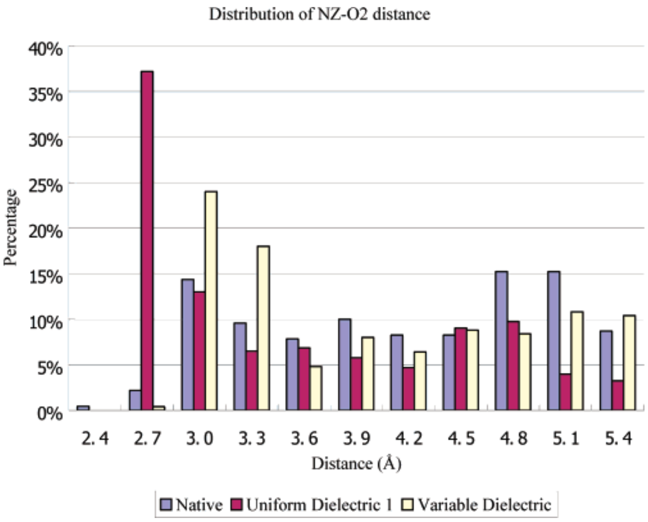


Figure 1. The distribution of distances between the lysine NZ atom and the carboxylic acid O2 atoms. The predictions of uniform dielectric 1 and the variable dielectric model are compared with native structures. The variable dielectric model eliminates the overprediction of salt bridges in the uniform dielectric model.

positive energy gap between the predicted structure and minimized native structures indicate that convergence is not, in fact, achieved using our current sampling algorithms.

While single side chain prediction tests are relatively straightforward to execute, it is far from trivial to attain robust prediction of experimental side chain geometries, particularly as the side chain becomes more solvent exposed. When the side chain is buried in the interior of the protein, geometrical constraints leave few alternatives with regard to configuring the side chain in a manner that is compatible with the remainder of the protein, which is kept rigid in the prediction. However, as the degree of solvent exposure increases, the number of plausible alternative conformations also increases. For example, a specific solvent exposed lysine can often form either a salt bridge or a charge-neutral hydrogen bond or remain free in solution, interacting closely with no other protein atoms. In many cases these configurations may be close in free energy and hence difficult for any energy model to discriminate; in other cases, the energy model may make large, systematic errors, incorrectly preferring one type of structure over another. Such solvent exposed cases provide a significant challenge to energy models, one that enables reliable assessment of the accuracy of the model for a wide variety of interactions.

The single side chain prediction results with the variable dielectric model are shown in Table 2 as well as a comparison with a variety of other possible dielectric models. We test our energy model on single side chain prediction of 11 polar and charged residues. The nonpolar residues have relatively small partial charges, and most of them are buried in the interior of the protein. The van der Waals interactions, instead of electrostatic interactions, are often the dominating forces for their conformations. Thus, nonpolar residues are less affected by the solvent model and hence are ignored in this study. We use the root-mean-square-deviation (RMSD) of all heavy side chain atoms as the accuracy measure (excluding the C^{β} atom which is largely fixed by backbone position). This measure accounts for the positions of the entire side chain and is more suitable for high-resolution comparison than the χ_1 and χ_{1+2} angles. 1.5 Å is chosen as the threshold for an accurate prediction. As Table 2 shows, compared with the default model (uniform dielectric of 1), the variable dielectric model improves all polar side chain accuracies, to varying degrees, except for arginine. The largest improvements come from lysine and two carboxylic acids predictions. The percentage of accurate predictions for lysine increases from 62.6% to 76.8%; this reduces the number of errors in lysine prediction from 74 to 46 or by a factor of 38%. For glutamate and aspartate, the accuracies increase from 67.4% and 83.9% to 73.6% and 87.8%, respectively. This is equivalent to the error reduction of 19% for glutamate and 24% for aspartate. The overall accuracy for these 2178 residues increases by a substantial amount, from 76.2% to 79.8%.

The single largest error eliminated by the variable dielectric model is the overstabilization of salt bridges. This overstabilization problem occurs on many other GB-type models and has been observed in various simulations.[46−49] In our single side chain predictions, a recurring scenario was that the solvent exposed lysines were often predicted to form a salt bridge instead of being free in solution as in the native structure. This clearly occurs because forming the salt bridges receives excessive stabilization energy in the energy model, as compared with being solvated in solution. The new variable dielectric model solves this problem. In Figure 1, we plot the distance distribution of lysine NZ atom and carboxylic O2 atom (glutamate and aspartate) in native structure and predicted structures. The NZ-O2 distances in native structures are relatively flat and show two maxima at 3.0 Å and 4.9 Å, which approximately correspond to the contact minimum and solvent-separated minimum in the

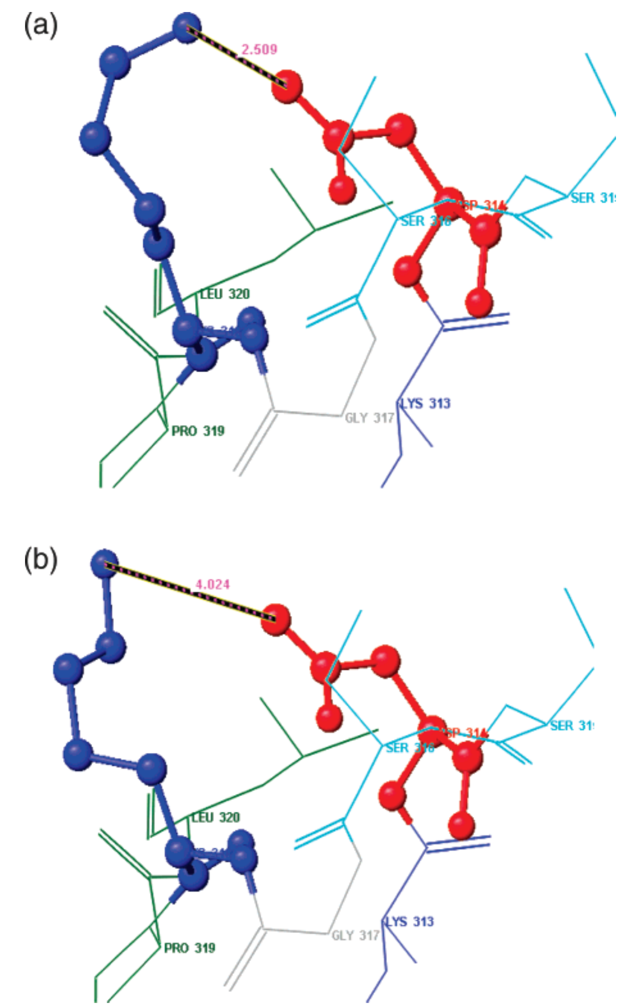


Figure 2. An example of using a variable dielectric model to improve the single side chain prediction. The single side chain prediction on 1ixh Lys318 with a uniform dielectric yields a structure with an erroneous salt bridge between Lys318 and Asp314. The RMSD is 2.82 Å (a). In the variable dielectric model, the lysine is correctly predicted with a RMSD of 0.74 Å (b).

potential of mean force (PMF) between NZ and O2 atoms. The predicted structures by the previously uniform dielectric energy model show a very high population at the distance around 2.8 Å. This overwhelmingly strong attraction between the NZ-O2 atoms leads to a collapse, which is held at a distance of 2.8 Å by the repulsion in the van der Waals term. The variable dielectric model prevents this collapse and greatly diminishes the population of salt bridges. Although not perfect, the NZ-O2 distribution shows two similar maxima as in the experimental structural data.

One specific example of the improvement the variable dielectric model produces is given Figure 2. The single side chain prediction on 1ixh Lys318 with uniform dielectric yields a bad structure with RMSD 2.82 Å. Lys318 and Asp314 form an erroneous salt bridge, and the distance between lysine NZ atom and carboxylic O2 atom is 2.51 Å. Using the variable dielectric model, the lysine is correctly predicted with a RMSD of 0.74 Å. The lysine ammonium group extends into the solvent, and the distance between lysine NZ atom and carboxylic O2 atom is 4.02 Å. Such a NZ-O2 distance is very commonly seen in the crystal structure and is favored because it allows for bridging waters between the two oppositely charged groups, maximizing the hydrogen bonding.

In addition to calculations employing a fixed dielectric of 1 and our new variable dielectric model, we also present results for fixed internal dielectric constants of 2 and 4 in Table 2. Values higher than 4 lead to significantly worse results and are not shown here. These results demonstrate that, as argued above, the use of any single alternative to unity as an internal dielectric does not improve the overall performance of side chain prediction. In particular, it is interesting to note that polar, uncharged side chains such as serine experience substantial degradation in performance as the single dielectric is increased. This is consistent with the hypothesis discussed above that for neutral−neutral hydrogen bond interactions the force field already has appropriate polarization included as a result of fitting to pure liquid-state simulations, and hence the use of a larger internal dielectric for these interactions is in effect double counting.

A complicating factor in using single side chain prediction to evaluate energy models is the dependence of prediction accuracy upon correct assignment of protonation states of the various side chains. For example, if a protonated histidine forms a salt bridge with a carboxylic acid in the native structure, a prediction performed with an unprotonated form of histidine may well prefer an alternative structure. To address this problem, we have developed a protonation state assignment methodology (referred to as Independent Cluster Decomposition Algorithm (ICDA), described in detail in ref 22), which already has been shown to provide substantial improvements in protonation state prediction given a crystal structure as a starting point. The ICDA infers the location of hydrogens in a high-resolution crystal structure based on the heavy atom positions obtained experimentally; it does not imply a complete search of conformational space, as the heavy atom positions are kept fixed during ICDA calculations. Hence, it is unsurprising that when hydrogens are assigned via the ICDA protocol, the native side chain conformer will in many cases be stabilized as compared to incorrect alternatives. However, the ICDA in and of itself is insufficient to produce perfectly accurate single side chain predictions; an accurate energy model is also essential; we illustrate this point by comparing side chain prediction results using the ICDA for both our old and new dielectric models.

In evaluating our new variable dielectric methodology, we perform comparisons with and without first assigning the protonation state by ICDA; as shown in Table 2, the combination of protonation state assignment and improved dielectric model yield substantially better results than either approach used by itself. For the 11 polar residues, the accuracy from the combination is 85.0%. The histidine accuracy increases from 59.8% to 86.4%. In this process, we simply take the structures generated by the ICDA assignment using the fixed dielectric of one and run the single side chain prediction with the variable dielectric model. It is possible to apply the new variable dielectric model into the ICDA algorithm; however, this would be involved with

Table 3. Classification of Prediction Errors for Four Charged Residues[a]

	sampling error (%)	energy error (%)	solution error (%)	hydrogen bond error (%)
Lys	10.9	90.1	43.9	56.1
Arg	23.3	76.7	12.1	87.9
Asp	19.4	80.6	4.0	96.0
Glu	35.3	64.7	21.2	79.8

[a] The results are based on the predictions with variable dielectric and ICDA assignment. The prediction error is defined as having a side chain heavy atom RMSD greater than 1.5 Å. If the energy of the predicted structure is higher than the directly minimized structure, then it is a sampling error, otherwise it is an energy error. All energy errors are further classified into two types. If both the native structure and the predicted structure do not form any hydrogen bond with other residues, then it is defined as a "solution error", otherwise it is defined as a "hydrogen bond error". The distance cutoff for a hydrogen bond is 3.1 Å between the acceptor and the donator, and no angle consideration is involved.

extensive reparametrization of ICDA algorithm, which we decided not to pursue at this point.

Having obtained such a significant improvement in the single side chain predictions, there is still a noticeable percentage of errors, especially for four charged residues: lysine, arginine, glutamate, and aspartate. We classify the errors into either energy or sampling errors, as shown in Table 3. If the energy of the final predicted structure is higher than the directly minimized native structure, then it indicates the sampling is not sufficient. If the energy of the final predicted structure is instead lower than the energy of the minimized native structure, then the error is attributed to the deficiency of energy model. Furthermore, we classify energy errors into two types: solution error and hydrogen bond error. If the side chain in both the experimental structure and the predicted structure does not form any hydrogen bond with the rest of the protein body, we call this type of misprediction a solution error. Otherwise, if the side chain is forming hydrogen bond(s) with the protein body in either the experimental structure or the predicted structure, then the error is designated as a hydrogen bond error. The hydrogen bond error represents a class of error that is more likely to be fixed by further improving the energy model because they are relatively easy to characterize. Table 3 shows that glutamate has the highest sampling error percentage of 35.3%, while arginine is second with 23.3%. This means that although the present sampling algorithm could produce accurate predictions for a majority of the cases, it stills needs to be improved for certain residues. Among the energy errors, the lysine has a very large percentage of solution errors in this study, at 43.9%. This of course is due to the fact that the lysine tends to be fully solvated in the solution. In contrast, most of energy errors of the aspartate are hydrogen bond errors. The characterization and correction of this type of error should have a high priority in order to further improve our energy model.

B. Loop Prediction. We apply the new variable dielectric model on a set of loop predictions ranging from 6, 8, 10, and 13 residue of loop length. We use the two-stage sampling in Jacobson et al.[2] paper for 6, 8, and 10 residue loops and a more powerful yet expensive multistage sampling algorithm for 13 residue loops.[4] The greatly improved accuracies are obtained on all length scales as Table 4 shows (the detailed results are in the Supporting Information) when we compare the variable dielectric model and the uniform dielectric model. The average loop backbone RMSDs (superimposing the rest of the protein) for 6, 8, 10, and 13 residue loops decrease from 0.48 Å, 0.84 Å, 1.27 Å and 2.73 Å to 0.40 Å, 0.79 Å, 0.73 Å, and 1.62 Å, respectively. In ref 4, we introduced a hydrophobic term into the SGB/NP model, which greatly improved the accuracy of long loop predictions. We attributed the success to the correction of absent hydrophobic interaction in the SGB/NP model, which is more prominent in the long loop prediction. Given the substantial advantage of a hydrophobic term on the SGB/NP model, it is important to verify whether it is compatible with the variable dielectric model. Table 4 shows that using the hydrophobic term improves the accuracy of loop prediction on both the variable dielectric model and the original SGB/NP model. The combination of both the variable dielectric model and the hydrophobic term yields the best accuracies of loop predictions with average backbone RMSDs of 0.41 Å, 0.74 Å, 0.76 Å, and 1.08 Å for 6, 8, 10, and 13 residue loops, respectively.

We define the energy gap (EGAP) as the energy of the predicted structure minus the energy of the directly minimized native structure. With the assumption that the native structure well represents the global minimum on the free energy surface, an ideal prediction should yield a reasonably good structure with a zero or slightly negative energy gap. A large negative energy gap with an incorrect structure indicates the energy function is flawed and thus has to be improved. A positive energy gap implies that the sampling is not sufficient; the status of the energy model for a test case of this type is unclear, although in practice such cases usually minimize to the native structure if one can locate it. Since our sampling method is a multiple stage process guided by the energy function, the energy function will bias the sampling to its favorable conformational space. A good energy function could bring the sampling region closer and closer to the native structure and finally find a nativelike structure, while a bad energy function would fail to do that. This difference is more prominent for long loops since sampling is more challenging in these cases. For example, there are a number of sampling errors in the predictions of 13 residue loops using the original SGB/NP model, while other improved energy models (hydrophobic term, variable dielectric, or a combination of both) eliminate the sampling errors, although the same sampling protocol is used (Table 4 shows the average energy gap. See the Supporting Information for detailed information.).

As Table 4 shows, the improvement due to the hydrophobic term when using the variable dielectric model is not as large as the improvement it provided with the original SGB/NP model. For example, when the hydrophobic term is introduced into the SGB/NP model, the RMSD for 13 residue loops decreases from 2.73 Å to 1.29 Å, while the combination of the hydrophobic term with the variable dielectric model reduces the RMSD from 1.62 Å to 1.08 Å. This is because sometimes both the hydrophobic term and the variable dielectric model fix the same problematic cases

Table 4. Average RMSDs and Energy Gaps for the Loop Prediction on 6, 8, 10, and 13 Residue Loops[a]

	uniform dielectric		variable dielectric		uniform dielectric + hydrophobic		variable dielectric + hydrophobic		variable dielectric + optHydrophobic	
	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP
6 residue	0.48	−4.09	0.40	−2.56	0.46	−4.09	0.41	−3.30	0.39	−3.50
8 residue	0.84	−6.48	0.79	−4.45	0.76	−7.50	0.74	−5.71	0.68	−5.09
10 residue	1.27	−4.96	0.73	−0.77	1.05	−4.38	0.76	−3.29	0.80	−6.23
13 residue	2.73	0.00	1.62	−1.17	1.29	−8.90	1.08	−3.65	1.00	−7.21

[a] The RMSD is the loop backbone RMSD while superimposing the rest of the protein. The energy gap (EGAP) is the energy of the predicted structure minus the energy of the directly minimized native structure. The units for RMSD and energy gaps are Å and kcal/mol, respectively. The first two columns show the results with a uniform dielectric model and a variable dielectric model. The next two columns show the results when these two models are combined with the hydrophobic term. The last column shows the results of our optimization of the hydrophobic term on the variable dielectric model by taking lysines out of the hydrophobic term. Hydrophobic and optHydrophobic represent the original hydrophobic term and the optimized hydrophobic term, respectively.

in the original SGB/NP model. They treat different physical phenomena: the hydrophobic term compensates for the absent hydrophobic interactions and variable dielectric screens for excessively strong charged interactions. However, they can still lead to duplicate effects in terms of generating the delicate balance among various forces that determinate the loop geometry. For example, the reduction of polar (electrostatic) interactions has somewhat similar effects as enhancing the nonpolar (hydrophobic) contributions. Thus sometimes the combination of the hydrophobic term and the variable dielectric does not work as well as either of them works alone. It should be possible to reoptimize the hydrophobic term with the new variable dielectric model to obtain better performance. In a preliminary effort, we exclude the lysine atoms from the hydrophobic energy term, which was originally defined as all heavy atoms with a partial charge less than 0.25 (absolute value) and thus contained some lysine atoms. The results are shown in the last column of Table 4. The average backbone RMSDs for 6, 8, 10, and 13 residue loops are further reduced down to 0.39 Å, 0.68 Å, 0.80 Å, and 1.00 Å, respectively. However, extensive reparametrization would require significant effort and is beyond the scope of this study.

C. Speed Comparison with the Previous Version. We compare the computational efficiency between the latest PLOP version and the version used in the previous publication[4,23] on a variety of tests. The first set of tests is the minimization of 35 13-residue loops. The minimizations start from the native structures and are converged until the norm of the gradient is below 0.001 kcal/mol/Å. We perform the minimization using both vacuum and generalized Born solvation conditions. The minimization using solvent conditions involves a self-consistent procedure in which the Born alphas are held fixed during the course of each minimization, then updated prior to the subsequent minimization, and so on until the energy ceases to decrease by more than 1 kcal/mol over one course of minimization. The 35 loop minimizations in vacuum are on average 3.1 times faster with the optimized code; all other variables such as processor and compiler are kept constant. This speedup comes from the removal of conditionals and function calls from the inner loops of energy and gradient evaluations as well as the reduction of short-range cutoffs due to the dipole-based cutoffs. For the minimization in the solvent, we separate the time spent on the minimization itself and the update of the Born alphas, as the latter requires significantly more time. The speedup factor for the minimization itself and for the update of the Born alphas are 6.5 and 2.6, respectively. The additional speedup factor for the minimization relative to the vacuum mainly comes from the elimination of SGB correction terms.

Table 5. Computational Costs for Loop Predictions[a]

	CPU time (h)			
	6 residue	8 residue	10 residue	13 residue
mean	4.7	14.4	91.0	333.6
median	3.0	11.3	85.4	277.7
min	0.5	3.1	21.2	87.9
max	71.2	77.6	198.9	1126.4

[a] The CPU time refers to the cumulative time counted as if on a single processor.

With the significant acceleration of minimizations and SGB calculations, the speed of loop predictions is also increased substantially. Single loop predictions for the 65 8-residue loops becomes, on average, 4.5 times faster. However, the multistage sampling protocol also involves steps of constrained loop buildup, which limits loop buildup within a certain distance of a given structure. This process often takes a longer time than the unconstrained buildup, because the effective resolution to sample the backbone library has to be reduced gradually to generate enough number of loop candidates that meet the distance constraint. Sometimes the buildup stage takes a significant percentage of the total time expense. The actual speedup of the full loop prediction is therefore smaller than the speedup of a single PLOP run without any constraint on the loop buildup. Table 5 shows the computational cost of our loop predictions. The average time cost of a 13-residue loop is 13.9 CPU days. Compared with the results in the previous paper,[4] where the average time for 13 residue loops is 31.4 CPU days, the speedup factor is 2.3. Since the multistage loop prediction protocol is highly parallel, the prediction of a 13-residue loop on a midsize cluster of around 32 nodes usually takes 1−2 days.

VI. Discussion and Conclusion

The results in the last column of Table 2 (variable dielectric plus ICDA model) represents a very substantial improvement in the accuracy of single side chain prediction as compared to our previous results, in the third column of Table 2. Comparison with the work of others is difficult because most papers do not report single side chain prediction results, and because those few that do typically do not incorporate the crystal environment into their predictions, making a fair comparison of the energy models problematic. Table 3 shows that the performance of the energy model is in fact substantially better than what is implied in the most conservative interpretation of the Table 2 data. That is, a nontrivial fraction of the errors reported in Table 2 are due to either sampling errors (which presumably could be fixed by the application of greater computational effort) or (primarily in the case of lysine) the inability to discriminate alternative structures in solution, a task that probably requires a considerably more accurate energy function than discriminating between hydrogen-bonded and non-hydrogen-bonded structures. It is quite possible that these alternative solution structures are very close in energy and hence both well populated in the native state of the protein; furthermore, for many practical applications, it may not matter very much which solution structures are used in the model.

Examination of the remaining hydrogen bond errors suggests further directions for improvement. Preliminary analysis indicates that residual errors in the molecular mechanics force field (such as torsional parameters) make significant contributions to the errors. Improving the force field will require performing suitable high level quantum chemical calculations, in conjunction with the side chain calculations presented here; work along these lines is currently in progress. Overall, it appears as though the goal of developing a robust, accurate side chain prediction method is within reach in the next few years.

The loop prediction results represent a nontrivial test of the side chain optimization effort, with no further adjustable parameters involved. The significant improvements observed in Table 4 validate the methodology independently and suggest that it is useful to adopt the protocol of revising the solvation model (and force field, if necessary) to fit the single side chain prediction data, driven by physical chemistry based models and calculations. The loop prediction results, to our knowledge, represent the best results reported in the literature to date. In combination, the loop and side chain prediction accuracy and robustness should now be sufficient for high-resolution tasks such as structural refinement of homology models, with the goal of enabling structure based drug design starting from the resulting refined active sites. There still remains a very significant challenge to put in place a global sampling algorithm that will enable progress in homology refinement, even assuming that the energy function is of the quality hypothesized above. In particular, loop and side chain prediction involve relatively localized structural changes, whereas homology models typically have errors distributed throughout the structure and hence a purely localized search strategy may not work. The speed improvements in the sampling algorithm reported here would be helpful in adapting our conformational search strategy to address this and other more realistic, delocalized problems.

The success of the variable dielectric model in improving side chain prediction, while useful in and of itself, particularly because of the simplicity of implementation, also suggests that an accurate treatment of polarization is important in achieving quantitative results in biomolecular modeling. A great deal of success has been obtained with fixed charge force fields which incorporate an average polarization as discussed above, but as one imposes more demanding criteria upon the model in terms of accuracy and robustness at the level of microscopic detail, it is unsurprising that this relatively crude approximation is increasingly problematic. The variable dielectric model is itself a crude approximation but does represent an improvement over taking no account of the difference in environment represented by charged (as opposed to neutral) polar groups. A model explicitly incorporating polarization presumably would be better still, but, as mentioned above, this is significantly work to develop and test. Nevertheless, the results presented here should provide strong motivation for efforts in this direction.

Acknowledgment. This work was supported in part by a grant to R.A.F. from the NIH (GM-52018). M.R.S. was supported by an NIH Ruth L. Kirschstein NRSA fellowship. The authors thank Barry Honig and Matt Jacobson for numerous useful discussions.

Supporting Information Available: Detailed loop prediction results. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597−608.

(2) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351−367.

(3) Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673−11680.

(4) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins* **2006**, *65*, 438−452.

(5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc*. **1996**, *118*, 11225−11236.

(6) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J. *J. Phys. Chem. B* **2001**, *105*, 6474.

(7) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem*. **2001**, *23* (5), 517−529.

(8) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(9) Monnigmann, M.; Floudas, C. A. *Proteins* **2005**, *61*, 748−762.

(10) Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. *Proteins* **2004**, *55*, 656−677.

(11) DePristo, M. A.; de Bakker, P.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51*, 41−55.

(12) de Bakker, P.; Depristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins* **2003**, *51*, 21−40.

(13) Xiang, Z. X.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432−7437.

(14) Fiser, A.; Do, R.; Sali, A. *Protein Sci.* **2000**, *9*, 1753−1773.

(15) Xiang, Z. X.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421−430.

(16) Xiang, Z.; Steinbach, P. J.; Jacobson, M. P.; Friesner, R. A.; Honig, B. *Proteins* **2007**, *66* (4), 814−823.

(17) Dunbrack, R. L. J.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543−574.

(18) Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V. *J. Comput. Chem.* **2004**, *25*, 712−724.

(19) Liang, S.; Grishin, N. V. *Protein Sci.* **2002**, *11*, 322−331.

(20) Desmet, J.; Spriet, J.; Lasters, I. *Proteins* **2002**, *48* (1), 31−43.

(21) Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539−542.

(22) Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S.; Friesner, R. A. *Proteins* **2007**, *66*, 824−837.

(23) Zhu, K.; Shirts, M. R.; Friesner, R. A.; Jacobson, M. P. *J. Chem. Theory Comput.* **2007**, *3*, 640−648.

(24) Nielsen, J. E.; Vriend, G. *Proteins* **2001**, *43*, 403−412.

(25) Yang, A.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins* **1993**, *15*, 252−265.

(26) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35* (24), 7819−7833.

(27) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44* (4), 400−417.

(28) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373−17387.

(29) Gilson, M. K.; Honig, B. *Biopolymers* **1986**, *25*, 2097−2119.

(30) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118* (35), 8452−8458.

(31) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem. B* **1994**, *98*, 1978−1988.

(32) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671−15682.

(33) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(34) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515−1531.

(35) Patel, S. A.; Brooks, C. L. *Mol. Simul.* **2006**, *32* (3−4), 231−249.

(36) Ponder, J. W.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27−85.

(37) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(38) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672−679.

(39) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. J. *J. Mol. Liq.* **1995**, *65−66*, 31.

(40) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101* (7), 6141.

(41) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237.

(42) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23* (16), 1497−1506.

(43) Maple, J. R.; Cao, Y. X.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1* (4), 694−715.

(44) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W. *J. Chem. Phys.* **2007**, *126*, 124114.

(45) Yu, Z.; Jacobson, M. P.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **2004**, *108*, 6643−6654.

(46) Zhou, R. *Proteins* **2003**, *53*, 148−161.

(47) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12777−12782.

(48) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56* (2), 310−321.

(49) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2005**, *2* (1), 115−127.

CT700166F

JCTC Journal of Chemical Theory and Computation

Group Polarizability Model for Molecular Mechanics Energy Functions

Kim Palmo[†] and Samuel Krimm*

Biophysics Research Division and Department of Physics, 930 North University Avenue, University of Michigan, Ann Arbor, Michigan 48109

Received May 1, 2007

Abstract: A polarization model for molecular mechanics energy functions is developed that is based on a local group paradigm, namely the polarizability of a rigid substructure of covalently connected atoms. Axes at a “diffuse” site within the group define an anisotropic local group polarizability as well as hyperpolarizability. The theoretical basis for this model is presented, and its performance is described through applications to water, alkanes, and *N*-methylacetamide. The excellent agreement with quantum mechanical electric potentials and molecular polarizabilities indicates that this model must be considered an important candidate for the inclusion of polarization into such force fields. The ab initio-based spectroscopically determined force field (SDFF) protocol for the calculation of parameters assures that, in addition to structures and energies, forces will be accurately modeled.

1. Introduction

Current standard force fields for computer simulation of macromolecular properties are mostly based on pairwise-additive interatomic interactions whose electrostatic component utilizes fixed atomic charges. Recent efforts to improve the quality of such force fields have focused on the need to include many-body, i.e., polarizability, effects in providing a more physical description of electrostatic interactions.[1] Although it is well-known that polarization is not an atomic but rather a group property[2,3] (i.e., that it is the overall electron cloud that is being deformed by the electric field at a site), various atom-based models for introducing polarizablity have been proposed.[4] Induced atomic dipoles, iterated to self-consistency, are perhaps the most commonly used method. This model has reached a high level of sophistication, notably also with respect to the difficult issue of intramolecular polarization.[5] Other methods, too, are currently being evaluated and further developed. Particular effort is devoted to a model based on the classical Drude oscillator.[6] The fluctuating charge model, which is based on electronegativity equalization, has received a lot of attention as well.[7] In the context of our spectroscopically determined force field (SDFF) methodology for developing molecular mechanics (MM) energy functions,[8] we described polarization using an iterative model consisting of induced atomic charges and anisotropically induced atomic dipoles.[9] However, such detailed approaches may be less than optimal from an efficiency point of view.

Efficiency is also an issue in another connection. In a previous paper,[10] we showed that the induction energy of a molecular system can be accurately computed by a non-iterative procedure suitable for MM calculations. Using waterlike test molecules, it was concluded that well over 90% (and always less than 100%) of the fully iterated induction energy will be retained at densities up to and beyond that of liquid water. This has also been confirmed in later tests. For example, using snapshots from a molecular dynamics (MD) simulation of a droplet of 216 water molecules, the one-step model yielded ~95% of the fully iterated induction energy. However, even with such a one-step model, inclusion of polarization is still very expensive because the electric field has to be calculated (once) at each polarizable site, and the induced quantities interact with each other. Reducing the number of polarizable sites is another way to increase efficiency, and since polarization is a group property, we

* Corresponding author phone: (734)763-8081; e-mail: skrimm@umich.edu.

[†] Present address: D.E. Shaw Research LLC, New York, NY 10036.

10.1021/ct700102g CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/14/2007

explore here the possibility of representing the polarizability of a relatively rigid substructure by only a few polarizable sites.

Before doing so, it is important to note that, although undoubtedly relevant, inclusion of polarization in the traditional manner may not be enough if the goal is to realistically reproduce variable charge distributions for intramolecular as well as intermolecular interactions. Although it has been speculated[7] that the fluctuating charge model may be able to account for geometry dependent charge variations, this may not be the case because of the atom-based approximations to the full electronic charge clouds in current MM force fields. Our studies indicate that the inclusion of geometry-dependent charges, implemented through properly balanced[11] charge fluxes,[12] may be even more important than polarizability in providing needed physical accuracy: such inclusion reproduced the opening of the water angle on going from the isolated molecule to the liquid, reproduced the ψ peptide torsion angle with only a single threefold (almost zero barrier) Fourier term, and reproduced the quantum mechanical (QM) MD φ,ψ map of a dipeptide analog.[13] We therefore believe that it is more useful to treat conventional polarization (through the electric field) separately from geometry induced readjustments in the electronic structure. (The latter effect can be very accurately accounted for in our SDFF because the valence charge flux terms can include contributions from all internal coordinates.)

Based on the induced dipole model, a scheme for defining group polarizabilities has been implemented in our modeling package (SPEAR, to be published), and parameters have been determined for a few test compounds. In this paper, we describe the methodology used, discuss some computational considerations, and apply the model to water, alkanes, and *N*-methylacetamide (NMA). In accordance with the SDFF optimization protocol[8] and generally accepted procedure,[2] the reproduction of QM electric potentials and molecular polarizabilities is used to provide a reliable test of the accuracy of the polarizability model.

2. Polarizability Model

2.1. Anisotropic Local Group Polarizability. In the new model, polarization groups are formed by sets of covalently connected atoms in molecular fragments. The location of a group is the weighted average of the positions of the atoms of the group, and the local polarizability axes are given by vectors between subsets of atoms in the group. The weighting factors and the axis vectors, and of course the polarizability parameters themselves, are specified for each group template in the electrostatic parameter file. Most other MM polarizability models are based on some implementation of isotropic atomic-level polarizability. Anisotropy in the molecular polarizability is then only produced in actual calculations by letting the induced charges or dipoles iteratively polarize one another to self-consistency. In our model, on the other hand, any anisotropy is explicitly included from the start. Up to three principal polarizability directions may be defined for a site. If the polarizability elements of a site i are α_{i1}, α_{i2}, α_{i3}, then the polarizability tensor for the site can be written as

$$\boldsymbol{\alpha}_i = \mathrm{diag}(\alpha_{i1},\alpha_{i2},\alpha_{i3}) \tag{1}$$

Under the influence of an electric field, $\mathbf{E}_i$, a dipole is then induced at the site according to

$$\boldsymbol{\mu}_i = \boldsymbol{\alpha}_i\mathbf{E}_i = \sum_{n=1}^{3} \alpha_{in}(\mathbf{e}_{in}\cdot\mathbf{E}_i)\mathbf{e}_{in} \tag{2}$$

where the $\mathbf{e}_{in}$ are unit vectors in the chosen directions. In practice, at most two such vectors need to be explicitly defined, the third direction being perpendicular to the plane formed by the first two vectors. The field includes balanced contributions from all permanent charges (and multipoles, if any) outside the group and any external electric field but not the field from induced quantities. Since SDFF atomic charges consist of bond charge increments (BCIs),[8] and charges inside a group are not allowed to contribute to the field at that site, charge balance is accomplished by excluding from the electric field calculation all BCIs that have at least one end inside the group.

Simplifications are often possible and should be utilized. In the case of cylindrical polarizability, only one vector $\mathbf{e}_i$ needs to be defined, since the polarizability is the same in all directions perpendicular to that vector. For totally isotropic polarizability, of course, no vectors need to be defined. Thus, in the general case, there are N_i vectors and N_i+1 polarizabilities given for a site i, with $N_i = 0$, 1, or 2. This notation can be used to establish a compact algorithm for calculating $\boldsymbol{\mu}_i$, applicable in all cases, as seen in the following way.

If $N_i = 0$ (isotropic polarizability), $\boldsymbol{\mu}_i$ is parallel to the field, i.e.,

$$\boldsymbol{\mu}_i = \alpha_{i1}\mathbf{E}_i \tag{3}$$

If $N_i = 1$ (cylindrical polarizability), the electric field component perpendicular to the given vector is obtained by subtracting the field component in the direction of the vector from the total field, giving the induced dipole as

$$\boldsymbol{\mu}_i = \alpha_{i1}(\mathbf{e}_{i1}\cdot\mathbf{E}_i)\mathbf{e}_{i1} + \alpha_{i2}[\mathbf{E}_i - (\mathbf{e}_{i1}\cdot\mathbf{E}_i)\mathbf{e}_{i1}] \tag{4}$$

Similarly, if $N_i = 2$ (totally anisotropic polarizability), then the field component in the direction of the third unit vector is obtained by subtracting the field components in the first two directions from the total field, i.e.,

$$\boldsymbol{\mu}_i = \alpha_{i1}(\mathbf{e}_{i1}\cdot\mathbf{E}_i)\mathbf{e}_{i1} + \alpha_{i2}(\mathbf{e}_{i2}\cdot\mathbf{E}_i)\mathbf{e}_{i2} + \alpha_{i3}[\mathbf{E}_i - (\mathbf{e}_{i1}\cdot\mathbf{E}_i)\mathbf{e}_{i1} - (\mathbf{e}_{i2}\cdot\mathbf{E}_i)\mathbf{e}_{i2}] \tag{5}$$

This can also be written as

$$\boldsymbol{\mu}_i = \sum_{n=1}^{N_i} \alpha_{in}(\mathbf{e}_{in}\cdot\mathbf{E}_i)\mathbf{e}_{in} + \alpha_{i,N_i+1}[\mathbf{E}_i - \sum_{n=1}^{N_i} (\mathbf{e}_{in}\cdot\mathbf{E}_i)\mathbf{e}_{in}] \tag{6}$$

and, rearranging the terms, we finally get

$$\boldsymbol{\mu}_i = \alpha_{i,N_i+1}\mathbf{E}_i + \sum_{n=1}^{N_i} (\alpha_{in} - \alpha_{i,N_i+1})(\mathbf{e}_{in}\cdot\mathbf{E}_i)\mathbf{e}_{in} \tag{7}$$

Although eq 7 was derived assuming totally anisotropic polarizability, it also holds in the other cases since setting

$N_i = 0$ yields eq 3 and setting $N_i = 1$ yields eq 4. Thus, eq 7 can be used to compute the induced dipole at any polarizable site.

In the one-step model, the first derivatives of the induced dipoles with respect to the atomic coordinates are explicitly needed to compute atomic forces. The derivatives are given by

$$\frac{\partial \boldsymbol{\mu}_i}{\partial x_{km}} = \alpha_{i,N_i+1}\frac{\partial \mathbf{E}_i}{\partial x_{km}} + \sum_{n=1}^{N_i}(\alpha_{in} - \alpha_{i,N_i+1})\left[\left(\frac{\partial \mathbf{E}_i}{\partial x_{km}}\cdot\mathbf{e}_{in}\right)\mathbf{e}_{in} + \left(\mathbf{E}_i\cdot\frac{\partial \mathbf{e}_i}{\partial x_{km}}\right)\mathbf{e}_{in} + (\mathbf{E}_i\cdot\mathbf{e}_{in})\frac{\partial \mathbf{e}_i}{\partial x_{km}}\right] \quad (8)$$

where x_{km}, $m = 1, 2, 3$, are the Cartesian coordinates of atom k. The major computational effort involved here is to calculate the electric field and its derivatives at each site.

2.2. Hyperpolarizability. With anisotropy in the polarizability explicitly taken into account, it is easy to include (limited) hyperpolarizability in the calculation of the induced dipoles. The most visible effect of hyperpolarizability is that an induced dipole may change significantly in magnitude when the electric field is reversed. In a water molecule, for example (using ab initio MP2/6-31++G**), the induced dipole moment is 0.65 D with an electric field of 0.04 au (∼2V/Å) pointing along the bisector from the oxygen toward the hydrogens but 0.69 D if the field is reversed. For compatibility with the linear polarizability of eq 2, and for convenient inclusion in eq 7, the following simplified forms can be used for the induced dipoles due to hyperpolarizability

$$\boldsymbol{\mu}_i^{\beta} = \sum_{n=1}^{N_i}\beta_{in}(\mathbf{e}_{in}\cdot\mathbf{E}_i)^2\mathbf{e}_{in}, \quad \text{quadratic} \quad (9)$$

$$\boldsymbol{\mu}_i^{\gamma} = \sum_{n=1}^{N_i}\gamma_{in}(\mathbf{e}_{in}\cdot\mathbf{E}_i)^3\mathbf{e}_{in}, \quad \text{cubic} \quad (10)$$

etc.

where β_{in} and γ_{in} are (scalar) quadratic and cubic polarizability parameters, respectively. Thus, hyperpolarizability is here limited to diagonal terms in the directions explicitly defined by the given N_i vectors, but inclusion of these nonlinear terms requires very little additional computational effort. For example, adding quadratic hyperpolarizability to the induced dipoles of eq 7 yields

$$\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^{\beta} = \alpha_{i,Ni+1}\mathbf{E}_i + \sum_{n=1}^{N_i}[\alpha_{in} - \alpha_{i,N_i+1} + \beta_{in}(\mathbf{e}_{in}\cdot\mathbf{E}_i)](\mathbf{e}_{in}\cdot\mathbf{E}_i)\mathbf{e}_{in} \quad (11)$$

Inclusion of hyperpolarizability in the first derivatives of the induced dipoles is likewise straightforward, with the addition of a few simple terms to the derivatives of the linearly induced dipoles.

2.3. One-Step vs Iterated Dipoles. Induced dipoles have somewhat nonintuitive properties and are not directly comparable to permanent dipoles. Similarly, one-step induced dipoles are not directly comparable to iterated ones. Iterated dipoles yield the induction energy by interactions with the electric field from permanent quantities only.[4,10] This is because the dipole−dipole interaction energy is implicitly included through the iterative calculation of the dipoles. In the one-step model, however, the induced dipoles behave more like the permanent ones, and the induction energy depends on the dipole−dipole interactions.

Using the electric field from the permanent quantities only, eqs 7 or 11 yields the induced dipoles needed in our noniterative polarization scheme to calculate the induction energy of a system. However, these dipoles cannot be directly used for nonpotential-energy calculations or comparisons. For such procedures, the $\boldsymbol{\mu}_i$ must be replaced by the *energy equivalent* (ee) induced dipoles

$$\boldsymbol{\mu}_i^{\mathrm{ee}} = \boldsymbol{\mu}_i + \boldsymbol{\alpha}_i\mathbf{E}_i^{(1)} \quad (12)$$

As indicated in eq 12, these are obtained by adding to the $\boldsymbol{\mu}_i$ at each site the incremental dipoles induced by the electric field, $\mathbf{E}_i^{(1)}$, from the primarily induced dipoles. This corrects the induced dipoles for their mutual interactions (which are explicitly included in the calculation of the potential energy) and is equivalent to making one iteration cycle (after the zeroth one). An important example of a nonpotential-energy property that depends directly on the induced dipoles, and where eq 12 has to be used, is the electric potential.

2.4. Diffuse Size. Because each polarizable site in the group polarizability model represents a substructure of some finite size, the induced dipoles may be defined to be ‘diffuse’ instead of singular points. This is done using a technique originating from fluid dynamics:[14,15] the distance r from the induced dipole to a point is replaced by the buffered distance

$$r_s = \sqrt{r^2 + s^2} \quad (13)$$

where s is the buffer size. This makes the dipole behave like a charge distribution[16] at close distances, with almost no computational overhead. Other models[17,18] are more expensive.

The electric potential U, generated by a dipole $\boldsymbol{\mu}$, is then calculated as

$$U = \frac{1}{4\pi\epsilon_0}\frac{\boldsymbol{\mu}\cdot\mathbf{r}}{r_s^{\,3}} \quad (14)$$

where $\mathbf{r}$ is an ordinary vector from the dipole to the point. The electric field from the dipole is determined by

$$\mathbf{E} = -\nabla U \quad (15)$$

in the usual way. The case of two interacting dipoles is equivalent to having one dipole interact with the field from the other. If the dipoles have sizes s_1 and s_2, then the square of the combined buffer size is taken as $s^2 = s_1^{\,2} + s_2^{\,2}$.

2.5. Group−Group Vector. A vector from a polarization group X to a group Y can be written as

$$\mathbf{r}_{xy} = \sum_{i=1}^{n_y} b_i\ \mathbf{y}_i - \sum_{i=1}^{n_x} a_i\mathbf{x}_i \quad (16)$$

where n_x and n_y are the number of atoms in each group, a_i and b_i are atomic weighting factors, and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$

and $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})$ are the Cartesian coordinates of the atoms in X and Y, respectively. Accordingly, the true distance between the sites is

$$r_{xy} = \sqrt{\mathbf{r}_{xy}\cdot\mathbf{r}_{xy}} \tag{17}$$

and the buffered distance is

$$r_{xy,s} = \sqrt{{r_{xy}}^2 + {s_x}^2 + {s_y}^2} \tag{18}$$

where s_x and s_y are the diffuse sizes of X and Y, respectively. The derivatives of eqs 16−18 are easily shown to be

$$\frac{\partial \mathbf{r}_{xy}}{\partial x_{km}} = -a_k\mathbf{e}_m, \frac{\partial \mathbf{r}_{xy}}{\partial y_{km}} = b_k\mathbf{e}_m \tag{19}$$

$$\frac{\partial r_{xy}}{\partial x_{km}} = \frac{\mathbf{r}_{xy}}{r_{xy}}\cdot\frac{\partial \mathbf{r}_{xy}}{\partial x_{km}}, \frac{\partial r_{xy}}{\partial y_{km}} = \frac{\mathbf{r}_{xy}}{r_{xy}}\cdot\frac{\partial \mathbf{r}_{xy}}{\partial y_{km}} \tag{20}$$

$$\frac{\partial r_{xy,s}}{\partial x_{km}} = \frac{\mathbf{r}_{xy}}{r_{xy,s}}\cdot\frac{\partial \mathbf{r}_{xy}}{\partial x_{km}}, \frac{\partial r_{xy,s}}{\partial y_{km}} = \frac{\mathbf{r}_{xy}}{r_{xy,s}}\cdot\frac{\partial \mathbf{r}_{xy}}{\partial y_{km}} \tag{21}$$

where $\mathbf{e}_m$ is a unit vector along the Cartesian axis m. Note that the use of a buffered distance does not add any terms to the derivatives.

Group−group vectors are also used to define the polarization directions inside a group. In this case, subgroups of atoms are formed as needed, and the polarization axes are given by unit vectors $\mathbf{e}_{xy} = \mathbf{r}_{xy}/r_{xy}$ (with no buffering). The derivatives of such a unit vector are

$$\frac{\partial \mathbf{e}_{xy}}{\partial x_{km}} = \left(r_{xy}\frac{\partial \mathbf{r}_{xy}}{\partial x_{km}} - \mathbf{r}_{xy}\frac{\partial r_{xy}}{\partial x_{km}}\right)\frac{1}{{r_{xy}}^2} \tag{22}$$

for atoms belonging to group X (and similarly for atoms of group Y).

2.6. Induction Energy. In the one-step polarization model, the induction energy, V^{ind}, consists of two terms

$$V^{\text{ind}} = -\frac{1}{2}\sum_i V_i + \frac{1}{4\pi\epsilon_0}\sum_{i<j} V_{ij} \tag{23a}$$

where i and j run over the polarizable sites, and

$$V_i = \boldsymbol{\mu}_i\cdot\mathbf{E}_i \tag{23b}$$

accounts for the interactions between the electric field and the induced dipoles as well as for the self-energy of the induced dipoles, and

$$V_{ij} = \frac{\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_j}{{r_{ij}}^3} - \frac{3(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_i)(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_j)}{{r_{ij}}^5} \tag{23c}$$

gives the dipole−dipole interaction energy. The induction energy can be efficiently computed in two stages. At each site i, the electric field $\mathbf{E}_i$ is first determined and used to compute $\boldsymbol{\mu}_i$ and V_i.

The $\boldsymbol{\mu}_i$ need to be stored, but the field can be discarded. In the second stage the V_{ij} are computed using the stored $\boldsymbol{\mu}_i$. Summing the V_i and V_{ij} according to eq 23a then yields the global induction energy V^{ind}.

2.7. First Derivatives. In molecular dynamics simulations, the first derivatives of the potential energy with respect to the Cartesian coordinates are needed for the calculation of atomic forces. Below we evaluate the derivatives of the two parts of V^{ind} and briefly indicate how the different terms that arise are processed.

The derivatives of V_i with respect to the atomic Cartesian coordinates are

$$\frac{\partial V_i}{\partial x_{km}} = \frac{\partial \boldsymbol{\mu}_i}{\partial x_{km}}\cdot\mathbf{E}_i + \boldsymbol{\mu}_i\cdot\frac{\partial \mathbf{E}_i}{\partial x_{km}} \tag{24}$$

The electric field $\mathbf{E}_i$ (from permanent quantities only) and its derivatives at each site are first calculated, from which the induced dipole and its derivatives as well as V_i and V_{ij} are obtained.

The electric field quantities can be discarded after use, but the $\boldsymbol{\mu}_i$ and $\partial\boldsymbol{\mu}_i/\partial x_{km}$ need to be stored to streamline the calculation of the $\partial V_{ij}/\partial x_{km}$ in the next step.

By writing eq 23c in the form

$${r_{ij}}^5 V_{ij} = {r_{ij}}^2\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_j - 3(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_i)(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_j) \tag{25}$$

the derivatives of V_{ij} are readily shown to be

$$\frac{\partial V_{ij}}{\partial x_{km}} = \frac{\partial r_{ij}}{\partial x_{km}}(-5{r_{ij}}^4 V_{ij} + 2r_{ij}\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_j)\frac{1}{{r_{ij}}^5} - \frac{\partial \mathbf{r}_{ij}}{\partial x_{km}}\cdot[\boldsymbol{\mu}_i(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_j) + \boldsymbol{\mu}_j(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_i)]\frac{3}{{r_{ij}}^5} + \frac{\partial \boldsymbol{\mu}_i}{\partial x_{km}}\cdot[{r_{ij}}^2\boldsymbol{\mu}_j - 3\mathbf{r}_{ij}(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_j)]\frac{1}{{r_{ij}}^5} + \frac{\partial \boldsymbol{\mu}_j}{\partial x_{km}}\cdot[{r_{ij}}^2\boldsymbol{\mu}_i - 3\mathbf{r}_{ij}(\mathbf{r}_{ij}\cdot\boldsymbol{\mu}_i)]\frac{1}{{r_{ij}}^5} \tag{26}$$

which are now straightforward to compute since all the quantities involved are known or can easily be determined. We have not implemented analytical second derivatives of the induction energy, but numerical differentiation of eqs 24 and 26 is used to derive very accurate Hessians for the calculation of vibrational frequencies and SDFF valence parameters.

3. Computational Considerations

In MD simulations, the most efficient way to incorporate polarizability is through the extended Lagrangian procedure,[4] which can be used with virtually any polarization model. However, in cases where the potential energy has to be explicitly calculated, inclusion of polarization requires nontrivial computational resources for large systems. In iterative models, the major drawback is that the electric field from the induced dipoles has to be calculated many times at each site for a single configuration while the dipoles converge toward self-consistency.

This process is difficult to parallelize efficiently because, in every iteration, the field at each polarizable site is influenced by the field from the induced dipoles at all other sites, which increases the need for communication and synchronization among the processors. Also, if a cutoff distance for nonbonded interactions is used, then the cutoff is not absolute for iterative polarization interactions, since the induced dipoles at all sites *j* within the cutoff sphere of a particular site *i* depend on the induced dipoles within the *j* (or secondary) cutoff spheres, the dipoles in which depend on the dipoles within the tertiary cutoffs, and so on. This may lead to slow convergence. If the induced dipoles are updated on the fly, iterative polarization also does not result in uniform quality of the induced dipoles across the system, unless a very large number of iterations are made. The first sites to be updated in, say, the second iteration, will only be subject to the field from the dipoles of the zeroth and first iterations, while the last sites to be updated will be subject to the zeroth, first, and second iteration fields from most other sites. The detailed result then also depends on the particular order in which the polarizable sites are processed.

In our one-step model, on the other hand, only the field from the permanent quantities is used, so no communication between the processors is needed to compute the induced dipoles. The calculation of the induction energy can also be very efficiently parallelized in this model. And if a cutoff distance is used, the one-step induced dipole at a site is only affected by the permanent charges inside its own cutoff sphere. In fact, the induced dipole at a site can be calculated without computing any other induced dipoles in the system, although the induction energy of course involves all induced dipole−induced dipole interactions. This makes the noniterative model ideal for Monte Carlo calculations or for other computations that do not require forces such as the calculation of solvation energies.

However, when forces have to be computed, the use of the derivatives of the induced dipoles requires memory and CPU resources not needed in iterative models. Efficient parallelization is still possible by dividing the polarizable sites evenly among the nodes so that each one calculates, stores, and uses only its share of the $\partial\boldsymbol{\mu}_i/\partial x_{km}$. But the calculation of the force in eq 26 is an $O(N^2)$ operation because every $\partial\boldsymbol{\mu}_i/\partial x_{km}$ consists of $3N$ vectors interacting with vectors at M polarizable sites (and $M\sim N/3$). The corresponding operation in an iterative model is $O(N)$, although a large number of iterations may be required in order to obtain near-analytical quality of the forces. Using analytically exact forces as given by the noniterative procedure provides many advantages, such as complete absence of the polarization catastrophe (or any spurious overpolarization), excellent accuracy and stability, and, therefore, unbeatable energy conservation, but such forces come at a price.

4. Applications

To illustrate the group polarization methodology described above, we present its application to a few important model systems, viz., water, alkanes, and NMA. For all of these systems, the electrostatic parameters (atomic and off-atom charges and polarizability) were optimized by least-squares fitting to the quantum mechanical (QM) electric potential outside the molecules. Using the GAMESS[19] software package, the potentials were calculated on CHELPG grids and on planes through the molecules, while applying electric fields ranging from −0.04 to +0.04 au in three mutually perpendicular directions. The MM parameter optimization was then done with SPEAR. Only very high-level QM methods yield close to experimental values for the geometry, dipole moment, and molecular polarizability of water. In order to maintain compatibility with our QM data on protein and other units, however, we have used the same method (MP2/6-31++G**) to calculate the electric potentials of all compounds.

Table 1. Water Group Polarizability Parameters

location: on HOH bisector 0.19 Å from oxygen toward hydrogens		
buffer size: 0.45 Å		
polarizability elements:	α_1 0.982 Å^3	along bisector
	α_2 1.188 Å^3	perpendicular to bisector, in plane
	α_3 0.976 Å^3	perpendicular to molecular plane
	β_1 −0.180[a]	along bisector, positive direction from oxygen to hydrogens

[a] The unit of β_1E is Å^3.

4.1. Water. Not surprisingly, one polarization group located near the oxygen was found to be sufficient for water. Our newly developed electrostatic model for water (to be published) was used as a basis to which polarization was simply added. However, for water at least, the optimized polarizability parameters are not very sensitive to the details of the basic electrostatic model.

The group polarizability properties are listed in Table 1. The linear polarizability tensor is almost cylindrical with the axis being parallel to a line through the H atoms. A small quadratic hyperpolarizability (β_1=−0.18) was determined for the direction along the bisector. With this hyperpolarizability included, the previously mentioned QM calculated change in magnitude of the induced dipole moment on reversal of the electric field is now accurately reproduced: the group polarizability model gives 0.656D and 0.691D compared to the QM values 0.652D and 0.687D, respectively. The fit to the electric potential is also excellent, the weighted relative root-mean-square (wrrms) deviation[20] of the MM electrostatic potential from the QM one being 1.14% (1.16% without β_1). Even with a totally isotropic and linear model, the fit to the electric potential is quite good, i.e., 1.28% (yielding α=0.991 Å^3).

4.2. Alkanes. Ethane, propane, *t*-butane, *g*-butane, and isobutane were chosen as model molecules for alkanes. Methane was also included but does not share any parameters with the other molecules. We first designed an appropriate static charge model for them. In CH_4, a CH BCI of 0.125e was obtained (H positive). However, in the CH_3 and CH_2 groups the atomic charges on the H atoms come out negative in a fit to the QM electric potential. This probably arises from not capturing the asymmetric charge distribution at the carbon atom by a point charge representation. Such a counterintuitive result was accommodated by adding an extra negative charge site near the carbon atom. In both cases the site is optimally located 0.375 Å from the carbon along the symmetry axis toward the hydrogens and carries a charge of −0.211e. No off-atom charge was needed for the CH group. The CH BCIs then came out as 0.0897e for CH_2 and CH_3 and 0.0558e for CH (all H atoms positive). No significant values were obtained for CC BCIs.

Polarizable sites were then placed near the carbon atoms, and parameters were optimized. Cylindrical symmetry was assumed for the CH_3 and CH groups, but full anisotropy was

Table 2. Alkane Group Polarizability Parameters

CH_4	location: on carbon			
	buffer size: 0.50 Å			
	polarizability element:	α_1	2.036 Å^3	isotropic
CH_3	location: on CH_3 symmetry axis, 0.12 Å from carbon toward hydrogens			
	buffer size: 0.65 Å			
	polarizability elements:	α_1	2.006 Å^3	along CH_3 symmetry axis
		α_2	1.883 Å^3	perpendicular to CH_3 symmetry axis
CH_2	location: on HCH bisector, 0.07 Å from carbon toward hydrogens			
	buffer size: 0.45 Å			
	polarizability elements:	α_1	2.125 Å^3	along carbon backbone chain (perpendicular to HCH plane)
		α_2	1.563 Å^3	perpendicular to backbone chain
CH	location: on carbon			
	buffer size: 0.80 Å			
	polarizability elements:	α_1	1.247 Å^3	along CH bond
		α_2	1.957 Å^3	perpendicular to CH bond

Table 3. NMA Group Polarizability Parameters

peptide group (OCNH)			
location: mean of O, C, N positions			
buffer size: 1.00 Å			
polarizability elements:	α_1	4.216 Å^3	along C═O bond
	α_2	5.292 Å^3	in OCN plane, perpendicular to C═O bond
	α_3	1.777 Å^3	perpendicular to OCN plane
C methyl			
location: on CH_3 carbon			
buffer size: 0.80 Å			
polarizability elements:	α_1	1.892 Å^3	along CH_3 symmetry axis
	α_2	1.959 Å^3	perpendicular to CH_3 symmetry axis
N methyl			
location: on CH_3 symmetry axis, 0.11 Å from carbon toward hydrogens			
buffer size: 0.40 Å			
polarizability elements:	α_1	1.157 Å^3	along CH_3 symmetry axis
	α_2	1.764 Å^3	perpendicular to CH_3 symmetry axis

initially allowed for the CH_2 group, with principal axis 1 along the imagined carbon backbone chain (perpendicular to the HCH plane), axis 2 parallel to a line though the H atoms, and axis 3 along the HCH bisector. However, the last two directions turned out to have equal polarizabilities within the statistical uncertainty limits, and cylindrical symmetry was therefore also imposed on the CH_2 group (axis 3 being the symmetry axis). The optimized polarizabilities are given in Table 2. The group polarizability properties turned out to be very well behaved, with excellent transferability. For example, the parameters of the CH_3 group optimized to ethane only (2.006 and 1.827 Å^3) do not differ much from the final values optimized to all of the compounds simultaneously (2.006 and 1.883 Å^3). The wrrms deviation from the QM electric potentials was 4.85% for all alkanes except methane, for which it was 11.24%.

A somewhat surprising result was obtained involving the interactions of intramolecular induced dipoles. It turned out, namely, that it is not beneficial for the fit to the electric potential, nor for the induction energy, to have (1,2), (1,3), or even (1,4) dipoles interact. This remains true even if screening is applied. However, for intermolecular interactions over distances similar to (1,4) it is clearly necessary to have the dipoles interact in order to reproduce QM induction energies. Thus, there is a difference between polarization interactions over space and those through covalently bonded structures. A possible reason is that intramolecular electron clouds deformed by an electric field cannot be treated in terms of isolated induced dipoles that interact normally. Rather, we should assume that the covalent interaction in a bond causes two local deformations (such as at the carbon atoms in ethane) to adjust immediately to one another, and when we fit induced dipoles to the electric potential, the coupling then becomes implicitly included in the polarizability parameters. Our results show that the coupling is the same whenever sp^3 carbons are involved. On the other hand, when two separate molecules are in an electric field, the electron clouds are deformed independently. Their interaction has not been implicitly included anywhere and therefore has to be explicitly taken into account.

4.3. NMA. Using our previously determined fixed atomic charges for NMA[20] as a starting point, we explored group polarizability models with one, two, and three sites for the peptide group, in addition to the C and N methyl sites. Although the fit to the electric potential was slightly better with more polarizable sites, the differences were not very significant. Additional virtual charge sites on the oxygen and on the methyl groups were also explored, but they did not improve the fit to the electric potentials enough to warrant their presence. The wrrms deviations were typically ∼4%, ∼3%, and ∼2.5% for the best one-, two-, and three-site models, respectively. Because of the simplicity it represented, we therefore chose to pursue the one-site model further. The optimum location of this site is at the mean of the O, C, and N positions. The site is anisotropic with (optimized) principal axis 1 parallel to the C═O bond, axis 2 in the molecular (OCN) plane perpendicular to axis 1, and axis 3 perpendicular to the plane. These directions are not obvious. We initially expected axis 1 to be in the direction of the NMA dipole moment, whose direction is approximately from the O to H(N) and makes an angle of 16° with the C═O bond, but this is not optimal. The optimized polarizability parameters are listed in Table 3.

Similarly to what was found for the alkanes, it is not beneficial to have any of the polarizable sites in NMA interact. This simple model was found to provide an excellent electrical component for the optimization of van der Waals parameters to NMA dimer interaction energies (to be published).

4.4. Molecular Polarizabilities. In Table 4 we compare the molecular polarizabilities given by the group polarizability parameters to those given by QM (calculated with Gaussian 03[21] using the MP2/6-31++G** level and basis set). The group model systematically overestimates the molecular polarizabilities by ∼3% on the average. Thus, by scaling the parameters by ∼97%, the discrepancies could be significantly reduced. However, we have not done so, mainly for two reasons. First, the parameters were optimized to the electric potentials around the molecules. The values are therefore those that are likely to best reproduce electro-

Table 4. Molecular Polarizability Tensor Elements (in Å^3)[a]

		QM	MM
water	xx	0.95	0.98
	yy	1.16	1.19
	zz	0.96	0.98
methane	isotropic	2.00	2.04
ethane	xx	3.61	3.77
	yy	3.61	3.77
	zz	3.90	4.01
propane	xx	5.12	5.33
	yy	5.84	6.06
	zz	5.34	5.41
t-butane	xx	7.03	7.24
	yy	7.84	7.92
	zz	6.57	6.89
	xy	−0.38	−0.52
g-butane	xx	6.88	7.12
	yy	7.43	7.64
	zz	6.94	7.28
	xy	0.40	0.39
isobutane	xx	7.44	7.77
	yy	7.44	7.77
	zz	6.69	6.93
NMA	xx	8.08	8.44
	yy	7.49	7.84
	zz	5.41	5.50
	xy	−0.36	−0.32

[a] Calculated using QM optimized geometries (MP2/6-31++G**) of the molecules in standard orientations as given by Gaussian 03. Only tensor elements whose absolute values are greater than 0.1 Å^3 are given.

static interactions with other molecules. The slightly too large molecular polarizabilities may come about in the optimization because the group sites are buried inside the molecules, and are also buffered, so that it takes somewhat larger induced dipoles to reproduce the electric potential when the fields are applied. Second, we combine the group polarizability model with our noniterative polarization protocol and thereby already scale the polarization energy by ∼95%. Further reduction of this energy may therefore not be warranted.

Aside from the systematic deviation, the molecular polarizability elements given by the model are quite good. Their magnitudes are always in the same order as the QM values, and even small off-diagonal elements are satisfactorily reproduced.

5. Conclusions

Our SDFF efforts to produce more physically accurate MM energy functions[8] have focused on the need to reproduce maximally correct forces as well as structures and energies.[13] The incorporation of polarizability is a significant component of this goal, although we also have noted that a necessary ingredient in point charge force fields is the contribution, through the nonelectrostatic terms, of conformation-dependent charges through charge fluxes[12,8] and polarizability fluxes.[13] Since the SDFF protocol is completely based on QM data, both the polarization and the fluxes are guaranteed to produce analytically exact forces.

This paper deals with the issue of an optimal polarization model and presents the theoretical foundation of a group polarizability paradigm. The group consists of a rigid substucture of covalently connected atoms, and local axes define an anisotropic local group polarizability, thus including anisotropy explicitly from the start. We also show that a simple form of hyperpolarizability can be incorporated. Provision is made for an equivalent nonsingular-point polarizability through a “diffuse” site. Various properties of the theoretical formulation are developed in detail.

The performance of our model is described through applications to water, alkanes, and NMA. The concurrently optimized charge and polarizability parameters give excellent fits to the QM electric potentials, and the QM molecular polarizabilities are well reproduced. We conclude that such a group polarizability model must be considered a very good candidate for the inclusion of polarization into MM force fields, and we note that the SDFF protocol assures that, in addition to structures and energies, forces will be accurately modeled.

Acknowledgment. This research was supported by NSF Grant No. MCB0212232. Additional funding was provided by the Michigan Universities Commercialization Initiative.

References

(1) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236−242.

(2) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27−85.

(3) Cho, K. H.; No, K. T.; Scheraga, H. A. *J. Mol. Struct.* **2002**, *641*, 77−91.

(4) Rick, S. W.; Stuart, S. J. In *Reviews in Computational Chemistry*; Lipkowitz, R. K., Boyd, D. B., Eds.; Wiley-VCH: New York, 2002; Vol. 18, pp 89−146.

(5) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497−1506.

(6) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587−1597.

(7) Patel, S.; Brooks, C. L. *Mol. Simul.* **2006**, *32*, 231−249.

(8) Palmo, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. *Biopolymers* **2003**, *68*, 383−394.

(9) Mannfors, B.; Palmo, K.; Krimm, S. *J. Mol. Struct.* **2000**, *556*, 1−21.

(10) Palmo, K.; Krimm, S. *Chem. Phys. Lett.* **2004**, *395*, 133−137.

(11) Palmo, K.; Mannfors, B.; Krimm, S. *Chem. Phys. Lett.* **2003**, *369*, 367−373.

(12) Palmo, K.; Krimm, S. *J. Comput. Chem.* **1998**, *19*, 754−768.

(13) Palmo, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. *Chem. Phys. Lett.* **2006**, *429*, 628−632.

(14) Krasny, R. personal communication.

(15) Lindsay, K.; Krasny, R. *J. Comput. Phys.* **2001**, *172*, 879−907.

(16) Jackson, J. D. *Classical Electrodynamics*; John Wiley & Sons, Inc.: New York, 1975; p 39.

(17) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556−7560.

(18) Guillot, B.; Guissani, Y. *J. Chem. Phys.* **2001**, *114*, 6720−6733.

(19) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(20) Mannfors, B.; Mirkin, N. G.; Palmo, K.; Krimm, S. *J. Comput. Chem.* **2001**, *16*, 1933−1943.

(21) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

CT700102G

JCTC Journal of Chemical Theory and Computation

On the Calculation of Atomic Forces in Classical Simulation Using the Charge-on-Spring Method To Explicitly Treat Electronic Polarization

Daan P. Geerke and Wilfred F. van Gunsteren*

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich (ETH), CH-8093 Zürich, Switzerland

Received June 29, 2007

Abstract: An expression for the atomic forces in simulations using the charge-on-spring (COS) polarizable model is rederived. In previous implementations of COS-based models, contributions arising from the dependence of the induced dipoles (i.e., the positions of the charges-on-spring) on the coordinates of the other sites in the system were not taken into account. However, from calculations on gas-phase dimers we found a significant contribution of these terms. Errors in the forces when neglecting these contributions in condensed-phase calculations can be significantly reduced by choosing an appropriately large value for the size of the charge-on-spring.

1. Introduction

In classical atomistic simulation, electronic polarization effects can be taken into account using polarizable force fields.[1−4] When using such a force field, atomic dipoles or molecular charge distributions can adapt to the electric field generated by the environment to induce a net dipole moment. Several methods have been described in the literature that explicitly treat electronic polarization.[1−4] In one of them, the point-polarizable dipole (PPD) model,[5−7] the polarizable centers i in the system are assigned an inducible point-dipole $\vec{\mu}_i$, which adapts size and direction according to its polarizability α_i and the electric field $\vec{E}_i$ at i (assuming isotropic α_i and linear dependence of $\vec{\mu}_i$ on $\vec{E}_i$, and using SI units)

$$\vec{\mu}_i = \alpha_i(4\pi\epsilon_0)\vec{E}_i \tag{1}$$

Additionally to the U^{qq} term to describe Coulomb interactions between the fixed point-charges (q_i), the induced dipoles enter the expression for the electrostatic part of the potential energy (U^{ele}) via U^{stat}, $U^{\mu\mu}$, and U^{self}. The first two terms account for induced dipole-fixed point-charge and induced dipole−induced dipole interactions, respectively[4]

$$U^{\text{ele}}(\vec{r},\vec{\mu}) = U^{\text{qq}} + U^{\text{stat}} + U^{\mu\mu} + U^{\text{self}}$$
$$= \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N-1}\sum_{j>i}^{N}\frac{q_iq_j}{|\vec{r}_i - \vec{r}_j|} - \sum_{i=1}^{N}\vec{\mu}_i\cdot\vec{E}_i^{\text{q}} - \frac{1}{2}\sum_{i=1}^{N-1}\sum_{j\neq i}^{N}\vec{\mu}_i\underline{\text{T}}_{ij}\vec{\mu}_j + \sum_{i=1}^{N}\frac{\vec{\mu}_i\cdot\vec{\mu}_i}{2\alpha_i(4\pi\epsilon_0)} \tag{2}$$

while U^{self} is the self-polarization term accounting for the energy cost of dipole induction,[8] $\vec{E}_i^{\text{q}}$ is the electric field at i from the fixed point-charges, and $\underline{\text{T}}_{ij}$ are the elements of the dipole tensor.[4] Because the $\vec{\mu}_i$'s depend via $\vec{E}_i$ on the positions of the other point-charges and the sizes of the induced dipoles in the system, the forces at the polarizable centers i are calculated from

$$\vec{f}_i = -\nabla_i U^{\text{ele}}(\vec{r},\vec{\mu}) = -\left(\frac{\partial U^{\text{ele}}}{\partial \vec{r}_i} + \sum_{k\neq i}^{N}\frac{\partial U^{\text{ele}}}{\partial \vec{\mu}_k}\cdot\frac{\partial \vec{\mu}_k}{\partial \vec{r}_i}\right) \tag{3}$$

For a given set of atomic positions, eq 1 can be satisfied by an instantaneous adaptation of the $\vec{\mu}_i$'s to the $\vec{E}_i$'s. Due to the mutual dependence of the $\vec{E}_i$'s and $\vec{\mu}_i$'s, an iterative scheme is usually employed[6] to minimize U^{ele}, following a Born−Oppenheimer-like approximation[9] to determine the induced dipoles. If the convergence criterion is chosen tightly enough, U^{ele} is minimized with respect to the $\vec{\mu}_i$'s and

* Corresponding author fax: (+41)-44-632-1039; e-mail: wfvgn@igc.phys.chem.ethz.ch.

10.1021/ct700164k CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/10/2007

$$\frac{\partial U^{\mathrm{ele}}}{\partial \vec{\mu}_i} = 0 \tag{4}$$

As a result, eq 3 reduces to

$$\vec{f}_i = -\frac{\partial U^{\mathrm{ele}}}{\partial \vec{r}_i} \tag{5}$$

Instead of employing an iterative scheme, simulation studies using the PPD model have been reported[7] in which fictitious masses are assigned to the $\vec{\mu}_i$'s. In this case, the $\vec{\mu}_i$'s enter the equations of motion via an extended Lagrangian (as in the Car−Parrinello approach[10] to treat electronic degrees of freedom). Thus, forces at the polarizable centers i can be evaluated as the sum of the term on the right in eq 5 and a term involving $-\partial U^{\mathrm{ele}}/\partial\vec{\mu}_i$, although simulation settings are often chosen such that it can be assumed that the components of the $\vec{\mu}_i$'s are close enough to their Born−Oppenheimer values to fulfill the condition of eq 4.

As an alternative to the PPD approach, the inducible dipole moments can be described by the charge-on-spring[11] (COS) (or Drude-oscillator[12] or shell[13]) model. In this case, an inducible dipole is modeled by attaching a massless, virtual site with a point-charge q_i^{ν} to the polarizable center i, via a spring with harmonic force constant k_i^{ho},[4]

$$k_i^{\mathrm{ho}} = \frac{(q_i^{\nu})^2}{\alpha_i(4\pi\epsilon_0)} \tag{6}$$

The charge at the polarizable center is then $(q_i - q_i^{\nu})$. Thus, the induced dipoles $\vec{\mu}_i$ are represented by

$$\vec{\mu}_i = q_i^{\nu}(\vec{r}_i{}' - \vec{r}_i) \tag{7}$$

where $\vec{r}_i{}'$ is the position of the charge-on-spring. In COS-based schemes in which the charges-on-spring are not explicitly treated as additional degrees of freedom, the sum of the forces acting on any charge-on-spring should be zero, and the virtual charge q_i^{ν} must be positioned such that

$$\vec{f}_i^{\mathrm{ho}\prime} + \vec{f}_i^{\mathrm{coul}\prime} = 0 \tag{8}$$

with the force $\vec{f}_i^{\mathrm{ho}\prime}$ due to the spring given by

$$\vec{f}_i^{\mathrm{ho}\prime} = -k_i^{\mathrm{ho}}(\vec{r}_i{}' - \vec{r}_i) = -\frac{(q_i^{\nu})^2}{\alpha_i(4\pi\epsilon_0)}(\vec{r}_i{}' - \vec{r}_i) \tag{9}$$

and $\vec{f}_i^{\mathrm{coul}\prime}$ due to the (Coulombic) electric field at the charge-on-spring ($\vec{E}_i{}'$) given by

$$\vec{f}_i^{\mathrm{coul}\prime} = q_i^{\nu}\vec{E}_i{}' \tag{10}$$

To satisfy eq 8, the $\vec{\mu}_i$'s ($\vec{r}_i{}'$'s) should be determined from the $\vec{E}_i{}'$'s. However, since the displacement $|\vec{r}_i{}' - \vec{r}_i|$ of the charge-on-spring from the polarizable center is nonzero upon polarization, a better approximation of the ideal inducible dipole $\vec{\mu}_i$ at site i would be to determine $\vec{r}_i{}'$ from the electric field $\vec{E}_i$ at the polarizable center itself. From eqs 1 and 7, the $\vec{r}_i{}'$'s are then determined from

$$\vec{r}_i{}' = \vec{r}_i + \frac{\alpha_i(4\pi\epsilon_0)\vec{E}_i}{q_i^{\nu}} \tag{11}$$

Equations 8−11 show that the total force acting on the charge-on-spring is only zero if

$$\vec{E}_i{}' = \vec{E}_i \tag{12}$$

which is usually not the case for the induced dipole due to the nonzero values for $|\vec{r}_i{}' - \vec{r}_i|$. By choosing q_i^{ν} large enough, $|\vec{r}_i{}' - \vec{r}_i|$ adopts relatively small values, resulting in small differences between $\vec{E}_i$ and $\vec{E}_i{}'$. However, the size of q_i^{ν} is limited to values for which $|\vec{r}_i{}' - \vec{r}_i|$ is significant enough with respect to interatomic distances such that numerical precision is ensured when calculating, e.g., interaction energies involving induced dipoles. Like the PPD model, the COS method has been employed in combination with iterative procedures[14] to energy-minimize for the $\vec{r}_i{}'$'s (i.e., solving eq 8) or with an extended Lagrangian in which a fictitious mass is assigned to the charge-on-spring and the charges-on-spring are treated as additional degrees of freedom.[15] Originally, a noniterative procedure was proposed,[11] to which the following discussion is of equal importance. In studies using the iterative procedure, q_i^{ν} was set to −8.0 e,[14,16,17] whereas less negative charges were chosen in simulations using a COS model in combination with an extended system Lagrangian (with typical values between −1.0 and −2.0 e).[15,18,19] The reason is that in simulations using an extended Lagrangian, the absolute size of q_i^{ν} is limited by the small simulation time steps that must be taken for large $|q_i^{\nu}|$, due to the high vibrational frequencies of springs with a large force constant.

In the current work, we quantitatively investigate the error made when calculating the atomic forces according to the version of the charge-on-spring model in which the $\vec{r}_i{}'$'s are determined from eq 11, while assuming eq 8 to be fulfilled. For this purpose, calculated values for the components of the atomic electrostatic forces in selected test systems are compared with numerical values for the corresponding finite differences in the electrostatic potential energy. To evaluate the accuracy of the different force calculations, we compare between including contributions from the second term in eq 13

$$\vec{f}_i = -\nabla_i U^{\mathrm{ele}}(\vec{r},\vec{r}\,') = -\left(\frac{\partial U^{\mathrm{ele}}}{\partial \vec{r}_i} + \sum_{k\neq i}^{N} \frac{\partial U^{\mathrm{ele}}}{\partial \vec{r}_k{}'} \cdot \frac{\partial \vec{r}_k{}'}{\partial \vec{r}_i}\right) \tag{13}$$

and completely neglecting these contributions (which is equivalent to assuming that eq 8 is satisfied). An expression for $\sum_{k\neq i}^{\mathrm{N}}(\partial U^{\mathrm{ele}}/\partial\vec{r}_k{}')\cdot(\partial\vec{r}_k{}'/\partial\vec{r}_i)$ is derived in section 2. Preferably, the calculation of the contributions from this term is to be avoided in simulations, because it involves tensors of rank two and higher (see section 2), and introducing these terms would reduce the appealing character of the charge-on-spring model when compared to the PPD approach. As an alternative, we investigate whether the size of $\sum_{k\neq i}^{\mathrm{N}}(\partial U^{\mathrm{ele}}/\partial\vec{r}_k{}')\cdot(\partial\vec{r}_k{}'/\partial\vec{r}_i)$ and, hence, the error in the forces when neglecting this term can be satisfactorily reduced by choosing the (absolute) size of the charges-on-spring appropriately large. Our prime interest is the effect of q_i^ν and the omission of terms involving $\partial U^{\rm ele}/\partial\vec r_k\,'$ in the expressions for $\vec f_i$ on the performance of iterative COS models that were recently implemented by us.[14,16,17,20] Therefore, we chose as test systems three gas-phase dimers in which strong dipole−dipole, weak dipolar, or ion-dipole interactions are present, representing interactions typically occurring in biomolecular simulation, in which our COS models are to be used. The effect of the strength of the electric field at the polarizable centers on the results of the calculations is investigated by varying the separation between the monomers, from hydrogen-bonding distance to the typical cutoff distance for electrostatic interactions. From our findings we comment on the optimal choice for q_i^ν and the expression of the atomic forces, thereby considering both the accuracy and consistency of the model and its computational efficiency.

2. Theory

Unlike the PPD model, the COS method treats the induced dipole moments via additional point charges only, which allows for an easy introduction of polarizability into schemes to compute long-range electrostatic forces, such as the reaction-field,[21,22] Ewald-summation,[23] Particle−Particle−Particle-Mesh (P3M),[24] and Particle-Mesh-Ewald (PME)[25] techniques. The electrostatic potential ϕ_i at the polarizable centers i due to the monopoles and dipoles in the system can be expressed using Coulombic terms only

$$\phi_i(\vec r,\vec r\,') = \frac{1}{4\pi\epsilon_0}\sum_{j\neq i}^{N}\left[\frac{(q_j-q_j^\nu)}{|\vec r_i-\vec r_j|}+\frac{q_j^\nu}{|\vec r_i-\vec r_j\,'|}\right] \tag{14}$$

Because of the dependence of the $\vec r_i\,'$'s on the $\vec r_j$'s (and $\vec r_j\,'$'s) via $\vec E_i$ in eq 11, the relation between ϕ_i and the electric field $\vec E_i$ is given by

$$\vec E_i = -\nabla_i\phi_i(\vec r,\vec r\,') = -\left(\frac{\partial\phi_i}{\partial\vec r_i}+\sum_{k\neq i}^{N}\frac{\partial\phi_i}{\partial\vec r_k\,'}\cdot\frac{\partial\vec r_k\,'}{\partial\vec r_i}\right) \tag{15}$$

When applying a Born−Oppenheimer-like iterative SCF procedure, however, the $\vec r_i\,'$'s are at every iteration step determined in the fixed electric field due to the other q_j's and q_j^ν's. When using a convergence criterion which minimizes the ϕ_i's with respect to the positions $\vec r_i\,'$, the second term in eq 15 is zero at convergence because $\partial\phi_i/\partial\vec r_k\,' = 0$. Thus

$$\vec E_i = -\frac{\partial\phi_i}{\partial\vec r_i} = \frac{1}{4\pi\epsilon_0}\sum_{j\neq i}^{N}\left[\frac{(q_j-q_j^\nu)(\vec r_i-\vec r_j)}{|\vec r_i-\vec r_j|^3}+\frac{q_j^\nu(\vec r_i-\vec r_j\,')}{|\vec r_i-\vec r_j\,'|^3}\right] \tag{16}$$

The electrostatic part $U^{\rm ele}$ of the potential energy can also be expressed in terms of Coulomb interactions. The only non-Coulombic term to be added to $U^{\rm ele}$ is the self-polarization energy $U^{\rm self}$, which in the COS model can be expressed in terms involving point charges as well[17]

$$U^{\rm ele}(\vec r,\vec r\,') = U^{\rm coul} + U^{\rm self} \tag{17}$$

with

$$U^{\rm coul}(\vec r,\vec r\,') = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N-1}\sum_{j>i}^{N}\left[\frac{(q_i-q_i^\nu)(q_j-q_j^\nu)}{|\vec r_i-\vec r_j|}+\frac{(q_i-q_i^\nu)q_j^\nu}{|\vec r_i-\vec r_j\,'|}+\frac{q_i^\nu(q_j-q_j^\nu)}{|\vec r_i\,'-\vec r_j|}+\frac{q_i^\nu q_j^\nu}{|\vec r_i\,'-\vec r_j\,'|}\right] \tag{18}$$

and

$$U^{\rm self}(\vec r,\vec r\,') = \frac{1}{2}\sum_{i=1}^{N}\frac{(q_i^\nu)^2}{\alpha_i(4\pi\epsilon_0)}|\vec r_i\,'-\vec r_i|^2 \tag{19}$$

Now we consider the expression for the forces $\vec f_i$ that act on (polarizable) atomic centers i

$$\vec f_i = -\nabla_i U^{\rm ele}(\vec r,\vec r\,') = -\left(\frac{\partial U^{\rm ele}}{\partial\vec r_i}+\sum_{k\neq i}^{N}\frac{\partial U^{\rm ele}}{\partial\vec r_k\,'}\cdot\frac{\partial\vec r_k\,'}{\partial\vec r_i}\right) \tag{20}$$

Note again the dependence of the $\vec r_k\,'$'s on the $\vec r_i$'s that appears in the second term on the right in eq 20, which might adopt nonzero values because $U^{\rm ele}$ not only contains terms due to the ϕ_i's (first two terms on the right in eq 18) but also due to the $\phi_i\,'$'s (last two terms on the right in eq 18) and $U^{\rm self}$, whereas when using eq 11 only the ϕ_i's have been minimized with respect to the $\vec r_i\,'$'s. When nevertheless using assumptions 8 and 12, eq 20 reduces to

$$\vec f_i^{\rm red} = -\frac{\partial U^{\rm ele}}{\partial\vec r_i} = -\left(\frac{\partial U^{\rm coul}}{\partial\vec r_i}+\frac{\partial U^{\rm self}}{\partial\vec r_i}\right) \tag{21}$$

From the assumptions in eq 8 and 12 we have

$$-\frac{\partial U^{\rm self}}{\partial\vec r_i} = \vec f_i^{\rm ho} = -\vec f_i^{\rm ho\prime} = \vec f_i^{\rm coul\prime} = -\frac{\partial U^{\rm coul}}{\partial\vec r_i\,'} \tag{22}$$

and the reduced expression $\vec f_i^{\rm red}$ for the atomic forces becomes[4,14]

$$\vec f_i^{\rm red} = \frac{1}{4\pi\epsilon_0}\sum_{j\neq i}^{N}\left[\frac{(q_i-q_i^\nu)(q_j-q_j^\nu)(\vec r_i-\vec r_j)}{|\vec r_i-\vec r_j|^3}+\frac{(q_i-q_i^\nu)q_j^\nu(\vec r_i-\vec r_j\,')}{|\vec r_i-\vec r_j\,'|^3}+\frac{q_i^\nu(q_j-q_j^\nu)(\vec r_i\,'-\vec r_j)}{|\vec r_i\,'-\vec r_j|^3}+\frac{q_i^\nu q_j^\nu(\vec r_i\,'-\vec r_j\,')}{|\vec r_i\,'-\vec r_j\,'|^3}\right] \tag{23}$$

In Appendix A, an expression for the second term on the right in eq 20 is derived up to first order in the many-body electrostatic interactions between polarizable centers and the charges-on-spring, to account for the contribution to the force at atomic center i originating from the change in the inducible dipoles $\vec\mu_k$'s ($\vec r_k\,'$'s) upon a change in $\vec r_i$. Contributions from higher-order many-body terms have not been calculated explicitly. The first-order terms are to be added to $\vec f_i^{\rm red}$ to obtain an expression for the first-order corrected force $\vec f_i^{(1)}$

$$\vec{f}_i^{(1)} = \vec{f}_i^{\mathrm{red}} + \frac{1}{4\pi\epsilon_0}\sum_{k\neq i}^{N}\left[\sum_{j\neq k}^{N}\left(\frac{(q_j - q_j^{\nu})(\vec{r}_j - \vec{r}_k{}')}{|\vec{r}_j - \vec{r}_k{}'|^3} + \frac{q_j^{\nu}(\vec{r}_j{}' - \vec{r}_k{}')}{|\vec{r}_j{}' - \vec{r}_k{}'|^3}\right) + \frac{q_k^{\nu}}{\alpha_k}(\vec{r}_k{}' - \vec{r}_k)\right]\bullet \alpha_k\left((q_i - q_i^{\nu})\left\{\frac{\overleftrightarrow{\mathrm{I}}}{|\vec{r}_i - \vec{r}_k|^3} - \frac{3(\vec{r}_i - \vec{r}_k)\otimes(\vec{r}_i - \vec{r}_k)}{|\vec{r}_i - \vec{r}_k|^5}\right\} + q_i^{\nu}\left\{\frac{\overleftrightarrow{\mathrm{I}}}{|\vec{r}_i{}' - \vec{r}_k|^3} - \frac{3(\vec{r}_i{}' - \vec{r}_k)\otimes(\vec{r}_i{}' - \vec{r}_k)}{|\vec{r}_i{}' - \vec{r}_k|^5}\right\}\right) \quad (24)$$

with $\overleftrightarrow{\mathrm{I}}$ the three-dimensional unit tensor of second rank. Equation 24 indicates that for infinitely large $|q_i^{\nu}|$, $\vec{f}_i^{(1)}$ will indeed reduce to $\vec{f}_i^{\mathrm{red}}$, because the difference between the two terms between curly brackets will vanish as $(\vec{r}_i{}' - \vec{r}_k) \rightarrow (\vec{r}_i - \vec{r}_k)$, and the term involving $q_k^{\nu}/\alpha_k\,(\vec{r}_k{}' - \vec{r}_k)$ does not diverge due to the inverse linear dependence of $(\vec{r}_k{}' - \vec{r}_k)$ on q_k^{ν} (eq 11).

3. Computational Details

Single-configuration calculations were performed on a water−water, a Na^+−water, a Na^+−Cl^-, an argon−water, and an argon−Na^+ dimer, using a version of the GROMOS96 code[26,27] adapted to the charge-on-spring (COS) model.[4,14] Water molecules were described by the polarizable COS/B2 model.[14] In the evaluations of the forces and energies, only nonbonded (and no covalent) interactions were taken into account. Bond constraints were not applied. Nonbonded parameters for Na^+, Cl^- and Ar were taken from the GROMOS 43A1 force field,[26] with Ar being polarizable with a polarizability of 1.6411 * 10^{-3} nm^3.[28] The polarizability of the Na^+ and Cl^- ions were either set to zero or to 1.0 * 10^{-3} nm^3. Unless stated otherwise, the charges-on-spring q_i^{ν} were set to −8 *e*.

For the water−water dimer, atomic coordinates corresponded to a C_s-symmetrical conformation, with covalent bond lengths and angles according to the COS/B2 model[14] (values for the O−H and H−H interatomic (bond) distances are r_{OH} = 0.1 nm and r_{HH} = 0.163299 nm). Dimer configurations were generated by placing the oxygen atoms on the *x*-axis and varying the distance between the two oxygen atoms from 0.25 to 1.4 nm (with increments of 0.05 nm). For one of the water molecules, one O−H bond was aligned along the *x*-axis, pointing to the oxygen of the other one. The other hydrogen of this water molecule was placed in the *xy*-plane. The angle between the *x*-axis and the bisector of the bond vectors of the other water molecule was set to 105.4°, and the H−H vector was orthogonal to the *xy*-plane, see Figure 1. Configurations for the other dimers were generated by replacing the oxygen of one or both of the water molecules by Ar, Na^+, or Cl^- and removing the hydrogens attached to the replaced oxygen atom(s). Single-configuration calculations of the electrostatic energies and forces were performed, in which except for the intramolecular interactions all interatomic interactions were taken into account. The convergence criterion[14] for the iterative procedure to determine the induced dipoles was

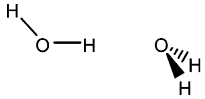


Figure 1. Geometry (C_s symmetry) of the water dimer in the finite-difference calculations. The geometry of the water molecules is described by the COS/B2 model.[14] The O−O distance (which is aligned along the *x*-axis) was varied; all other degrees of freedom were kept fixed. The water molecule on the left and the bisector of the water molecule on the right are in the *xy*-plane. The angle of this bisector with the *x*-axis is 105.4°. The H−H vector of the water molecule on the right is orthogonal to the *xy*-plane.

$$\max_{i,x,y,z}(|\Delta E_{i,x}|,|\Delta E_{i,y}|,|\Delta E_{i,z}|)|q_{\mathrm{O}}||d| < \Delta U \quad (25)$$

with ΔU set to 1 * 10^{-9} kJ mol^{-1}. In eq 25, *d* is a measure for typical interatomic distances determining the electric field at the polarizable center *i* (here we set *d* arbitrarily to 1 nm), q_{O} was set to the charge of the COS/B2 oxygen (−0.746 *e*), and $|\Delta E_{i,k}|$ is the change between consecutive iteration steps in the electric field component *k* at site *i*. The applied criterion not only ensures convergence of the calculated electric field within machine precision but also of the electrostatic potential ϕ_i.

Calculations were performed using expression 23 or 24 for the electrostatic forces at the atomic centers. Components of the electrostatic forces were compared to numerical values obtained from a finite difference in the electrostatic potential energy as calculated after shifting the coordinates of any of the atoms in the *x*-, *y*-, or *z*-direction by Δx, Δy, or Δz (only the expression for the *x*-component is given here)

$$f_{i,x} = \frac{U_{-\Delta x}^{\mathrm{ele}} - U_{+\Delta x}^{\mathrm{ele}}}{2\Delta x} \quad (26)$$

$U_{+\Delta x}^{\mathrm{ele}}(U_{-\Delta x}^{\mathrm{ele}})$ is the electrostatic energy after applying the shift in the positive (negative) direction. Estimated errors in the components of the calculated atomic forces (eqs 23 and 24) are defined as the deviation from the value obtained from eq 26. A value of 0.5 * 10^{-5} nm was chosen for the size of the shifts ($|\Delta x|$, $|\Delta y|$, or $|\Delta z|$).

4. Results and Discussion

From finite-difference calculations on the nonpolarizable Na^+−Cl^- dimer and the nonpolarizable water dimer with the COS/B2 partial charges scaled by a factor of 0.01 and the atomic polarizabilities of the oxygens (α_{O}) set to zero, we concluded that the length of the shift in the atomic coordinates of 0.5 * 10^{-5} nm is a proper choice for our test systems in order to evaluate whether machine precision can be obtained for the components of the electrostatic forces when compared to finite differences in the Coulombic potential energy. In the case of close contact between the ions (large electrostatic forces and potential-energy values) and of the water molecules with the scaled partial charges at an interatomic distance of 1.4 nm (weak Coulombic interactions), absolute values for the differences between the *x*-, *y*-, or *z*-component of the calculated atomic gradient and

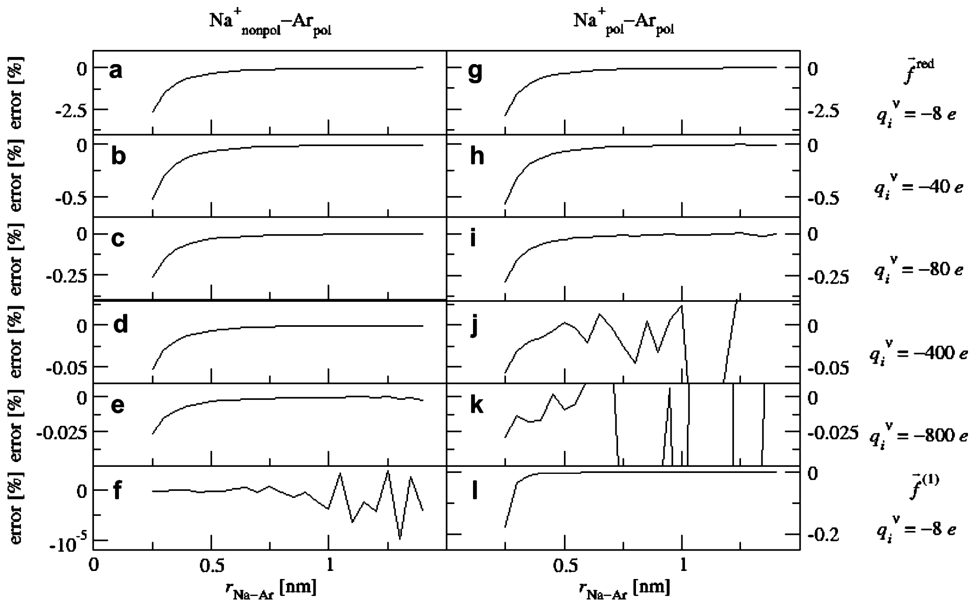


Figure 2. Relative error in a single component of the electrostatic atomic force with respect to the corresponding finite difference in the electrostatic potential energy (eq 26) for a gas-phase dimer consisting of a polarizable argon probe and either a nonpolarizable (panels (a)−(f)) or a polarizable (panels (g)−(l)) Na$^+$ ion, separated by an interatomic distance $r_{\mathrm{Na-Ar}}$, where the atomic forces are calculated using eq 23 ($\vec{f}_i^{\,\mathrm{red}}$, panels (a)−(e) and panels (g)−(k)) or using eq 24 ($\vec{f}_i^{\,(1)}$, panels (f) and (l)). In panels (a)−(e), and (g)−(k), the size of the charges-on-spring is varied and set to $q_i^v = -8\ e$, $-40\ e$, $-80\ e$, $-400\ e$, and $-800\ e$, respectively. In panels (f) and (l), $q_i^v = -8\ e$. Note the different scales on the *y*-axes.

the corresponding finite difference in the energy were maximally in the order of 10^{-8}% and 10^{-6}%, respectively. Also for the nonpolarizable COS/B2 ($\alpha_{\mathrm{O}} = 0$) and Na$^+$−water dimers, machine precision in the calculated force components was obtained, with maximum errors of 10^{-7}% at any separation distance between the monomers.

In contrast, when considering the electrostatic atomic gradients in the dimers with one or more nonzero polarizabilities, discrepancies with the finite differences in the electrostatic potential energy were found to be larger by several orders of magnitude than for the nonpolarizable dimers. For the Na$^+$−Ar system in which only argon has an inducible dipole moment, e.g., the atomic force $\vec{f}_i^{\,\mathrm{red}}$ as calculated from eq 23 (having only one component) differs by -0.015% from the negative of the finite-difference in the electrostatic energy when the distance between the sodium ion and the argon atom ($r_{\mathrm{Na-Ar}}$) is 1.4 nm, and by no less than -2.7% for $r_{\mathrm{Na-Ar}} = 0.25$ nm, see Figure 2a. Machine precision is recovered by evaluating $\vec{f}_i^{\,(1)}$ (using eq 24) instead of $\vec{f}_i^{\,\mathrm{red}}$: Figure 2f shows that the errors in $\vec{f}_i^{\,(1)}$ are typically 10^{-6} or 10^{-5}%. The significant error in the calculated forces when using the reduced expression 23 can be explained from the neglect of the contribution to the atomic forces that is due to the change in the electric field at argon and, hence, its inducible dipole upon a displacement of Na$^+$ or Ar. This effect is accounted for by the first-order correction terms in eq 24. The decay of the relative error in $\vec{f}_i^{\,\mathrm{red}}$ with increasing $r_{\mathrm{Na-Ar}}$ (Figure 2a) can be explained from the corresponding decrease in the size of the correction factor in eq 24 which scales with $(r_{\mathrm{Na-Ar}})^{-3}$, whereas $\vec{f}_i^{\,\mathrm{red}}$ scales with $(r_{\mathrm{Na-Ar}})^{-2}$.

In the presence of a single polarizable center, the first-order correction on $\vec{f}_i^{\,\mathrm{red}}$ suffices to obtain machine precision in the calculated electrostatic gradients (Figure 2f). However, when considering a Na$^+$−Ar dimer in which both the argon and the ion are polarizable (from hereon referred to as Na$^+_{\mathrm{pol}}$−Ar$_{\mathrm{pol}}$), we observe discrepancies not only between the finite differences in the electrostatic potential energies with calculated values for $\vec{f}_i^{\,\mathrm{red}}$ (see Figure 2g) but also with calculated values for $\vec{f}_i^{\,(1)}$, see Figure 2l. The reason is that in the presence of more than one polarizable center, higher-order many-body effects come into play. A change in dipole induction of Ar (Na$^+$) due to a displacement of Na$^+$ (Ar) leads again to a change in polarization of the ion (argon probe) itself, and terms accounting for this second-order many-body effect have been neglected in deriving $\vec{f}_i^{\,(1)}$ (Appendix A). The error in $\vec{f}_i^{\,(1)}$ calculated for the Na$^+_{\mathrm{pol}}$−Ar$_{\mathrm{pol}}$ dimer at short $r_{\mathrm{Na-Ar}}$ is an order of magnitude smaller than the error in $\vec{f}_i^{\,\mathrm{red}}$, but its maximal value of 0.2% is still significant (see Figure 2l). Because the second-order correction terms on $\vec{f}_i^{\,\mathrm{red}}$ that are missing in $\vec{f}_i^{\,(1)}$ inversely scale with a larger exponent in $r_{\mathrm{Na-Ar}}$ than the terms in $\vec{f}_i^{\,(1)}$, errors in $\vec{f}_i^{\,(1)}$ for the Na$^+_{\mathrm{pol}}$−Ar$_{\mathrm{pol}}$ dimer with $r_{\mathrm{Na-Ar}}$ approximating 1.4 nm come close to the limit of machine precision, with values on the order of 10^{-4}%. However, in condensed-phase simulations, close neighbor interactions will screen long-range interactions and will dominate the size and error of the atomic gradients. Note that for Na$^+_{\mathrm{pol}}$−Ar$_{\mathrm{pol}}$, errors in $\vec{f}_i^{\,\mathrm{red}}$ are of similar magnitude when compared to the case in which the sodium ion was not polarizable (compare parts a and g of Figure 2), because of the small contribution of the induced dipole at Na$^+$ to the total electrostatic interactions which are dominated by the (fixed) net charge of the ion.

For the Na$^+_{\mathrm{pol}}$−Ar$_{\mathrm{pol}}$ dimer, observed differences between the components of $\vec{f}_i^{\,\mathrm{red}}$ and $\vec{f}_i^{\,(1)}$ and the corresponding finite

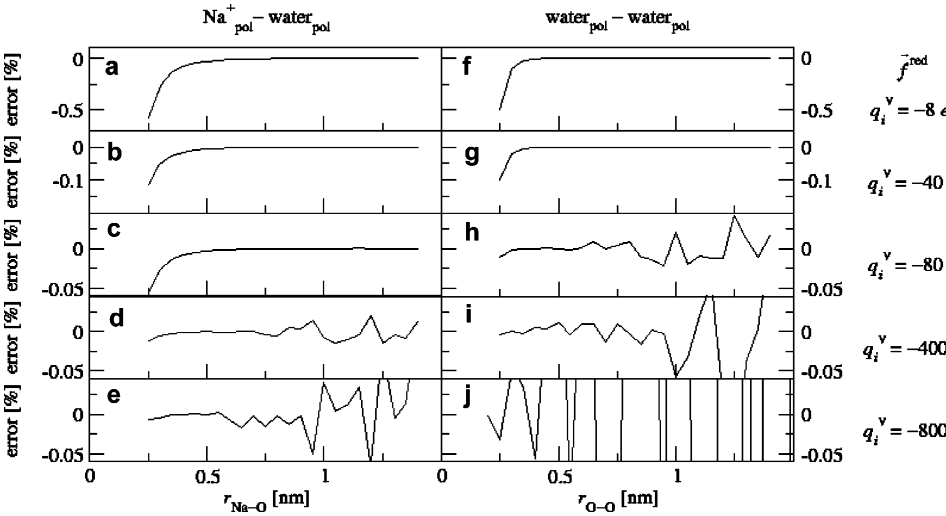


Figure 3. Relative error in the *x*-component of the electrostatic force at Na$^+$ or at the water oxygen, with respect to the corresponding finite difference in the electrostatic potential energy (eq 26) for a gas-phase dimer consisting of a COS/B2 water molecule and a polarizable Na$^+$ ion (panels (a)−(e)) or two COS/B2 water molecules (panels (f)−(j)), separated by an interatomic distance (r_{Na-O} and r_{O-O}, respectively), where the atomic forces are calculated using eq 23 ($\vec{f}_i^{red}$). In panels (a)−(e), and (f)−(j), the size of the charges-on-spring is varied and set to $q_i^v = -8$ *e*, −40 *e*, −80 *e*, −400 *e*, and −800 *e*, respectively. Note the different scales on the *y*-axes.

differences in the electrostatic energies are relatively large when compared to the situation of the Na$^+$−water or Na$^+$−Cl$^-$ dimers in which both monomers are polarizable as well. Because electrostatic interactions between Na$^+$ and argon are only due to interactions involving the induced dipole on argon, relative contributions from the terms due to $\sum_{k\neq i}^{N}(\partial U^{ele}/\partial \vec{r}'_k)\cdot(\partial \vec{r}'_k/\partial \vec{r}_i)$ in eq 13 sum up to the total error in the electrostatic forces when calculated for $\vec{f}_i^{red}$. When going from the Na$^+$−Ar system to the Na$^+$−water or the Na$^+$−Cl$^-$ dimer, in which apart from ion-induced dipole interactions, ion-fixed dipole or ion−ion interactions are present, we observe a significant decrease in the relative errors in calculated values for $\vec{f}_i^{(1)}$ and $\vec{f}_i^{red}$ at Na$^+$ (which have again a single (*x*-)component). At close distance between the (polarizable) monomers, the error in $\vec{f}_i^{red}$ at Na$^+$ is one or several orders of magnitude smaller for Na$^+$−water (−0.6%) or Na$^+$−Cl$^-$ (−0.003%) than for Na$^+_{pol}$−Ar$_{pol}$ (−2.9%). Moreover, at maximal separation between the monomers, machine precision in the components of the atomic forces when calculated from eq 23 is obtained for Na$^+$−water (Figure 3a) and Na$^+$−Cl$^-$ (results not shown). The reason is that the contribution to the forces from the fixed charge distributions is correctly computed (as is apparent from our finite-difference calculations on the nonpolarizable dimers), and the contribution to the total forces from the fixed point charge distribution is usually larger than the contribution from the inducible dipoles, especially in case of large separation distance between the monomers (due to small local electric fields and accordingly small induced dipoles) or ion−ion interactions.

When considering dimers such as the water−argon and water−water systems, in which interactions involving ionic species are not present, the smaller electric fields at the polarizable centers compared to when ions are present cause a smaller induction of the inducible dipoles, and the assumption in eq 12 (and in eq 8) is more justified. On the other hand, in the presence of an ion, the many-body contributions from the induced dipole moment on the neutral monomer is the main source of errors in $\vec{f}_i^{red}$ (compare parts a and g of Figure 2). When going to dimers with neutral monomers solely, there is a doubling of the number of the error sources. For the fully polarizable water−argon (results not shown) and water−water dimer (Figure 3f), these effects apparently counteract such that deviations of the components of the $\vec{f}_i^{red}$'s at oxygen and argon from the corresponding finite differences in the electrostatic potential energy are comparable to the errors in $\vec{f}_i^{red}$ in the case of Na$^+$−argon (Figure 2g) and Na$^+$−water (Figure 3a), respectively.

From the above it is clear that contributions from $\sum_{k\neq i}^{N}(\partial U^{ele}/\partial \vec{r}'_k)\cdot(\partial \vec{r}'_k/\partial \vec{r}_i)$ in eq 13 can adopt significant values. Furthermore, if a first-order correction on the reduced expression for the atomic forces in eq 23 is taken into account (using expression 24), errors in the calculated forces decrease but are still found to be significant in the presence of more than one polarizable center in the system. Including the first-order correction terms in eq 24 makes the calculation of the forces already more expensive compared to the evaluation of $\vec{f}_i^{red}$, and the number of terms in the higher-order corrections rapidly increases due to the third term on the right in eq A2 and the second term on the right in eq A3, which contain tensors of third rank and higher. With an eye to the computational efficiency of the charge-on-spring model, it is not an option to take these terms into account in molecular dynamics simulations. As an alternative, we investigate here to which extent errors in the atomic forces (as a result of completely neglecting $\sum_{k\neq i}^{N}(\partial U^{ele}/\partial \vec{r}'_k)\cdot(\partial \vec{r}'_k/\partial \vec{r}_i)$ in eq 13) can be minimized by choosing an appropriately large (absolute) value for the charge-on-spring q_i^ν.

Figures 2 and 3 show errors in the calculated values for the atomic gradients in the polarizable Na$^+$−Ar, Na$^+$−water,

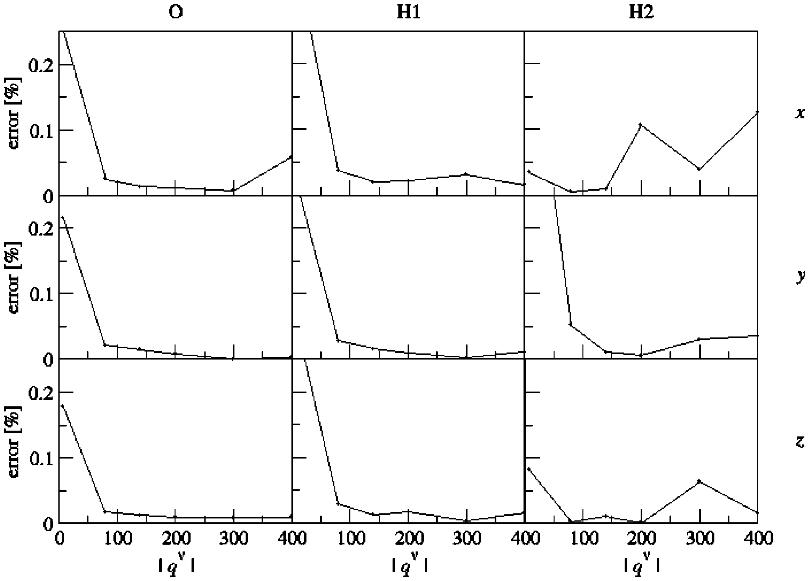


Figure 4. Absolute value of the relative error in the *x*-, *y*-, and *z*-component (upper, middle, and lower panels, respectively) of the electrostatic force (calculated using eq 23) at the oxygen (left panels), first hydrogen (middle panels), or second hydrogen atom (right panels) of a randomly chosen water molecule in a periodic box filled with 1000 COS/B2 water molecules, with respect to the corresponding finite differences in the electrostatic potential energy (eq 26), for different absolute values for the charge-on-spring q_i^ν.

and water−water dimers with respect to the finite difference in the electrostatic energy for different values of the charges-on-spring, varying from −8 *e* to −800 *e*. For the polarizable dimers considered, the discrepancy between the components of $\vec{f}_i^{\rm red}$ and the corresponding finite differences in the electrostatic potential energy is found to inversely scale down with $|q_i^\nu|$ in case of close contact between the monomers (see Figures 2 and 3). However, at large separation distance between the monomers and from $q_i^\nu = -80\ e$ onward, the relative error in $\vec{f}_i^{\rm red}$ rapidly increases with q_i^ν adopting more negative values. The reason is that at large separation between the monomers, the $\vec{E}_i$'s adopt small values and the displacements $|\vec{r}_i{}' - \vec{r}_i|$ of the charges-on-spring become relatively small compared to interatomic distances $|\vec{r}_i - \vec{r}_j|$. As a result, $|\vec{r}_i{}' - \vec{r}_i|$ values are not significant enough anymore, resulting in an inaccurate determination of $\vec{f}_i^{\rm red}$ (and $\vec{f}_i^{(1)}$ as well, results not shown). This effect is most pronounced for the water dimer (Figure 3f−j), because the dipole−dipole interactions within this dimer scale with $|\vec{r}_i - \vec{r}_j|^{-3}$ instead of $|\vec{r}_i - \vec{r}_j|^{-2}$ as the ion-dipole interactions do in the case of Na^+ being present (Figures 2 and 3a−e). The error in $\vec{f}_i^{\rm red}$ for large $|q_i^\nu|$ and separation distance between the monomers is also more pronounced for $Na^+_{\rm pol}$−$Ar_{\rm pol}$ and $Na^+_{\rm pol}$−water than for the Na^+−Ar dimer in which only the argon probe is polarizable. In the latter case, dipole induction is only due to the electric field generated by a net charge which only scales with $|\vec{r}_i - \vec{r}_j|^{-2}$, resulting in values for $|\vec{r}_i{}' - \vec{r}_i|$ that are even significant at maximal separation between the monomers.

From the above, effectively reducing errors in $\vec{f}_i^{\rm red}$ (and $\vec{f}_i^{(1)}$) through choosing large values for $|q_i^\nu|$ is limited in the situation of the gas-phase dimers, because of inaccuracies introduced in evaluating long-range forces due to dipolar interactions. However, our COS models[14,16,17,20] were developed for use in condensed-phase simulations, and solvent screening will significantly affect long-range interactions. To quantify the effect of solvent screening on errors in the reduced force $\vec{f}_i^{\rm red}$, we repeated the calculations on our test systems for a randomly picked water molecule out of configurations consisting of 1000 COS/B2 water molecules in a cubic periodic box with a volume of 29.616 nm^3. (The system was equilibrated in a NVT molecular dynamics simulation at 298.15 K. Nonbonded interactions were calculated for water molecules which were within 1.4 nm, and no long-range correction for the long-range nonbonded interactions was applied.) Whereas machine precision was obtained in the evaluation of the forces when all atomic polarizabilities are set to zero (results not shown), significant errors are observed for the components of $\vec{f}_i^{\rm red}$ at the atoms from calculations on the polarizable system. Figure 4 shows trends in the error in the *x*-, *y*-, and *z*-components of $\vec{f}_i^{\rm red}$ for the oxygen and hydrogen atoms for the selected water molecule upon varying q_i^ν from values of −8 *e* up till −400 *e*. As in the case of the gas-phase dimer consisting of two COS/B2 water molecules at close contact (see Figure 3f), the error in $\vec{f}_i^{\rm red}$ is a few tenths of a percent for $q_i^\nu = -8\ e$. Also the finding that this error reduces by increasing $|q_i^\nu|$ to values of up to −200 *e* (see Figure 4) indicates that short-range interactions determine the total force acting on the particle. We found the error in the long-range forces to increase for the dimeric water$_{\rm pol}$−water$_{\rm pol}$ system when going, for example, from $q_i^\nu = -40\ e$ to $q_i^\nu = -80\ e$. From Figure 4, the error in the components of $\vec{f}_i^{\rm red}$ at the atoms is in most cases minimal at $q_i^\nu = -200\ e$ or $q_i^\nu = -300\ e$. In many cases, an increase in the error is observed when going from $q_i^\nu = -300\ e$ to $q_i^\nu = -400\ e$, probably due to too small $|\vec{r}_i{}' - \vec{r}_i|$ values relative to interatomic distances. Similar trends were observed when calculating $\vec{f}_i^{\rm red}$ at Na^+ or Ar after removing the selected water molecule from the system and

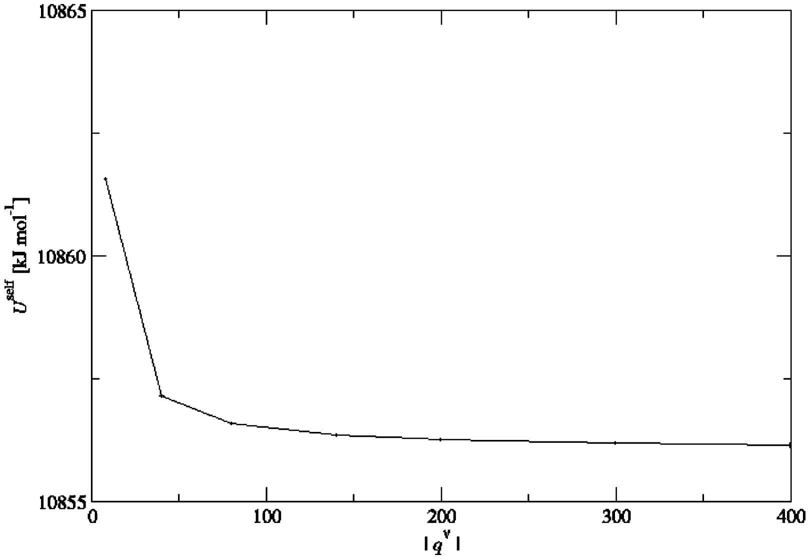


Figure 5. Total self-polarization energy $U^{\rm self}$ (eq 19) for a periodic box filled with 1000 COS/B2 water molecules for different absolute values for the charge-on-spring q_i^ν.

placing a polarizable Na$^+$ or argon probe at the position of its oxygen atom (results not shown). Only the error in the *x*-component of $\vec{f}_i^{\rm red}$ at one of the hydrogen atoms of the water molecule is minimal at a less negative value for q_i^ν than -200 *e*. The reason is that this component of $\vec{f}_i^{\rm red}$ is by 1 or 2 orders of magnitude smaller than the other components in the water molecule, making its determination intrinsically less accurate. Upon choosing $q_i^\nu = -200$ *e* or $q_i^\nu = -300$ *e*, relative errors in $\vec{f}_i^{\rm red}$ at Na$^+$, argon, and the atoms in water are typically up to 2 orders of magnitude smaller when compared to using a value of -8 *e* for q_i^ν. Note that not only the error in $\vec{f}_i^{\rm red}$ is affected by q_i^ν but also the total electrostatic energy, as reflected by the self-polarization energy $U^{\rm self}$ (see Figure 5 for the pure water box). $U^{\rm self}$ is (from linear response theory) a direct measure for the total contribution of the induced dipoles to the potential energy of the system and is for relatively small values for $|q_{\rm i}^\nu|$ not converging to the value that would be obtained in the case of treating the induced dipoles as being infinitely small. However, the total variation of $U^{\rm self}$ with respect to q_i^ν for $|q_i^\nu| \geq 40$ *e* is small when compared to $k_{\rm B}T$ (=2.5 kJ mol^{-1} at 298.15 K), see Figure 5. It would be interesting to see if using a value of $q_i^\nu = -200$ *e* or -300 *e* instead of -8 *e* in combination with expression 23 for the forces leads not only to a better description of the (reduced) atomic forces and self-polarization energy but also to an improvement in describing other relevant properties of condensed-phase systems, such as the dielectric permittivity of polar liquids which was found to be significantly off from experiment for COS-based solvent models that were recently developed in our group.[14,16,17,20] This will be the subject of a future molecular dynamics simulation study by us. In addition, the effect on various liquid properties of using expression 24 for $\vec{f}_i^{(1)}$ instead of expression 23 for the atomic forces in simulations of polarizable liquids will be evaluated. From calculations on $\vec{f}_i^{\rm red}$ and $\vec{f}_i^{(1)}$ for the randomly picked water molecule in the COS/B2 water box with q_i^ν set to -200 *e*, use of the expression for $\vec{f}_i^{(1)}$ was found to additionally reduce deviations with finite differences in the potential energy by 1 order of magnitude (results not shown).

In simulations using an extended Lagrangian in combination with the COS model, the first-order correction in eq 24 (due to $\sum_{k\neq i}^{\rm N}(\partial U^{\rm ele}/\partial\vec{r}_k')\cdot(\partial\vec{r}_k'/\partial\vec{r}_i)$ in eq 13) does in principle not have to be considered due to the explicit treatment of the charges-on-spring as degrees of freedom. However, simulation settings are usually[15,29] chosen such that the charges-on-spring follow the Born−Oppenheimer dipole interaction energy surface according to the atomic coordinates, making the $\vec{r}_i'$'s implicitly depending on the set of $\vec{r}_i$'s. It would be interesting to evaluate contributions from this dependence of $\vec{r}_i'$ on $\vec{r}_i$ to the forces and electric fields in simulations using an extended Lagrangian, which might be large because q_i^ν is usually set to relative small values to reduce the vibrational frequencies of the charges-on-spring. For the Na$^+$−argon dimer, for example, the difference between $\vec{f}_i^{\rm red}$ and the finite difference in the electrostatic potential energy was found to add up to −12% when $q_i^\nu = -2$ *e* is chosen, as in ref 15. From the discussion in section 1, the contribution of $\sum_{k\neq i}^{\rm N}(\partial U^{\rm ele}/\partial\vec{r}_k')\cdot(\partial\vec{r}_k'/\partial\vec{r}_i)$ will be zero upon replacing $\vec{E}_i$ in eq 11 by $\vec{E}_i'$ as Lamoureux and Roux did in their implementation of the charge-on-spring model using an extended Lagrangian.[15] Indeed, with an adapted version of the polarizable GROMOS96 code in which the inducible dipoles were determined from $\vec{E}_i'$ instead of $\vec{E}_i$, machine precision was obtained in the reduced force $\vec{f}_i^{\rm red}$ when compared to the finite difference in the electrostatic energy (results not shown).

Finally, we note that in the PPD model, the correction term to be added to the 'reduced' atomic forces (due to the second term on the right in eq 3) is zero when applying an iterative SCF procedure, because of the infinitely small size of the induced dipoles. However, as pointed out by Rick and Stuart already,[2] this does not make the PPD model physically more realistic than the COS model: treating electron polarization effects by inducible dipoles of finite size might even be considered a better representation of the (noninfinitely small) electron clouds which in quantum-mechanically treated systems cause the corresponding induced dipoles.

5. Conclusions

When using the charge-on-spring model to explicitly account for electronic polarization and calculating the induced dipole at a polarizable center from the electric field at the polarizable center itself (i.e., using eq 11), a contribution to the atomic forces arises due to the dependence of the positions of the charges-on-spring on the positions of all other point charges (second term on the right in eq 13). For a system with a single polarizable center, this contribution can be accounted for by adding a first-order correction to the reduced atomic force $\vec{f}_i^{\text{red}}$ which neglects this effect (eq 23). However, we found that higher-order corrections are to be included for systems with more than one charge-on-spring in order to obtain machine precision in the calculated atomic forces. These higher-order corrections contain second- and higher-rank tensors and make the evaluation of the forces cumbersome and time-consuming. As an alternative, the error in calculating the atomic forces introduced via neglecting the second term on the right in eq 13 can be minimized by choosing the size of the charge-on-spring q_i^ν appropriately large, thereby minimizing the finite size of the induced dipoles. From the evaluation of the force at atoms of a water molecule, at argon, or at Na$^+$ solvated in a box of water, with the system described by a polarizable force field, we found that components of $\vec{f}_i^{\text{red}}$ are closest to the finite difference in the Coulombic energy, upon setting q_i^ν to -200 *e* or -300 *e*. Work is under way to quantify the effect of using this value for q_i^ν (instead of -8 *e* as in previous molecular dynamics simulation studies) on the properties of condensed-phase systems as determined from simulation using the charge-on-spring model.

Appendix A

To derive an expression for $\sum_{k\neq i}^{N}(\partial U^{\text{ele}}/\partial\vec{r}_k{}')\cdot(\partial\vec{r}_k{}'/\partial\vec{r}_i)$, we consider in the following the two partial derivatives separately. First

$$\frac{\partial U^{\text{ele}}}{\partial\vec{r}_k{}'} = \frac{\partial U^{\text{coul}}}{\partial\vec{r}_k{}'} + \frac{\partial U^{\text{self}}}{\partial\vec{r}_k{}'} \tag{A1}$$

with, from eq 18 and by taking the dependence of the other $\vec{r}_m{}'$'s on $\vec{r}_k{}'$ into account

$$\frac{\partial U^{\text{coul}}}{\partial\vec{r}_k{}'} = \frac{1}{4\pi\epsilon_0}\sum_{j\neq k}^{N}\left(\frac{(q_j - q_j^\nu)q_k^\nu(\vec{r}_j - \vec{r}_k{}')}{|\vec{r}_j - \vec{r}_k{}'|^3} + \frac{q_j^\nu q_k^\nu(\vec{r}_j{}' - \vec{r}_k{}')}{|\vec{r}_j{}' - \vec{r}_k{}'|^3} + \right. \\ \left. \sum_{m\neq k}^{N}\frac{\partial U^{\text{coul}}}{\partial\vec{r}_m{}'}\cdot\frac{\partial\vec{r}_m{}'}{\partial\vec{r}_k{}'}\right) \tag{A2}$$

and from eq 19

$$\frac{\partial U^{\text{self}}}{\partial\vec{r}_k{}'} = \frac{(q_k^\nu)^2}{\alpha_k}\frac{(\vec{r}_k{}' - \vec{r}_k)}{4\pi\epsilon_0} + \sum_{m\neq k}^{N}\frac{\partial U^{\text{self}}}{\partial\vec{r}_m{}'}\cdot\frac{\partial\vec{r}_m{}'}{\partial\vec{r}_k{}'} \tag{A3}$$

Second, using eq 11 we find for $i \neq k$

$$\frac{\partial\vec{r}_k{}'}{\partial\vec{r}_i} = \frac{\alpha_k(4\pi\epsilon_0)}{q_k^\nu}\left(\frac{\partial\vec{E}_k}{\partial\vec{r}_i} + \sum_{m\neq k}^{N}\frac{\partial\vec{E}_k}{\partial\vec{r}_m{}'}\cdot\frac{\partial\vec{r}_m{}'}{\partial\vec{r}_i}\right) \tag{A4}$$

where using eq 16 and for $i \neq k$

$$\frac{\partial\vec{E}_k}{\partial\vec{r}_i} = -\frac{(q_i - q_i^\nu)}{4\pi\epsilon_0}\left\{\frac{\overleftrightarrow{\mathrm{I}}}{|\vec{r}_i - \vec{r}_k|^3} - \frac{3(\vec{r}_i - \vec{r}_k)\otimes(\vec{r}_i - \vec{r}_k)}{|\vec{r}_i - \vec{r}_k|^5}\right\} \tag{A5}$$

and

$$\frac{\partial\vec{E}_k}{\partial\vec{r}_i{}'} = -\frac{q_i^\nu}{4\pi\epsilon_0}\left\{\frac{\overleftrightarrow{\mathrm{I}}}{|\vec{r}_i{}' - \vec{r}_k|^3} - \frac{3(\vec{r}_i{}' - \vec{r}_k)\otimes(\vec{r}_i{}' - \vec{r}_k)}{|\vec{r}_i{}' - \vec{r}_k|^5}\right\} \tag{A6}$$

Neglecting the third term on the right in eq A2 and the second term on the right in eq A3 and keeping only the terms $m = i \neq k$ in the summation in eq A4, eqs 20, 23, and A1–A6 yield the first-order corrected expression for the atomic force on *i*, $\vec{f}_i^{(1)}$, as given in eq 24.

Acknowledgment. The authors thank Anna-Pitschna Kunz and Merijn Schenk for helpful discussions. Financial support from the National Center of Competence in Research (NCCR) in Structural Biology and from grant number 200021-109227 of the Swiss National Science Foundation is gratefully acknowledged.

References

(1) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236.

(2) Rick, S. W.; Stuart, S. J. Potentials and algorithms for incorporating polarizability in computer simulations. In *Reviews in Computational Chemistry;* 2002; Vol. 18, p 89.

(3) Ponder, J. W.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27.

(4) Yu, H. B.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.

(5) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(6) Vesely, F. J. *J. Comput. Phys.* **1977**, *24*, 361.

(7) Van Belle, D.; Couplet, I.; Prevost, M.; Wodak, S. J. *J. Mol. Biol.* **1987**, *198*, 721.

(8) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.

(9) Born, M.; Oppenheimer, R. *Ann. Phys. (Leipzig)* **1927**, *84*, 457.

(10) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(11) Straatsma, T. P.; McCammon, J. A. *Mol. Simul.* **1990**, *5*, 181.

(12) Drude, P. *The Theory of Optics*; Longmans, Green, and Co.: New York, 1902.

(13) Born, M.; Huang, K. *Dynamic Theory of Crystal Lattices*; Oxford University Press: Oxford, U.K., 1954.

(14) Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221.

(15) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

(16) Yu, H. B.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *121*, 9549.

(17) Yu, H. B.; Geerke, D. P.; Liu, H. Y.; van Gunsteren, W. F. *J. Comput. Chem.* **2006**, *27*, 1494.

(18) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.

(19) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**, *1*, 153.

(20) Geerke, D. P.; van Gunsteren, W. F. *Mol. Phys.* **2007**, DOI: 10.1080/00268970701444631.

(21) Barker, J. A.; Watts, R. O. *Mol. Phys.* **1973**, *26*, 789.

(22) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451.

(23) Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253.

(24) Hockney, R. W.; Eastwood, J. W. *Computer Simulation using Particles*, 2nd ed.; Institute of Physics Publishing: Bristol, U.K., 1988.

(25) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(26) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf Hochschulverlag: ETH Zürich, Switzerland, 1996.

(27) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596.

(28) *CRC Handbook of Chemistry and Physics*, 82th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2001−2002.

(29) Sprik, M. *J. Phys. Chem.* **1991**, *95*, 2283.

CT700164K

JCTC Journal of Chemical Theory and Computation

Results from an Early Polarization Model Based on Maxwell’s Invariant Multipole Form

Mihaly Mezei*

Department of Structural and Chemical Biology, Mount Sinai School of Medicine, NYU, New York, New York 10029

Received May 30, 2007

Abstract: This paper reviews the cooperative water model of Campbell and Mezei based on the Maxwellian form of multipole interaction. The Maxwellian form is described, and the algorithms and software for their implementation in both disordered and ordered phases are presented, followed by the specifics of the model. The model has been used in a number of calculations on various water clusters, liquid, and crystal models. The results of these calculations are briefly summarized, and their implications, relevant to polarization model in general, are discussed.

Introduction

Standard statistical mechanics offers a systematic treatment of cooperative interactions by partitioning the total energy into sums of two-body, three-body, etc. terms. This approach is quite general and does not take advantage of the specific nature of interactions. Terms beyond the two-body represent the cooperativity of the interaction.

For interacting molecules, the cooperativity is due to the deformability of the electron density upon interaction. For water clusters, Del Bene and Pople[1] demonstrated that such cooperativity is indeed significant. This led to the idea of representing the cooperativity of water−water interactions with interactions of induced moments albeit at first in a negative manner: the idea was first discarded out of hand based on the fact that dipoles represent cylindrical symmetry but that the charge distribution of water has only planar symmetry.[2] As the results reviewed here on Ice Ih calculations show, this skepticism is not unfounded in the sense that the contributions beyond dipole polarizability are not negligible. However, the use of dipole polarizability has proven to be very useful in modeling the cooperativity of water, as witnessed by subsequent work in the Stillinger Laboratory[3] as well as the model, contemporary to Stillinger’s, reviewed here.

Background

The Maxwellian Form of Multipole Interaction. The energy of the electrostatic interaction of two nonoverlapping charge distributions A and B can be expressed through a double Taylor series of $1/|\mathbf{r}_B - \mathbf{r}_A|$ about the two origins $\mathbf{O}_A$ and $\mathbf{O}_B$

$$E_{AB} = \sum_{N=0}^{\infty} \sum_{0 \le N_A + N_B \le N}^{\infty} E_{N_A,N_B} \tag{1}$$

with

$$E_{N_A,N_B} = \sum_{n_{A1}+n_{A2}+n_{A3}=N_A} \sum_{n_{B1}+n_{B2}+n_{B3}=N_B}$$

$$\prod_{k=1}^{3} (\nabla_{Ak})^{n_{Ak}} \prod_{k=1}^{3} (\nabla_{Bk})^{n_{Bk}} I_A(\mathbf{n}_A) * I_B(\mathbf{n}_B)(1/(|\mathbf{r}_B - \mathbf{r}_A|))|_{r_\gamma = O_\gamma} \tag{2}$$

where $\gamma = A$ or B

$$\nabla_{\gamma k} = (\partial/\partial x_{\gamma k}) \tag{3}$$

and

$$I_\gamma(\mathbf{n}_\gamma) = \int \rho_\gamma \prod_{i=1}^{3} x_i^{n_{\gamma i}}\, d\mathbf{x} / \prod_{k=1}^{3} n_{\gamma k}! \tag{4}$$

where ρ_γ is the charge density of system γ.

* Corresponding author e-mail: Mihaly.Mezei@mssm.edu.

10.1021/ct700130v CCC: $37.00
Published on Web 10/06/2007

The Maxwellian formalism is based on the fact[4] that for each center and for each N, there exists a unique set of real pole vectors (called characteristic directions) $\mathbf{s}_1^N$, ..., $\mathbf{s}_N^N$ and scalar multipole moments $p^{(N)}$, such that the sum of directional derivatives in eq 2, involving only vectors along the Cartesian axes, can be replaced by a single directional derivative along a general direction:

$$E_{N_A,N_B} = \{p_A{}^{(N_A)} * p_B{}^{(N_B)}/(N_A! * N_B!)\}* \prod_{i=1}^{N_A} (\mathbf{s}_i{}^{N_A}\bullet\nabla_A) * \prod_{i=1}^{N_B} (\mathbf{s}_i{}^{N_B}\bullet\nabla_B)(1/(|\mathbf{r}_B - \mathbf{r}_A|))|_{\mathbf{r}_\gamma=O_\gamma} \quad (5)$$

Use of eq 5 not only reduces the number of directional derivatives per charge distribution from $(N + 1)(N + 2)/2$ to N but also leads naturally to an extension where the calculation of the electric field generated by the multipoles (required for the calculation of induced dipoles), field gradients, etc. involves the same computational procedure as the calculation of the interaction between the multipoles: adding unit vectors as additional characteristic directions.[5] A formalism for the calculation of torques has been also developed[6] that includes higher order induced moments as well.

Use of eq 5, however, requires first the determination of the poles $p(N)$ and characteristic directions $\{\mathbf{s}_i^N\}$. For the case of $N = 2$ an explicit formula has been developed.[7] For $N > 2$ it was shown that the characteristic directions can be obtained from the roots of a polynomial of order $2N$, and the poles can be calculated based on calculating the directional derivatives with the newly derived characteristic directions at selected points.[8]

The calculation of the characteristic directions and of the calculation of the scalar poles from the moments $I(n)$ is implemented in the program **chardir**, that is part of the package **Maxwell**.[9] The same package also includes the program **moments** that evaluates $I(\mathbf{n})$ for any $\mathbf{n}$ from single-determinant wave functions generated by the software **POLYATOM**[10] or **Gaussian**.[11] The calulated moments $I(\mathbf{n})$ can be translated and/or rotated by the program **momtrnsf** of **Maxwell**.

Since eq 5 involves only derivatives its evaluation is, in principle, simple. However, the number of terms increases exponentially with N since derivation of each term in a fraction results in two terms. Fortunately, the exponential complexity can be reduced to polynomial order since eq 5 can also be evaluated recursively, resulting in an efficient algoritm.[12] This recursion has been implemented in the program **multipol** of the Maxwell package.[9]

Periodic Systems. Calculation of the electrostatic energy of a crystal presents a nontrivial mathematical problem. Even when the unit cell is neutral but has a finite dipole moment (which is the case in most systems), the infinite sum of dipole−dipole energies not only is slow to converge but also is conditionally convergent, i.e., dependent on the order of summation which can be interpreted as dependent on the crystal's shape. The classic solution to the problem is the one presented by Ewald[13] who represented the lattice sums with two different, fast converging sums, one in real space and the other in reciprocal space. A detailed analysis of the question of which shape does the Ewald sum correspond to has been presented by Campbell.[14] Subsequently, Campbell derived the Ewald summation formulas for multipoles of arbitrarily high order, using the Maxwellian formalism of multipole expansion.[15] The general form of the electrostatic energy of a crystal consisting of a set of simple translation lattices $\{T_i\}$ containing a set of charge distributions centered at $\{\mathbf{X}_c\}$ is given as

$$U_p = \frac{1}{2} \sum_{\{\mathbf{X}_c\}} \sum_{\{T_i\}} U(\mathbf{X}_c, T_i) \quad (6)$$

Each $U(\mathbf{X}_c,T_i)$ is obtained from a multipole sum over the multipole tensors of order N_X and N_T that can be written in the form[15]

$$U(\mathbf{X}_c, T_i, \langle N_X, N_T\rangle) = \frac{(P^{N_X})(P^{N_T})}{N_X!N_T!} \sum_{\{\nu\}} K(\nu, \mathbf{X}_c - \mathbf{X}_i, \langle N_X, N_T\rangle)\sigma(\mathbf{S}(\mathbf{X}_c), \mathbf{S}(\mathbf{X}_T)) \quad (7)$$

where the summation is over the set of non-negative integer triples $\nu = \langle \nu_1, \nu_2, \nu_3\rangle$ with $\nu_1 + \nu_2 + \nu_3 = N_X + N_T$ and $P(^{N_X})$, $\mathbf{S}(\mathbf{X}_c)$, and $P(^{N_T})$, $\mathbf{S}(\mathbf{X}_T)$ are the poles and characteristic directions at the site $\mathbf{X}_c$ and the lattice site T_i, respectively. The formulas for the so-called 'crystal constants' K and the directional derivatives σ are given in ref 15. Furthermore, for the calculation of the σ's a recursion, analogous to those used to evaluate eq 5, has been developed.[12]

The salient feature of this expression is that all geometric information about the crsytal is incorporated into the crystal constants K and that all information about the charge distributions is separated into the factor σ. The lattice sums (both direct and reciprocal space) contribute only to the crystal constants. This means that once the crystal constants are calculated for a given lattice, the calculation using different charge distributions or just different orientations of the same distribution can proceed without the need for additional lattice summation. Calculation of the crystal constants K and the recursion calculating the σ's have been implemented into the programs **cryscon** and **crysten**, respectively.[16] The calculation of the electrostatic energy of a crystal can be supplemented by the direct summation of r^{-k} ($k \geq 4$) terms with the program **cryspot**. These programs are also part of the **Maxwell** package.

Density Partitioning. There are two important facts worth remembering concerning the Taylor expansion represented by eqs 1−4 or, equivalently, eq 5. First, the series is only convergent if the charge distributions ρ_A and ρ_B do not overlap. Second, if a charge distribution ρ_γ includes only basis functions centered on the same point (usually a nucleus), then the multipole expansion of order $2n$ is exact where n is the highest order term in the wave function representing the density.[17] The nonoverlapping requirement suggests that the convergence can be improved if the molecular density is split up. This improvement comes, however, at the expense of increasing the number of interacting multipoles. The tradeoff between the two has been examined for water−water interactions using different density partition schemes.[17] The most efficient scheme took advantage of the chance to get exact (partial) results with a finite order: it assigned all overlap densities to the oxygen atom of the water molecule and assigned to each hydrogen only the density that comes from the basis functions centered on it—a partition called ‘very extreme split’.

Campbell−Mezei (CM) Water Model. The combination of the algorithms evaluating multipole interactions in the Maxwellian formalism to arbitrary high order and the (at that time) extensive data set of water dimer Hartree−Fock (HF) energies published by the Clementi Laboratory[18,19] led to the development of a fully ab initio cooperative model for water−water interactions.[20] The energy of a system of *n* water molecules was assumed to be of the form

$$U(n) = U_p(n) + U_i(n) + U_r(n) + U_d(n) \tag{8}$$

where $U_p(n)$ is the electrostatic energy of the *n* charge distributions representing the water molecules in their respective orientation, $U_i(n)$ is the additional electrostatic energy due to the induced moments, and $U_r(n)$ and $U_d(n)$ are the repulsion and dispersion contributions, respectively.

$U_p(n)$ represents the electrostatic interaction as approximated by the multipole expansion of the static wave function of Clement et al.[18,19] The density was partitioned according to the ‘very extreme split’[17] technique described above, resulting in second-order expansion of the density on the hydrogens and 10th-order expansion on the oxygen.

$U_i(n)$ represents the interaction energy due to the induced moments. The static contribution to electric field was calculated from the multipolar representation of the charge distribution by adding a unit vector to the characteristic directions, allowing the use of the algorithm calculating the interaction energy of multipoles to calulate the fields as well. $U_i(n)$ was calculated in the dipole approximation (i.e., induced moments of order higher than dipole were neglected). The polarizibility tensor α required for this calculation was obtained from the ab initio calculations of Liebmann and Moskowitz.[21] The induced dipoles were calculated using the method of Campbell.[5] This calculation involves the solution of a system of linear equations instead of the customarily employed iteration. The fact that the induced dipoles are obtained from such an unequivocal fashion suggests that the so-called ‘polarization catastrophy’ where the iteration diverges for centers too close is only an artifact of the iteration process and does not represent a physical phenomenon. Indeed, it is known that the iterative solution of a system of linear equations is not necessarily convergent.[22]

The terms $U_r(n) + U_d(n)$ represented the nonelectrostatic contributions to the interaction, resulting from exchange and dispersion effects. Since the HF energies do not include dispersion effects, the difference between our calculated electrostatic energy, $U_p(n) + U_i(n)$, and the HF energy calculated for the same conformation represents only the exchange repulsion term, $U_r(n)$. In the CM model the repulsion term $U_r(n)$ was represented by r_{AB}^{-k} terms whose coefficients were fitted to reproduce the difference between the $U_p(n) + U_i(n)$ terms and the corresponding HF energy. The exponent set was arrived at by testing several different values—the best fit was obtained with the exponent set {9, 12}. As expected of repulsion contributions, they were indeed positive for all dimer conformations in the Clementi dataset.

This left open the determination of $U_d(n)$. In subsequent work (vide infra) several empirical expressions were tested on ice lattice energies, and the one giving results closest to the experimenta data was selected.[23] The poles, characteristic directions, elements of the polarizability tensor, the parameters of the terms representing $U_r(n)$, and the dispersion function $U_d(n)$ found the best are also part of the **Maxwell** package.

Figure 1 shows the CM potential and its individual terms as a function of the O−O distance for a water dimer in linear hydrogen bonded orientation. It shows that the contribution of both the induced moments and the nonelectrostatic terms become negligible beyond ca. 3.5 Å. Furthermore, these two terms largely cancel in the 2.8−3.5 Å range, resulting in a remarkable good representation of the total energy with just the permanent electrostatic energy alone in the 2.8 Å−∞ range.

The computational effort required to evaluate the energy of an assembly of waters is quite high when compared to the widely used simple central-force models. While, in principle, the model could be incorporated into molecular dynamics simulations, this high cost excludes it from consideration. It is feasible, however, to evaluate the energy of a limited set of configurations (under periodic boundary conditions, if required). Such calculations were performed to show the feasibility of deriving a so-called ‘effective cooperative’ potential approximating a cooperative one by fitting the parameters of the pairwise additive effective cooperative potential to the cooperatively calculated *total* energy of a set of condensed-phase conformations.[24] As for incorporating the CM model into a Monte Carlo simulation, an additional, more fundamental difficulty, common to all cooperative models based on polarization, arises: the calculation of the polarization energy is an $O(N^2) - O(N^3)$ process, while the calculations at a usual Monte Carlo step without polarization requires only $O(N)$ effort. As a result, the incorporation of polarizability slows the calculation by an order of magnitude. Possible solutions to this problem are discussed by Mahoney and Jorgensen.[25]

Test of the CM Model. The HF nonadditivity given by Kistenmacher et al.[18] for the optimal closed trimer, −1.13 kcal/mol, is reproduced very well with the polarization model that gave −1.12 kcal/mol. Comparisons were also made with the trimer data set of Hankins et al.[2] The results, given in Table 1, show that the general trends are well represented by the polarization model. Since the basis set, hence the wave function, used in this trimer data set was different from the one used to build the model, the lack of quantitative agreement is understandable.

Results

Water Clusters. With a near-exact representation of the electrostatic interactions in terms of multipole interaction as well as a reasonable representation of the cooperativity through the calculation of induced dipoles several important questions can be answered: convergence of the multipole

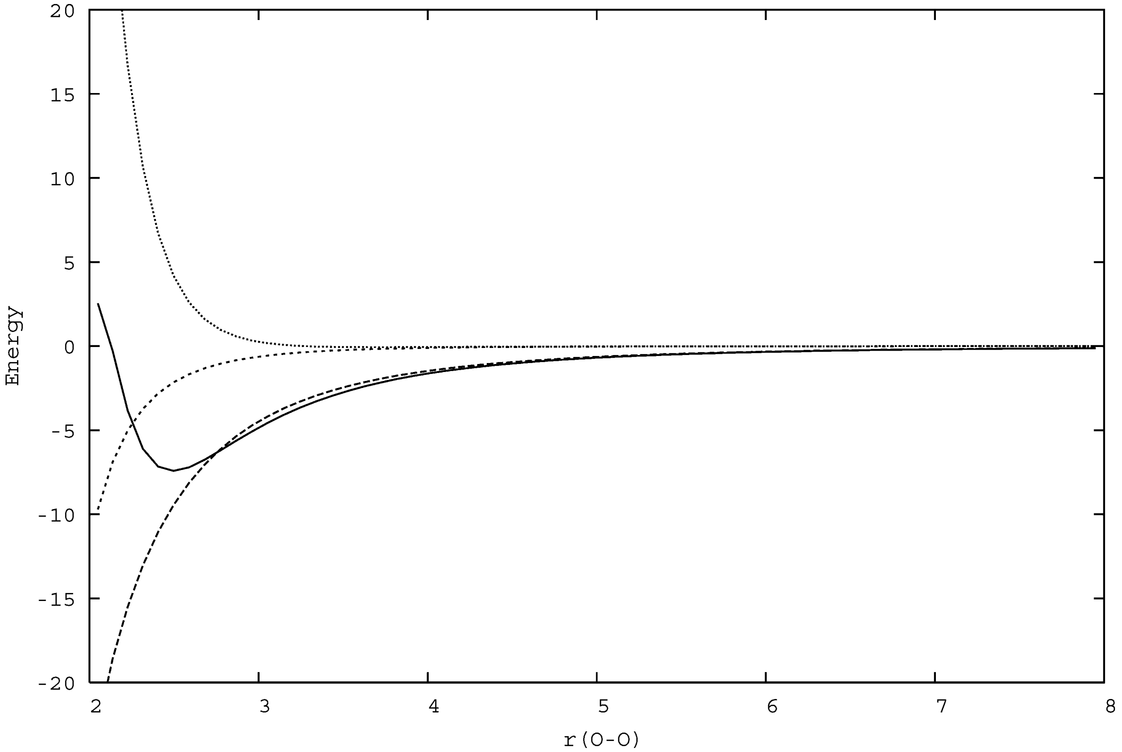


Figure 1. The contributions to the CM potential (in kcal/mol) for a linear dimer (orientation VI decribed in ref 30) as the function of the oxygen−oxygen distance r(O−O) (in Å). Full line: total energy; long-dashed line: permanent electrostatic energy ($U_p(n)$); short-dashed line: induced electrostatic energy ($U_i(n)$); dotted line: exchange and dispersion energy ($U_r(n) + U_d(n)$).

Table 1. Comparison of the HF and CM Model Nonadditivities[a−c]

R(O−O)	type	θ_{23}	E(H−F)	E(CM model)
2.76	sequential	−54.7°	−0.94	−0.77
3.15	sequential	−54.7°	−0.64	−0.44
2.76	double donor	−54.7°	1.30	1.09
3.15	double donor	−54.7°	0.38	0.42
2.76	double acceptor	−54.7°	0.77	1.26
3.00	double acceptor	−54.7°	0.37	0.68
3.39	double acceptor	−54.7°	0.10	0.28
3.00	double acceptor	−25.7°	0.49	0.29
3.00	double acceptor	−70.0°	0.36	0.85

[a] Energies are in kcal/mol. [b] Distances are in Å. [c] θ_{23} is the angle between the third water dipole and the O−O line between the second and third waters.

expansion, effect of neglecting cooperativity on the minimum energy geometries, and the estimate of the relative magnitude of three-body, four-body, etc., terms compared to the total cooperative energy and to the total energy. In a series of papers Campbell and Belford studied optimized water clusters[26,27] with $n = 4, 5, 6$ and clusters in conformations corresponding to the ones seen in Ice Ih[28] for $n \leq 33$. The highlights of these papers are summarized below.

The good performance of the CM model in reproducing the HF trimer energy was repeated in a study of the optimal tetramer,[26] where the HF and CM energies agreed within 3%. The magnitude of the induced dipole was found to be 0.25 D—this falls between the corresponding value for the dimer (0.12 D) and the average value observed in Ice Ih (0.55 D). It is also observed that the relative contribution of cooperativity (i.e., inluding the effects of the induced dipoles to the electric field) increases with cluster size: 1.7%, 2.9%, 5.2% for the optimal dimer, trimer, and tetramer, respectively. This compares with 15−20% for the different ice forms treated (vide infra).

Subsequently, clusters of 48 waters in Ice Ih configurations were studied.[28] Among the several results of this study was the partitioning of the calculated total energy into two-body, three-body, etc. contributions, allowing an estimate of the convergence of the alternative approach to multibody effects. Extrapolating the results to infinite lattice, the three- and four-body terms were found to contribute to the total lattice energy 21% and 3%, respectively. The contributions of five and higher order multibody terms were found to be 0.6% or less, with the exception of two conformations where the three- and four-body terms partially cancelled.

Still another study[27] determined a number of local minima in small water clusters containing three, four, and six waters. A major goal of that work was to find out if inclusion of cooperativity would affect the order of energies of the local minima found—the answer was affirmative for hexamers: the additive approximation favored a nearly planar ring, while the cooperative approximation favored an ice Ih-like staggered ring. Also, the optimal oxygen−oxygen distance was found to decrease with cluster size. For hexamers, it already fell into the range of vibrationally averaged oxygen−oxygen distances seen in condensed phases.

A Trifurcated Water Dimer. Ab initio calculations identified a low-energy water dimer conformation[29] that involves three hydrogen bonds: the O−H bond of one water is roughly antiparallel to the dipole vector of the other. Subsequently, several pairwise additive potentials and the CM model were used to compare the calculated dimer

Table 2. Comparison of the Ab Initio, Empirical, and Polarization Model Energies with Quantum-Mechanical Energies

		configuration					
		I	II	III	IV	V	VI
		3[a]	3[a]	2[a]	1[a]	1[a]	1[a]
ab initio models:	MCY[33]	0.41	−3.33	−3.72	−5.50	−5.25	−5.24
	YMD[34]	0.92	−3.59	−3.75	−5.34	−4.93	−5.16
empirical models:	ST2[35]	1.80	−3.05	−2.97	−6.44	−5.66	−5.99
	SPC[36]	5.16	−3.10	−3.91	−5.59	−5.36	−5.11
	TIP3P[37]	3.86	−3.45	−3.91	−5.48	−5.25	−5.10
	TIP4P[37]	5.07	−3.05	−4.29	−5.57	−5.60	−5.14
polarizable model:	CM	−3.70	−5.09	−5.89	−6.03	−6.06	−5.76
ab initio energy MP4SDQ/6-311G**		−3.08	−6.02	−5.84	−6.23	−6.14	−5.40

[a] Number of H bonds.

Table 3. Lattice Energy Contribution for Ice Forms[a]

form	U_p	U_i	$U_p + U_i$	U_r	U_d^a	U_d^b	U_d^c	U_t^a	U_t^b	U_t^c
Ih	−20.2	−7.0	−27.2	15.9	−3.8	−4.1	−6.6	−15.1	−15.5	−17.9
II	−17.5	−7.7	−25.2	12.6	−3.3	−3.9	−7.3	−15.9	−16.5	−19.9
IX	−18.3	−7.2	−25.5	13.8	−3.2	−3.9	−7.1	−15.0	−15.7	−18.9

[a] U_p, U_i, U_r, U_d: see eq 6; $U_t = U_p + U_i + U_r + U_d$; dispersion term parameters for U_d^a, U_d^b, and U_d^c from refs 40−42, respectively.

energies with the ab initio values at different conformations, ranging from the trifurcated to the dimer in the 'classical' linear hydrogen bond conformation.[30]

Table 2 shows the results for six conformations. Conformations I and II are both trifurcated; conformations IV−VI are linear dimers optimized with different levels of theory; and conformation III is an intermediate between linear and trifurcated. For each conformation the CM model is the closest to the ab initio values among the models tested. The difference between the CM model and the rest is particularly large for the trifurcated conformations. The poor performance of the models fitted to ab initio energies is understandable since trifurcated dimers were not in any of the data sets used for the fit. The poor performance of the empirical models, on the other hand, indicates that trifurcated dimers do not occur with significant probability in normal aqueous systems. This implies that this shortcoming is not affecting seriously calculations on aqueous systems. However, it was pointed out[30] that in situations where interactions with individual water molecules are important, these empirical potentials should be used with caution. The good performance of the CM model is particularly impressive since it was also derived without using any trifurcated structure. This supports the notion that the polarization approach can be effective in the modeling of intermolecular interactions. At a more general level, the comparison highlights the point that, in some situations, explicit calculation of the cooperativity is necessary.

It should be mentioned that a comment to this work questioned its validity since no counterpoise correction was applied to correct for the basis-set superposition error.[31] In answering this comment,[32] it was pointed out that the no correction was applied for the neglect of zero-point vibration energy either, and it was shown that the two corrections work in the opposite direction, thus reinforcing our conclusion about the potential significance of such trifurcated conformations.

Calculations on Disordered Ice Ih. The lattice energy of Ice Ih was calculated by Campbell in the dipolar approximation as a tool to assess the electrostatic nature of hydrogen bond.[38] The work was later extended to multipoles of order six.[39]

With the development of the recursion algorithm[12] to evaluate eqs 4 and 5 the permanent multipole energies were calculated up to multipole order 14, using $r_o = 2.741$ Å. The calculations were performed on Ice Ih crystals with disordered water orientations. The disorder was represented by all the possible tetrahedral orientations of waters that still satisfy the Bernal−Fowler rule (exactly one hydrogen between neighboring oxygens). All possible arrangements of a 16-site unit cell were considered, resulting in 55 classes of conformations with distinct permanent multipole energy; 10 of these had zero total dipole.[23,39]

From the convergence of the series it was estimated that the truncation error is about 0.03 kcal/mol. The spread in the permanent multipole energy for the 55 classes was 0.14 kcal/mol that was reduced to 0.1 kcal/mol for the zero-dipole subclass. While this spread was steadily decreasing as the multipole order was increased toward 14, it remained higher than the estimated trunction error. This indicates that there indeed is a residual energy difference among the different water orientations—a question that was debated at that time.

The induced dipoles and the induced energies were also calculated.[5] This increased the spread in the electrostatic energy to 1.1 kcal/mol and 0.6 kcal/mol for the whole and nonpolar class, respectively. This increase in the spread serves as an indicator of the fact that the truncation of the induced multipoles at the dipole level introduced a non-negligible error and points out the importance of considering higher order polarizibilities.

Comparison of the Energies of Different Ice Forms. The ice crystal energy calculations[23] have been extended to two additional ice forms with ordered hydrogen positions: Ice II and Ice IX. The dispersion contribution was calculated with three different approximations.[40−42] The comparison with experiment, however, is only valid if the zero-point energy, that is neglected in these calculation, is known, and the experimental value can be adjusted accordingly. This was the case only for Ice Ih, giving −14.1 kcal/mol. Comparison of the three different dispersion approximations shows that that of Zeiss and Meath[40] gave the best approximation. Subsequent work with the CM model used this dispersion contribution throughout.

Calculation of the Energies of Ice Ih Bjerrum Defects. Hassan and Campbell performed a series of calculations[43] to study the energetic penalty of a Bjerrum defect in an Ice Ih crystal: either there are two hydrogens between neigboring oxygens or there are none. They considered both 'formal defects', i.e., the waters were not allowed to relax due to the repulsion caused by the defect and allowed relaxation of the waters in involve in the defect as well as their neighbors. The system studied involved altogether 27 water molecules placed in an arrangement corresponding to the Ice Ih lattice with all but the central water pair satisfying the Bernal−Fowler rule. Energy calculations involved both a pairwise additive potential fit to ab initio data[18] and the cooperative CM model.

Optimization of the molecular orientations and positions using the additive model was found to reduce the defect energy by ∼40% The optimization involved the defect pair and their neighbors, while the rest provided the boundary effect. It was found that relaxing the orientational degrees of freedom contributes significantly more to lowering the defect energy than relaxing the positions.

The optimization with the additive potential was followed by orientational optimization with the cooperative CM model. This yielded an additional 10% lowering of the defect energy.

These calculations also brought into focus the lower quality of the repulsive contributions in most analytical potentials. There are two major reasons for this. The simpler and easier to remedy source is the limited sampling of the repulsive conformations in the database used to fit the potentials. More difficult is the establishment of an adequate objective function to the fit since the energy surface can vary by orders of magnitude if strongly repulsive orientations are considered.

Charge Transfer. Molecular mechanics force fields that are used in most large-scale computer simulations use fixed charges on the interaction centers and thus are not equipped to handle charge transfer. Thus, it is of interest to examine the magnitude of this neglected contribution.

A set of free energy simulations was performed for monovalent cations in water and chloroform using ab initio derived parameters[44] for the ions.[45] The calculated solvation free energies were strongly underestimated when compared with experiment. This was expected, with the reasoning that polarization is neglected in the calculations.[44] To test if polarization can indeed account for the shortfall, the CM model was used to calculate the additional contribution to the solvation free energy, by evaluating the induced dipole energy of a selected set of conformations from the simulation. For this application, the program **multipol** has been extended to handle periodic boundary conditions and to use a lower order of expansion for more distant pairs of waters. The calculation showed that including polarization can indeed reduce the discrepancy between calculation and experiment, but a significant gap still remained.

It was proposed that the source of the remaining discrepancy between calculation and experiment is the neglect of charge transfer. To support this claim, ab initio calculations were performed on selected configurations containing the ion and its first solvation shell. A Mulliken population analysis showed that significant charge-transfer exists: Li^+ lost 0.32 e and Na^+ lost 0.27 e to the waters surrounding it. Note, that the Mulliken population analysis is known to be a rather simple approach to charge density partitioning. However, it was used in the present work only to demonstrate the *presence* of charge transfer and not to quantitate it.

Subsequently, van der Vaart and Merz have published analogous calculations with similar results; the conclusion held even when a more sophisticated charge partitioning scheme was used.[46] They also found that even for a hydrogen bond there is a measurable amount of charge transfer.[47]

Representation of the Exchange Repulsion. Another ingrained property of the widely used molecular mechanics force fields is the representation of the nonbonded interactions as the sum of eletrostatic and Lennard-Jones terms, i.e., terms of the form r^{-1}, r^{-6}, and r^{-12}. The r^{-6} term is usually identified with the dispersion energy and the r^{-12} term with the repulsion. While the dispersion term has its physical justification, the repulsive term's *raison d'être* is the fact that $r^{-12} = (r^{-6})^2$ and thereby it is easy to compute. Also, as discussed above, the quality of these contributions is generally lower than that of the rest.

There are two problems with this approach. First, the functions r^{-1}, r^{-6}, and r^{-12} are close to be linearly *dependent*. This was seen from least-square fit calculations aiming at obtaining the best fitting coefficients: the matrix of the resulting system of linear equation is usually very ill-conditioned.[24] This is not just a technical problem that can be simply overcome with better numerical algorithms or higher precision arithmetics—it means that a wide range of coefficient sets can give a virtually identical fit, making the identification of individual terms with physical meaning unreliable. This, in turn, is not just a question of ‘esthetics’ since the generally assumed transferability of atomic parameters from one molecule to another largely relies on the fact that these parameters have a physical interpretation.

The other problem is the well-known fact that the exhange repulsion is an exponential function of r (see, e.g., ref 48); the r^{-12} term results in too steep a repulsion. For simulations around room-temperature it is not a significant problem. However, one of the main justifications of the explicit inclusion of cooperativity into computer simulations is that such representation can remain valid over a wide range of thermodynamic conditions as opposed to the force fields where the effect of cooperativity is mapped to pairwise additive terms that are only valid in the thermodynamic vicinity of the state the functions were parametrized. This advantage, however, is only realized if the rest of the potential functions are parametrized well enough to represent the full range of thermodynamic conditions under consideration.

The fact that the r^{-12} repulsion is inadequate for this task was highlighted by simulations at high temperature. A comparison of (T, V, N) and (T, P, N) ensemble simulations on three polarizable and two nonpolarizable water models found that the polarizable models underestimate the density by 10−50%, leading to the suggestion that the functional form of the repulsive term has to be changed.[49] Subsequently, the comparison was extended to more models and to simulation in the Gibbs ensemble[50] to determine the critical point of each model.[51] While in this test several of the polarizable models gave critical densities close to the experimental value, the overall comparison between the pairwise additive and polarizable models failed to show the expected superiority of the polarizable models. These comparisons reinforce the suggestion that, for optimum performance, the repulsion term has to be revised concurrently with the development of polarizable models.

Summary and Conclusions

Work in the Campbell Laboratory has showed that the Maxwellian formalism[4] is an elegant and efficient way to treat electrostatic interactions in the multipole expansion approximation. The necessary formulas[15] and algorithms[8,12] for their use to describe intermolecular interactions in clusters as well as crystals have been developed and implemented in the software package **Maxwell**.[9]

The formalism was also used to derive an ab initio cooperative water potential based on Hartree−Fock energies and representing cooperativity with dipole polarizability.[20] Subsequently, the model was used in a variety of studies on water clusters[24,26−28,30,43] and ices.[23] These calculations showed that dipole polarizability can treat the cooperative contribution to water−water interactions reasonably well and also quantitated the limitations inherent in this approximation.

Calculations on Ice Ih showed that the orientational disorder results in a finite energy range even when the orientations obey the Bernal−Fowler rule and even when the unit cell dipole is zero. The calculation of induced dipole energies showed that the dipole approximation to the cooperative contributions is not fully converged.[23]

An important result from calculations with the CM cooperative model is the recognition of the significance of charge transfer. Other calculations led to the recognition of the importance of adequate treatment of the repulsion contribution.

Acknowledgment. The author is deeply indebted to Professor Edwin S. Campbell for providing him with an opportunity to start his scientific carrier in the United States and for guidance in this work as well as for his continued much cherished friendship. The original development of the model was supported by NIH Grant 1 ROl GM 20436-01 to Prof. E. S. Campbell.

References

(1) Del Bene, J.; Pople, J. A. Theory of molecular interactions. I. Molecular orbital studies of water polymers using a minimal Slater-type basis *J. Chem. Phys.* **1970**, *52*, 4858−4866.

(2) Hankins, D.; Moskowitz, J. W.; Stillinger, F. S. Water molecule interactions. *J. Chem. Phys.* **1970**, *53*, 4544−4554. Erratum: **1973**, *59*, 995.

(3) Stillinger, F. H.; David, C. W. Polarization model for water and its ionic dissociation products. *J. Chem. Phys.* **1978**, *69*, 1473−1484.

(4) (a) Hobson, E. W. *The theory of spherical and ellipsoidal harmonics*; Cambridge University Press: London/New York, 1931; p 119. (b) Hobson, E. W. *The theory of spherical and ellipsoidal harmonics*; Cambridge University Press: London/New York, 1931; pp 135−137.

(5) Campbell, E. S. Method for the calculation of the induced dipole moment and the permanent multipole−induced moment interaction in crystals. *Helv. Phys. Acta* **1967**, *40*, 387−388.

(6) Mezei, M.; Campbell, E. S. Torque algorithms: the permanent multipole and induced dipole vector contributions in a set of charge distributions. *J. Comput. Phys.* **1982**, *47*, 245−257.

(7) Campbell, E. S. Determination of a single quadrupole for a charge distribution. *J. Chem. Phys.* **1952**, *20*, 666−667.

(8) Mezei, M.; Campbell, E. S. A Computational procedure for obtaining the poles of a spherical harmonics of order N; application to the multipole expansion of electrostatic interaction. *J. Comput. Phys.* **1976**, *20*, 110−116.

(9) Campbell, E. S.; Mezei, M. Maxwell: multipole expansion for condensed phases. URL: http://inka.mssm.edu/∼mezei/maxwell (accessed 09/01/2007).

(10) Barnett, M. P. Mechanized molecular calculations − the POLYATOM system. *Rev. Mod. Phys.* **1963**, *35*, 571−572.

(11) Gaussian. URL: http://www.gaussian.com/ (accessed 09/01/2007).

(12) Efficient construction of directional derivatives of a function of a vector magnitude and Maxwell's invariant multipole form. *J. Comput. Phys.* **1976**, *21*, 114−122.

(13) Ewald, P. P. Evaluation of optical and electrostatic lattice potentials. *Ann. Phys. (Berlin)* **1921**, *64*, 253−287.

(14) Campbell, E. S. Existence of a "well defined" specific energy for an ionic crystal; justification of Ewald's formulae and of their use to deduce equations for multipole lattices. *J. Phys. Chem. Solids* **1963**, *24*, 197−208.

(15) Campbell, E. S. Computer calculations for a perfect crystal of multipoles. *J. Phys. Chem. Solids* **1965**, *26*, 1395−1408.

(16) Mezei, M.; Campbell, E. S. Computer algorithms and programs for permanent multipole and induced dipole vectors in crystals. *J. Comput. Phys.* **1978**, *29*, 297−301.

(17) Mezei, M.; Campbell, E. S. Efficient multipole expansion: Choice of order and density partitioning techniques. *Theor. Chim. Acta* (*Berlin*) **1977**, *43*, 227−237.

(18) Kistenmacher, H.; Lie, G. C.; Popkie, H.; Clementi, E. Study of the structure of molecular complexes. VI. Dimers and small clusters of water molecules in the Hartree−Fock approximation. *J. Chem. Phys.* **1974**, *61*, 546−561.

(19) Popkie, H.; Kistenmacher, H.; Clementi, E. Study of the structure of molecular complexes. IV. The Hartree−Fock potential for the water dimer and its application to the liquid state. *J. Chem. Phys.* **1973**, *59*, 1325−1336.

(20) Campbell, E. S.; Mezei, M. Use of a non−pair−additive intermolecular potential function to fit quantum−mechanical data on water molecule interactions. *J. Chem. Phys.* **1977**, *67*, 2338−2344.

(21) Liebmann, S. P.; Moskowitz, J. W. Polarizabilities and hyperpolarizabilities of small polyatomic molecules in the uncoupled Hartree−Fock approximation. *J. Chem. Phys.* **1971**, *54*, 3622−3631.

(22) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes in C: The art of scientific computing*; Cambridge Unversity Press: Cambridge, U.K., 1988.

(23) Campbell, E. S.; Mezei, M. A cooperative calculation and analysis of electric fields, induced dipole vectors and lattice energies for rotationally ordered ices IX, II and disordered Ih. *Mol. Phys.* **1981**, *41*, 883−905.

(24) Mezei, M. On the possibility of obtaining an effective pairwise additive intermolecular potential via an ab initio route by fitting to a cooperative model of condensed phase configurations. *J. Phys. Chem.* **1991**, *95*, 7043−7049.

(25) Mahoney, M. W.; Jorgensen, W. L. Rapid estimation of electronic degrees of freedom in Monte Carlo calculations for polarizable models of liquid water. *J. Chem. Phys.* **2000**, *114*, 9337−9349.

(26) Campbell, E. S.; Belford, D. A cooperative calculation of geometries, energetics and electrical properties of water trimers and tetramers. *Theor. Chim. Acta* **1982**, *61*, 295−301.

(27) Campbell, E. S.; Belford, D. Geometries, energies, and electrostatic properties of nonadditively optimized small water clusters. *J. Chem. Phys.* **1987**, *86*, 7013−7024.

(28) Belford, D.; Campbell, E. S. Multibody energy components for clusters of water molecules and ice Ih. *J. Chem. Phys.* **1984**, *80*, 3288−3296.

(29) Dannenberg, J. J. An AM1 and ab initio molecular orbital study of water dimer. *J. Phys. Chem.* **1988**, *92*, 6869−6871.

(30) Mezei, M.; Dannenberg, J. J. An evaluation of water−water analytical potentials in the region of low energy trifurcated structures. *J. Phys. Chem.* **1988**, *92*, 5860−5861.

(31) Aastrand, P. O. A.; Wallqvist, A.; Karlstroem, G. On the basis set superposition error in the evaluation of water dimer interactions. *J. Phys. Chem.* **1991**, *95*, 6395−6396.

(32) Dannenberg, J. J.; Mezei, M. Reply to the comment on the application of basis set superposition error to ab initio calculation of water dimer. *J. Phys. Chem.* **1991**, *95*, 6396−6398.

(33) Matsuoka, O.; Clementi, E.; Yoshimine, M. CI study of the water dimer potential surface. *J. Chem. Phys.* **1976**, *64*, 1351−1361.

(34) Yoon, B. J.; Morokuma, K.; Davidson, E. R. Structure of ice Ih. Ab initio two- and three-body water−water potentials and geometry optimization. *J. Chem. Phys.* **1985**, *83*, 1223−1231.

(35) Stillinger, F. H.; Rahman, A. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.* **1974**, *60*, 1545−1557.

(36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: 1981.

(37) Jorgensen, W. L.; Chandrashekar, J.; Madura, J. D.; Impey, R.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(38) Campbell, E. S. Hydrogen bonding and the interaction of water molecules. *J. Chem. Phys.* **1952**, *20*, 1411−1420.

(39) Campbell, E. S.; Gelernter, G.; Heinen, H.; Moorti, V. R. G. Interpretation of the energy of hydrogen bonding; permanent multipole contribution to the energy of ice as a function of the arrangement of hydrogens. *J. Chem. Phys.* **1967**, *46*, 2690−2707.

(40) Zeiss, G. D.; Meath, W. J. The H2O-H2O dispersion energy constant and the dispersion of the specific refractivity of dilute water vapour. *Mol. Phys.* **1975**, *30*, 161−169.

(41) Lie, G. C.; Clementi, E. Study of the structure of molecular complexes. XII. Structure of liquid water obtained by Monte Carlo simulation with the Hartree−Fock potential corrected by inclusion of dispersion forces. *J. Chem. Phys.* **1975**, *62*, 2195−2199.

(42) Jeziorski, B.; van Hemert, M. Variation−perturbation treatment of the hydrogen bonds between water molecules. *Mol. Phys.* **1976**, *31*, 713−729.

(43) Campbell, E. S.; Hassan, R. The energy and structure of Bjerrum defects in ice Ih determined with an additive and a nonadditive potential. *J. Chem. Phys.* **1992**, *97*, 4326−4335.

(44) Chandrasekhar, J.; Spellmeyer, D. C.; Jorgensen, W. L. Energy component analysis for dilute aqueous solutions of lithium$^+$, sodium$^+$, fluoride$^-$, and chloride$^-$ ions. *J. Am. Chem. Soc.* **1984**, *106*, 903−910.

(45) Maye, P. V.; Mezei, M. Calculation of the free energy of solvation of the Li^+ and Na^+ ions in water and chloroform. *J. Mol. Struct.* (*THEOCHEM*) **1996**, *362*, 317−324.

(46) van der Vaart, A.; Merz, K. M., Jr. Charge transfer in biologically important molecules: comparison of high−level ab initio and semiempirical methods. *J. Comput. Chem.* **2000**, *77*, 27−43.

(47) van der Vaart, A.; Merz, K. M., Jr. Charge transfer in small hydrogen bonded clusters. *J. Chem. Phys.* **2002**, *116*, 7380−7388.

(48) Stone, J. A. *The theory of intermolecular forces*; Oxford University Press: Oxford, 1996.

(49) Jedlovszky, P.; Richardi, J. Comparison of different water models from ambient to supercritical conditions: A Monte Carlo and molecular Orstein−Zernike Study. *J. Chem. Phys.* **1999**, *110*, 8019−8031.

(50) Panagiotopoulos, A. Z. Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble. *Mol. Phys.* **1987**, *61*, 813−826.

(51) Jedlovszky, P.; Vallauri, R. Liquid−vapor and liquid−liquid phase equilibria of the Brodholt−Sampoli−Vallauri polarizable water model. *J. Chem. Phys.* **2005**, *122*, 081101.

CT700130V

JCTC Journal of Chemical Theory and Computation

Coarse-Grained Protein Model Coupled with a Coarse-Grained Water Model: Molecular Dynamics Study of Polyalanine-Based Peptides

Wei Han*[,†] and Yun-Dong Wu*[,†,‡]

Department of Chemistry, The Hong Kong University of Science & Technology, Kowloon, Hong Kong, China, and State Key Lab of Molecular Dynamics and Stable Structures, College of Chemistry, Peking University, Beijing, China

Received June 24, 2007

Abstract: The coupling of a coarse-grained (CG) protein model with the CG water model developed by Marrink et al. (*J. Phys. Chem. B* **2004**, *108*, 750) is presented. The model was used in the molecular dynamics studies of Ac-(Ala)$_6$-Xaa-(Ala)$_7$-NHMe, Xaa = Ala, Leu, Val, and Gly. A Gly mutation in the middle of polyalanine is found to destabilize the helix and stabilize the hairpin by favoring a type-II′ turn and probably to speed up hairpin folding. The simulations allow us to derive thermodynamic parameters of, in particular, the helical propensities (*s*) of amino acids in these polyalanine-based peptides. The calculated *s* values are 1.18 (Ala), 0.84 (Leu), 0.30 (Val), and <0.02 (Gly) at 291 K, in excellent agreement with experimental values (R^2=0.970). Analyses using a structural approach method show that the helical propensity difference of these amino acids mainly comes from solvation effect. Leu and Val have lower helical propensities than Ala mainly because the larger side chains shield the solvation of helical structures, while Gly has a much poorer helical propensity mainly due to the much better solvation for the coil structures than for the helical structures. Overall, the model is at least about 10^2 times faster than current all-atom MD methods with explicit solvent.

Introduction

The problem of protein folding is an active area of experimental and theoretical research.[1,2] Computer simulations, as indispensable tools in this area, provide microscopic insights to complement the interpretation of experimental observations.[3−5] All-atom simulations, which explicitly represent every atom of proteins and solvent molecules, can reveal the maximum details but are computationally demanding. The shortest time scale of the folding of a small protein is about tens of microseconds, but most all-atom simulations can only be carried out up to microseconds.[6−8] Thus, all-atom simulations are currently impractical in describing the protein folding completely.

An alternative to all-atom simulations is coarse-grained (CG) simulation, in which a group of atoms is reduced to a single particle.[9] In addition, solvent molecules are usually implicitly represented. The gain in simulation speed comes from a large reduction in the number of particles and smoother interparticle potentials. Thus, coarse-grained simulations have been useful in the study of protein folding. There have been many successful applications of coarse-grained simulations that characterize proteins at different levels of details.[10−19] The minimalist model characterizes each amino acid with one single particle or one particle for the backbone and one for the side chain.[9−13] This model has provided insights into general folding mechanisms, but it has less predictive values since it relies on preknowledge of native structures to impose Gō-type biases onto interparticle potentials.[10,11] Models between the minimalist level and the all-atom level have been devised to reproduce reasonable Ramachandran maps and anisotropic hydrogen bond (HB)

* Corresponding author e-mail: hwer@ust.hk (W.H.) and chydwu@ust.hk (Y.-D.W.).

[†] The Hong Kong University of Science & Technology.

[‡] Peking University.

10.1021/ct700151x CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/02/2007

potentials.[14−18] These models allow coarse-grained simulations to qualitatively study the protein folding or aggregation without any biased potential.

HB and hydrophobic interactions play essential roles in protein stability and dynamics. They strongly depend on the local environment, such as the solvation level, and have a many-body character.[20] These interactions in some CG models remain pair wise additive.[15,16] Interestingly, Takada et al.[14] introduced HB and hydrophobic potentials that depend on local densities of protein particles. These densities indicate the extent of exposure of particles involved in HB or hydrophobic interaction.

Another way to model the solvent effect is to use a coarse-grained (CG) solvent model. Shelley et al. proposed a CG model to study lipid aggregation in solutions.[21] Their model basically has two types of CG particles (waterlike and oil-like). A group of water molecules can be represented by one CG waterlike particle. Water−oil interactions are unfavorable. The interactions are favored with like particles. The solvation effect can be explicitly taken into account. This model has been further developed and calibrated by Marrink et al.[22] to enable semiquantitative or quantitative comparisons with experiments. The simulation speed is increased by at least 10^3-fold. This model has been successfully applied to the study of membrane fusion.[23,24] Bond et al.[25] and Shih et al.[26] have shown that some behaviors of membrane proteins can be studied by using minimalist models of CG proteins with this CG solvent/lipid model. The current limit of the CG solvent/lipid model is that it cannot deal with atomistic details of proteins.[22,26] However, recent studies by Voth's group[27,28] indicated that all-atom force fields and CG models can work together well when each water molecule is represented by one particle.

In this paper, we present a CG protein model at an intermediate resolution level. This model is developed in tandem with the CG solvent model by Marrink et al.[22] and has been applied to the molecular dynamics studies of Ac-(Ala)$_6$-Xaa-(Ala)$_7$-NHMe (Xaa = Ala, Leu, Val, and Gly). We show that (1) parameters can be optimized in a systematic way that is compatible to the procedures by Marrink et al.; with such procedures, it is easy to incorporate new models and parameters into our CG model in a consistent manner; (2) structures in atomistic details, such as α-helix, β-hairpin, and β-turn, can fold and/or refold properly in polypeptides without biased potentials; (3) the simulations are fast enough to obtain the equilibrium thermodynamic and kinetic properties of polypeptides; and (4) the properties from optimized parameters are comparable to those from experiments.

Models and Methods

The CG Protein Model. As shown in Figure 1a, in the backbone of our CG protein, each heavy atom with its attached hydrogen(s) is explicitly represented by one CG particle, resembling the model by Ding et al.[16] "R" is the side-chain group that determines the identity of the amino acid (aa). Four kinds of amino acids are studied in this paper, including Ala, Val, Leu, and Gly. Their representations are pictured in Figure 1b. All heavy atoms of Ala and Val are explicitly considered, while the isopropyl group of Leu, which connects to C_β carbon, is replaced by one CG particle at its centroid position. The potential energy of this CG model is described by eq 1

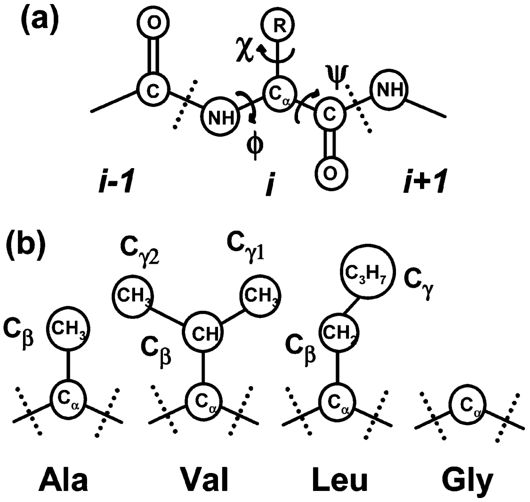


Figure 1. The scheme of the CG protein model.

$$V_{\mathrm{Total}} = V_{\mathrm{Angle}} + V_{\mathrm{Improper}} + V_{\mathrm{Torsion}} + V_{\mathrm{loc-cdW}} + V_{\mathrm{vdW}} + V_{\mathrm{HB}} \quad (1)$$

where V_{Total} can be partitioned into the bonded terms, $V_{\mathrm{Angle}} + V_{\mathrm{Improper}} + V_{\mathrm{Torsion}} + V_{\mathrm{loc-vdW}}$, and the nonbonded terms, which are the remaining parts.

Bonded Interactions. The bonded interactions are defined as the interactions between the particles connected by the direct bonds or separated by less than four chemical bonds. The direct bonds are constrained by the LINCS algorithm[29] with the bond length of r_0. The interactions between particles i and j, which both connect to particle k, are described by eq 2 with $K_{\mathrm{Angle}} = 72$ kcal mol^{-1} rad^{-2}. The equilibrium value of $\angle(i\text{-}k\text{-}j)$ is θ_0.

$$V_{\mathrm{Angle}} = K_{\mathrm{Angle}}(\theta - \theta_0)^2/2 \quad (2)$$

In order to maintain the planarity of a carbonyl group or the chirality of an sp^3 carbon, eq 3 is used

$$V_{\mathrm{Improper}} = K_{\mathrm{Improper}}(\xi - \xi_0)^2/2 \quad (3)$$

where ξ is the dihedral of the four particles involving in planar or the chiral groups. K_{Improper} is 72 kcal mol^{-1} rad^{-2}.

The torsional angle is defined as the dihedral angle, $\angle$-$(i\text{-}j\text{-}k\text{-}l)$, of four particles, i, j, k, and l, which are connected by three successive chemical bonds. They are known to be critical to determining the local conformational features of amino acids. Following Takada et al.,[14] we use the combination of two kinds of potentials (eqs 4 and 5) to describe the torsion.

$$V_{\mathrm{Torsion}} = K_{\mathrm{Torsion}}(1 + \cos(n\phi - \phi_0)) \quad (4)$$

$$V_{\mathrm{loc-vdW}} = \sum_{1-4\ \mathrm{relationship}} 4\epsilon_{\mathrm{loc}}\left(\frac{\delta_{\mathrm{loc},ij}^{12}}{r^{12}} - \frac{\delta_{\mathrm{loc},ij}^{6}}{r^{6}}\right) \quad (5)$$

In eq 4 ϕ is $\angle(i\text{-}j\text{-}k\text{-}l)$, and n is the multiplicity of the periodic potential. Equation 5 describes the van der Waals (vdW)

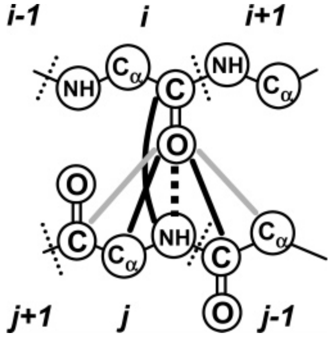


Figure 2. The scheme of the HB interaction between amide units. The *i* and *j* are residue numbers.

overlap between particles *i* and *l*. Such an interaction is atomic in nature and is weaker than nonlocal vdW interactions, where atoms are separated by more than three bonds.[14] In our model, $\epsilon_{\rm loc}$ is set to be 0.22 kcal/mol, smaller than that for nonlocal vdW interactions, which will be introduced later. The vdW radii, $r_{\rm vdW-loc}$, are also somewhat smaller than the nonlocal vdW radii.

Nonbonded Interactions. Nonbonded interactions in our model comprise nonlocal vdW interactions (separated by >3 bonds) and HB interactions.

Lennard-Jones (LJ) potentials (eq 6) are used to describe nonlocal vdW interactions. Following the treatment by Marrink et al.,[22] electrostatic interactions between groups with partial charges are implicitly incorporated into nonlocal vdW interactions.

$$V_{\rm vdW} = \sum_{i<j} 4\epsilon_{ij}\left(\frac{\delta_{ij}^{12}}{r^{12}} - \frac{\delta_{ij}^{\;6}}{r^{6}}\right) \quad (6)$$

Another important nonbonded interaction, the HB interaction between amide units (Figure 2), is known to be crucial for α-helix, β-sheet, and β-turn conformations of proteins. The HB interaction is normally described with both dipole−dipole interactions and vdW interactions in current all-atom force fields. Takada et al. devised a simplified way to model this anisotropic interaction in their CG model.[14] The basic idea is to allow not only the attraction between the carbonyl (carbonyl is represented as one particle in their model) of one peptide unit and the nitrogen of another but also the auxiliary repulsions between particles adjacent to the carbonyl/nitrogen of one peptide unit to the nitrogen/carbonyl of the other. Our HB model is similar but slightly different since a carbonyl group is represented by C and O particles. This representation has been argued to be better for HB geometry.[16] Indeed Takada et al. reported that polyalanine folded into helices and also fold an experimentally designed peptide[30] folded into a helix bundle using the model.[14] In addition, Ding et al. also used the HB potential to successfully fold a miniprotein of 20 aas, namely Trp-cage,[31] into its NMR conformation.[16] These examples reflect the applicability of the HB potential.

Figure 2 illustrates the HB potential described by eq 7, where the black bold dotted lines indicate the attraction, and the black bold solid lines indicate the auxiliary repulsion. Nevertheless, because our HB model adopts the repulsive interaction between O particles and C/C$_\alpha$ particles, an extra repulsion exists (gray lines in Figure 2) although this repulsion is irrelevant to the HB interaction. This can weaken the expected HB interaction, especially in β-sheets, where the distances between particles associated with the extra repulsion are quite short. To counteract this extra repulsion, we increased the C−C, C−C$_\alpha$, and C$_\alpha$−C$_\alpha$ attraction ($\epsilon_{\rm vdw}$) between residues *i* and *j* if |*i*-*j*| > 2.

$$V_{\rm HB} = \sum_{|i-j|>2}\left[4\epsilon_{\rm attr}\left(\frac{\delta^{12}_{{\rm O}i-{\rm NH}j}}{r^{12}_{{\rm O}i-{\rm NH}j}} - \frac{\delta^{6}_{{\rm O}i-{\rm NH}j}}{r^{6}_{{\rm O}i-{\rm NH}j}}\right) + 4\epsilon_{\rm rep}\frac{\delta^{12}_{{\rm O}i-{\rm C}\alpha j}}{r^{12}_{{\rm O}i-{\rm C}\alpha j}} + 4\epsilon_{\rm rep}\frac{\delta^{12}_{{\rm O}i-{\rm C}j-1}}{r^{12}_{{\rm O}i-{\rm C}j-1}} + 4\epsilon_{\rm rep}\frac{\delta^{12}_{{\rm C}i-{\rm NH}j}}{r^{12}_{{\rm C}j-{\rm NH}j}}\right] \quad (7)$$

Parameter Optimization and Optimized Parameters. The optimization procedures for all parameters are shown in Figure 3. Our strategy is to optimize the parameters for the backbone (Ala, Gly) first by molecular dynamics (MD) simulations of polyalanine. The parameters for Ala have enough details to describe Gly. Once the backbone parameters are obtained, they remain unchanged. For other amino acids, such as Leu and Val, that are presented in this paper, we need to optimize the parameters for the part of the side chains that are attached on C$_\beta$ of Ala.

Parameters for Bonded Interactions. The values for all bonded interactions are listed in Table 1. Specifically, they are obtained as follows:

The parameters of bond length (r_0) and bond angle (θ_0) related to backbone atoms, such as C, C$_\alpha$, NH, O, and C$_\beta$, and the side-chain atoms of Val are available in the average results of X-ray crystal structures,[32] which are basically the same with the parameters used in previous studies.[14−18] The improper (ξ_0) parameters are set here to ensure that the C$_i$, C$_{\alpha i}$, N$_{i+1}$, and O$_i$ are placed in the same plane and that the amino acid has an L-configuration.

The torsional potentials ($K_{\rm Torsion}$, ϕ_0, and n) and the local vdW parameters ($\delta_{{\rm loc},ij}$) for backbone atoms were optimized as follows: (1) An alanine dipeptide was simulated with varying torsional and local vdW parameters so that its Ramanchadran (ϕ, ψ) map achieved separated regions, such as α, β, and PPII regions, and the population deviation in each region was less than 10−15% compared to previous studies (see the next paragraph). (2) Ac-(Ala)$_{14}$-NHMe in the CG water was simulated repeatedly (about 150 times) with varying local vdW parameters. Each trial simulation lasted for about 1 μs. Parameters were optimized so that (a) the full helical structure could refold for more than five times and the full hairpin structure (defined in Data Analyses) can refold twice from fully extended conformation and that (b) the two secondary structures, once formed, could last for nanoseconds before they were broken again.

With these parameters, the (ϕ, ψ) plots of Ac-Xaa-NHMe in CG water at 300 K are shown in Figure 4. The Ala dipeptide has about 38% β conformation ((ϕ, ψ) at (−135°± 45°,135°±45°)), about 16% PPII conformation ((ϕ, ψ) at (−45°±45°,135°±45°)), and about 27% total helical conformation ((ϕ, ψ) at (−90°±90°,−45°±45°)). This is in good

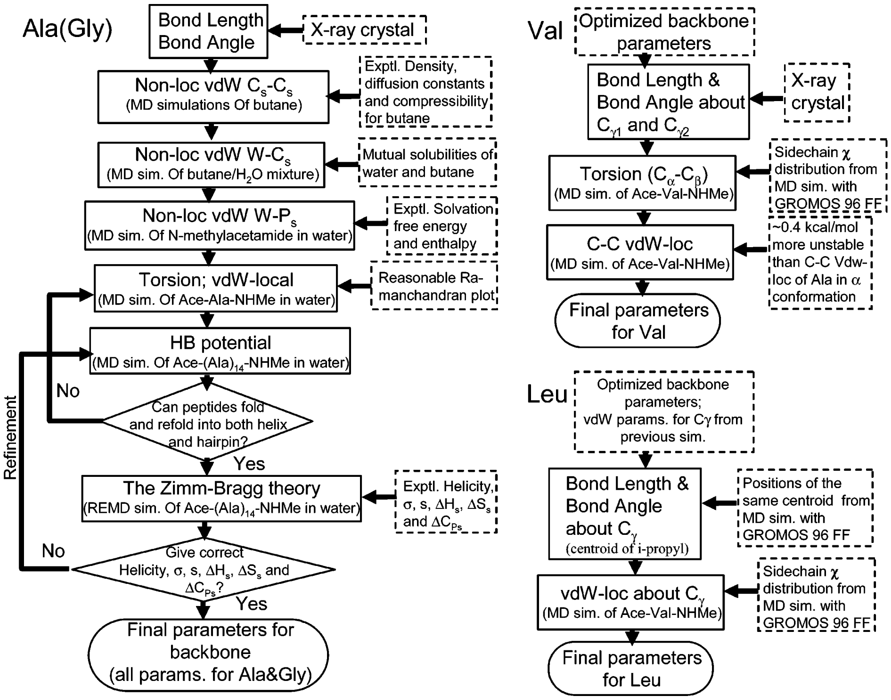

Ala(Gly)
Bond Length Bond Angle
X-ray crystal
Non-loc vdW C_s-C_s (MD simulations Of butane)
Exptl. Density, diffusion constants and compressibility for butane
Non-loc vdW W-C_s (MD sim. Of butane/H_2O mixture)
Mutual solubilities of water and butane
Non-loc vdW W-P_s (MD sim. Of N-methylacetamide in water)
Exptl. Solvation free energy and enthalpy
Torsion; vdW-local (MD sim. Of Ace-Ala-NHMe in water)
Reasonable Ramanchandran plot
HB potential (MD sim. Of Ace-$(Ala)_{14}$-NHMe in water)
Refinement
No
Can peptides fold and refold into both helix and hairpin?
Yes
The Zimm-Bragg theory (REMD sim. Of Ace-$(Ala)_{14}$-NHMe in water)
Exptl. Helicity, σ, s, ΔH_s, ΔS_s and ΔC_{Ps}
No
Give correct Helicity, σ, s, ΔH_s, ΔS_s and ΔC_{Ps}?
Yes
Final parameters for backbone (all params. for Ala&Gly)
Val
Optimized backbone parameters
Bond Length & Bond Angle about $C_{\gamma1}$ and $C_{\gamma2}$
X-ray crystal
Torsion (C_α-C_β) (MD sim. of Ace-Val-NHMe)
Sidechain χ distribution from MD sim. with GROMOS 96 FF
C-C vdW-loc (MD sim. of Ace-Val-NHMe)
~0.4 kcal/mol more unstable than C-C Vdw-loc of Ala in α conformation
Final parameters for Val
Leu
Optimized backbone parameters; vdW params. for Cγ from previous sim.
Bond Length & Bond Angle about C_γ (centroid of i-propyl)
Positions of the same centroid from MD sim. with GROMOS 96 FF
vdW-loc about C_γ (MD sim. of Ace-Val-NHMe)
Sidechain χ distribution from MD sim. with GROMOS 96 FF
Final parameters for Leu

Figure 3. Flowcharts of parameter optimization. Each rectangle with a solid outline contains one optimization step. Each rectangle with a dotted outline contains the data used for optimization and/or their sources. Each optimization step is performed based on the parameters optimized from previous steps. W indicates CGW particles; C indicates CG nonpolar particles; C_s indicates a small $CH_{x\ (x=0-3)}$ group or carbonyl carbon; P_s indicates an O atom or NH group.

Table 1. Parameters of r_0, θ_0, and ξ_0

bond	r_0 (nm)	bond	r_0 (nm)
C_α−C	0.152	C_β−$C_{\gamma1/\gamma2}$	0.153
C−N	0.133	C_β−C_γ (Leu)	0.194
C_α−N	0.145	C−O	0.123
C_α−C_β	0.153		
angle	**θ_0 (deg)**	**angle**	**θ_0 (deg)**
N−C_α−C	111.6	C_α−C−O	121.0
C_α−C−N	117.5	N−C−O	124.0
C−N−C_α	120.0	C_α−C_β−$C_{\gamma/\gamma2}$	111.0
N−C_α−C_β	110.0	$C_{\gamma/1}$−C_β−$C_{\gamma/2}$	111.0
C−C_α−C_β	110.0	C_α−C_β−C_γ (Leu)	124.0
improper	**ξ_0 (deg)**	**improper**	**ξ_0 (deg)**
$C_{\alpha i}$−N_i−C_i−$C_{\beta i}$	35.3	C_i−$C_{\alpha i}$−N_{i+1}−O_i	0.0

agreement with the results from the OPLS/AA/L force fields,[33] which produce slightly more β (31%) conformation than PPII (25%) conformation.[34] It is in moderate agreement with a recent experiment based on PR/FTIR, which found that $(Ala)_3$ can dominantly adopt β and PPII conformations at a 1:1 ratio.[35]

The torsional and pair interactions involving the side chains of Val were optimized by matching the distribution of N−C_α−C_β−$C_{\gamma1}$ of the CG model with that from the GROMOS96 force field[36] through the simulations of Val dipeptide in water. The matching results are shown in Figure S1 of the Supporting Information (SI).

For Leu, a similar way was used to optimize the parameters about bond C_β−C_γ, angle C_α−C_β−C_γ, and the dihedral angles involved with the side chain, where C_γ represents the centroid of the isopropyl group. The r_0 of C_β−C_γ and the θ_0 of C_α−C_β−C_γ (Table 1) are just the positions of the single narrow peaks in the distributions of C_β−C_γ and C_α−C_β−C_γ of the Leu dipeptide from the GROMOS96[36] simulations (Figure S2a, Supporting Information), respectively. The detailed matching results about the distribution of the N−C_α−C_β−C_γ dihedral angle are given in Figure S2b, Supporting Information.

In addition, a local C_i−C_{i+1} vdW interaction of β-branched aa, such as Val, is modified (Table 2). A model building reveals that when ϕ is about −60° (Figure 5a), hydrogen atoms of C_γs are too close to hydrogen atoms of backbone amides in the three side-chain rotamers. The distance is about 0.21 nm, shorter than previously reported repulsive vdW diameters (0.24−0.26 nm) for hydrogen.[37] Such repulsion is absent in extended conformations ($\phi < -120°$). This effect cannot be explicitly considered if amide hydrogen is ignored. We therefore handled the repulsion by enlarging δ_{loc} of the local vdW C_i−C_{i+1} interaction. With molecules A and B (Figure 5b) used to model the ϕs of Val and Ala, respectively, quantum mechanics calculations at the B3LYP/6-311++G** level show that the energy difference between A with $\phi = -80°$ to −60° and A in its global minimum is about 0.4−0.6 kcal/mol higher than that for B. This gives an estimation of the repulsive effect for Val compared to Ala. According to dipeptide simulations, our modified $\delta_{\text{loc},Ci-Ci+1}$ on average leads to about 0.4−0.5 kcal/mol more C_i−C_{i+1} repulsion for Val than it does for Ala when ϕ is −80° to −60°.

Parameters for Nonlocal vdW Interactions. One of the purposes of this work is to make our CG model compatible with the CG solvent model by Marrink et al.[22] In the original CG model, ϵ_{ij} (eq 6) of the vdW interactions has discrete

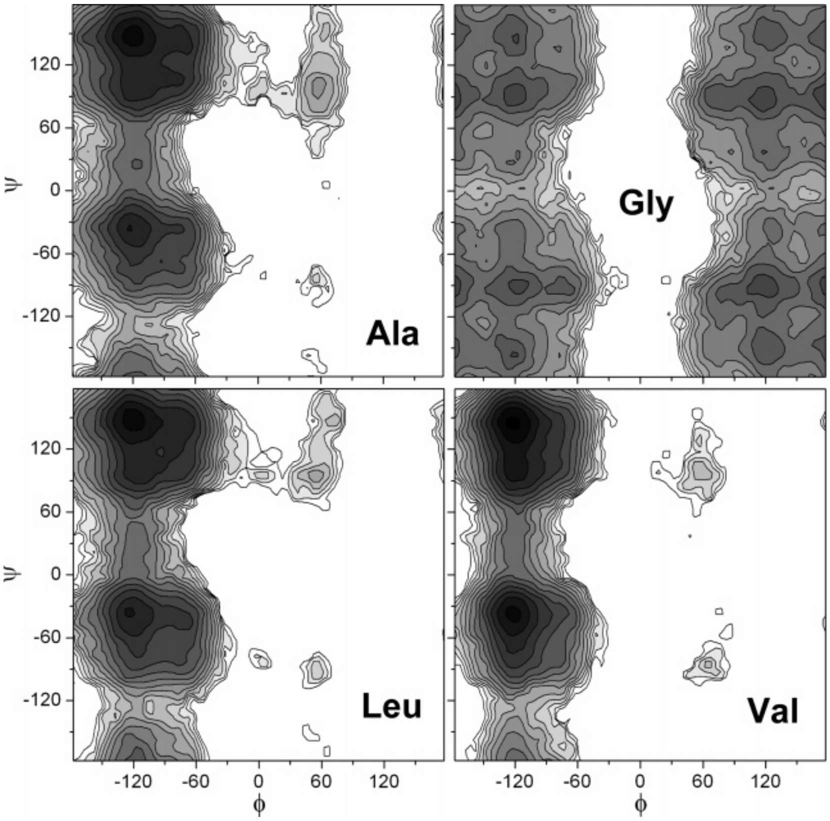


Figure 4. The Ramachandran plots of Ala, Gly, Leu, and Val. The free energy interval for the contours is 0.25 kcal/mol. The darker region has lower free energy.

Table 2. K_{Torsion}, ϕ_0, n, and $r_{\mathrm{vdW-loc}}$

torsional angle	K_{Torsion} (kcal/mol)	ϕ_0 (deg)	N
$C_{\alpha i-1}-C_{i-1}-N_i-C_{\alpha i}$	10.00	180.0	2
$C_{i-1}-N_i-C_{\alpha i}-C_i$	0.20	180.0	6
$N_i-C_{\alpha i}-C_i-N_{i+1}$	0.20	0.0	6
$N_i-C_{\alpha i}-C_{\beta i}-C_{\gamma 1,i/\gamma,i}$	1.20	0.0	3

atom type	$r_{\mathrm{vdW-loc}}$[a] (nm)
O	0.130
NH	0.145
C	0.165/0.155/0.168[b]
$CH/CH_2/CH_3$	0.165
C_3H_7	0.185

[a] $\delta_{\mathrm{loc},ij}$ is equal to $r_{\mathrm{vdw-loc},i}+r_{\mathrm{vdw-loc},j}$, where i and j are atom types. [b] 0.165 is for the interaction between C and other particles with different kinds of atom types, 0.155 is for the C−C interaction in the non-β-branched aas, and 0.168 is $^1/_2\delta_{\mathrm{loc},Ci-1-Ci}$ for β-branched aa i.

levels, which are 1.20, 1.01, 0.82, 0.63, and 0.44 kcal/mol in the metric of 0.19 kcal/mol. δ_{ij} is uniformly 0.47 nm. These parameters were optimized for CG particles that represent four atoms (butane or equivalent).[25,26] However, the values of ϵ_{ij} and δ_{ij} of the original model seem to be too large for our CG particles, some of which represent only one atom. Therefore, the vdW radii of our CG particles are taken from the statistical survey of crystal structures.[38] The isopropyl particle in Leu is composed of three carbon atoms, and its r_{vdw} is taken from Shelley et al.[21] In addition, we used a method similar to that of Marrink et al.[22] to reoptimize the ϵ_{ij} of interactions between the CG solvent and our CG protein. As shown in Figure 3, nonlocal vdW parameters are optimized so that simulations can reproduce important physical properties of pure liquid and liquid mixtures. All nonlocal vdW parameters are listed in Table 3.

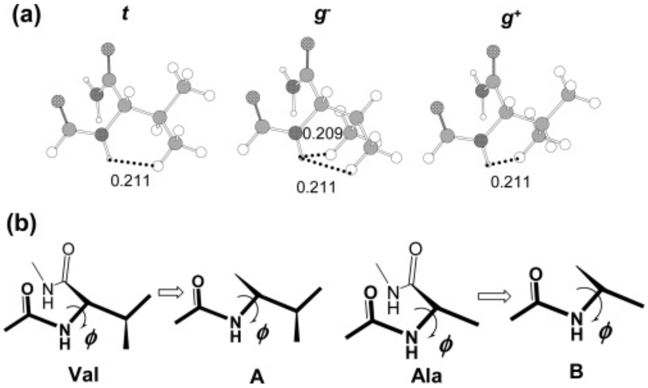


Figure 5. (a) Val in α conformation with its side chain in three rotamers and (b) model molecules for Val (A) and Ala (B).

The ϵ_{ij} for interactions between small hydrophobic particles ($CH_{x(x=0-3)}$) (called C_s particles) is obtained by simulating a system of 400 CG butane molecules, each of which is composed of four C_s particles. ϵ_{ij} between CG water (CGW) and C_s particles is obtained by simulating mixtures of 400 CGW molecules and 400 CG butane molecules (900 ns for each simulation). The simulations are conducted at $T = 300$ K and $P = 1$ atm with a Nose-Hoover thermostat[39,40] and a Parrinello-Rahman pressure bath.[41] The resulting values for ϵ_{ij} of both C_s-C_s and CGW$-C_s$ interactions are 0.25 kcal/mol.

With these parameters, the density, compressibility, and self-diffusion constant of butane as well as free energies to

Table 3. ϵ_{ij} and r_{vdW}

ϵ_{ij} (kcal/mol)	W[a]	$CH_x(x=0-3)$	O	NH	C_3H_7
W(P)	1.20				
$CH_{x(x=0-3)}$	0.25	0.25			
O	1.20	0.25	0.25		
NH	1.20	0.25	0.25[a]	0.25	
C_3H_7	0.44	0.44	0.25	0.25	0.82

type	r_{vdW}[b] (nm)	type	r_{vdW}[b] (nm)
$CH_{x(x=1-3)}$	0.185	W	0.235
C	0.165	NH	0.165
C_3H_7	0.220	O	0.140

[a] The ϵ_{ij} of O_i and NH_j is valid only if $|i\text{-}j| < 3$, where i and j are residue numbers. [b] δ_{ij} in eq 6 is equal to $r_{vdW,i} + r_{vdW,j}$.

Table 4. Experimental and Calculated Physical Properties of Butane and Mutual Transfer Free Energies between Butane and CG Water at 300 K

	exptl	I[e]	II
density/g·cm^{-3}	0.58[a]	0.68	0.74
compressibility/10^{-5} bar^{-1}	>17[a]	28	8
diffusion[b]/10^{-5} cm^2·s^{-1}	>5[b]	1.9	1.8
$\Delta G_{But(But\rightarrow W)}$/kcal·mol^{-1}	5.5[c]	5.4	7.2
$\Delta G_{W(W\rightarrow But)}$/kcal·mol^{-1}	6.0[d]		6.5

[a] Measured at 293 K.[43] [b] Obtained from the slope of the mean squared displacement (MSD) curve in the long time limit. Experimental values are from ref 44. [c] Measured at 298 K.[44] [d] Measured at 294 K.[46] [e] Obtained by using one CG particle to represent one butane molecule (ref 22). [f] No detectable mutual solubility.

transfer water into butane, $\Delta G_{W(W\rightarrow But)}$, or butane into water $\Delta G_{But(But\rightarrow W)}$, which is derived from mutual solubilities of water and butane,[42] were calculated from simulations. The simulated properties are reasonably comparable to those from experiments and from the work by Marrink et al. (Table 4[43−46]), except that compressibility is about half of the experimental value and $\Delta G_{But(But\rightarrow W)}$ is 1.7 kcal/mol higher than the experimental value, indicating that $\epsilon_{ij,Cs-Cs}$ may need to be further reduced. The vdW interactions for other small particles are here supposed to be similar to those of C_s particles. Therefore, for simplicity, we add one more energy level, $\epsilon_{ij} = 0.25$ kcal/mol, which satisfies the metric of 0.19 kcal/mol in the original energetic system,[22] for interactions between small particles except for the interacting particles participating in HB interactions.

In order to parametrize ϵ_{ij} of the interactions between O and NH of amide and CGW, we investigated the solvation of N-methylacetamide, which is a model compound widely used in the study of peptide hydration.[47−49] An N-methylacetamide molecule was placed in a box with 350 CG water particles. Its solvation free energy, ΔG_{sov}, enthalpy ΔH_{sov}, and entropy $T\Delta S_{sov}$ were calculated with the thermal integration (TI) method[50] as described in the Appendices. The TI can normally compute solvation free energy accurately if that the sampling is adequate.[51] ΔG_{sov}, ΔH_{sov}, and $T\Delta S_{sov}$ with different ϵ_{ij} from the TI are shown in Table 5. Compared to experimental values (−10.1 kcal/mol), ΔG_{sov} is the best (−9.2 kcal/mol)) when $\epsilon_{ij} = 1.20$ kcal/mol.

Although TI is good for accurate calculation of the solvation free energy, it can only be applied to small or rigid molecules.[52] TI becomes impractical for polypeptides that are highly flexible. For this reason, an approximate method based on the scaled particle theory (SPT) together with the assumption of a solvent accessible surface (SAS) for polyatomic molecules was used to estimate the solvation of polypeptide in our study (Appendices).[53−57] To examine the difference between the TI and SPT approaches for polypeptides, the solvation properties of N-methylacetamide were also computed by the SPT method and are listed in Table 5. The results reveal that the SPT based approach overestimates ΔG_{sov} by about 4−5 kcal/mol compared with the TI method. The difference is from the calculation of the free energy, ΔG_{cav}, to make a solute-sized cavity. It is calculated to be about 10 kcal/mol by the TI method, which is close to the value of 9.2 kcal/mol derived from a theoretic treatment of experimental data.[58] But the ΔG_{cav} values by the SPT approach is only 5.6 kcal/mol, about half of the TI value. Indeed, a previous study found that the SPT method could underestimate the work to create cavity.[59] Similarly, the calculated $T\Delta S_{cav}$ (−2.1 kcal/mol) and ΔH_{cav} (3.5 kcal/mol) by the SPT method are also about twice the $T\Delta S_{cav}$ (−4 to −5 kcal/mol) and ΔH_{cav} (5−6 kcal/mol) by the TI method. Consequently, when the solvation of polypeptides was calculated by the SPT with SAS assumption, $T\Delta S_{cav}$ and ΔH_{cav} were increased by 1-fold as a rough correction for the SPT approach.

Table 5. Experimental (298 K) and Calculated (300 K) ΔG_{sov}, ΔH_{sov}, $T\Delta S_{sov}$, ΔG_{cav}, and $<U_{int}>$ of N-Methylacetamide (kcal/mol)

	ΔG_{sov} −10.1[a]		ΔH_{sov} −17.1[c]		$-T\Delta S_{sov}$ 7.0		ΔG_{cav}[d]		
exptl	TI[b]	SPT[b]	TI	SPT	TI	SPT	TI	SPT	$<U_{int}>$[e]
$\epsilon = 1.39$	−11.6	−16.5	−15.6	−18.7	4.0	2.1	10.5	5.6	−22.1
$\epsilon = 1.20$	**−9.2**	**−13.8**	**−14.2**	**−15.9**	**5.0**	**2.1**	**10.2**	**5.6**	**−19.4**
$\epsilon = 1.01$	−6.7	−11.0	−11.3	−13.2	4.5	2.1	9.8	5.6	−16.6
$\epsilon = 0.82$	−4.4	−8.4	−8.4	−10.5	4.0	2.1	9.6	5.6	−14.0

[a] From ref 47. [b] "TI" means thermal integration; "SPT" means the scaled particle theory with SAS assumption. [c] From refs 48 and 49. [d] Free energy to create a cavity in solvent, which is approximately equal to $\Delta G_{sov} - <U_{int}>$. [e] Average interaction energy between solvent and solute.

It should be noted that in our present model, we used the discrete energy levels of nonbonded parameters that were used by Marrink et al. for simplicity. In our further work, we will remove this restriction so that solvation free energy may be better calculated by fine-tuning parameters such as $\epsilon_{ij,Cs-Cs}$ and $\epsilon_{ij,W-O/NH}$.

Parameters for HB Interactions. The optimized HB parameters are as follows: $\epsilon_{attr} = 3.35$ kcal/mol; $\epsilon_{rep} = 1.08$ kcal/mol; $\delta_{Oi-NHj} = 0.24$ nm; $\delta_{Oi-C\alpha j} = \delta_{Oi-Cj-1} = 0.29$ nm; $\delta_{Ci-NHj} = 0.338$ nm. In addition, the enhanced $\epsilon_{vdW,ij}$ for C and C_α particles is 0.63 kcal/mol.

Finally, during the optimization, we found that π-helices ($i\rightarrow i+5$ HB) were significantly sampled in the simulations, and sometimes their population even overwhelms that of α-helices ($i\rightarrow i+4$ HB). In real polypeptides, however, they should be rare. This may be because that the current HB model cannot differentiate between the two helical structures very well. To differentiate these two helices, their structures were inspected in detail. We found that the distance between

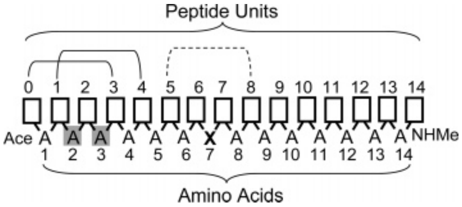


Figure 6. The scheme of Ac-(Ala)$_6$-Xaa-(Ala)$_7$-NHMe. The squares represent peptide units, and the arches indicate the helical hydrogen bonds between the peptide units.

$C_{\alpha i}$ and $C_{\beta i+4}$ in the π-helices in our model is quite short (~0.37 nm), which is not the case for α-helices. We therefore can increase δ_{vdW} for $C_{\alpha i}$ and $C_{\beta j}$ ($|i\text{-}j| \geq 2$) to selectively destabilize the π-helices. It turns out that when σ_{vdW} is 0.435 nm, the π-helices are significantly weakened and the α-helices are strengthened, while the β-hairpins are not affected.

Models. The polypeptide models used in the simulations were Ac-(Ala)$_6$-Xaa-(Ala)$_7$-NHMe, where Xaa is Ala if the peptide is a polyalanine (polyA) and Leu, Val, and Gly if the peptide is a Xaa mutant of polyA. The peptide chain includes two kinds of groups (Figure 6), a peptide unit (−CO−NH−) (PU, represented by squares) and an amino acid unit (C_α and side chain) (AU, represented by letters). Each PU/AU has two neighboring AUs/PUs. The numbering scheme for PUs and AUs is illustrated in Figure 6.

Simulation Setup. The simulations were performed with the GROMACS 3.3.1 package.[60] A peptide with a helical conformation was placed into a dodecahedron box with ~1100 CG water particles. The shortest distance between the peptide and the edges of the box was 1.5 nm. The vdw interaction had a cutoff of 1.2 nm, and it was smoothed to 0 from 0.9 to 1.2 nm. The temperature and pressure were controlled by a thermostat and a pressure bath, with coupling constants of 0.1 and 0.5 ps, respectively.[61] The time interval to integrate the Newton equations was 6 fs, and the neighboring list was updated every 10 steps. The whole system was subjected to 5000 steps of steep descent optimization and then to a 200 ps of pre-equilibrium at 300 K and 1 atm with the peptide constrained. The system was then heated at 340 K with the peptide relaxed for 100 ns. The generated conformations with no apparent helical or hairpin structures were used as starting points for the long simulations.

Replica exchange molecular dynamics (REMD) simulations provide an efficient platform to perform equilibrium simulations,[62,63] which is beneficial to parameter optimization. Our REMD simulations contained 14 replicas with temperatures ranging from 291 to 436 K at 1 atm. Each replica started with a different conformation generated from the heating simulation at 340 K. Exchanges were attempted every 1 ps. Each REMD simulation lasted for 200 ns, while the results of the last 150 ns of the simulations were analyzed.

Mass Scaling in the REMD Simulations. In the original CG solvent model,[22] the time interval to integrate the Newton equations is 25−45 fs. Since some particles in our model represent only a single atom and are connected directly by strong covalent bonds, they are so light that the time interval can only be up to 6−10 fs in order to keep the simulations from crashing. This crashing problem is especially serious for REMD since high-temperature (up to 436 K) simulations are involved. To avoid the problem we quadrupled the masses of peptide particles. This allows the interval for integration to be kept at 10 fs. Although this action changes the dynamics of those motions heavily dependent on mass, it may be feasible for the REMD simulations, whose dynamics have lost their physical meaning. The effect of mass scaling on the thermodynamics of the simulations at physiological temperatures is reported in the Results and Discussion.

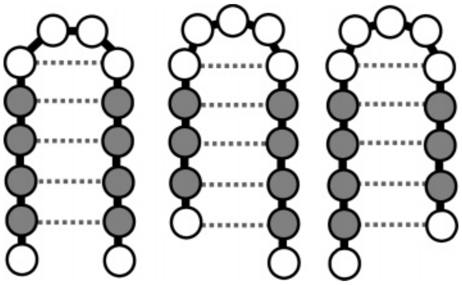


Figure 7. Three possible hairpin topologies with turns in the middle of the peptide. Dotted lines denote the conditions (<0.65 nm) between C_α atoms of different aas used to identify hairpin topologies.

Data Analyses

Definition of Helical Structures. Helical structures are formed by at least three successive aas in the α conformation, which is defined as an aa conformation with its (ϕ, ψ) at (−60°±30°,−47°±30°), as suggested by García et al.[64] The middle residue of the three aas is considered to be in the helical state in our study. As shown in Figure 6, if aas 1, 2, and 3 form a helical structure, aa 2 is in a helical state and the CO group of PU 0 will form a HB with the NH group of PU 3 (the solid arch), which is called as a helical HB. For a given peptide conformation, its helical content, h_{HLX}, is the ratio of the number of helical HBs of this conformation to the maximum number of helical HBs allowed for this peptide, which is 12. It represents the extent of the formation of helical structures.

Definition of Hairpin Structures. An aa conformation with (ϕ, ψ) at (−135°±45°,135°±45°) is defined as a β conformation. If an aa and its two neighbors all have β conformations, this aa is considered to be in a β-strand. h_{HP} is a score (0−1) to measure the extent of hairpin formation, which is defined as the ratio of the number of aas in β-strands to the maximum possible number (eight for hairpins, gray circles in Figure 7) if the peptide can have any hairpin topology shown in Figure 7.

Besides, the reverse turn of hairpins is defined by the backbone dihedrals of two aas *i* and *i+1*, with $(\phi, \psi)_i$ and $(\phi, \psi)_{i+1}$ at (−60°±45°,−30°±45°) and (−90°±45°,0°±45°) for Type I, (60°±45°,30°±45°) and (90°±45°,0°±45°) for Type I′, (−60°±45°,120°±45°) and (80°±45°,0°±45°) for Type II, and (60°±45°,−120°±45°) and (−80°±45°,0°±45°) for Type II′.[65] If aas *i* or *i+1* is in a helical structure, it will not be considered for turns.

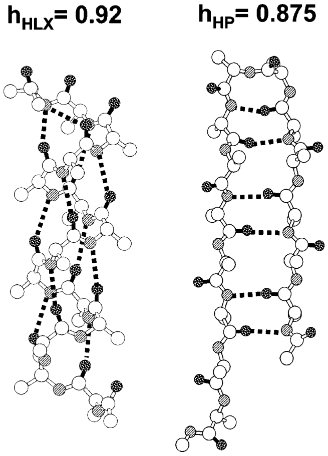


Figure 8. Typical α-helix and β-hairpin in the polyA simulation.

Table 6. Structural Properties of Peptides in Long Simulations at 310 K for PolyA and Its G Mutant

	polyA	G mutant	polyA′	G′ mutant
simulation time^a (ns)	5000	4000	4900	5400
$<h_{HLX}>$	0.238^b	0.072^b	0.229^b	0.087^b
$<h_{HP}>$	0.007	0.024	0.007	0.032
refold (HLX)^c	40	48	40	50
refold (HP)^c	7	12	8	16

[a] Due to the coarse-graining, a nanosecond here does not mean the real time. [b] 0.242 for polyA and 0.053 for G mutant from AGADIR.[68] [c] Occurrence of the refolding of helices or hairpins.

Other Analyses. The extraction of the helical parameters s and σ of the Zimm-Bragg theory[66,67] is given in Appendix A. The method to calculate solvation effect is in Appendix B. The description about structural approaches is in Appendix C.

Results and Discussion

The α-Helix and β-Hairpin in Long Simulations. To examine the quality of the current CG model in reproducing structures and thermodynamics of peptides, long simulations were performed with Ac-$(Ala)_{14}$-NHMe (polyA) and Ac-$(Ala)_6$-Gly-$(Ala)_7$-NHMe (G mutant). This allowed us to examine the ability of our model to discern sequence-dependent properties of peptides.

The α-helix and β-hairpin are our major targeted peptide structures. The h_{HLX} and h_{HP} (see Methods and Models) are taken as the indicators of these structures. The typical conformations with high h_{HLX} or h_{HP} are indeed α-helix and β-hairpin (Figure 8), suggesting that these indicators are good for our purpose.

All relevant results are listed in Table 6. At 310 K, the polyA and its G mutant have 23.8% and 7.2% helices (columns “polyA” and “G mutant”). Predictions by AGADIR,[68] an accurate algorithm to predict the helical content of a special sequence based on the statistical analyses of a great number of sequences with known helical content, gave 24.2% and 5.3% for these two peptides, respectively. This agrees with the notion that Gly is a strong helix breaker.[69] As expected, β-hairpin is scarce in both peptides. Interestingly, the chance of β-hairpin formation in a G mutant is about three times that of the polyA. This implies that our model can also recognize the sequence-dependent stability of hairpins.

To examine if helices or hairpins can refold in our simulations, we monitored the change of h_{HLX} or h_{HP} with simulation time. We roughly define a refolding event of helices or hairpins as the recovery of h_{HLX} or h_{HP} to 0.5 or above after h_{HLX} or h_{HP} has been 0 for over 10 ns. This corresponds to the reformation of these structures after peptides have fully unfolded. The results in Table 6 clearly show that in these long time simulations, our CG model is capable of refolding peptides dozens of times. This guarantees statistical meaning for the calculation results of thermodynamic properties. Helices fold and refold much more frequently than do hairpins (40/48 vs 7/12), as can be expected generally. Furthermore, hairpins of the G mutant fold and refold more frequently than those of the PolyA (12 vs 7). This indicates that changing the turn sequence of a hairpin may alter the folding speed of the hairpin in our model, which coincides with the proposed zipper mechanism for hairpin folding.[70]

As demonstrated above, the turns in the middle of peptides are crucial. We therefore computed the probabilities of double aa units AG, GA, and AA in the middle of peptides to form four types of β-turns (see Methods and Models) in the long time simulations. The relative stabilities of the turns are in the following descending order: AA_I (0.0/0.0), AG_{II} (0.6/0.9), $GA_{II'}$ (0.7/1.0), AG_I (1.0/1.1), GA_I (1.0/1.1), $AA_{II'}$ (2.2/2.7), AA_{II} (2.8/1.7), $AG_{II'}$ (3.3/3.8), $AG_{I'}$ (3.8/2.1), GA_{II} (4.1/2.8), $GA_{I'}$ (4.7/2.9), and $AA_{I'}$ (6.0/3.0). The values in the parentheses are destabilization energies (kcal/mol) relative to AA_I for the formation of a given type of turn. The values before the slashes are from relative probabilities of formation of turns in simulations. The ones after the slashes are from the free energy perturbation calculations with an all-atom force field and an explicit solvent.[71] The order and magnitudes of the relative turn stabilities by the two calculation methods compare quite favorably. It is interesting that although our CG model is optimized for helices and hairpins, it can even capture many other structural features. During the simulations, 20.1% of the $GA_{II'}$ turn can occur in hairpins, which is the highest for all turn sequences. Although the AA_I turn has the highest probability to occur, only about 0.9% of the AA_I turn is found in hairpin. For polyA, the $AA_{II'}$ turn has the highest chance (6.1%) for the hairpin. This is consistent with the discovery that the type II′ turn is primarily found in hairpins.[72] Therefore, the G mutant favors a hairpin more because it can have a II′ turn with considerable stability.

To examine the effect of mass scaling on thermodynamic properties of peptides, we also carried out long simulations of the two peptides with quadrupled masses for peptide sites and with increased step size of 18 fs. These are defined as polyA′ and G′ mutant simulations, and their results are given in Table 6. They give very similar $<h_{HLX}>$ and $<h_{HP}>$ values to those of poly A and G mutant simulations. The calculated turn stabilities without (before slash) and with (after slash) mass scaling are also very similar: AA_I (0.0/0.0), AG_{II} (0.6/0.6), $GA_{II'}$ (0.7/0.5), AG_I (1.0/0.8), GA_I (1.0/

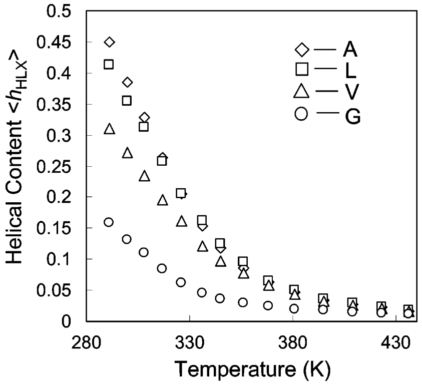


Figure 9. The helical contents of polyA and its mutants in the REMD simulations.

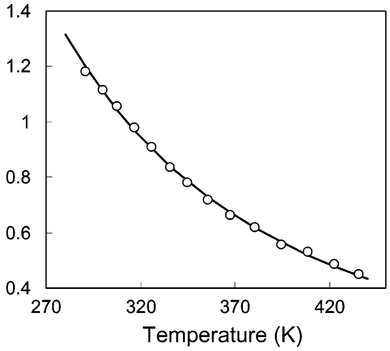


Figure 10. The fitted results for polyA. Empty circles indicate the *s* values calculated at each temperature, and the solid line is the trend line of these *s* values by eq A5.

0.9), $AA_{II'}$ (2.2/1.9), AA_{II} (2.8/3.1), $AG_{II'}$ (3.3/3.0), $AG_{I'}$ (3.8/3.8), GA_{II} (4.1/4.2), $GA_{I'}$ (4.7/4.8), and $AA_{I'}$ (6.0/5.3). These suggest that mass scaling has a negligible effect on thermodynamics of peptides. It is also interesting that the helical and hairpin structures also refold many times in these mass-scaled simulations.

Helical Propensities of Our CG Amino Acids. In the long time simulations, the polyA and its G mutant possess different <h_{HLX}>. This inspired us to perform a series of equilibrium REMD simulations on polyA and its Xaa mutants (X = L, V, and G). These simulations allow us to obtain helical contents of these peptides at different temperatures. As shown in Figure 9, the CG model clearly gives a sequence-dependent formation of helices for the four peptides.

In standard analyses, it is necessary to extract probabilities in helices (helical propensity, *s*) and probabilities in initiating helices (σ) (Appendix A) for different aas from these simulations in order to compare our results with the available experimental and theoretical results. The Zimm-Bragg (ZB) theory provides a way to obtain s and σ values from the average properties of a system.[66] For homopolymers like polyA, as applied by Garcia et al.[64] and Sorin et al.,[73] we used the average helical content, <h_{HLX}>, and the mean number of helical fragments, <n_s> (eqs A2 and A4 in Appendix A), to obtain *s* and σ for Ala at each temperature. The ΔG, ΔH, and ΔS associated with *s* during the coil-helix transition of Ala were computed by fitting *s* values at different temperatures to eq A5 in Appendix A. The fitted results are shown in Figure 10. There is no experimental study of the polyA peptide. For comparison, the average helical contents of our peptides at 273−395 K were estimated by AGADIR. These data were fitted by eqs A2 and A5, assuming that $\sigma = 0.004$ which is measured by Yang et al.[74]

The fitted ΔH and $T\Delta S$ of coil-helix transition are, respectively, −1.44 and −1.34 kcal/mol at 291 K for our Ala model. These are very close to the fitted ΔH and $T\Delta S$ of Ala from AGADIR, which are −1.43 and −1.24 kcal/mol, respectively, at 291 K. Our ΔH *is also close to the values of −1.3 kcal/mol per residue for polyalanine helices measured by Scholtz et al.* with calorimetric methods.[75] Besides, the ΔC_P (−0.004 kcal·mol^{-1}·K^{-1}) of our model and the ΔC_P (0.007 kcal·mol^{-1}·K^{-1}) from AGADIR are both within the acceptable range of ±0.008 kcal·mol^{-1}·K^{-1} for helix formation as measured by Lopez et al.[76] The helix initiation parameter, σ, of Ala in our model is 0.033 at 291 K. Such σ is larger than the σ *(0.004) obtained by Yang et al.*[74] in their CD measurements, the σ (0.004) derived from simulations with a modified AMBER-94 force field,[64] and the σ (0.007) obtained with the OPLS/AA/L force field.[34] However, our value is closer to the σ *value of 0.01−0.025 from the T-jump experiments by Thompson et al.* who intended to measure helix initiation kinetics more accurately[77] and the σ (0.027) obtained with a modified AMBER-99 force field, which showed an improved agreement of helix thermodynamics and kinetics of Fs peptide ($Ac\text{-}A_5(A_3R^+A)_3A\text{-}NHMe$) with experimental measurements.[73] Finally, we can also derive *s* and σ values for Ala from our long-time simulation of the polyA with the same procedure. They are 1.00 for *s* and 0.019 for σ at 310 K. The *s* and σ values from REMD at the same temperature are 1.02 and 0.029, respectively. Thus, the long time simulation and REMD give similar results.

Luo et al. obtained the s values of various aas (X) from the helix contents of polypeptides $Ac\text{-}KA_4XA_4KGY\text{-}NH_2$ at 273 K by fitting with the ZB theory.[78] Since their peptide models are similar to ours, it is desirable to compare our s_x values of aas with their experimental values. To obtain s_x values of aas other than Ala, we adopted a similar fitting procedure that was used by Luo et al.[78] and Myers et al.[79] who extracted the *s* values of single mutants in host peptides. We assumed that each Ala residue in the Xaa mutants takes the same *s* and σ values that were in polyA. The σ_x of Xaa was also assumed to be that of Ala, while its s_x was fitted by eq A3 in Appendix A.[80] The same procedures were also used to obtain the s_x values of Xaa based on the helix contents at 291 K predicted by AGADIR. The fitted s_x values (s_{sim}) are listed in Table 7. The results reveal that our *s* values for Ala, Leu, Val, and Gly at low temperature (291 K) are in good agreement with those from the AGADIR prediction (R^2=0.98), the single mutation experiments by Luo et al.[78] (R^2=0.99), and the measurements based on vast peptides with various sequences (R^2=0.97) (Table 7).[69]

Theoretically, it should be possible to derive ΔS and therefore ΔH at 291 K for aas other than Ala through the temperature dependence of s_x around 291 K. However, in practice, it is difficult because of large uncertainties of fitted s_x for Xaa when the temperature increases. This problem has

Table 7. *s* Values of Ala, Leu, Val, and Gly from Simulations, AGADIR Prediction and Experiments as Well as Linear Correlation of s_{sim} Values against Other *s* Values with $s_{\mathrm{other}} = a{\cdot}s_{\mathrm{sim.}}$

	A	L	V	G	a/R^2
$s_{\mathrm{sim}}{}^{a}$	1.18	0.84	0.30	0.02	
$s_{\mathrm{AGADIR}}{}^{a}$	1.38	0.81	0.25	0.03	1.09/0.98
$s_{\mathrm{exptl(A)}}{}^{b}$	1.5	1.0	0.3	0.05	1.23/0.99
$s_{\mathrm{exptl(B)}}{}^{c}$	1.54	0.92	0.22	0.05	1.21/0.97

[a] Derived at 291 K. [b] Measured at 273 K.[78] [c] Measured at 278 K.[69]

already been addressed by Luo et al.[78] For this reason, we used another method, as described in the following section, to estimate Δ*S* and Δ*H* of mutants at 291 K.

Structural Approaches to Helical Propensities. Although the ZB theory does not allow us to derive the Δ*H* and Δ*S* values for the coil-to-helix transition other than Ala using the current peptide models, it would still be desirable and informative to approximately calculate these components for different amino acids. This would allow us to examine the performance of the CG model in greater detail by comparing the available experimental and theoretical studies and to qualitatively analyze the factors that cause the differences of Δ*G*. For this purpose, we applied an approximate method, the so-called structural approach, which has been applied to decompose Δ*G* in other studies,[81−84] to estimate the Δ*H* and Δ*S* at 291 K.

Basically, the Δ*G* of various aas in our case is roughly considered as part of the free energy difference that is contributed from the aa 7 (Figure 6), between the ensemble of conformations with the aa 7 in helix (HE) and the ensemble with the aa 7 in coil (CE). The HE in our analyses is defined as the conformations with aas 5−9 in helical states, and the CE is the conformations with none of the aas 5−9 in helical states. It would be convenient to separate the overall free energy change due to the coil-to-helix transition into two parts: the free energy change of the peptide conformation, $\Delta G_{\mathrm{V,CE-HE,}}$ and the free energy change of hydration, $\Delta G_{\mathrm{W,CE-HE}}$:

$$\Delta G_{\mathrm{CE-HE}} = \Delta G_{\mathrm{V,CE-HE}} + \Delta G_{\mathrm{W,CE-HE}} \tag{8}$$

$$\Delta G_{\mathrm{V,CE-HE}} = \Delta\langle U_{\mathrm{HB}}\rangle_{\mathrm{CE-HE}} + \Delta\langle U_{\mathrm{vdW}}\rangle_{\mathrm{CE-HE}} + \Delta\langle U_{\mathrm{loc}}\rangle_{\mathrm{CE-HE}} - T\Delta S_{\mathrm{loc,CE-HE}} \tag{9}$$

$$\Delta G_{\mathrm{W,CE-HE}} = \Delta\langle H_{\mathrm{W}}\rangle_{\mathrm{CE-HE}} - T\Delta S_{\mathrm{W,CE-HE}} \tag{10}$$

The enthalpy change of the peptide conformation during the coil-helix transition, $\Delta H_{\mathrm{V,CE-HE}}$, includes three components: the local torsional energy change, $\Delta < U_{\mathrm{loc}} >_{\mathrm{CE-HE}}$, of the aa 7, the vdW interaction energy change $\Delta < U_{\mathrm{vdW}} >_{\mathrm{CE-HE}}$ between AU 7, which belongs to aa 7, and the rest of the peptide (Figure 6), and the HB interaction energy change, $\Delta < U_{\mathrm{HB}} >_{\mathrm{CE-HE}}$, involving both PUs 6 and 7, which surround AU 7 (Figure 6). The entropy change, $\Delta S_{\mathrm{loc,CE-HE}}$, includes both the backbone and/or the side-chain torsional entropies of the aa 7. The details on how to calculate these quantities are given in Appendix C.

Table 8. Fitted Δ*G* and All Energy Components (kcal/mol) from the Structural Approaches for Simulations at 291 K

	A	L	V	G
$\Delta G_{\mathrm{fitting}}$	−0.1	0.1	0.7	2.3
$\Delta < U_{\mathrm{HB}} >_{\mathrm{CE-HE}}$	−1.7	−2.0	−2.2	−2.2
$\Delta < U_{\mathrm{vdW}} >_{\mathrm{CE-HE}}$	−2.0	−2.5	−2.9	−2.3
$\Delta < U_{\mathrm{loc}} >_{\mathrm{CE-HE}}$	0.0	0.1	0.3	0.4
$-T\Delta S_{\mathrm{loc,CE-HE}}$	1.5	1.7	1.4	1.9
$\Delta G_{\mathrm{V,CE-HE}}{}^{a}$	−2.2	−2.7	−3.4	−2.2
$\Delta < H_{\mathrm{w}} >_{\mathrm{CE-HE}}$	2.4	3.4	4.6	5.5
$-T\Delta S_{\mathrm{W,CE-HE}}$	−0.3	−0.5	−0.5	−0.8
$\Delta G_{\mathrm{W,CE-HE}}{}^{b}$	2.1	2.9	4.1	4.7
$\Delta H_{\mathrm{CE-HE}}{}^{c}$	−1.3(−1.44) [f]	−1.0	−0.2	1.4
$-T\Delta S_{\mathrm{CE-HE}}{}^{d}$	1.2(1.34) [f]	1.2	0.9	1.1
$\Delta G_{\mathrm{CE-HE}}{}^{e}$	−0.1	0.2	0.7	2.5

[a] $\Delta G_{\mathrm{V,CE-HE}} = \Delta < U_{\mathrm{HB}} >_{\mathrm{CE-HE}} + \Delta < U_{\mathrm{vdW}} >_{\mathrm{CE-HE}} + \Delta < U_{\mathrm{loc}} >_{\mathrm{CE-HE}} - T\Delta S_{\mathrm{loc,CE-HE}}$. [b] $\Delta G_{\mathrm{W,CE-HE}} = \Delta < H_{\mathrm{W}} >_{\mathrm{CE-HE}} - T\Delta S_{\mathrm{W,CE-HE}}$. [c] $\Delta H_{\mathrm{CE-HE}} = \Delta < U_{\mathrm{HB}} >_{\mathrm{CE-HE}} + \Delta < U_{\mathrm{vdW}} >_{\mathrm{CE-HE}} + \Delta < U_{\mathrm{loc}} >_{\mathrm{CE-HE}} + \Delta < H_{\mathrm{W}} >_{\mathrm{CE-HE}}$. [d] $\Delta S_{\mathrm{CE-HE}} = \Delta S_{\mathrm{loc,CE-HE}} + \Delta S_{\mathrm{W,CE-HE}}$. [e] $\Delta G_{\mathrm{CE-HE}} = \Delta H_{\mathrm{CE-HE}} - T\Delta S_{\mathrm{CE-HE}}$. [f] Values in parentheses are from fitting with the ZB theory.

To evaluate the applicability of the structural approach to the current peptide systems, we first compare the thermodynamic properties of the polyA, which have been derived from fitting with the ZB theory and by other experimental and theoretical methods. As shown in the first column of Table 8, the enthalpy difference, $\Delta H_{\mathrm{CE-HE}}$ (−1.3 kcal/mol), and the entropy difference, $-T\Delta S_{\mathrm{CE-HE}}$ (1.2 kcal/mol), between HE and CE are comparable to Δ*H* (−1.44 kcal/mol) and −*T*Δ*S* (1.34 kcal/mol) derived by the fitting with the ZB theory. The resulting $\Delta G_{\mathrm{CE-HE}}$ for Ala agrees well with the fitted Δ*G* and the Δ*G* derived by other methods.[69,78,64,34,73] The structural approach gives a hydration enthalpy of about 2.4 kcal/mol. This is in good agreement with a recent calculation by Avbelj with the finite difference Poisson−Boltzmann (PB) method, which gives a range of 1.6−3.2 kcal/mol for solvation change to transfer an aa from *β*-strands to the middle of a helix.[85] In addition, the unfavorable $-T\Delta S_{\mathrm{CE-HE}}$ in our model is mainly due to the loss of local conformational entropy $-T\Delta S_{\mathrm{loc,CE-HE}}$ (1.5 kcal/mol), consistent with the estimated values of 1.5 kcal/mol by Wang et al.[81] and 1.37 kcal/mol by D'Aquino et al.[86]

For other aas, the structural approach derives a free energy change, $\Delta G_{\mathrm{CE-HE}}$ of 0.2, 0.7, and 2.5 kcal/mol for the coil-to-helix transition of Leu, Val, and Gly, respectively, of our model. These are in good agreement with the values deduced from the fitting with the ZB theory ($\Delta G_{\mathrm{fitting}}$ in Table 8), which has already been shown to agree with experimental results well (Table 7). Thus, the structural approach is also able to estimate the helix propensities of these aas well: Leu and Val have lower helical propensities than Ala, while Gly has a much lower helical propensity. The structural approach also indicates that the entropic contributions ($-T\Delta S_{\mathrm{CE-HE}}$) of the four amino acids are quite similar, ranging in 1.2−0.9 kcal/mol, with Val slightly more favorable, and the relative helix propensities are mainly determined by the enthalpy change ($\Delta H_{\mathrm{CE-HE}}$) of the coil-helix transition, with Leu, Val, and Gly being less favorable by about 0.3, 1.1, and 2.7 kcal/mol with respect to Ala, respectively. This suggests that in our model, the enthalpy factor $\Delta\Delta H_{\mathrm{CE-HE}}$

is the major determinant for the helical propensities ($\Delta\Delta G_{\mathrm{CE-HE}}$) of different aas, while the entropy factor plays a minor role.

It has been suggested that the entropy loss from the local conformation may determine the difference in helical propensities of different aas.[87] The previous calculations showed that the relative local entropy losses ($-T\Delta\Delta S_{\mathrm{loc,CE-HE}}$) with respect to Ala are about -0.06, -0.18, and 0.72 kcal/mol for Leu, Val, and Gly, respectively.[86,87] These are close to our calculation results of 0.2, -0.1, and 0.4 kcal/mol for Leu, Val, and Gly, respectively. However, as pointed out by Luo et al.,[78] the magnitude of this relative local entropy loss only accounts for less than one-third of $\Delta\Delta G$ obtained from their experiments on polyalanine-based peptides. They suggested that enthalpy should play a major role. The structural approach indicates that our model agrees better with the experiments by Luo et al.[78]

The structural approach allows the analysis of various factors that contribute to the free energy change in the coil-to-helix transition for our polyalanine-based model. As shown in Table 8, the coil-to-helix transition is favored by the formation of hydrogen bonds in the HE ($\Delta\langle U_{\mathrm{HB}}\rangle_{\mathrm{CE-HE}} = -1.7-2.2$ kcal/mol) and the vdW interaction ($\Delta\langle U_{\mathrm{vdW}}\rangle_{\mathrm{CE-HE}} = -2.0-2.5$ kcal/mol) between the aa 7 and the rest of of the peptide. It is disfavored by a solvation enthalpy change ($\Delta\langle H_{\mathrm{w}}\rangle_{\mathrm{CE-HE}} = 2.4-5.5$ kcal/mol) due to a poorer solvation of the helix, and the loss of local conformational entropy ($-T\Delta S_{\mathrm{loc,CE-HE}} = 1.4-1.9$ kcal/mol). The change in local conformational enthalpy ($\Delta\langle U_{\mathrm{loc}}\rangle_{\mathrm{CE-HE}} = 0.0-0.4$ kcal/mol) and hydration entropy ($-T\Delta S_{\mathrm{W,CE-HE}} = -0.3-0.8$ kcal/mol) is less significant. Overall, the coil-to-helix transition is favored by the free energy change of the peptide conformation ($\Delta G_{\mathrm{V,CE-HE}} = -2.2$ to -3.4 kcal/mol), with Leu and Val more favorable, but it is disfavored by the free energy change of hydration or solvation ($\Delta G_{\mathrm{W,CE-HE}} = 2.1-4.7$ kcal/mol). The lower helical propensities of Leu, Val, and Gly with respect to Ala are mainly caused by relative poorer hydration of the helical structures than the coil structures. In particular, the relative difference of hydration enthalpy, $\Delta\Delta\langle H_{\mathrm{W}}\rangle_{\mathrm{CE-HE}}$, in reference to Ala, which is the sum[88] of the relative difference of hydration enthalpy for PUs 4$-$9 and AUs 5$-$9 (Figure 6) between the mutant peptides and polyA (Appendix C), contributes about 1.0, 2.2, and 3.1 kcal/mol for Leu, Val, and Gly, respectively. This result is in agreement with the experimental observations by Luo et al.,[78] who suggest that hydration enthalpy plays a major role in determining helical propensities of different aas in polyalanine based-host/guest systems.

To further reveal the origin of the hydration effect on helical propensity, we analyzed the detailed solvation contributions from each of PUs 4$-$9 and AUs 5$-$9 to $\Delta\Delta\langle H_{\mathrm{W}}\rangle_{\mathrm{CE-HE}}$. Since the solute$-$solvent interaction $\langle U_{\mathrm{int}}\rangle$ accounts for the major part of solvation enthalpy of our peptide model (Table 5), we calculated the average interaction energy $\langle U_{\mathrm{int}}\rangle$ between each of these PUs or AUs and the solvent for HE or CE as well as the corresponding $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ during the coil-to-helix transition (Figure 11).

As shown in Figure 11a, $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ of nonpolar AUs is much smaller than that of polar PUs for all aas. The difference in $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ between Ala and its mutant is also smaller for AUs than for PUs. Therefore, the hydration of polar PUs is important for different helical propensities of aas in our model. In addition, a significant change of $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ not only occurs at the site of mutation (PUs 6 and 7) but also at the neighboring aas (PUs 5 and 8, Figure 11a), in accord with previous calculation results.[88]

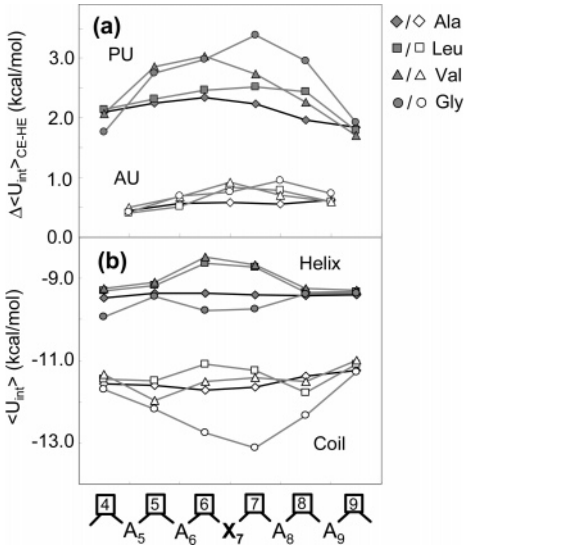


Figure 11. (a) The solvation energy change $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ of coil-helix transition for PUs 4$-$9 (top) and AUs 5$-$9 (bottom). (b) The solvation energy of PUs 4$-$9 for the helix ensemble (top) and coil ensemble (bottom). Ala (diamond); Leu (square); Val (triangle); Gly (circle).

Figure 11b shows the solute$-$solvent interaction energy for each of the PUs in the HE ($\langle U_{\mathrm{int}}\rangle_{\mathrm{HE}}$) and CE ($\langle U_{\mathrm{int}}\rangle_{\mathrm{CE}}$), from which $\Delta\langle U_{\mathrm{int}}\rangle_{\mathrm{CE-HE}}$ can be derived. It shows that in the HE, the side chains of Val and Leu shield the solvent-PU interaction that is available to Ala. The side chain of Val shields slightly more than does the side chain of Leu. The PU of the Gly mutant is somewhat better hydrated than the PUs of Ala. The difference in $\langle U_{\mathrm{int}}\rangle_{\mathrm{HE}}$ among our aas is expected since the more particles attached to the C_{β}, the less accessible the polar backbone is, as suggested by Makhatadze[89] and by Avbelj et al.,[88] and Gly has no C_{β} carbon at all, allowing a greater exposure of PUs. In the CE, while its solvation is apparently better than the HE for all four aas, Leu and Val have similar solvation as Ala. On the other hand, Gly is found to have a much better solvation for its PUs. Thus, the structural approach reveals that in our model, Leu and Val have lower helical propensities than Ala mainly because the larger side chains shield the solvation of helical structures, while Gly has a much poorer helical propensity mainly due to the much better solvation for the coil structures than for the helical structures.

Simulation Speed-Up. Computational speed-up of the current CG model compared to the all-atom model comes from two reasons.

As was pointed out by Marrink et al.[22] and Shih et al.,[26] the most important reason is that the four-water-one-particle mapping greatly reduces the number of interaction sites, and the number of interaction pairs between the interaction sites therefore decreases further. To demonstrate this, we performed a CG simulation of a polyA peptide in about 1000 CG solvent particles. We also carried out an all-atom simulation of a polyA peptide in about 4000 all-atom water molecules, which corresponds to the CG simulation. Both the simulations have a time interval of 2 fs and a cutoff of 0.12 nm. With the same computational power and in the same amount of time, the CG simulation samples 65-time more steps than the all-atom simulation. In addition, our CG model can use a step size of 18 fs, which gives a further speed-up by several folds. As a result, the CG model has a 200−300-fold speed-up compared to the all-atom model in explicit solvent.

The second reason is that the time scale in the CG simulation could be increased by the coarse-graining, which has been pointed out in other studies.[19,90,91] Marrink et al. found that the time scale of the CG water model, which is used in our CG model, is increased by four folds based on the comparison of the self-diffusion coefficients of water from their CG simulations and from all-atom simulations and experiments.[22] Interestingly, this time scale factor is found in all the dynamics in their CG water and CG lipid system.[22] Since most heavy atoms of our peptide model are explicitly represented, this time scale factor may only be applied to the diffusive motions in peptides.

For the above two reasons, our CG model may be about 10^3 faster than the all-atom model in the best situation where peptide motions are controlled by diffusion and may be 10^2 faster in the worst situation where local motions are dominant in peptides.

Summaries and Conclusions

We have constructed a CG protein model (for Gly, Aly, Leu, and Val) at an intermediate level coupled with the CG solvent model developed by Marrink et al.[22] A systematic method has been used to optimize parameters for protein potentials and protein−solvent interactions. The optimized CG model can fold polyalanine and its mutants into both helix and hairpin conformations without biased potentials. The calculated stabilities and dynamics of the peptides are sequence-dependent and compare very favorably with available experimental data. In particular, the helical propensities of Ala, Leu, Val, and Gly calculated by the CG model are very close to experimental values. Structural analysis indicates that the helical-forming propensities of different amino acids are mainly determined by solvation effects. Although some fine-tuning is still needed, we expect that a full development of this coarse-grained protein model for the remaining residue side chains will provide a promising tool for the study of fast folding of small proteins in aqueous solutions and in membrane environments.

Acknowledgment. We are grateful to RGCHK (HKUST6083/02M, N-HKUST 623/04) and NSFC (20225312) for financial support of the research.

Appendix A

The seminal work by Zimm and Bragg[66] (ZB) developed two parameters, σ and s, to describe helices. In their work, σ is defined as the probability of two peptide units to initiate a helical turn, and s is the probability of a peptide unit in the helical HB. In this study, we define s as the probability of an aa in the helical structures, where this aa and its two neighboring aas are all in helical conformations, and define σ as the probability of two aas at two ends of a helical sequence to initiate helices. These probabilities are relative to the coil structures. Such definition of s and σ should be equivalent to the original one as suggested by Schellman et al.[67]

As shown in Figure 6, if aas 1−4 are in helical conformations, aas 2 and 3 are in helical structures and aas 1 and 4 are at two ends. The probability of such structure is σs^2. If Xaa is mutated from Ala to another aas, it will contribute s_x and σ_x. If Xaa is at either end of a helical sequence, it contributes the weight of σ_x. The weight of a specific peptide structure can be $\sigma^i s^j$ for the polyalanine and $\sigma^i s^j \sigma_x{}^k s_x{}^n$ for its mutants. The weighted sum of all possible structures for the polyalanine (Q) and its mutants (Q') reads

$$Q = \sum_{i=0}^{3} \sum_{j=0}^{12} C_{ij} \sigma^i s^j$$

$$Q' = \sum_{i=0}^{3} \sum_{j=0}^{11} \sum_{k=0}^{1} \sum_{n=0}^{1} C_{\mathrm{i,j,k,n}} \sigma^i s^j \sigma_x{}^k s_x{}^n \qquad \text{(A1)}$$

where C is the number of structures with the same weight. It can be obtained by computer enumeration. “3” and “12” indicate the maximum numbers of helical sequences and helical HBs in a given structure according to the ZB theory.[66] Two important properties, the average helical content $<h_{\mathrm{HLX}}>$ or $<h_{\mathrm{HLX}}'>$ and the average helical fragment $<n_s>$, can read

$$<h_{\mathrm{HLX}}> = \frac{1}{12Q} \sum_{i=0}^{3} \sum_{j=0}^{12} j C_{\mathrm{i,j}} \sigma^i s^j \qquad \text{(A2)}$$

$$<h_{\mathrm{HLX'}}> = \frac{1}{12Q'} \sum_{i=0}^{3} \sum_{j=0}^{11} \sum_{k=0}^{1} \sum_{n=0}^{1} (j+n) C_{\mathrm{i,j,k,n}} \sigma^i s^j \sigma_x{}^k s_x{}^n \qquad \text{(A3)}$$

$$<n_s> = \frac{1}{Q} \sum_{i=0}^{3} \sum_{j=0}^{11} i C_{i,j} \sigma^i s^j \qquad \text{(A4)}$$

These equations are used to obtain σ, s, σ_x, and s_x by fitting the results from simulations. The relevant ΔH, ΔS, and ΔC_p can be obtained from s at different temperatures by eq A5.

$$\begin{aligned} \Delta G(T) &= -RT\, 1\mathrm{n}\, s \\ &= \Delta H(T) - T\Delta S(T) \\ &= \Delta H_0 + \Delta C_p(T - T_0) - T\Delta S_0 - \Delta C_p T\, 1\mathrm{n}\left(\frac{T}{T_0}\right) \end{aligned} \qquad \text{(A5)}$$

In eq A5, ΔH_0 and ΔS_0 are the changes of enthalpy and entropy at the reference temperature ($T_0 = 291$ K).

Appendix B

The Thermal Integration (TI) Approach. The solvation free energy ΔG_{sov} is defined as the free energy difference between the state where the solute is immersed in solvent and the state where the solute is isolated from solvent. The TI method can calculate the free energy difference between two states, for instance, A and B by introducing a coupling parameter (λ). As λ gradually varies from zero to unity, state A transforms to state B. During this process, $\Delta G_{A \to B}$ can be calculated as[50]

$$\Delta G_{A-B} = G_B - G_A = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \qquad \text{(A6)}$$

where U(λ) is the total energy when the system is in the intermediate state λ. In our calculation of ΔG_{sov}, the interaction between solute and solvent is gradually switched off during the state transformation. In addition, a soft-core Lennard-Jones potential is applied to avoid the singularity problem when λ is close to unity or zero.[92]

To calculate ΔS, and therefore ΔH, we apply the finite difference as Smith et al. [51]

$$-\Delta S_{A-B} \approx \frac{\Delta G_{A-B}(T + \Delta T) - \Delta G_{A-B}(T - \Delta T)}{2\Delta T} \qquad \text{(A7)}$$

where $\Delta T = 5$ K in our calculation. Each simulation for the TI totally lasts for 40 ns, which is long enough for accurate calculation of both ΔG_{sov} and ΔS_{sov}.[52]

The Approach Based on the Scaled Particle Theory (SPT). The solvation free energy of a sphere in CGW can be divided into two parts:[53] (1) the work ΔG_{cav} to create a cavity and (2) the free energy change ΔG_{int} by turning on solvent–solute interactions. Since the solvent molecule is essentially Lennard-Jones sphere, ΔG_{cav} can be computed from the scaled particle theory (SPT)[54,55]

$$\Delta G_{\text{cav}} = K_0 + K_1 \alpha_{12} + K_2 \alpha_{12}^{\;2} + K_2 \alpha_{12}^{\;3} \qquad \text{(A8)}$$

with

$$K_0 = RT\left\{-1\text{n}(1 - y) + \frac{9}{2}[y/(1 - y)]^2\right\} - (\pi P a_1^{\;3})/6$$

$$K_1 = -\,(RT/a_1)\{6y/(1 - y) + 18[y/(1 - y)]^2\} + \pi P a_1^{\;2}$$

$$K_2 = (RT/a_1^{\;2})\{12y/(1 - y) + 18[y/(1 - y)]^2\} - 2\pi P a_1$$

$$K_3 = \left(\frac{4}{3}\right)\pi P \qquad \text{(A9)}$$

where R is the gas constant, T is the temperature, P is the pressure, and $y = (\pi a_1^3 n_s)/6$ with n_s the number density of solvent molecules. The a_1 and a_2 are diameters of solvent and solute molecules, and the $a_{12} = (a_1+a_2)/2$ is the diameter of cavity. As suggested by Pierotii et al.,[56] the a_1 and a_2 can effectively be vdW diameters. ΔH_{cav} can read[56]

$$\Delta H_{\text{cav}} = \alpha_P RT^2[y/(1 - y)]\Big\{[6/(1 - y)][2(a_{12}/a_1)^2 - (a_{12}/a_1)] - [36y/(1 - y)^2]\left[(a_{12}/a_1)^2 - (a_{12}/a_1) + \frac{1}{4}\right] + 1\Big\} \qquad \text{(A10)}$$

where α_P is thermal expansion coefficient of the CG water. α_P(291 K) is calculated as 0.00089 K^{-1} by finite difference between two simulations at 290 and 292 K with $P = 1$ atm. ΔG_{int} can be approximately $<U_{\text{int}}>$, the average interaction energy between solvent and solute, if $<U_{\text{int}}>$ comes from the vdW interaction.[53] Since we have

$$\begin{aligned} \Delta G_{\text{sov}} &= \Delta G_{\text{cav}} + \Delta G_{\text{int}} \\ &\approx \Delta H_{\text{cav}} - T\Delta S_{\text{cav}} + \langle U_{\text{int}} \rangle \\ &= \Delta H_{\text{sov}} - T\Delta S_{\text{sov}} \end{aligned} \qquad \text{(A11)}$$

therefore, $\Delta H_{\text{sol}} \approx \Delta H_{\text{cav}} + <U_{\text{int}}>$ and $\Delta S_{\text{sov}} \approx \Delta S_{\text{cav}}$. According to Claverie et al.,[57] ΔG_{cav} or ΔH_{cav} of the solute with complex shape can be roughly estimated by

$$\Delta G_{\text{cav}} = \sum_i \frac{A_i}{4\pi a_{1i}^{\;2}} \Delta G_{\text{cav}}(a_{1i})$$

$$\Delta H_{\text{cav}} = \sum_i \frac{A_i}{4\pi a_{1i}^{\;2}} \Delta H_{\text{cav}}(a_{1i}) \qquad \text{(A12)}$$

where a_{1i} is the cavity diameter of the ith particle composing the solute, and A_i is its accessible surface area.

Appendix C

In order to estimate the free energy difference ΔG between the conformations with Xaa (the aa 7 in our peptide model) in helical states and the conformations with Xaa in coil states, we define helical ensemble (HE) and coil ensemble (CE) for Xaa as peptide conformations where aas 4–10 are in helical conformations ((ϕ, ψ) in (−60°±30°,−47°±30°)), and where no three aas can be in helical conformations for aas 4–10, respectively. By such definition, the aa 7 is embedded in the middle of a helical stretch, and each of the PUs 6 and 7 (Figure 6) that connect to the aa 7 involves in two HB interactions. The ΔG is therefore roughly estimated by $\Delta G_{\text{CE-HE}} = G_{\text{HE}} - G_{\text{CE}}$, which is the contribution from the aa 7 to the free energy difference between the HE and the CE. The enthalpy part of $\Delta G_{\text{CE-HE}} = \Delta H_{\text{CE-HE}} - T\Delta S_{\text{CE-HE}}$ can be calculated as

$$\begin{aligned} \Delta H_{\text{CE-HE}} &= H_{\text{HE}} - H_{\text{CE}} \\ &\approx U_{\text{HE}} - U_{\text{CE}} \\ &= \langle U_{\text{V}}(r) + H_{\text{W}}(r) \rangle_{\text{HE}} - \langle U_{\text{V}}(r) + H_{\text{W}}(r) \rangle_{\text{CE}} \end{aligned} \qquad \text{(A13)}$$

where the last equality is from Wang et al.[81] $U_{\text{v}}(\boldsymbol{r})$ is the conformational energy of the Xaa mutant in conformation $\boldsymbol{r}$, and $H_{\text{w}}(\boldsymbol{r})$ is the solvation enthalpy of Xaa of this conformation. "$\langle\rangle_{\text{HE/CE}}$" indicates the average over HE or CE. For the conformational energy in the HE or CE

$$\langle U_{\text{v}}(r) \rangle = \langle U_{\text{loc}}(r) \rangle + \langle U_{\text{vdW}}(r) \rangle + \langle U_{\text{HB}} \rangle \qquad \text{(A14)}$$

$U_{\text{loc}}(\boldsymbol{r})$ is the local torsional energy of Xaa. $U_{\text{vdW}}(\boldsymbol{r})$ is half of the vdW energy between Xaa and other parts of the protein. Because AU 7 belongs to aa 7 and both PUs 6 and 7 can be considered as parts of aa 7 (Figure 6 and Methods and Models), for Ala in polyA, $U_{\text{HB}}(\boldsymbol{r})$ reads

$$\langle U_{\mathrm{HB}}(r)\rangle = \langle {}^{1}/_{2}[\phi(6) + \phi(7)]\rangle \qquad (A15)$$

where $\phi(i)$ is half of the HB energy of PU i with the other part of the protein. For the Xaa mutant, we assume that changes of $\phi(6)$ and $\phi(7)$ are all induced by the mutation of Ala into Xaa. The difference of $U_{\mathrm{HB}}(\boldsymbol{r})$ between Xaa and Ala is calculated as

$$\Delta\langle U_{\mathrm{HB}}(r)\rangle^{\mathrm{Ala-Xaa}} = \langle U_{\mathrm{HB}}(r)\rangle^{\mathrm{Xaa}} - \langle U_{\mathrm{HB}}(r)\rangle^{\mathrm{Ala}}$$

$$= \langle \phi(6) + \phi(7)\rangle^{\mathrm{Xaa}} - \langle \phi(6) + \phi(7)\rangle^{\mathrm{Ala}} \qquad (A16)$$

The $U_{\mathrm{HB}}(\boldsymbol{r})$ for Xaa can therefore be derived from eq A16.

The hydration enthalpy $H_{\mathrm{w}}(\boldsymbol{r})$ was calculated with the SPT based method with the SAS assumption (Appendix B). For Ala in polyA, $H_{\mathrm{w}}(\boldsymbol{r})$ is computed by

$$\langle H_{\mathrm{W}}(r)\rangle = \langle {}^{1}/_{2}[H_{\mathrm{W}}^{\mathrm{PU}}(6) + H_{\mathrm{W}}^{\mathrm{PU}}(7)] + H_{\mathrm{W}}^{\mathrm{AU}}(7)\rangle \qquad (A17)$$

where $H_{\mathrm{W}}{}^{\mathrm{PU}}(i)$ and $H_{\mathrm{W}}{}^{\mathrm{AU}}(j)$ are the solvation enthalpy of PU i and AU j, respectively. In the case that Ala is mutated into Xaa, we follow the suggestion by Avbelj et al.[88] that the mutation induces not only the solvation change of aa 7 but also the solvation change of its several neighboring aas. Therefore, PUs 4−9 and AUs 5−9 are considered in the calculation of solvation enthalpy. The difference of $H_{\mathrm{W}}(\boldsymbol{r})$ between Xaa and Ala reads

$$\Delta\langle H_{\mathrm{W}}\rangle^{\mathrm{Ala-Xaa}} = \langle H_{\mathrm{W}}(r)\rangle^{\mathrm{Xaa}} - \langle H_{\mathrm{W}}(r)\rangle^{\mathrm{Ala}}$$

$$= \langle \sum_{i=4}^{9} H_{\mathrm{W}}^{\mathrm{PU}}(i) + \sum_{j=5}^{9} H_{\mathrm{W}}^{\mathrm{AU}}(j)\rangle^{\mathrm{Xaa}}$$

$$- \langle \sum_{i=4}^{9} H_{\mathrm{W}}^{\mathrm{PU}}(i) + \sum_{j=5}^{9} H_{\mathrm{W}}^{\mathrm{AU}}(j)\rangle^{\mathrm{Ala}} \qquad (A18)$$

The entropy part $\Delta S_{\mathrm{CE-HE}}$ of $\Delta G_{\mathrm{CE-HE}}$ is estimated by

$$\Delta S_{\mathrm{CE-HE}} \approx \Delta S_{\mathrm{W,CE-HE}} + \Delta S_{\mathrm{loc,CE-HE}}$$

$$\approx \Delta S_{\mathrm{W,CE-HE}} + \Delta S_{\mathrm{bb,CE-HE}} + \Delta S_{\mathrm{sc,CE-HE}} \qquad (A19)$$

$\Delta S_{\mathrm{W,CE-HE}}$ is solvation entropy difference of S_{W} between the HE and the CE. It can be obtained in the same way as described in eqs A17 and A18 as well as in Appendix B. $\Delta S_{\mathrm{loc,CE-HE}}$ is the local conformational entropy change including backbone $\Delta S_{\mathrm{bb,CE-HE}}$ and side-chain $\Delta S_{\mathrm{sc,CE-HE}}$ entropy loss. The backbone entropy in the HE or the CE is calculated by dividing (ϕ, ψ) of Xaa into 36 × 36 states and summing $-R\Sigma[p_i\ln(p_i)]$, where p_i is the probability of state i. The side-chain entropy is similarly obtained by dividing χ of Xaa (Figure 1a) into 36 states.

Finally, in the calculation of solvation of peptides, we have increased both ΔH_{cav} and ΔS_{cav} from the SPT methods by 1-fold as a rough correction, which has been demonstrated in Models and Methods. Without the correction, $\Delta G_{\mathrm{CE-HE}}$ (kcal/mol) for Ala is 0.3, and the relative $\Delta\Delta G_{\mathrm{CE-HE}}$ of the mutants in reference to Ala is 0.5 for Leu, 1.0 for Val, and 3.1 for Gly. Compared to the corresponding values in Table 8, the removal of the correction should not make a qualitative difference from the results with correction and the results from fitting.

Supporting Information Available: Simulated structural information of valine and leucine dipeptides. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dobson, C. M. *Nature* **2003**, *426*, 884-890

(2) Hartl, F. U.; Hayer-Hartl, M. *Science* **2002**, *295*, 1851−1858.

(3) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430.

(4) Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638−642.

(5) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(6) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.

(7) Burton, R. E.; Huang, M. A.; Daugherty, M. A.; Fullbright, P. W.; Oas, T. G. *J. Mol. Biol.* **1996**, *263*, 311.

(8) Ding, F.; Dokholyan, N. V. *TRENDS Biotechnol.* **2005**, *23*, 450−455.

(9) Nymeyer, H.; García, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5921.

(10) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183.

(11) Abe, H.; Gō, N. *Biopolymers* **1980**, *20*, 1013.

(12) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860.

(13) Thirumalai, D.; Guo, Z. *Biopolymers* **1995**, *35*, 137.

(14) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616−11629.

(15) Ding, F.; Borreguero, J. M.; Buldyrey, S. V.; Stanley, H. E.; Dokholyan, N. V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 220−228.

(16) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147−155.

(17) Smith, A. V.; Hall, K. C. *J. Mol. Biol. Proteins: Struct., Funct., Genet.* **2001**, *44*, 344−360.

(18) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Fold Des.* **1998**, *3*, 577−587.

(19) Sharma, S.; Ding, F.; Dokholyan, N. V. *Biophys. J.* **2007**, *92*, 1457−1470.

(20) Honig, B.; Yang, A. *Adv. Protein Chem.* **1995**, *46*, 27.

(21) Shelley, J. C.; Shelley, M.; Reeder, R.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464.

(22) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750−760.

(23) de Vries, A. H.; Mark, A. E.; Marirnk, S. J. *J. Am. Chem. Soc.* **2004**, *126*, 4488−4489.

(24) Kasson, M. P.; Kelly, N. W.; Singhal, N.; Vrljic, M.; Brunger, A. T.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11916−11921.

(25) Bond, P. J.; Sansom, M. S. P. *J. Am. Chem. Soc.* **2006**, *128*, 2697.

(26) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674−3684.

(27) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.

(28) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.

(29) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(30) Regan, L.; DeGrado, W. F. *Science* **1988**, *241*, 976.

(31) Neidigh, J. W.; Fesinmeyer, R. M.; Anderson, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425−430.

(32) Marsh, R. E.; Donohue, J. *Adv. Protein Chem.* **1967**, *22*, 249.

(33) Kaminski, G.; Friesner, R.; Tirado-Rives, J.; Jorgensen, W. *J. Phy. Chem. B* **2001**, *105*, 6474−6487.

(34) Gnanakaran, S.; Garcia, A. E. *Proteins* **2005**, *59*, 773−782.

(35) Eker, F.; Cao, X.; Nafie, L.; Schweitzer-Stenner, R. *J. Am. Chem. Soc.* **2002**, *124*, 14330−14341.

(36) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P. *Biomolecular simulation: the GROMOS96 manual and user guide*; Hchschulverlag AG an der ETH: Zurich, 1996.

(37) Williams, D. E.; Craycroft, D. J. *J. Phys. Chem.* **1987**, *91*, 6365−6373.

(38) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253−266.

(39) Nosé, S. *Mol. Phys.* **1984**, *52*, 255−268.

(40) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(41) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(42) Ben-Naim, A. *Solvation Thermodynamics*; Plenum Press: New York, 1987.

(43) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 72nd ed.; CRC Press: Boca Raton, FL, 1992.

(44) Douglass, D. C.; Mccall, D. W. *J. Phys. Chem.* **1958**, *62*, 1102.

(45) Kresheck, G. C.; Schneider, H.; Scheraga, H. A. *J. Phys. Chem.* **1965**, *69*, 3132−3144.

(46) Black, C.; Joris, G. G.; Taylor, H. S. *J. Chem. Phys.* **1948**, *16*, 537.

(47) Wolfenden, R. *Biochemistry* **1978**, *17*, 201−204.

(48) Wolfenden, R.; Anderson, L.; Cullis, P. M.; Southgate, C. C. *Biochemistry* **1981**, *20*, 849−855.

(49) Della Gatta, G. D.; Barone, G.; Elia, V. *J. Solution Chem.* **1986**, *15*, 157−167.

(50) Mezei, M.; Beveridge, D. L. *Ann. N. Y. Acad. Sci.* **1986**, *482*, 1.

(51) Smith, D. E.; Haymet, A. D. J. *J. Chem. Phys.* **1993**, *98*, 6445−6454.

(52) Wan, S.-Z.; Stote, R. H.; Karplus, M. *J. Chem. Phys.* **2004**, *121*, 9539−9548.

(53) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027−2094.

(54) Reiss, H.; Frisch, H. L.; Lebowitz, J. L. *J. Chem. Phys.* **1959**, *31*, 369.

(55) Reiss, H.; Frisch, H. L.; Helfand, E.; Lebowitz, J. L. *J. Chem. Phys.* **1960**, *32*, 119.

(56) Pierotti, R. A. *J. Phys. Chem.* **1965**, *69*, 281−288.

(57) Claverie, P. In *Intermolecular Interactions: from Diatomics to Biomolecules*; Pullman, B., Ed.; J. Wiley: Chichester, 1978.

(58) Graziano, G. *J. Phys. Soc. Jpn.* **2000**, *69*, 3720−3725.

(59) Pohorille, A.; Pratt, L. R. *J. Am. Chem. Soc.* **1990**, *112*, 5066.

(60) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91* 43−56.

(61) Berendsen, H. J. C.; Pstma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(62) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(63) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435−439.

(64) García, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782−2787.

(65) Hutchinson, G.; Thornton, J. M. *Protein Sci.* **1994**, *3*, 2207−2216.

(66) Zimm, B. H.; Bragg, J. K. *J. Chem. Phys.* **1959**, *31*, 526−535.

(67) Qian, H.; Schellman, J. A. *J. Phys. Chem.* **1992**, *96*, 3987−3994.

(68) Munoz, V.; Serrano, L. *Nature Struct. Biol.* **1994**, *1*, 399−409.

(69) Chakrabartty, A.; Kortemme, T.; Baldwin, R. L. *Protein Sci.* **1994**, *3*, 843−852.

(70) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196−199.

(71) Yan, Y. B.; Erickson, B. W.; Tropsha, A. *J. Am. Chem. Soc.* **1995**, *117*, 7592−7599.

(72) Sibanda, B. L.; Thornton, J. M. *Nature* **1985**, *316*, 170−174.

(73) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472−2493.

(74) Yang, J.; Zhao, K.; Gong, Y.; Vologodskii, A.; Kallenbach, N. R. *J. Am. Chem. Soc.* **1998**, *120*, 10646−10647.

(75) Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2854−2858.

(76) Lopez, M. M.; Chin, D. H.; Baldwin, R. L.; Makhatadze, G. I. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1298−1302.

(77) Thompson, P. A.; Eaton, W. A.; Hofrichter, J. *Biochemistry* **1997**, *36*, 9200−9210.

(78) Luo, P. Z.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4930−4935.

(79) Myers, J. K.; Pace, C. N.; Scholtz, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2833−2837.

(80) In the fitting for Gly, a s_x value of −0.07 was obtained assuming $\sigma_x = 0.033$. The negative s_x value is mathematically possible but physically meaningless. It indicates that the σ_x of 0.033 is too large for Gly. We found that $s_x < 0.02$ is required for $\sigma_x > 0$ and $\sigma_x < 0.0045$ is required for $s_x > 0$. However, for Val and Leu, variation of σ_x from 0.01 to 0.1 has little influence on the fitted s_x. Thus, the assumption of $\sigma_x = 0.033$ should be valid in the cases of Val and Leu.

(81) Wang, J.; Purisima, E. O. *J. Am. Chem. Soc.* **1996**, *118*, 995−1001.

(82) Hermans, J.; Anderson, A. G.; Yun, R. H. *Biochemistry* **1992**, *31*, 5646−5653.

(83) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35*, 13681−13688.

(84) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1995**, *252*, 351−365.

(85) Avdelj, F. *J. Mol. Biol.* **2000**, *300*, 1335−1359.

(86) D'Aquino, J. A.; Gómez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E. *Proteins: Struct., Funct., Genet.* **1996**, *25*, 143−156.

(87) Creamer, T. P.; Rose, G. D. *Proteins: Struct., Funct., Genet.* **1994**, *19*, 85−97.

(88) Avbelj, F.; Luo, P. Z.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10786−10791.

(89) Makhatadze, G. I. *Adv. Protein Chem.* **2006**, *72*, 199−226.

(90) Nielson, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Phys. Condens. Matter*. **2004**, *16*, R481−R512.

(91) Zhou, Y.; Karplus, M. *J. Mol. Biol.* **1999**, *293*, 917−951.

(92) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Greber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529−539.

CT700151X

JCTC Journal of Chemical Theory and Computation

Improving the Accuracy of the Linear Interaction Energy Method for Solvation Free Energies

Martin Almlöf,[†] Jens Carlsson,[†] and Johan Åqvist*

Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, Box 596, SE-751 24 Uppsala, Sweden

Received May 3, 2007

Abstract: A linear response method for estimating the free energy of solvation is presented and validated using explicit solvent molecular dynamics, thermodynamic perturbation calculations, and experimental data. The electrostatic contribution to the solvation free energy is calculated using a linear response estimate, which is obtained by comparison to the free energy calculated using thermodynamic perturbation. Systematic deviations from the value of $^1/_2$ in the potential energy scaling factor are observed for some types of compounds, and these are taken into account by introducing specific coefficients for different chemical groups. The derived model reduces the rms error of the linear response estimate significantly from 1.6 to 0.3 kcal/mol on a training set of 221 molecules used to parametrize the model and from 3.7 to 1.3 kcal/mol on a test set of 355 molecules that were not used in the derivation of the model. The total solvation free energy is estimated by combining the derived model with an empirical size dependent term for predicting the nonpolar contribution. Using this model, the experimental hydration free energies for 192 molecules are reproduced with an rms error of 1.1 kcal/mol. The use of LIE in simplified binding free energy calculations to predict protein−ligand binding free energies is also discussed.

1. Introduction

Understanding the solvation properties of molecules in different environments is of tremendous importance for the understanding of various biological processes such as ligand binding, protein folding, and enzyme catalysis. Since a majority of these processes involve molecules in the aqueous phase, reliable estimates of hydration energies are crucial in order to accurately estimate the involved thermodynamic properties. Microscopic simulations can provide a detailed description of molecular interactions and are an efficient means of estimating the free energy changes of these processes. The most rigorous approaches to estimate solvation energies using molecular dynamics (MD) or Monte Carlo (MC) simulations are the free energy perturbation (FEP) and thermodynamic integration (TI) methods. The absolute free energies of solvation can be accurately estimated using these methods in combination with appropriate thermodynamic cycles. MD or MC simulations are carried out in both gas and aqueous phase and in each calculation the free energy is typically evaluated in two separate steps. First, the nonpolar part of the solvation energy is obtained by creating the van der Waals cavity formed by the solute. Then, the electrostatic contribution is calculated by gradually turning on the partial charges of the solute atoms. The transformation process is divided into several intermediate steps, and the total change in free energy is evaluated as the sum of these. FEP and TI calculations have shown that force fields are able to reproduce experimental solvation energies quite accurately, in particular those which have been specifically parametrized for this purpose.[1−10] However, application of these methods to more complex problems, such as estimating protein−ligand binding free energies,[11−13] has been shown to be difficult due to convergence and sampling problems mainly associated with creation/annihilation of particles and the large number of simulations that has to be carried out. For these reasons, a faster method for obtaining free energies of solvation would be extremely valuable.

* Corresponding author phone: +46 18 471 4109; fax: +46 18 53 69 71; e-mail: aqvist@xray.bmc.uu.se.

[†] These authors contributed equally to this work.

10.1021/ct700106b CCC: $37.00
Published on Web 09/28/2007

The nonpolar contribution to the free energy of solvation is often estimated using linear relationships between solute size measures, e.g., molecular or solvent accessible surface area (MS or SASA), and free energies of solvation that have been observed for nonpolar compounds.[6,14−16] For the electrostatic contribution, there are a number of approaches that are based on a solvent linear response (LR) assumption. A classic example of this is the Born equation, which predicts solvation free energies of ions from a quadratic dependence on the ion charge and an inverse dependence on its radius.[17] Other continuum dielectric approaches employing LR to estimate solvation free energies are the Poisson−Boltzmann[18,19] and Generalized Born[20] methods. Another useful approach to estimate the electrostatic solvation free energy contribution from microscopic simulations, which also is derived from a LR assumption, is based on collecting the average electrostatic solute−solvent interaction energies from MD or MC simulations. The total free energy of turning on the electrostatic solute−solvent interactions can be approximated as

$$\Delta F_{\mathrm{el}} = \beta[\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle_{\mathrm{on}} + \langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle_{\mathrm{off}}] \tag{1}$$

where $\beta = {}^1/_2$ and the $\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle$ terms represent the average value of the solute−solvent electrostatic interaction energy evaluated by sampling with (on) and without (off) these interactions turned on.[21] The second term in eq 1 can further be neglected if the solute and solvent become randomly oriented with respect to each other in the absence of electrostatic interactions (see below). The above type of relationship has been used for estimating protein−ligand binding free energies with the linear interaction energy (LIE) method[22−24] and to calculate pK_a values of protein residues.[25,26] A similar approach has also been successfully used to predict experimental solvation energies of small organic compounds.[27−29]

Åqvist and Hansson performed a detailed investigation of the accuracy of eq 1 for estimating the free energy of turning on electrostatic solute−solvent interactions for various small solutes in different polar solvents. It was found that generally $\beta < {}^1/_2$ for neutral molecules and that systematic deviations from the theoretical value could be identified for some chemical groups, e.g., for monoalcohols $\beta = 0.37$.[21] In this work we consider an alternate thermodynamic cycle, which includes both the electrostatic intra- and intermolecular interactions in the gas and aqueous phase, and investigate how the LR approximation can be used to predict the electrostatic component of solvation free energies from microscopic simulations. First, a general formula for the total electrostatic part of the solvation free energy, which includes both intra- and intermolecular contributions, is derived. Second, the scaling factor, β, is estimated for 211 small molecules that represent common neutral and ionic chemical groups. Different models for predicting β are then discussed, and, in particular, we again identify systematic deviations in the coefficient for specific chemical groups. The derived models are validated on a test set of 361 compounds. These molecules are more flexible and contain combinations of different chemical groups, for which the validity of LR was not extensively studied by Åqvist and Hansson.[21] The total free energy of hydration is estimated by combining the calculated charging free energies with an empirical term for the nonpolar contribution. The use of the LR approximation in the LIE method to predict protein−ligand binding free energies is also discussed.

2. Theory

A Linear Response Approximation for Estimation of the Electrostatic Contribution to Solvation Free Energies. The interactions in a system will here be described using two classical potentials U_A and U_B, which represent states A and B of a solute. In state A all electrostatic interactions involving the solute are turned off, while B represents the state where these terms are turned on (compare eq 1). In addition, the two potentials have exactly the same force field parameters for bonded and van der Waals (Lennard-Jones) terms. Hence, the transformation from state A to state B will represent the turning on of solute electrostatic interactions and the difference between the two potentials can be expressed as

$$\Delta U = U_B - U_A = U^{\mathrm{el}}_{\mathrm{r-s}} + U^{\mathrm{el}}_{\mathrm{r-r}} \tag{2}$$

where $U^{\mathrm{el}}_{\mathrm{r-s}}$ and $U^{\mathrm{el}}_{\mathrm{r-r}}$ are the electrostatic solute−solvent (r−s) and solute−solute (r−r) interaction energies. One approach to derive eq 1 is to start with Zwanzig's expression for the free energy difference between two states[30]

$$\Delta F_{\mathrm{el}} = -kT\ln\langle e^{-\Delta U/kT}\rangle_A \tag{3}$$

where T is the temperature, k is Boltzmann's constant, and $\langle ...\rangle_A$ is an ensemble average on state A. Herein, we will make no distinction between the Helmholtz and Gibbs free energies since the difference between the two quantities has negligibile effects on the results. The cumulant expansion[31] of eq 3 yields

$$\Delta F_{\mathrm{el}} = \langle \Delta U\rangle_A - \frac{1}{2kT}\langle(\Delta U - \langle\Delta U\rangle_A)^2\rangle_A + \frac{1}{6(kT)^2}\langle(\Delta U - \langle\Delta U\rangle_A)^3\rangle_A + ... \tag{4}$$

The corresponding expression utilizing a configurational average on state B is

$$\Delta F_{\mathrm{el}} = \langle \Delta U\rangle_B + \frac{1}{2kT}\langle(\Delta U - \langle\Delta U\rangle_B)^2\rangle_B + \frac{1}{6(kT)^2}\langle(\Delta U - \langle\Delta U\rangle_B)^3\rangle_B + ... \tag{5}$$

Adding eqs 4 and 5 and discarding terms of order three and higher yields

$$\Delta F_{\mathrm{el}} = \frac{1}{2}[\langle\Delta U\rangle_A + \langle\Delta U\rangle_B] - \frac{1}{4kT}\langle(\Delta U - \langle\Delta U\rangle_A)^2\rangle_A + \frac{1}{4kT}\langle(\Delta U - \langle\Delta U\rangle_B)^2\rangle_B \tag{6}$$

Furthermore, if equal fluctuations of the energy gaps are assumed (i.e., the parabolic free energy functions corresponding to states A and B have equal curvatures), the free energy can be evaluated from the averages of ΔU sampled on states A and B, which gives us the linear response estimate of the free energy of charging

$$\Delta F_{\mathrm{el}} = \frac{1}{2}\,[\langle \Delta U\rangle_A + \langle \Delta U\rangle_B] \tag{7}$$

To estimate the electrostatic contribution to the free energy of hydration, the difference between the water and gas-phase free energy must be taken

$$\Delta F^{\mathrm{el}}_{\mathrm{sol}} = \Delta F^{\mathrm{w}}_{\mathrm{el}} - \Delta F^{\mathrm{g}}_{\mathrm{el}} = \frac{1}{2}\,[\langle \Delta U\rangle^{\mathrm{w}}_A + \langle \Delta U\rangle^{\mathrm{w}}_B - \langle \Delta U\rangle^{\mathrm{g}}_A - \langle \Delta U\rangle^{\mathrm{g}}_B] \tag{8}$$

where superscripts w and g indicate that the averages are taken in aqueous and gas phase, respectively. Separation of eq 8 into intra- (r−r) and intermolecular (r−s) energies gives

$$\Delta F^{\mathrm{el}}_{\mathrm{sol}} = \frac{1}{2}\,[\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_A + \langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_B + \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{w}}_A - \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{g}}_A + \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{w}}_B - \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{g}}_B] \tag{9}$$

Equation 9 can be further simplified by noting that $\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_A$ is close to zero.[21] This average is calculated from a simulation carried out on potential A, where the electrostatic solute interactions are switched off. The solvent will therefore in general be randomly oriented with respect to the solute, and the net electrostatic contribution to the solute−solvent interaction energy will be close to zero (note, however, that ionic solutes are an exception that will be further addressed below[21,32−34]). In addition, it could be expected that the solute conformations will be similar in both phases when sampling is carried out on state A, i.e. $\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{w}}_A - \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{g}}_A \approx 0$. Introducing these two approximation results in a LR approximation of the electrostatic contribution to the free energy of solvation

$$\Delta F^{\mathrm{el}}_{\mathrm{sol}} = \frac{1}{2}\,(\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_B + \Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B) = \beta(\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_B + \Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B) \tag{10}$$

where $\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B$ is the difference between the intramolecular energies in aqueous and gas phase and $\beta = {}^1/_2$. For rigid molecules, eq 10 can be further simplified. In this case $\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B = 0$ and gives

$$\Delta F^{\mathrm{el}}_{\mathrm{sol}} = \beta\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_B \tag{11}$$

While this expression requires the approximation $\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B = 0$ using the thermodynamic cycle employed in the present work, this is not the case for the cycle used by Åqvist and Hansson.[21] With their approach, in which only solute−solvent interactions are perturbed, a different thermodynamic cycle can be used to calculate the free energy of solvation. A combined thermodynamic cycle, which illustrates the difference between the two approaches for calculating the total solvation free energy, is shown in Figure 1. The vertical legs in Figure 1 can be estimated using the linear response approximation and, depending on if the intramolecular energies are included in the perturbation, the two cycles differ slightly in how the nonpolar contribution would be calculated. In the upper cycle, the nonpolar contribution is the free energy of transferring the solute between the phases with its intermolecular electrostatic interactions turned off, while in the complete cycle (used here) it is the free energy of transferring the solute with all electrostatic interactions involving the solute turned off.

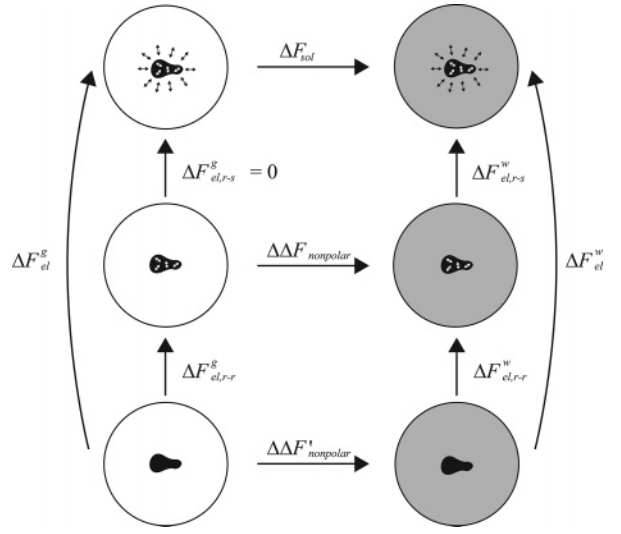

ΔF_{sol}
$\Delta F^{g}_{el,r\text{-}s} = 0$
$\Delta F^{w}_{el,r\text{-}s}$
ΔF^{g}_{el}
$\Delta\Delta F_{nonpolar}$
ΔF^{w}_{el}
$\Delta F^{g}_{el,r\text{-}r}$
$\Delta F^{w}_{el,r\text{-}r}$
$\Delta\Delta F'_{nonpolar}$

Figure 1. A thermodynamic cycle describing how the free energy of hydration can be obtained from the linear response approximation derived in this work and the approach devised by Åqvist and Hansson.[21] The upper cycle corresponds to the approach of Åqvist and Hansson, from which the total free energy of hydration can be estimated from $\Delta F_{\mathrm{sol}} = \Delta F^{\mathrm{w}}_{\mathrm{el,r-s}} + \Delta\Delta F_{\mathrm{nonpolar}}$, where $\Delta F^{\mathrm{w}}_{\mathrm{el,r-s}}$ is the free energy of turning on the intermolecular electrostatic interactions in water. In the derivation presented in this work the entire cycle is used, which yields the hydration energy $\Delta F_{\mathrm{sol}} = \Delta F^{\mathrm{w}}_{\mathrm{el,r-s}} + \Delta F^{\mathrm{w}}_{\mathrm{el,r-r}} - \Delta F^{\mathrm{g}}_{\mathrm{el,r-r}} + \Delta\Delta F'_{\mathrm{nonpolar}}$. The top row represents states in which all interactions are turned on, while the middle row represents states in which the electrostatic solute−solvent interactions are turned off and all other interactions are on. The bottom row represents states where all electrostatic interactions involving the solute are turned off.

Models for Including Systematic Deviations in β. By calculating the free energy of turning on the electrostatic solute interactions and also extracting the corresponding average electrostatic solute−solvent and solute−solute interaction energies in gas and aqueous phase, the value of the β coefficient can be evaluated from

$$\beta_{\mathrm{FEP}} = \frac{\Delta F^{\mathrm{el}}_{\mathrm{sol}}}{\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_B + \Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_B} \tag{12}$$

where the subscript FEP has been added to indicate that the β coefficient is obtained from FEP calculations.

In order to investigate and account for systematic deviations in the β_{FEP} coefficient, several different models were investigated. The first model (A) uses the theoretically derived value of $\beta = {}^1/_2$. The second model (B) was proposed by Hansson et al.[23] and is based on β values calculated for a small series of model compounds. In this model, $\beta = {}^1/_2$ is only used for ionic compounds. For neutral molecules without any hydroxyl groups $\beta = 0.43$ is used, while 0.37 or 0.33 is used for molecules with one or several hydroxyl groups, respectively. Models C−E are new models presented in this work, and these are parametrized on a set of training compounds. Model C assumes that a single value of β, which is parametrized on the training set, can be used for all molecules. In models D and E, β can assume different values depending on the chemical nature of the compound. This is somewhat similar to model B, but models D and E take into account more chemical groups than model B (i.e., not just hydroxyls) and provide a method for determining β's for compounds containing a mixture of chemical groups.

In the derivation of a β value for solutes containing several chemical groups we will assume that the contributions from each group are additive, and the total electrostatic hydration free energy can then be written as

$$\Delta F_{\rm sol}^{\rm el} = \beta_1[\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_1 + \beta_2[\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_2 + \beta_3[\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_3 + ... \quad (13)$$

where $[\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_i$ is the change in electrostatic solute energy, and β_i is the FEP derived scaling factor for group i. If each scaling factor is rewritten as $\beta_i = \beta_0 + \Delta\beta_i$, eq 13 can be expressed as

$$\Delta F_{\rm sol}^{\rm el} = \left(\beta_0 + \frac{\sum_i [\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_i \Delta\beta_i}{\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} - \Delta\langle U_{\rm r-r}^{\rm el}\rangle}\right)(\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} - \Delta\langle U_{\rm r-r}^{\rm el}\rangle) = \beta(\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} - \Delta\langle U_{\rm r-r}^{\rm el}\rangle) \quad (14)$$

From this expression, a β for the total molecule can be estimated if the electrostatic energies for each group are known. In order for this expression to be useful, however, weighting factors predetermined from FEP calculations are introduced. The molecular β coefficient can now be identified as

$$\beta = \beta_0 + \frac{\sum_i w_i \Delta\beta_i}{\sum_i w_i} \quad (15)$$

where w_i, β_0, and $\Delta\beta_i$ are derived from explicit solvent FEP calculations of single chemical groups. To test this approach two different models are investigated. Model D uses weighting factors that are equal to the average solute potential energies of each chemical group in the training set. The second model (E) uses $w_i = 1.0$ for all neutral groups and one single weighting factor for anions and cations. The latter ($w_{\rm anion/cation}$) is determined as the ratio between the average $[\langle U_{\rm r-s}^{\rm el}\rangle^{\rm w} + \Delta\langle U_{\rm r-r}^{\rm el}\rangle]_i$ of the ionic and neutral groups in the training set and is found to be 11.0 ± 1.7.

Prediction of the Total Free Energy of Solvation. In order to make a comparison to experiment, the nonpolar contribution to the free energy of solvation has to be added to eq 10. In approximations of the nonpolar free energy contribution it is, in most cases, assumed that there is a linear relationship between size measures and the free energy. This is based on experimental observations that solvation free energies of hydrophobic compounds, for which there should be a negligible electrostatic contribution, generally depend linearly on size measures such as surface area.[14−16] Hence, these observations indicate that the nonpolar contribution to the solvation free energy can be estimated from

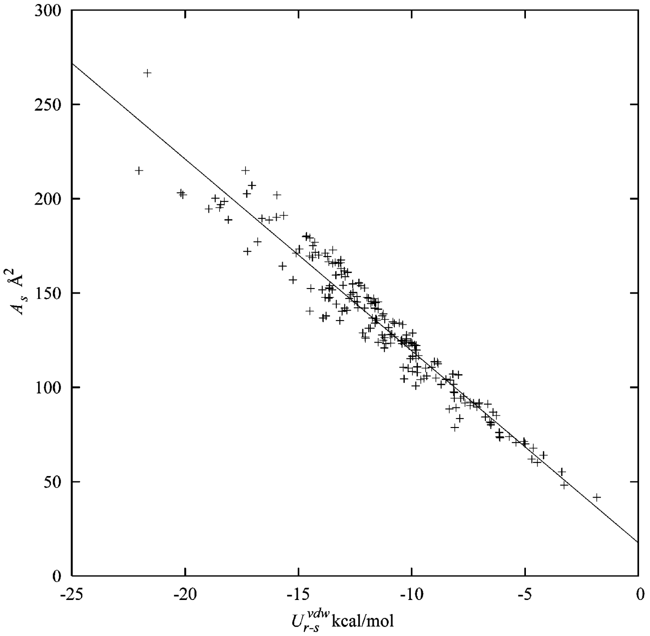


Figure 2. The average van der Waals solute−solvent interaction energy (kcal/mol) sampled on state A and the molecular surface ($A_{\rm S}$, Å^2) for the molecules in the parametrization set. Linear regression yields the relation $A_{\rm S} = -10.2$-$\langle U_{\rm r-s}^{\rm vdW}\rangle_A + 17.7$ (solid line).

$$\Delta F_{\rm sol}^{\rm np} = a_{\rm w}\,\sigma + b_{\rm w} \quad (16)$$

where np denotes that it is the nonpolar contribution, σ is a size measure, and $a_{\rm w}$ and $b_{\rm w}$ are empirically derived parameters. By extrapolating eq 16 to zero solute size, $b_{\rm w}$ can be interpreted as the free energy of inserting a point particle into the solvent.[35] The first term in eq 16 corresponds to the free energy change of increasing the size of the point particle. A less empirical way of formulating eq 16 is to introduce the (macroscopic) surface tension, that determines the cost of cavity creation, together with the van der Waals or dispersion energy associated with introducing a nonpolar solute into the cavity[36,37]

$$\Delta F_{\rm sol}^{\rm np} = \gamma_{\rm S}\, A_{\rm S} + \langle U_{\rm r-s}^{\rm vdW}\rangle_A \quad (17)$$

where $\gamma_{\rm S}$ is the surface tension, and $A_{\rm S}$ the molecular surface area. Note, however, that eq 17 formally neglects the free energy of inserting a point particle into the solvent. The surface area is also very strongly correlated with the van der Waals interaction energy (cf. Figure 2), making the latter an equally useful size measure.[22] The regression equation obtained between molecular surface area and van der Waals energy in water (see below and Figure 2) is $A_{\rm S} = -10.2\,\langle U_{\rm r-s}^{\rm vdW}\rangle_A^{\rm w} + 17.7$ in kcal/mol and Å^2, which together with the experimental value for the surface tension of water of 73 mN/m = 105 cal/(mol Å^2) predicts that $\Delta F_{\rm sol}^{\rm np} = -0.07\,\langle U_{\rm r-s}^{\rm vdW}\rangle_A^{\rm w} + 1.9$ kcal/mol. We will return to this prediction later on but can conclude that using the solute−solvent van der Waals energy as a size measure should give us the nonpolar contribution to the hydration energy as

$$\Delta F_{\mathrm{sol}}^{\mathrm{np}} = \alpha_{\mathrm{w}}^{\mathrm{vdW}} \langle U_{\mathrm{r-s}}^{\mathrm{vdW}} \rangle_{B}^{\mathrm{w}} + \gamma_{\mathrm{w}}^{\mathrm{vdW}} \tag{18}$$

where $\alpha_{\mathrm{w}}^{\mathrm{vdW}}$ and $\gamma_{\mathrm{w}}^{\mathrm{vdW}}$ are parameters that could either be derived empirically or be obtained as above. In accordance with the thermodynamic cycle of Figure 1, the nonpolar contribution to the hydration free energy (bottom row) should be estimated using $\langle U_{\mathrm{r-s}}^{\mathrm{vdW}} \rangle_{A}^{\mathrm{w}}$ instead of $\langle U_{\mathrm{r-s}}^{\mathrm{vdW}} \rangle_{B}^{\mathrm{w}}$ as a size measure. However, by using $\langle U_{\mathrm{r-s}}^{\mathrm{vdW}} \rangle_{B}^{\mathrm{w}}$ in eq 18 the total hydration free energy can be estimated using a single simulation, and as will be shown in the Results and Discussion these two alternatives yield similar results.

A semiempirical expression for the total free energy of hydration can now be obtained by combining eqs 10 and 18

$$\Delta F_{\mathrm{sol}} = \alpha_{\mathrm{w}}^{\mathrm{vdW}} \langle U_{\mathrm{r-s}}^{\mathrm{vdW}} \rangle_{B}^{\mathrm{w}} + \beta(\langle U_{\mathrm{r-s}}^{\mathrm{el}} \rangle_{B}^{\mathrm{w}} + \Delta \langle U_{\mathrm{r-r}}^{\mathrm{el}} \rangle_{B}) + \gamma_{\mathrm{w}}^{\mathrm{vdW}} \tag{19}$$

Here, the free energy of hydration can be estimated using two free parameters ($\alpha_{\mathrm{w}}^{\mathrm{vdW}}$ and $\gamma_{\mathrm{w}}^{\mathrm{vdW}}$), while β is determined according to one of the above derived models.

3. Method

Training and Test Set. The 211 compounds listed in Table 1 of the Supporting Information, hereafter referred to as the "training set", were used to parametrize the β_0's, $\Delta\beta_i$'s, and w_i's of models C–E. These compounds consist of hydrocarbon groups and at most one non-hydrocarbon moiety. For the set, 19 different groups were defined: alcohols, 1° amides, 2° amides, 3° amides, 1° amines, 2° amines, 3° amines, ketones/aldehydes, thiols, ethers, esters, carboxylic acids, nitriles, nitros, sulfides, anions, and cations. The parametrization of β_0's and $\Delta\beta_i$'s for the different models (C–E) was performed in a least-squares fashion by minimizing the squared error between the free energy of charging as calculated from the FEP simulations (in gas and water) and as predicted by eq 10. A second data set, referred to as the "test set", was created by combining fragments (Figure 3) consisting of the 19 types of moieties used in the training set ($19 \cdot 19 = 361$ molecules in total). These were used to validate the models parametrized on the training set.

Molecular Dynamics and Free Energy Calculations. All molecular dynamics (MD) simulations were carried out with the program Q[38] using the OPLS all atom (OPLS-AA) force field.[39] The simulations were carried out at 300 K in an 18 Å sphere centered on the geometrical center of the solute, and the system was solvated with TIP3P[40] water molecules. Water molecules were subjected to radial and polarization restraints according to the SCAAS method.[38,41] A nonbonded cutoff of 10 Å was used, and long-range electrostatic interactions were treated with the local reaction field multipole expansion method,[42] except for the solute interactions that were calculated without any cutoff. The time step was set to 1 fs, and nonbonded pair lists were updated every 25 steps.

Electrostatic free energies of solvation were determined using the FEP method. In this method the free energy of charging the solute is determined by simulating several intermediate states of the charged and uncharged solute. The potentials governing the intermediate states are defined by $U_m = \lambda_m U_A + (1 - \lambda_m) U_B$ where A and B represent the uncharged and charged solute, respectively, and λ_m is a mapping parameter which varies from $\lambda_1 = 0$ to $\lambda_n = 1$. The free energy difference between state A and B can then be calculated by summing up the free energy differences of the intermediate states as calculated using the Zwanzig expression (eq 3).

$$\Delta F_{A \rightarrow B} = -kT \sum_{m=1}^{n-1} \ln \langle e^{-(U_{m+1} - U_m)/kT} \rangle_m \tag{20}$$

Each calculation comprised a 16 ps heating scheme followed by 50 ps of equilibration and then 41 data collection simulations at evenly spaced λ-values. The trajectories at the FEP endpoints (λ=0 and 1) were 200 ps in length, while at intermediate λ-values they were 20 ps in length. Energies were extracted every 25 steps, and the simulations were carried out for the charging and uncharging process in both gas and aqueous phases. Energies from the first 5 ps of each λ-simulation were discarded in the FEP calculations. Due to slower convergence for the zwitterionic compounds these simulations were run considerably longer (1 ns at each λ-value).

Standard errors of the FEP calculations were estimated as half the difference between the free energy of charging and uncharging the compound. The standard errors of the electrostatic potential energies were estimated as half the difference in average potential energy between the charging and uncharging simulations of the appropriate endpoint.

4. Results and Discussion

Parametrization of β. The electrostatic contribution to the free energy of solvation, electrostatic solute energies, and the calculated value of β_{FEP} for each molecule in the training set are presented in Table 1 of the Supporting Information. For the alkanes, the β_{FEP} values vary widely from −1.1 to 2.3, but this is simply due to the small electrostatic energies in these cases, which results in large errors when the quotient in eq 12 is evaluated. The β_{FEP} values for the other compounds are very similar to those obtained by Åqvist and Hansson.[21] In order to evaluate the LR approximation, several models were tested, and these are summarized in Table 1. The models differ from each other in that they have different rules to determine what value of β to use in eq 10. The model based on the theoretically derived $\beta = {}^1/_2$ (model A) consistently overpredicts the absolute charging energy for anions and neutral solutes (rms = 1.64 kcal/mol), which indicates that lower values of β are necessary for these compounds. For cations, on the other hand, the estimated free energies are more positive than those calculated using FEP, which indicates that higher values of β are required to reproduce the FEP calculations. It should be noted that the deviations from $\beta = {}^1/_2$ for ionic and neutral molecules have different origins. For the neutral molecules, the lower values of β were found to arise from nonquadratic free energy functions with unequal curvatures[21] (approximation in eqs 6 and 7). For the ionic molecules, on the other hand, the deviations from $\beta = {}^1/_2$ are primarily a result of neglecting the contribution from $\langle U_{\mathrm{r-s}}^{\mathrm{el}} \rangle_{A}^{\mathrm{w}}$ [21,32,34] (one of the approximations made in going from eq 9 to 10). The results for model A are shown in Figure 4.

In the work of Hansson et al. it was suggested that $\beta = 0.43$ was appropriate for most neutral molecules except

Table 1. Models for Predicting β Studied in This Work

model	treatment of β and w_i	rms training[a]	rms test[a]
A	$\beta = 0.5$	1.64	3.72
B	β dependent on number of hydroxyls and net charge as in ref 23	1.21	3.29
C	one β parametrized for entire training set ($\beta = 0.48$)	1.52	3.68
D	$\beta = \beta_0 + \dfrac{\sum_i w_i \Delta b_i}{\sum_i w_i}$ $w_i \propto \langle U_i^{\mathrm{el}}\rangle_B$	0.32	1.22
E	$\beta = \beta_0 + \dfrac{\sum_i w_i \Delta b_i}{\sum_i w_i}$ $w\begin{Bmatrix}1 \text{ for net charge} = 0\\ 11 \text{ for net charge} = \pm 1\end{Bmatrix}$	0.32	1.26

[a] In kcal/mol.

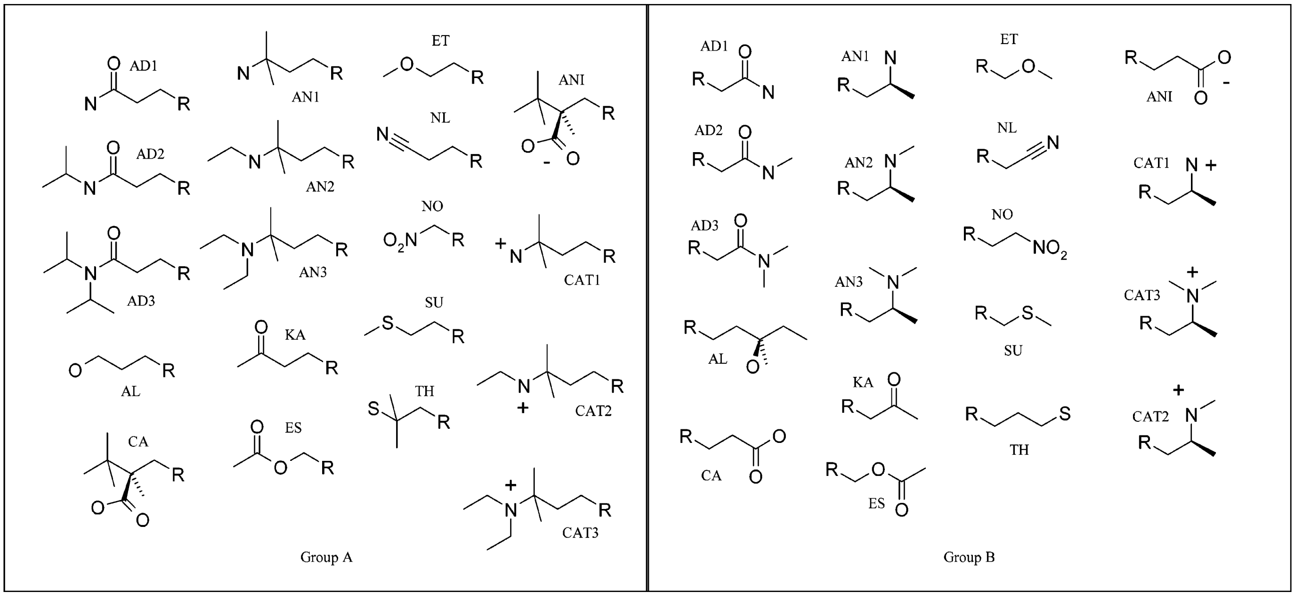


Figure 3. The fragments used to build the test set. Fragments from group A are combined with fragments from group B.

alcohols ($\beta = 0.33$ or 0.37), cations ($\beta = 0.5$), and anions ($\beta = 0.5$).[21,23] This scheme (model B) fares slightly better (rms = 1.21 kcal/mol) than using the theoretically derived $\beta = {}^1/_2$. It also performs slightly better compared to using a single β for the entire set (model C). For model C an optimized value of $\beta = 0.48$ was obtained, and the estimated charging free energies are again in reasonable agreement with the FEP calculations (rms = 1.52 kcal/mol).

In an attempt to improve model B, the compounds in the training set were grouped depending on their chemical nature, i.e., alcohols, 1° amides, 2° amides, 3° amides, 1° amines, 2° amines, 3° amines, ketones/aldehydes, thiols, ethers, esters, carboxylic acids, nitriles, nitros, sulfides, anions, and cations. The average signed error of using $\beta = 0.43$ in eq 10 compared to the FEP calculations was then calculated for each group of compounds. The result of using $\beta = 0.43$ is shown in Figure 5, where each group of compounds is indicated by using different symbols. The groups deviating by more than 0.5 kcal/mol were alcohols, 1° amines, 2° amines, 1° amides, carboxylic acids, cations, and anions. The charging free energies for the alcohols, 1° amines, 2° amines, 1° amides, and carboxylic acids are all underestimated, which indicates that lower values of β are required for these groups. For the anions and cations the effect of charging the solutes is underestimated, suggesting $\beta > 0.43$ is necessary. Åqvist and Hansson showed that the lower values of β obtained for alcohols derives from their ability to form hydrogen bonds to the solvent.[21] Therefore it was not surprising that other hydrogen bond donating groups (amines, amides, and acids) displayed similar properties. While the values of β that were obtained for neutral molecules have been shown to reflect deviations from LR, the $\Delta\beta_i$'s for the anions and cations are due to the fact that $\langle U_{\mathrm{r-s}}^{\mathrm{el}}\rangle_A^{\mathrm{w}}$ is not negligible.[21] However, compared to $\langle U_{\mathrm{r-s}}^{\mathrm{el}}\rangle_B^{\mathrm{w}}$ the contribution from $\langle U_{\mathrm{r-s}}^{\mathrm{el}}\rangle_A^{\mathrm{w}}$ is relatively small, and since the contribution is of a systematic nature,[21,32−34] it can be taken into account by adjusting $\Delta\beta_i$ for these groups. For

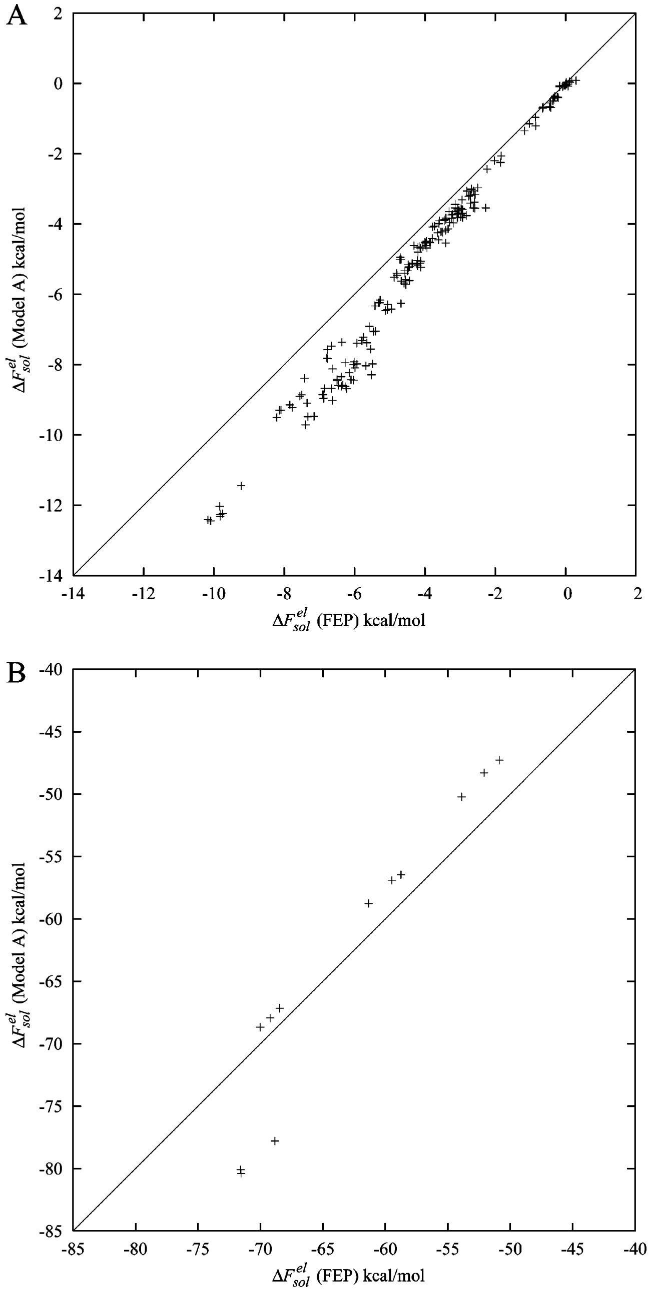


Figure 4. FEP calculated and estimated electrostatic components of the free energy of hydration ($\Delta F^{\mathrm{el}}_{\mathrm{sol}}$) for model A for the neutral (A) and ionic (B) molecules in the training set. All values are in kcal/mol.

anions $\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_{A} = 8.4 \pm 0.2$ kcal/mol, and this results in a β value lower than 0.5, while the opposite effect is observed for cations ($\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_{A} = -8.7 \pm 0.2$ kcal/mol). Thus models D and E have $\beta_0 = 0.43$ and $\Delta\beta_i \neq 0$ for six chemical groups: alcohols ($i = 1$), 1° and 2° amines ($i = 2$), 1° amides ($i = 3$), carboxylic acids ($i = 4$), anions ($i = 5$), and cations ($i = 6$). Optimization of $\Delta\beta_i$ for these six groups yields $\Delta\beta_1 = -0.06$, $\Delta\beta_2 = -0.04$, $\Delta\beta_3 = -0.02$, $\Delta\beta_4 = -0.03$, $\Delta\beta_5 = 0.02$, and $\Delta\beta_6 = 0.09$ (Table 2). For models D and E (Figure 6) the rms for the full data set is in remarkably good agreement with the FEP calculations (rms = 0.32 kcal/mol) and is significantly better than models A−C. Note that weighting factors are not necessary for the parametrization set since each molecule only contains one of the defined chemical groups, i.e., models D and E are equivalent for the training set. The a priori assumption of assigning $\beta_0 = 0.43$[23] was tested by optimizing all parameters ($\Delta\beta_1$, $\Delta\beta_2$, $\Delta\beta_3$, $\Delta\beta_4$, $\Delta\beta_5$, $\Delta\beta_6$, and β_0) and yielded an optimal β_0 of 0.43,

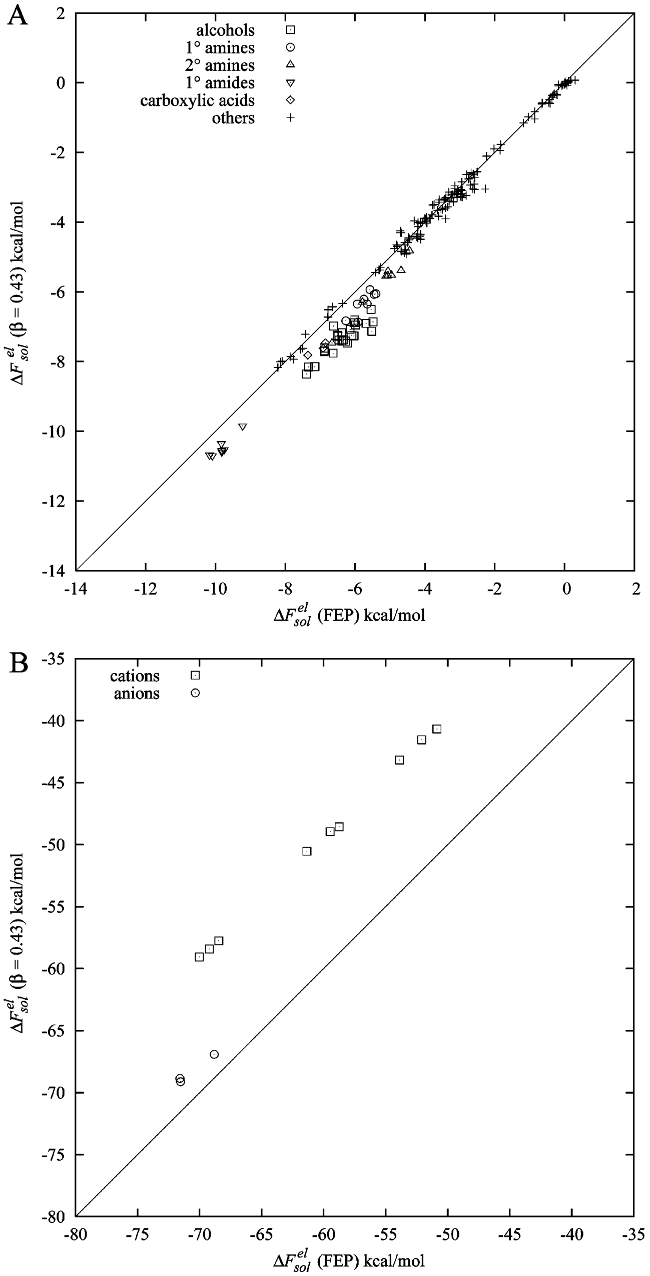


Figure 5. FEP calculated and estimated electrostatic components of the free energy of hydration ($\Delta F^{\mathrm{el}}_{\mathrm{sol}}$) using $\beta = 0.43$ for the neutral (A) and ionic (B) molecules in the training set. All values are in kcal/mol.

Table 2. Obtained Parameters for Models D/E

β_0	0.43
$\Delta\beta_1$(alcohols)	−0.06
$\Delta\beta_2$(1°, 2°-amines)	−0.04
$\Delta\beta_3$(1°-amides)	−0.02
$\Delta\beta_4$(COOH)	−0.03
$\Delta\beta_5$(anions)	0.02
$\Delta\beta_6$(cations)	0.09
$\Delta\beta_7$(other)	0

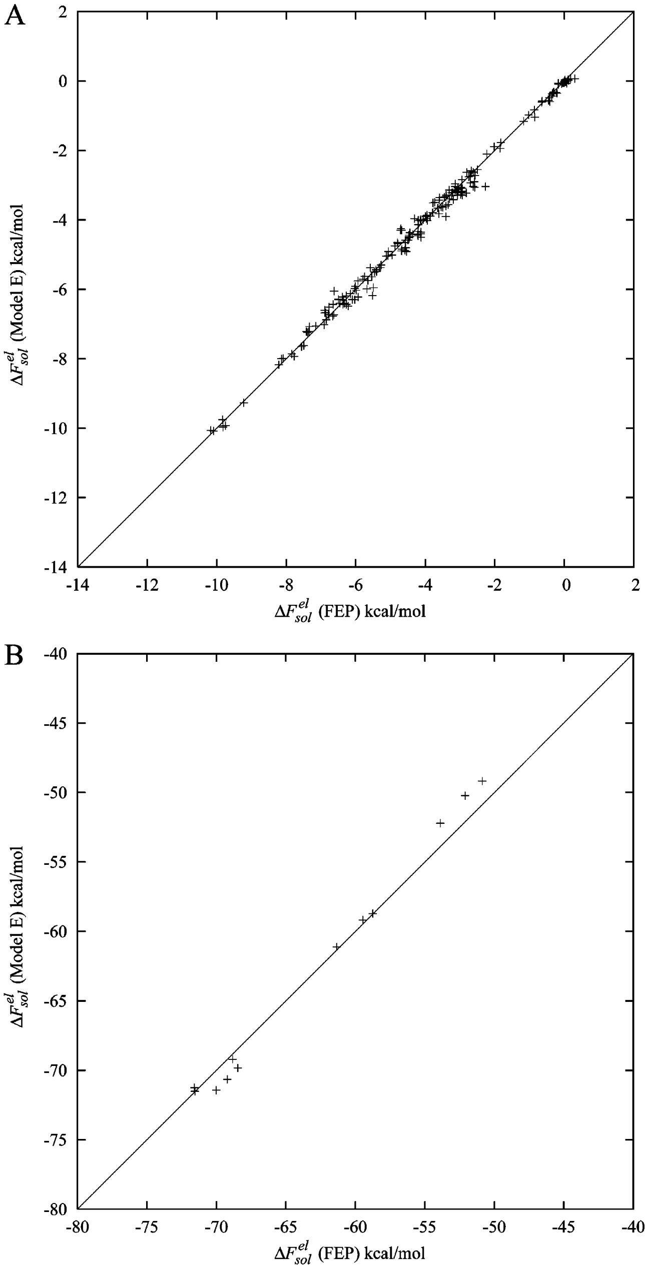


Figure 6. FEP calculated and estimated electrostatic components for the free energy of hydration (ΔF^{el}_{sol}) for model E for the neutral (A) and ionic (B) molecules in the training set. All values are in kcal/mol.

confirming the results of Åqvist and Hansson.[21] Using $\beta_0 = 0.43$ instead of $\beta_0 = 0.50$ also has its practical reasons since it allows for a smaller number of $\Delta\beta_i$'s.

Validation of the Parametrization. In order to test the proposed schemes to estimate β_{FEP} for molecules containing mixed chemical groups we calculated charging free energies for a test set of compounds, which consisted of all possible pairwise combinations of group A and B in Figure 3. The results from these calculations are shown in Table 2 of the Supporting Information. In total there were 361 test compounds comprising combinations of all chemical groups in the training set. To elucidate which model is most useful and predictive, models A−E (as parametrized on the training set) were tested. The results for the different models are summarized in Table 1. In this case models D and E use the values of β_0 and $\Delta\beta_i$ parametrized on the training set, but the models have different weighting factors (w_i). In model D the weighting factors were taken as the average solute electrostatic energies for each group. This is compared to using $w_i = 1.0$ for all neutral groups and $w_i = 11.0$ for the anions and cations in model E. Note that even though models D and E only have nonzero $\Delta\beta_i$'s for six types of groups, the other chemical groups will still influence the estimation of β through their weighting. For example, a compound containing only one alcohol moiety will receive an estimated β of 0.37 ($0.37 = 0.43 + (-0.06)/1$, using $w_i = 1$) in model E, while a compound containing an alcohol and keto moiety will receive an estimated β of 0.40 ($0.40 = 0.43 + (-0.06 + 0.00)/2$). A problematic case is the six zwitterions in the test set. The additivity assumed in eq 13 might not hold for these combinations due to the field canceling effect of the opposite charges and/or strong electrostatic intramolecular interactions. Therefore the results for these molecules will be presented separately at the end of this section.

The rms errors for models A−E on the test set (355 combinations in total, zwitterions excluded) are 3.72, 3.29, 3.68, 1.22, and 1.26 kcal/mol, respectively. The differences in rms between the models tested on the test set should not be seen as a statistical effect of increasing the number of free parameters since the parametrizations were performed on a different set of compounds. Thus it is clear that models D and E outperform the other models and hence 1°, 2° amine, 1° amide, alcohol, carboxylic acid, anion, and cation moieties need different β coefficients than other compounds. The introduced complexity of using specific weighting factors for each group of the neutral moieties does not seem justified considering that model E yields very similar accuracy. The best model is hence considered to be model E, and the result of using this approach is shown in Figure 7.

Westergren et al. recently suggested that the deviations in LR observed by Åqvist and Hansson[21] were due to neglect of the change in solute−solvent van der Waals interactions upon charging of the solute.[36] Instead they proposed that ΔF^{el}_{sol} should be approximated by the expression

$$\Delta F^{el}_{sol} = \frac{1}{2}\langle U^{el}_{r-s}\rangle^{w}_{B} + \Delta\langle U^{vdW}_{r-s}\rangle^{w} \quad (21)$$

where $\Delta\langle U^{vdW}_{r-s}\rangle^{w}$ is the difference between the intermolecular van der Waals energies in states *B* and *A*. However, as can be seen from Zwanzig's expression (eq 3) and from the derivation of the LR approximation, all explicit contributions from the change in van der Waals interactions cancel when the difference between U_A and U_B is taken (while they, of course, implicitly affect the Boltzmann factor in all ensemble averages). Thus, there appears to be no theoretical support for additionally including van der Waals energy differences upon charging as was done by Westergren et al. (eq 21). Nevertheless, this ad hoc approximation of the electrostatic contribution is somewhat better than the strict LR result, yielding an rms of 3.53 kcal/mol compared to 3.72 kcal/mol using model A.

For the zwitterions in the test set, which were excluded from the above analysis, $\beta = 0.48$ is obtained using model

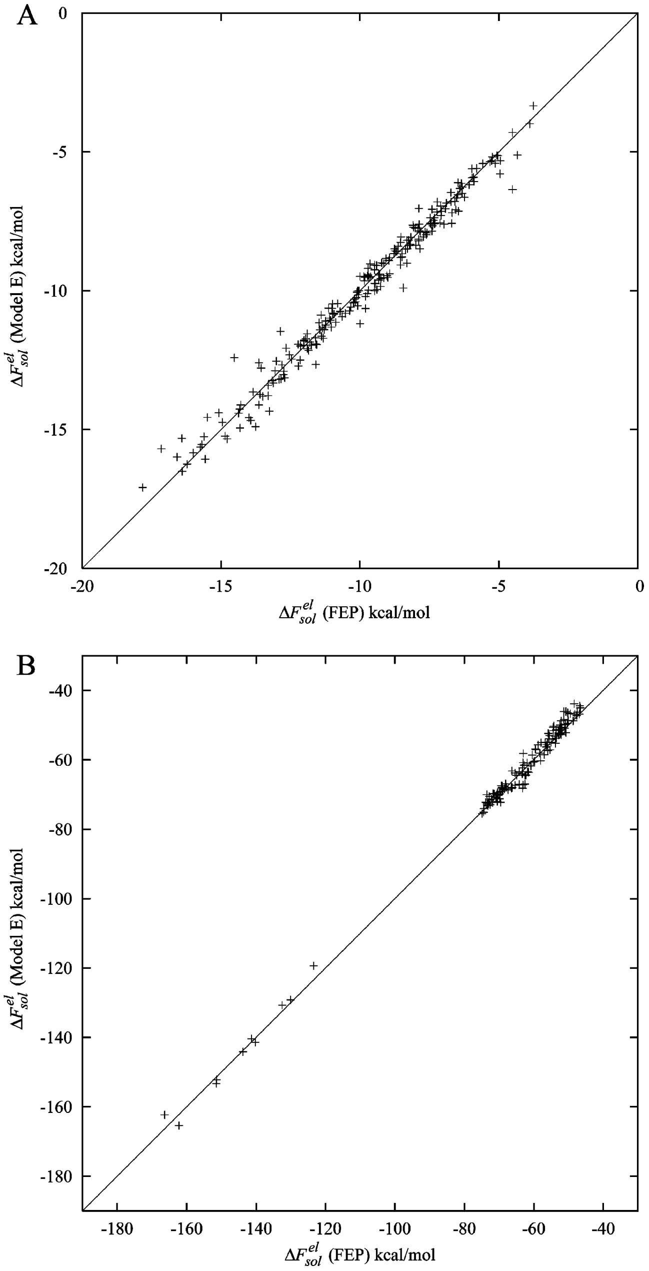


Figure 7. FEP calculated and estimated electrostatic components of the free energy of hydration ($\Delta F^{\mathrm{el}}_{\mathrm{sol}}$) for model E for the neutral (A) and ionic (B) molecules in the test set. All values are in kcal/mol.

E, and this yields an rms error of 13.1 kcal/mol. The predicted absolute electrostatic component of the solvation energy is overestimated in each case and shows that lower β values are appropriate for these molecules.[21] A possible explanation is that $\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{w}}_{A} - \langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle^{\mathrm{g}}_{A}$ is not negligible, which was assumed in the derivation of eq 10. However, including this term in eq 10 does not improve the results significantly (rms = 12.5 kcal/mol). Optimizing the coefficient gives $\beta = 0.39$, which is similar to that obtained for neutral compounds, and reduces the rms error significantly (rms = 3.8 kcal/mol). Further analysis, however, indicates that the zwitterions do not behave like the other neutral solutes. A plausible explanation is the inability of LR to properly describe the free energy differences associated with the large conformational changes upon charging flexible zwitterions (to some extent this probably also applies to FEP calculations in general, where charging of multiply ionized flexible compounds can be associated with severe convergence problems). In order to test this hypothesis, the electrostatic contribution to the free energies of solvation was calculated for a series of rigid zwitterions (Figure 8), for which the intramolecular energy contribution will be negligible. In excellent agreement with model E, the calculated β_{FEP} values for the rigid zwitterions are all 0.47 (Table 3). The observed differences in β_{FEP} values for flexible and rigid zwitterions suggests an alternative formulation of the LR approximation for solvation free energies, with intra- and intermolecular contributions scaled separately using the expression

$$\Delta F^{\mathrm{el}}_{\mathrm{sol}} = \beta^{\mathrm{inter}}\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_{B} + \beta^{\mathrm{intra}}\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_{B} \qquad (22)$$

where β^{inter} and β^{intra} are scaling factors for inter- (r−s) and the change in intramolecular (r−r) energies, respectively. For the flexible zwitterions, this yields $\beta^{\mathrm{inter}} = 0.48$, in excellent agreement with model E, and a $\beta^{\mathrm{intra}} = 0.66$, with an rms of 1.9 kcal/mol. This clearly shows that the anomalous β_{FEP} values obtained for the flexible zwitterions originate from inaccuracies in the LR assumption for solute−solute energies and not from solute−solvent energies. Adding another scaling factor to eq 10 would, of course, add more complexity to the model, and, as shown above, model E reproduces the

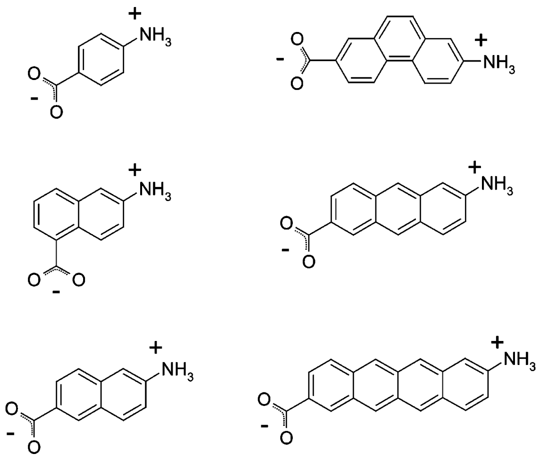


Figure 8. The compounds used to investigate the linear response approximation for rigid zwitterions.

Table 3. Electrostatic Solute Energies ($\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_{B}$ and $\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_{B}$) and Electrostatic Contribution to the Free Energies of Hydration from FEP Calculations ($\Delta F^{\mathrm{el}}_{\mathrm{sol}}$(FEP)) for the Rigid Zwitterions (Figure 6)

solute	$\Delta F^{\mathrm{el}}_{\mathrm{sol}}$(FEP)[a,b]	$\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle^{\mathrm{w}}_{B}$ [a,c]	$\Delta\langle U^{\mathrm{el}}_{\mathrm{r-r}}\rangle_{B}$ [a,d]	β_{FEP}
ZW1	−101.8	−220.2	1.7	0.47
ZW2	−116.3	−249.2	1.2	0.47
ZW3	−108.7	−234.0	1.6	0.47
ZW4	−125.3	−266.1	1.0	0.47
ZW5	−125.6	−267.3	0.9	0.47
ZW6	−131.9	−280.0	0.8	0.47

[a] kcal/mol. [b] Average uncertainties are 0.0 and 0.4 kcal/mol for the gas and water phase, respectively. [c] Average uncertainties are 0.7 kcal/mol. [d] Average uncertainties are 0.1 and 0.0 kcal/mol for the gas and water phase, respectively.

FEP calculated values rather well and suggests that such a model is not necessary for nonzwitterionic compounds. In fact, using $\beta^{\mathrm{inter}} = \beta_{\mathrm{model\ E}}$ and parametrizing β^{intra} on the test set (excluding zwitterions) yields $\beta^{\mathrm{intra}} = 0.48$, but this approach does neither improve nor diminish the agreement with the electrostatic component of the solvation free energies calculated using FEP. Hence, eq 10 is preferable to eq 22 for all molecules except flexible zwitterions, for which separate scaling factors appear necessary in order to obtain accurate results.

All the models investigated in this work are parametrized using eq 10 and the complete thermodynamic cycle in Figure 1, which differs from the approach used by Åqvist and Hansson[21] who used eq 11 along with the upper thermodynamic cycle of Figure 1. As described above, the change in intramolecular energies in going from gas to water for the training set are close to zero for all compounds, and thus the derived β values can be expected to be identical using these two approaches. The test set, however, contains compounds which may change significantly in intramolecular energies when going from gas to water phase. In order to test the benefit, if any, of using the thermodynamic cycle in which all electrostatic solute interactions are turned on as compared to the cycle of only turning on the solute−solvent electrostatic interactions, we have performed FEP calculations using both these approaches for all the nonionic compounds in the test set. Using eq 10 along with the full thermodynamic cycle on the nonionic compounds in the test set yields an rms error of 0.20 kcal/mol, while Åqvist and Hansson's approach[21] using eq 11 yields an rms error of 0.80 kcal/mol (data not shown). In particular, large errors are observed using Åqvist and Hansson's approach[21] for molecules which can form internal hydrogen bonds. Hence, for flexible molecules, the full thermodynamic cycle employed in this work seems to perform better.

Semiempirical Prediction of the Total Free Energy of Solvation. The total solvation free energy is estimated here by combining the predicted electrostatic part of the solvation energy with an empirical term for the nonpolar contribution. Experimental hydration free energies[43] were available for 194 of the molecules in the training set, and net neutral and ionic molecules are presented separately.

Using eq 19 combined with β values from model E to calculate hydration free energies of the neutral compounds results in a parametrization of $\alpha_{\mathrm{w}}^{\mathrm{vdW}} = 0.01$ and $\gamma_{\mathrm{w}}^{\mathrm{vdW}} = 1.18$ kcal/mol with an rms of 1.1 kcal/mol (Figure 9A) (using $\langle U_{\mathrm{r-s}}^{\mathrm{vdW}}\rangle_A^{\mathrm{w}}$ as a size measure in eq 19 yields similar results (rms = 1.1 kcal/mol)). The largest individual errors are obtained for secondary and tertiary amides, for which OPLS-AA and many other force fields are known to have problems reproducing experimental hydration free energies.[44,45] Excluding secondary and tertiary amides yields an rms of 0.94 kcal/mol and $r^2 = 0.86$ for eq 19. The low value of the $\alpha_{\mathrm{w}}^{\mathrm{vdW}}$ coefficient suggests that the $\alpha_{\mathrm{w}}^{\mathrm{vdW}}\langle U_{\mathrm{r-s}}^{\mathrm{vdW}}\rangle_B^{\mathrm{w}}$ term is almost insignificant in improving the prediction of hydration free energies, and exclusion of the $\alpha_{\mathrm{w}}^{\mathrm{vdW}}\langle U_{\mathrm{r-s}}^{\mathrm{vdW}}\rangle_B^{\mathrm{w}}$ term does, in fact, yield similar results ($\Delta F_{\mathrm{sol}} = \beta\Delta\langle U^{\mathrm{el}}\rangle + \gamma$, rms = 1.1 kcal/mol). Optimizing all three coefficients in eq 19 yields $\beta = 0.42$, $\alpha_{\mathrm{w}}^{\mathrm{vdW}} = 0.05$, and $\gamma_{\mathrm{w}}^{\mathrm{vdW}} = 1.65$ and gives a slightly worse agreement with experiment (rms = 1.2 kcal/mol), which emphasizes the importance of group specific β-values. Excluding the change in intramolecular energies, i.e., using eq 11 instead of eq 10 in eq 19, also gives similar results (rms = 1.1 kcal/mol), which shows that there are only small differences in solute conformations in gas and aqueous phase for the training set. Although several studies[46,47] have shown that a combination of $\langle U_{\mathrm{r-s}}^{\mathrm{vdW}}\rangle_B$ and surface area gives the best description of the nonpolar contribution to hydration, it is of questionable statistical merit to introduce free scaling factors for both these terms because they are strongly correlated (Figure 2).[27] An alternative approach is to replace the nonpolar contribution in eq 19 ($\alpha_{\mathrm{w}}^{\mathrm{vdW}}\langle U_{\mathrm{r-s}}^{\mathrm{vdW}}\rangle_B^{\mathrm{w}} + \gamma_{\mathrm{w}}^{\mathrm{vdW}}$)

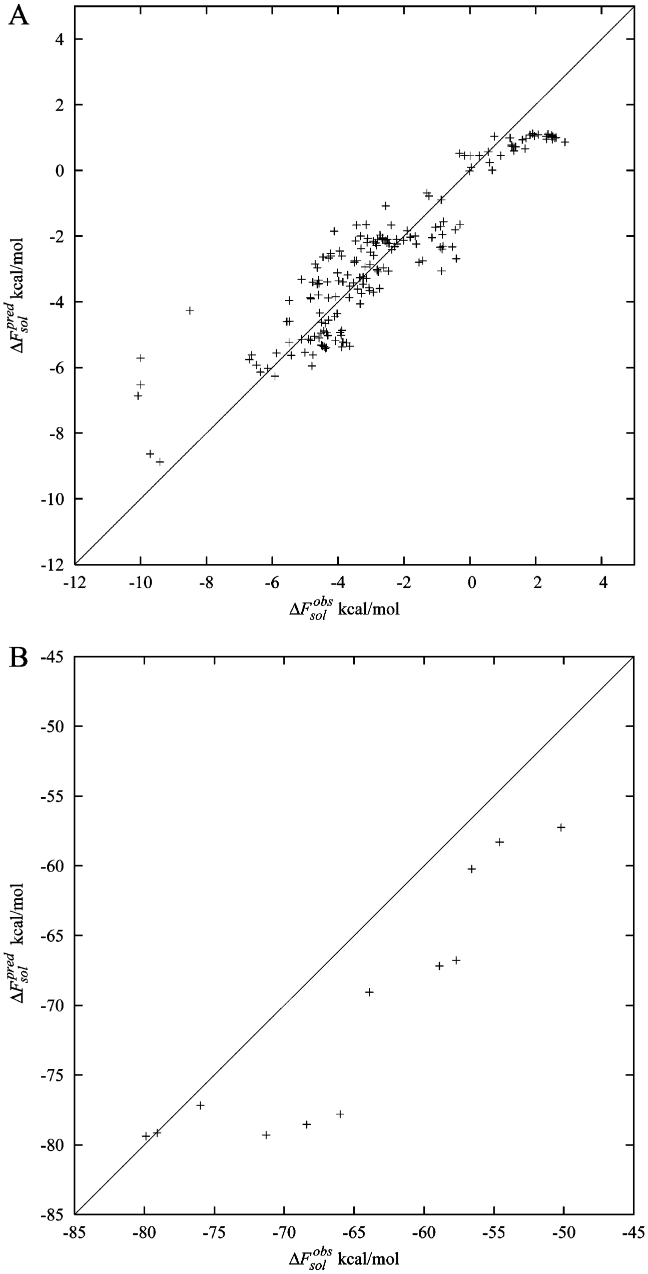


Figure 9. Experimental[43] and predicted absolute hydration free energies (kcal/mol) for the neutral (A) and ionic (B) molecules using eq 19 with $\alpha = 0.01$, $\beta = \beta_{\mathrm{model\ E}}$ and $\gamma = 1.18$ kcal/mol.

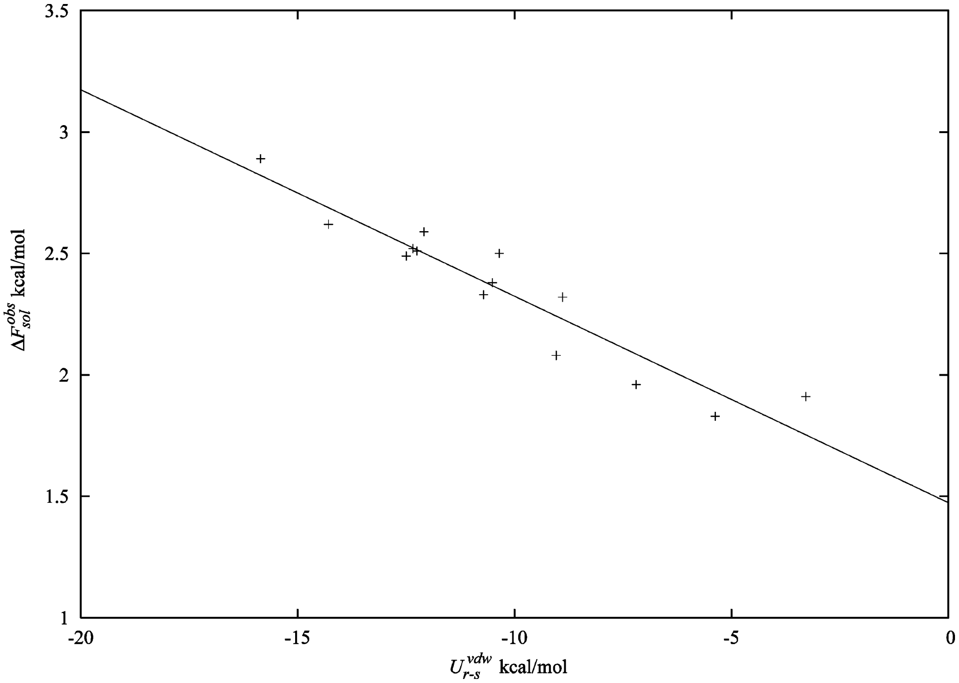


Figure 10. Correlation between hydration free energies,[43] $\Delta F_{\text{sol}}{}^{\text{obs}}$, (kcal/mol) for linear and branched alkanes and the solute−solvent van der Waals energy, $\langle U_{\text{r-s}}^{\text{vdW}}\rangle_A^{\text{w}}$ (kcal/mol).

with $\gamma_{\text{S}}{\cdot}A_{\text{S}} + \langle U_{\text{r-s}}^{\text{vdW}}\rangle_B + \gamma$.[37] Parametrization of γ_{s} and γ yields an rms error of 1.4 kcal/mol and $\gamma_{\text{s}} = 0.10$ kcal mol^{-1} Å^{-2} and $\gamma = -2.1$ kcal mol^{-1} which is slightly worse than the model using eq 19. Overall, the best model found in the course of this work, which, however, requires an additional simulation of the solute in the uncharged state, is

$$\Delta F_{\text{sol}} = \gamma_{\text{S}}{\cdot}A_{\text{S}} + \langle U_{\text{r-s}}^{\text{vdW}}\rangle_A^{\text{w}} + \beta(\langle U_{\text{r-s}}^{\text{el}}\rangle_B^{\text{w}} + \Delta\langle U_{\text{r-r}}^{\text{el}}\rangle_B) \quad (23)$$

Equation 23 incorporates the nonpolar approximation of eq 17, treating γ_{S} as a free parameter, along with the polar approximation of model E. Parametrizing eq 23 on experimental hydration free energies for the neutral compounds yields an rms error of 0.82 kcal/mol with $\gamma_{\text{S}} = 0.09$ kcal mol^{-1} Å^{-2}. Adding a constant to eq 23 does not improve the results significantly (rms = 0.82 kcal/mol). Interestingly the parametrized value of γ_{S} agrees very well with the experimental surface tension of 0.105 kcal mol^{-1} Å^{-2}.

For the ionic molecules, the $\alpha_{\text{w}}^{\text{vdW}}$ and $\gamma_{\text{w}}^{\text{vdW}}$ coefficients as parametrized above were used in eq 19. For each molecule the free energy contribution arising from interactions with water outside the simulation sphere boundary were calculated using the Born equation[17] which yields −9.1 kcal/mol. While reasonable agreement was obtained for anions (rms = 0.75 kcal/mol for eq 19), large errors compared to experiment were observed for the cations (rms = 9.66 kcal/mol for eq 19) using model E in eq 19 (Figure 9B). Since the calculated FEP energies were very well reproduced by model E (Figure 7), it must be the actual nonbonded force field parameters that do not reproduce experimental absolute solvation energies in this case. This has also been observed previously for ammonium.[48] Hence, in order to reproduce experimental solvation energies, reparameterization of the OPLS-AA force field for charged amines seems necessary.

The weak correlation between the nonpolar contribution to the hydration free energy of these compounds and the size descriptors was surprising considering how well this relationship has been documented experimentally.[16,49] The same observation was also made for SASA in a recent study using PB and GB electrostatic hydration energies.[50] One possible explanation is that although there is a clear correlation between nonpolar hydration energies and size, the difference in nonpolar hydration energy between the molecules in our set is relatively small. This combined with inaccuracies of force field parameters and experimental values makes it difficult to accurately predict this quantity. For example, in the case of linear and branched alkanes, where the electrostatic contribution is negligible, there is a strong correlation between $\langle U_{\text{r-s}}^{\text{vdW}}\rangle_A^{\text{w}}$ and the experimental hydration free energy. The slope is, however, relatively small, $\Delta F_{\text{sol}} = -0.08\langle U_{\text{r-s}}^{\text{vdW}}\rangle_A^{\text{w}} + 1.5$ kcal/mol, and the difference in experimental hydration energy between the largest (octane) and the smallest (methane) molecule is only 1.1 kcal/mol (Figure 10).[43] For larger solutes or other solvents, e.g., n-hexane, where the corresponding relation between solute−solvent van der Waals interactions and nonpolar solvation energies is associated with considerably steeper slopes, a constant term would clearly not be accurate.[16] It is also noteworthy here, that our estimate in section 2 ($\Delta F_{\text{sol}}^{\text{np}} = -0.07\langle U_{\text{r-s}}^{\text{vdW}}\rangle_A^{\text{w}} + 1.9$ kcal/mol) of the above relationship, from the experimental surface tension of water together with the correlation between molecular surface and van der Waals energy for nonpolar solutes, is surprisingly accurate.

Jorgensen and co-workers have proposed another LR variant to estimate free energies of hydration from microscopic simulations.[27−29] In their approach, eq 11 is combined with both SASA and the solute−solvent van der Waals interaction energy as in eq 24

$$\Delta F^{\mathrm{solv}} = \alpha'\langle U^{\mathrm{vdW}}_{\mathrm{r-s}}\rangle_B + \beta'\langle U^{\mathrm{el}}_{\mathrm{r-s}}\rangle_B + \gamma'\langle \mathrm{SASA}\rangle_B \quad (24)$$

but with α', β', and γ' treated as free scaling factors (the use of SASA instead of MS is not very important, although the latter quantity is more compatible with the experimental surface tension). Using eq 24 for the neutral compounds in our set yields $\alpha' = 0.38$, $\beta' = 0.49$, and $\gamma' = 0.02$ kcal mol^{-1} Å^{-2} with rms = 1.1 kcal/mol, which is the same quality as obtained using eq 19, and the obtained coefficients agree nicely with the coefficients obtained by Jorgensen and co-workers[28] using OPLS-AA charges (they obtained $\alpha' = 0.42$, $\beta' = 0.49$, and $\gamma' = 0.02$ kcal mol^{-1} Å^{-2}). The α', β', and γ' coefficients derived by Jorgensen and co-workers also appear to vary depending on the chosen scheme to assign partial charges for the molecules used in the parametrization when these are carried out on experimental hydration free energies.[27−29] It is important to note that this does not imply that there is a force field dependence of the β coefficient. The varying β parameters obtained by Jorgensen and co-workers rather reflect force field deficiencies, i.e., the chosen charge scheme does not reproduce experimental hydration free energies. For example, when EPS charges are used,[51] which are more polarized than OPLS-AA charges, smaller values of the β coefficient will be required to reproduce experiment. Since different values of the α' and γ' coefficients are obtained for the different charge schemes, they not only reflect nonpolar contributions to the solvation free energy but also provide compensation for possible force field errors.[27−29,51] Our models were instead optimized to reproduce FEP results, and thus the accuracy of the models when compared to experiment will be limited by the accuracy of the force field.

Implications for Calculations of Protein−Ligand Binding Free Energies. The scheme derived by Hansson et al. for predicting β_{FEP} (model B)[23] has successfully been used in the LIE method for predicting binding free energies of ligands binding to their receptors.[52−57] In the LIE method, the binding free energy is estimated in analogy with solvation energies as the free energy of transfer between water and protein environments. Simulations are carried out for the ligand in water and the solvated protein system, and the Gibbs free energy of binding is calculated from the ligand-surrounding (l-s) (the ligand's interactions with both protein and solvent atoms) electrostatic (el) and van der Waals (vdW) interaction energies

$$\Delta G^{\mathrm{LIE}}_{\mathrm{bind}} = \alpha\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle + \beta\Delta\langle U^{\mathrm{el}}_{\mathrm{l-s}}\rangle + \gamma \quad (25)$$

where the Δ's refer to differences in protein and water simulations.[22] The standard parametrization of the model was derived with β values according to model B using a set of 18 protein−ligand complexes, and the optimal value of α was found to be 0.18.[23] For this data set the constant γ was found to be 0.0, which is not always the case.[58] Note that the derived values of α and γ in eq 25 cannot be directly compared to $\alpha^{\mathrm{vdW}}_{\mathrm{w}}$ and $\gamma^{\mathrm{vdW}}_{\mathrm{w}}$ in eq 19, as attempted by Almlöf et al.[58] That is, if eq 18 is used to express a relation between nonpolar solvation free energies in protein and solvent environments, we obtain

$$\Delta\Delta G^{\mathrm{np}}_{\mathrm{sol}} = \Delta G^{\mathrm{np,p}}_{\mathrm{sol}} - \Delta G^{\mathrm{np,w}}_{\mathrm{sol}} = \alpha_{\mathrm{p}}\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle - \alpha_{\mathrm{w}}\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle + \gamma_{\mathrm{p}} - \gamma_{\mathrm{w}} \quad (26)$$

which cannot be rewritten in terms of $\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle$ to identify α and γ in eq 25. Rather, α and γ can be derived from relations relating size to the change in ligand-surrounding van der Waals interactions ($\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle$) and the nonpolar free energy of solvation ($\Delta\Delta G^{\mathrm{np}}_{\mathrm{sol}}$) between protein and water environments[59]

$$\Delta\Delta G^{\mathrm{np}}_{\mathrm{sol}} = a\sigma + b$$

$$\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle = c\sigma + d$$

$$\Rightarrow \Delta\Delta G^{\mathrm{np}}_{\mathrm{sol}} = \frac{a}{c}(\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle - d) + b = \alpha\Delta\,\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle + \gamma \quad (27)$$

where σ is a size measure, such as MS, SASA, or the number of heavy atoms in the ligand, and a, b, c, and d are empirically derived parameters. From eq 27, the contributions from nonpolar solvation to α and γ in eq 25 can be identified as *a/c* and *b-ad/c*, respectively. Since the parametrization of LIE was performed using experimental binding free energies, $\alpha = 0.18$ takes into account van der Waals interactions and all other size dependent contributions to binding, e.g., the hydrophobic effect and relative translational and rotational entropies.[59] As noted above, the nonpolar contribution to the hydration free energy was small and could be well represented by a constant term for our data set. In contrast to hydration free energies however, the nonpolar term in eq 25 often makes a significant contribution to the binding free energy and cannot be represented by a constant term.

In our development of the LIE method, the ligand-surrounding electrostatic energies in both the protein and water simulations are scaled by the same factor. The original idea of LIE was that the β coefficient would not be used as a free parameter and, even though the β coefficient for the protein environment is to some extent uncertain and deserves further investigation, it is somewhat questionable to optimize β in eq 25 freely. In several published attempts at reproducing binding free energies using LIE, the electrostatic scaling factor in the LIE method is sometimes found to be very small and in a few cases even negative.[60−63] In the present work, it is clear that for ligands in aqueous phase a value of $\beta = 0.37-0.52$ is appropriate, and, therefore, in the above-mentioned problematic cases, it would hence make more sense to scale the electrostatic ligand-surrounding energies of the water and protein simulations separately and only treat the scaling of the electrostatics in the protein simulation (β_{prot}) as a free parameter

$$\Delta G^{\mathrm{LIE}}_{\mathrm{bind}} = \alpha\Delta\langle U^{\mathrm{vdW}}_{\mathrm{l-s}}\rangle + \beta_{\mathrm{prot}}\langle U^{\mathrm{el}}_{\mathrm{l-s}}\rangle_{\mathrm{p}} - \beta_{\mathrm{wat}}\langle U^{\mathrm{el}}_{\mathrm{l-s}}\rangle_{\mathrm{w}} + \gamma \quad (28)$$

This approach was actually suggested initially,[23] but to this date there has been no reason to introduce the increased amount of complexity into the model.

The results presented here also show how intramolecular energies can be explicitly included in the LIE method. In analogy to solvation free energies, this would lead to the introduction of an intramolecular term in eq 25, yielding

$$\Delta G_{\mathrm{bind}}^{\mathrm{LIE}} = \alpha\Delta\langle U_{\mathrm{l-s}}^{\mathrm{vdW}}\rangle + \beta(\Delta\langle U_{\mathrm{l-s}}^{\mathrm{el}}\rangle + \Delta\langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle) + \gamma \quad (29)$$

$\Delta\langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle = \langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle_{\mathrm{p}} - \langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle_{\mathrm{w}}$, where $\langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle_{\mathrm{p}}$ and $\langle U_{\mathrm{l-l}}^{\mathrm{el}}\rangle_{\mathrm{w}}$ are the intramolecular ligand−ligand energies in the bound and free state, respectively. It should be noted, in analogy with the two different cycles used for solvation free energies (Figure 1), that eqs 25 and 29 are both rigorously derived. The accuracy of either equation will ultimately depend on the appropriateness of the approximations used in eqs 25 and 29 to predict the relevant legs of each thermodynamic cycle.

5. Conclusions

In this work, a LR approach to estimate the electrostatic component of the free energy of solvation has been presented. The main result is that derivation of scaling factors for specific chemical groups yields remarkable agreement with the exact results calculated using the FEP method. For molecules containing several chemical groups, a scheme for deriving specific values of β for each compound was proposed, and this was shown to yield impressive results on a large data set not included in the parametrization. For estimates of the total hydration free energy, the electrostatic component was combined with an empirical size dependent treatment of the nonpolar contribution to the free energy, and the results are in good agreement with experiment. The results reported herein should be useful for predicting free energies of solvation and also to improve the accuracy of simplified binding free energy calculations.

Abbreviations. MD, molecular dynamics; FEP, free energy perturbation; rms, root mean square; LIE, linear interaction energy; MC, Monte Carlo; TI, thermodynamic integration; LR, linear response.

Acknowledgment. Support from the Swedish Foundation for Strategic Research (SSF/Rapid) and the Swedish Research Council (VR) is gratefully acknowledged. The computations were performed on UPPMAX resources under project p2004003.

Supporting Information Available: Electrostatic and Lennard-Jones interaction energies, free energies of charging, predicted free energies of charging using models D/E, calculated β_{FEP} values, and experimental hydration free energies for the training set and electrostatic interaction energies, free energies of charging, and predicted free energies of charging for the test set. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050−3054.

(2) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564−568.

(3) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876−5886.

(4) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021−8024.

(5) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161−185.

(6) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.

(7) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740−5761.

(8) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656−1676.

(9) Oostenbrink, C.; Juchli, D.; van Gunsteren, W. F. *ChemPhysChem* **2005**, *6*, 1800−1804.

(10) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616−17626.

(11) Mezei, M.; Beveridge, D. L. *Ann. N. Y. Acad. Sci.* **1986**, *482*, 1−23.

(12) Åqvist, J. *J. Comput. Chem.* **1996**, *17*, 1587−1597.

(13) Hermans, J.; Wang, L. *J. Am. Chem. Soc.* **1997**, *119*, 2707−2714.

(14) Abraham, M. H. *J. Am. Chem. Soc.* **1979**, *101*, 5477−5484.

(15) Abraham, M. H. *J. Am. Chem. Soc.* **1982**, *104*, 2085−2094.

(16) Blokzijl, W.; Engberts, J. B. F. N. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1545−1579.

(17) Born, M. *Z. Phys.* **1920**, *1*, 45−48.

(18) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671−679.

(19) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144−1149.

(20) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(21) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512−9521.

(22) Åqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385−91.

(23) Hansson, T.; Marelius, J.; Åqvist, J. *J. Comput-Aided. Mol. Des.* **1998**, *12*, 27−35.

(24) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. *Adv. Prot. Chem.* **2003**, *66*, 123−158.

(25) Del Buono, G. S.; Figueirido, F. E.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 85−97.

(26) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458−4472.

(27) Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.

(28) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667−10673.

(29) McDonald, N. A.; Carlson, H. A.; Jorgensen, W. L. *J. Phys. Org. Chem.* **1997**, *10*, 563−576.

(30) Zwanzig, R. *J. Chem. Phys.* **1954**, *22*, 1420−1426.

(31) Kubo, R. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100−1120.

(32) Åqvist, J.; Hansson, T. *J. Phys. Chem. B* **1998**, *102*, 3837−3840.

(33) Vorobjev, Y. N.; Hermans, J. *J. Phys. Chem. B* **1999**, *103*, 10234−10242.

(34) Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 124106.

(35) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717−726.

(36) Westergren, J.; Lindfors, L.; Hoglund, T.; Luder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872−1882.

(37) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. *J. Chem. Theory Comput.* **2007**, *3*, 256−277.

(38) Marelius, J.; Kolmodin, K.; Feierberg, I.; Åqvist, J. *J. Mol. Graphics Modell.* **1998**, *16*, 213-+.

(39) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(40) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(41) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647−3661.

(42) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100−3107.

(43) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563−595.

(44) Morgantini, P. Y.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 6057−6063.

(45) Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322−1332.

(46) Zacharias, M. *J. Phys. Chem. A* **2003**, *107*, 3000−3004.

(47) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523−9530.

(48) Luzhkov, V. B.; Almlöf, M.; Nervall, M.; Åqvist, J. *Biochemistry* **2006**, *45*, 10807−10814.

(49) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016−2027.

(50) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128−139.

(51) Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. *J. Comput. Chem.* **1993**, *14*, 1240−1249.

(52) Marelius, J.; Graffner-Nordberg, M.; Hansson, T.; Hallberg, A.; Åqvist, J. *J. Comput-Aided. Mol. Des.* **1998**, *12*, 119−131.

(53) Graffner-Nordberg, M.; Kolmodin, K.; Åqvist, J.; Queener, S. F.; Hallberg, A. *J. Med. Chem.* **2001**, *44*, 2391−2402.

(54) Luzhkov, V. B.; Åqvist, J. *FEBS Lett.* **2001**, *495*, 191−196.

(55) Ersmark, K.; Feierberg, I.; Bjelic, S.; Hultén, J.; Samuelsson, B.; Åqvist, J.; Hallberg, A. *Bioorg. Med. Chem.* **2003**, *11*, 3723−3733.

(56) Leiros, H. K. S.; Brandsdal, B. O.; Andersen, O. A.; Os, V.; Leiros, I.; Helland, R.; Otlewski, J.; Willassen, N. P.; Smalås, A. O. *Protein Sci.* **2004**, *13*, 1056−1070.

(57) Ersmark, K.; Feierberg, I.; Bjelic, S.; Hamelink, E.; Hackett, F.; Blackman, M. J.; Hultén, J.; Samuelsson, B.; Åqvist, J.; Hallberg, A. *J. Med. Chem.* **2004**, *47*, 110−122.

(58) Almlöf, M.; Brandsdal, B. O.; Åqvist, J. *J. Comput. Chem.* **2004**, *25*, 1242−1254.

(59) Carlsson, J.; Åqvist, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385−5395.

(60) Lamb, M. L.; Tirado-Rives, J.; Jorgensen, W. L. *Bioorg. Med. Chem.* **1999**, *7*, 851−860.

(61) Tounge, B. A.; Reynolds, C. H. *J. Med. Chem.* **2003**, *46*, 2074−2082.

(62) Stjernschantz, E.; Marelius, J.; Medina, C.; Jacobsson, M.; Vermeulen, N. P. E.; Oostenbrink, C. *J. Chem. Inf. Model.* **2006**, *46*, 1972−1983.

(63) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517−529.

CT700106B

JCTC Journal of Chemical Theory and Computation

Stability of $N_{10}C_{10}H_{10}$ and $N_{12}C_{12}H_{12}$ Cages and the Effects of Endohedral Atoms and Ions

DeAna McAdory,[‡] Jacqueline Jones,[†] Ami Gilchrist,[†] Danielle Shields,[†]
Ramola Langham,[‡] Kasha Casey,[‡] and Douglas L. Strout*,[‡]

Departments of Biological Sciences and Physical Sciences, Alabama State University, Montgomery, Alabama 36101

Received April 20, 2007

Abstract: Cages of carbon and nitrogen have been studied by theoretical calculations to determine the potential of these molecules as high-energy density materials. Following previous theoretical studies of high-energy $N_6C_6H_6$ and $N_8C_8H_8$ cages, a series of calculations on several isomers of the larger $N_{10}C_{10}H_{10}$ and $N_{12}C_{12}H_{12}$ is carried out to determine relative stability among a variety of three-coordinate cage isomers with four-membered, five-membered, and/or six-membered rings. Additionally, calculations are carried out on the same molecules with atoms or ions inside the cage. Calculations are carried out with the B3LYP and PBE1PBE density functional (DFT) methods, with MP2 and MP4 calculations carried out to evaluate the accuracy of the DFT results. Trends in stability with respect to cage geometry and arrangements of atoms are calculated and discussed. Stability effects caused by the endohedral atoms and ions are also calculated and discussed.

Introduction

Nitrogen molecules have been the subjects of many recent studies because of their potential as high-energy density materials (HEDM). An all-nitrogen molecule N_x can undergo the reaction $N_x \rightarrow (x/2)N_2$, a reaction that can be exothermic by 50 kcal/mol or more per nitrogen atom.[1,2] To be a practical energy source, however, a molecule N_x would have to resist dissociation well enough to be a stable fuel. Theoretical studies[3−7] have shown that numerous N_x molecules are not sufficiently stable to be practical HEDM, including cyclic and acyclic isomers with 8−12 atoms. Cage isomers of N_8 and N_{12} have also been shown[7−10] by theoretical calculations to be unstable. Experimental progress in the synthesis of nitrogen molecules has been very encouraging, with the N_5^+ and N_5^- ions having been recently produced[11,12] in the laboratory. More recently, a network polymer of nitrogen has been produced[13] under very high-pressure conditions. Experimental successes have sparked theoretical studies[1,14,15] on other potential all-nitrogen molecules. More recent developments include the experimental synthesis of high-energy molecules consisting predominantly of nitrogen, including azides[16,17] of various molecules and polyazides[18,19] of atoms and molecules, such as 1,3,5-triazine. Future developments in experiment and theory will further broaden the horizons of high-energy nitrogen research.

The stability properties of N_x molecules have also been extensively studied in a computational survey[20] of various structural forms with up to 20 atoms. Cyclic, acyclic, and cage isomers have been examined to determine the bonding properties and energetics over a wide range of molecules. A more recent computational study[21] of cage isomers of N_{12} examined the specific structural features that lead to the most stable molecules among the three-coordinate nitrogen cages. Those results showed that molecules with the most pentagons in the nitrogen network tend to be the most stable, with a secondary stabilizing effect due to triangles in the cage structure. A recent study[22] of larger nitrogen molecules N_{24}, N_{30}, and N_{36} showed significant deviations from the pentagon-favoring trend. A computational study[23] of the even larger cylindrical cage N_{72} has been carried out to elucidate the bonding properties of cylindrical nitrogen. Each of these molecule sizes has fullerene-like cages consisting solely of

* Corresponding author e-mail: dstrout@alasu.edu.
[†] Department of Biological Sciences.
[‡] Department of Physical Sciences.

10.1021/ct700090h CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/29/2007

pentagons and hexagons, but a large stability advantage was found for molecules with fewer pentagons, more triangles, and an overall structure more cylindrical than spheroidal. Studies[24,25] of intermediate-sized molecules N_{14}, N_{16}, and N_{18} also showed that the cage isomer with the most pentagons was not the most stable cage, even when compared to isomer containing triangles (which have 60° angles that should have significant angle strain). For each of these molecule sizes, spheroidally shaped molecules proved to be less stable than elongated, cylindrical ones.

However, while it is possible to identify in relative terms which nitrogen cages are the most stable, it has been shown[7] in the case of N_{12} that even the most stable N_{12} cage is unstable with respect to dissociation. The number of studies demonstrating the instability of various all-nitrogen molecules has resulted in considerable attention toward compounds that are predominantly nitrogen but contain heteroatoms that stabilize the structure. In addition to the experimental studies[16−18] cited above, theoretical studies have been carried out that show, for example, that nitrogen cages can be stabilized by oxygen insertion[26,27] or phosphorus substitution.[28]

A study[29] of carbon−nitrogen cages showed that carbon substitution into an N_{12} cage results in a stable $N_6C_6H_6$, but the only isomer considered was one in which the six carbon atoms replaced the nitrogen atoms in the two axial triangles of the original N_{12}. A further study[30] of several isomers of $N_6C_6H_6$ showed that, for substitutions of carbon atoms into an N_{12} cage, the most stable isomers were the ones with the largest number of C−N bonds. Also, the isomers with the highest number of C−N bonds also had the highest dissociation energies in the N−N bonds, which is significant because the N−N were weaker than other bonds in the cage. The strength of the N−N bonds, therefore, plays a key role in the overall stability of the molecules with respect to dissociation. A similar study[31] of numerous cage isomers of $N_8C_8H_8$ further illustrated the stabilizing effects of heteronuclear bonds. That study also showed that the N−N bonds in the $N_8C_8H_8$ cages can be strengthened by carbon atoms in the local environment.

In the current study, three isomers of $N_{10}C_{10}H_{10}$ and six isomers of $N_{12}C_{12}H_{12}$ are examined by theoretical calculations to determine their relative stability. The cages are also studied with endohedral atoms and ions. The noble gases helium, neon, and argon are studied. The ions Li^+, Be^{2+}, Na^+, Mg^{2+}, and Al^{3+} are also studied to determine their impact on the stability of the cage molecules. These molecules are also used to test the relative accuracy of density functional theory methods (specifically B3LYP and PBE1PBE). Trends of molecular stability are calculated and discussed.

Computational Methods

Geometries are optimized with two density functional theory (DFT) methods, the B3LYP method[32,33] and the PBE1PBE method.[34] Optimizations of selected molecules are carried out with second-order perturbation theory[35] (MP2). Single energy points are calculated with fourth-order perturbation theory[35] (MP4(SDQ)). The basis set is the polarized valence double-ζ (cc-pVDZ) set of Dunning.[36] Atomic charges referred to in this work are Mulliken charges. Geometry optimizations with endohedral atoms or ions are full optimizations, with the cage permitted to relax structurally. The Gaussian03 computational chemistry software[37] (along with Windows counterpart Gaussian03W) has been used for all calculations in this study.

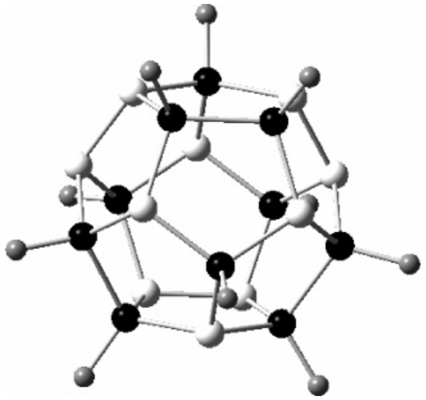


Figure 1. $N_{10}C_{10}H_{10}$ cage isomer A (C_s point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

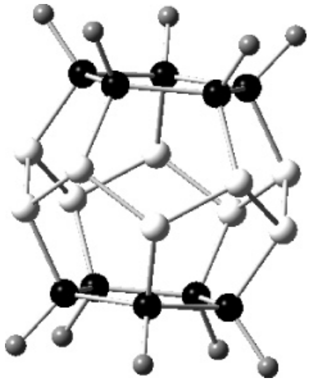


Figure 2. $N_{10}C_{10}H_{10}$ cage isomer B (D_{5d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

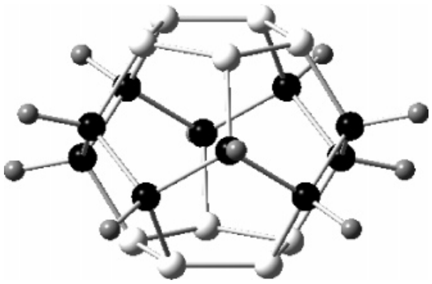


Figure 3. $N_{10}C_{10}H_{10}$ cage isomer C (D_{5d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

Results and Discussion

Three isomers of $N_{10}C_{10}H_{10}$ in this study are shown in Figures 1−3 and designated as isomers A−C. These are all based on carbon substitution on an N_{20} dodecahedron. Six isomers of $N_{12}C_{12}H_{12}$ are examined in this study, and they are shown in Figures 4−9. Each is named according to the polygons that make up the cage. The molecules in Figures 4−7 are

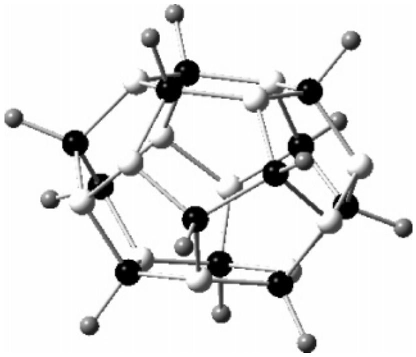


Figure 4. $N_{12}C_{12}H_{12}$ cage isomer 56A (D_{3d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

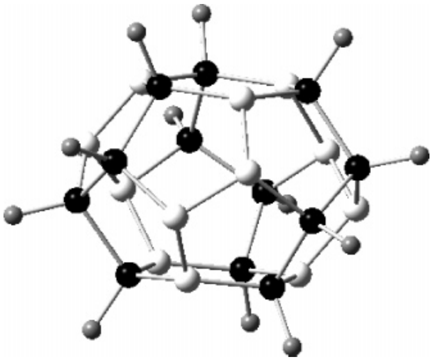


Figure 5. $N_{12}C_{12}H_{12}$ cage isomer 56B (D_{3d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

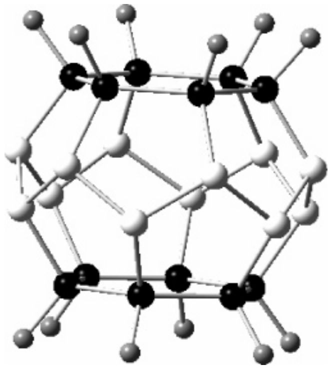


Figure 6. $N_{12}C_{12}H_{12}$ cage isomer 56C (D_{6d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

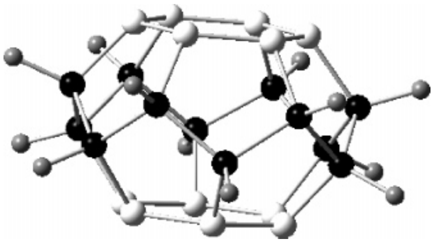


Figure 7. $N_{12}C_{12}H_{12}$ cage isomer 56D (D_{6d} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

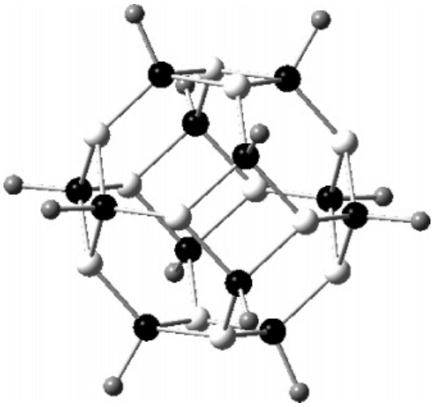


Figure 8. $N_{12}C_{12}H_{12}$ cage isomer 46 (T_h point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

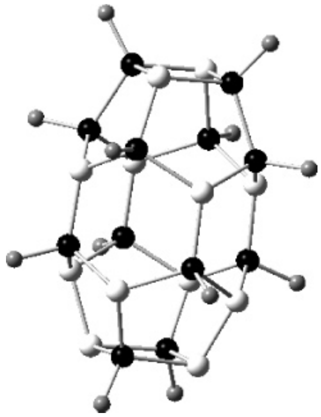


Figure 9. $N_{12}C_{12}H_{12}$ cage isomer 456 (C_{2v} point group symmetry). Nitrogen atoms are shown in white, carbon atoms in black, and hydrogen atoms in gray.

called 56A, 56B, 56C, and 56D because they are composed of five- and six-membered rings. Figures 8 and 9 show molecules 46 and 456, respectively, and they are so named because they incorporate four-membered rings. The relative energies of the $N_{10}C_{10}H_{10}$ isomers, calculated with B3LYP/cc-pVDZ and PBE1PBE/cc-pVDZ, are shown in Tables 1 and 2, respectively, and the energies of the $N_{12}C_{12}H_{12}$ cages are shown in Tables 3 and 4. Energies are shown for empty cages as well as for cages with endohedral atoms and ions. The following general trends are evident in the data.

Empty Cages. The two primary structural features that tend to destabilize these molecules are homonuclear bonds and four-membered rings. Homonuclear bonds are destabilizing for these systems because a pair of C−N bonds has higher bond enthalpy than a C−C bond and an N−N bond. Therefore, increasing the number of heteronuclear bonds increases the stability of the molecules. This is evident for the $N_{10}C_{10}H_{10}$ cages, in which isomer A has only three pairs of homonuclear bonds, whereas isomers B and C have ten such pairs. As a result, isomer A is more stable (by over 100 kcal/mol) than isomers B and C. Also, four-membered rings are destabilizing because of ring strain from the 90° (approximately) angles. The two most stable $N_{12}C_{12}H_{12}$ cages, namely 56A and 456, are the most stable because they have small numbers of homonuclear bonds and four-membered rings. Isomer 56A has three pairs of homonuclear bonds (the minimum for the 56 architecture) and zero four-membered

Table 1. Relative Energies of $N_{10}C_{10}H_{10}$ Cage Isomers Calculated with the B3LYP/cc-pVDZ Method[a]

interior	isomer A	isomer B	isomer C
empty	0.0	+119.2	+129.0
He	0.0	+110.9	+121.0
Ne	0.0	+96.2	+107.5
Ar	0.0	+67.4	+78.9
Li$^+$	0.0	+121.7	+130.3
Be^{2+}	0.0	+152.7	+156.7
Na$^+$	0.0	+105.2	+116.1
Mg^{2+}	0.0	+121.4	+130.5
Al^{3+}	0.0	+152.0	+158.4

[a] Results are shown for empty cages and cages with endohedral atoms and ions. Energies are in kcal/mol.

Table 2. Relative Energies of $N_{10}C_{10}H_{10}$ Cage Isomers Calculated with the PBe1PBE/cc-pVDZ Method[a]

interior	isomer A	isomer B	isomer C
empty	0.0	+124.3	+134.1
He	0.0	+116.5	+126.6
Ne	0.0	+102.2	+113.6
Ar	0.0	+74.9	+86.7
Li$^+$	0.0	+127.3	+136.1
Be^{2+}	0.0	+159.0	+163.0
Na$^+$	0.0	+111.3	+122.3
Mg^{2+}	0.0	+127.6	+137.0
Al^{3+}	0.0	+158.3	+165.2

[a] Results are shown for empty cages and cages with endohedral atoms and ions. Energies are in kcal/mol.

Table 3. Relative Energies of $N_{12}C_{12}H_{12}$ Cage Isomers Calculated with the B3LYP/cc-pVDZ Method[a]

interior	56A	56B	56C	56D	46	456
empty	0.0	+107.3	+170.6	+217.4	+76.9	+5.2
He	0.0	+101.7	+154.9	+208.3	+64.2	+28.9
Ne	0.0	+93.7	+136.5	+199.1	+50.2	+75.3
Ar	0.0	+74.4	+97.0	+173.2	+25.5	+146.1
Li$^+$	0.0	+108.7	+192.0	+219.2	+72.2	+23.0
Be^{2+}	0.0	+93.1	+301.3	+272.6	+118.9	+33.6
Na$^+$	0.0	+99.9	+159.8	+205.7	+50.9	+64.5
Mg^{2+}	0.0	+109.8	+208.2	+223.7	+59.3	+51.5
Al^{3+}	0.0	+102.1	+308.3	+272.9	+87.7	+61.5

[a] Results are shown for empty cages and cages with endohedral atoms and ions. Energies are in kcal/mol.

Table 4. Relative Energies of $N_{12}C_{12}H_{12}$ Cage Isomers Calculated with the PBE1PBE/cc-pVDZ Method[a]

interior	56A	56B	56C	56D	46	456
empty	0.0	+111.4	+178.6	+197.3	+78.8	+4.9
He	0.0	+106.1	+162.1	+217.2	+65.4	+27.4
Ne	0.0	+102.0	+143.4	+209.7	+51.5	+76.4
Ar	0.0	+80.6	+102.9	+188.1	+27.2	+153.2
Li$^+$	0.0	+112.8	+200.4	+228.3	+73.4	+22.6
Be^{2+}	0.0	+96.3	+309.7	+281.1	+119.1	+35.9
Na$^+$	0.0	+104.6	+166.6	+216.4	+51.5	+66.9
Mg^{2+}	0.0	+114.7	+215.3	+234.3	+59.2	+54.2
Al^{3+}	0.0	+108.0	+312.9	+282.3	+85.0	+63.7

[a] Results are shown for empty cages and cages with endohedral atoms and ions. Energies are in kcal/mol.

rings, whereas isomer 456 has two pairs of homonuclear bonds and two four-membered rings. The other isomers in this study have either at least nine pairs of homonuclear bonds (isomers 56B, 56C, and 56D) or six four-membered rings (isomer 46). Those isomers are all much higher in energy than isomers 56A and 456.

Geometric Effects and Noble Gases. In terms of structure, all three isomers of $N_{10}C_{10}H_{10}$ are based on the dodecahedron and have very similar shape. However, Tables 1 and 2 show a systematic variation in energy with the size of the noble gas atoms. With increasing atom size, isomers B and C become more stable. The answer lies in a more detailed analysis of structure. For isomers B and C, the atoms are on average 2.09 Å from the center (PBE1PBE/cc-pVDZ). Isomer A is slightly smaller, with an average distance of 2.06 Å from the center. A few hundredths of an angstrom may not seem significant, but in terms of crowding a noble gas atom (especially an argon atom with a radius of 1.74 Å), the energetic effect is substantial.

Regarding the isomers of $N_{12}C_{12}H_{12}$, the 56 isomers all have more or less the same structure. Isomer 46 is unique in that it is the most spherical of all isomers in this study. The spherical interior of isomer 46 is the best suited to enclose the progression of increasingly large noble gases (He, Ne, Ar). Tables 3 and 4 show that the relative energy of isomer 46 decreases rapidly as endohedral noble gas size is increased. Conversely, isomer 456 has a narrow, crowded center (16 atoms close to the molecule's center). Therefore, this molecule is severely strained by the inclusion of endohedral noble gases. Tables 3 and 4 show that the energy of isomer 456 increases greatly with increasing noble gas size.

Isomeric Structure and Endohedral Cations. The data in Tables 3 and 4 also show interesting variations among the $N_{12}C_{12}H_{12}$ cages regarding the endohedral metal cations. The four isomers of type 56 have very similar geometric structure, but they vary in the number of nitrogen atoms in the axial hexagons, as opposed to the equatorial belt between the pentagons. Isomers 56A and 56B have three nitrogens in each hexagon (six total), whereas 56C has all 12 nitrogens in the equatorial belt, and isomer 56D has all 12 nitrogens in the axial hexagons. The arrangement of the hexagons is important because the C−N bonds in the cage structures are polar bonds. The nitrogen atoms take on a partial negative charge because of their greater electronegativity relative to carbon.

Because of the oblate structure of the 56 framework, the atoms in the axial hexagons are closer to the molecule's center than the atoms in the equatorial belt. This explains why isomer 56D becomes more stable than 56C in the presence of highly charged endohedral cations. The cations, especially the highly charged Be^{2+} and Al^{3+}, experience a highly negatively charged environment in isomer 56D because of their proximity to the 12 nitrogens. The interaction between metal cations and negatively charged nitrogen stabilizes the entire structure. Tables 3 and 4 show that, for Be^{2+}@$N_{12}C_{12}H_{12}$ and Al^{3+}@$N_{12}C_{12}H_{12}$, isomer 56D is lower in energy than isomer 56C.

56A and 56B are stabilized relative to 56C and 56D in the presence of endohedral cations because the polarity of the C−N bonds in the hexagons causes the partial negative

Table 5. Comparison of B3LYP/cc-pVDZ and PBE1PBE/cc-pVDZ Results for $N_{12}C_{12}H_{12}$ Cages[a]

molecule/interior	B3LYP	PBE1PBE	MP2	MP4//MP2
456/empty	+5.2	**+4.9**	−2.9	−0.7
456/He	+28.9	**+27.4**	+21.0	+24.0
456/Ne	**+75.3**	+76.4	+70.6	+73.8
456/Ar	+146.1	**+153.2**	+148.8	+155.7
56B/empty	+107.3	**+111.4**	+117.1	+116.4
56B/He	+101.7	**+106.1**	+110.5	+110.6
56B/Ne	+93.7	**+102.0**	+102.4	+103.3
56B/Ar	+74.4	**+80.6**	+81.3	+84.2
46/empty	+76.9	**+78.8**	+85.3	+82.0
56C/empty	+170.6	**+178.6**	+183.7	+182.6
56D/empty	**+217.4**	+197.4	+235.1	+233.3

[a] Energies are in kcal/mol, relative to the energy of isomer 56A. For each molecule, the DFT result closest to MP4 is shown in bold.

Table 6. Energy Release Properties of $N_{12}C_{12}H_{12}$ Isomer 56A, Calculated with the PBE1PBE Method and the Cc-pVDZ Basis Set[a]

X	energy (kcal/mol)	energy (kcal/g)
none	157	0.49
He	194	0.59
Ne	250	0.73
Ar	465	1.28
Li^+	142	0.43
Be^{2+}	131	0.39
Na^+	225	0.65
Mg^{2+}	51	0.16
Al^{3+}	333	0.95

[a] The reaction $X@N_{12}C_{12}H_{12} \rightarrow 6N_2 + 2C_6H_6 + X$ is used to model the decomposition. X = endohedral atom or ion.

charge on the axial nitrogens to be greater in isomers 56A and 56B. At the PBE1PBE/cc-pVDZ level of theory, for example, the axial nitrogens of isomers 56A and 56B have charges of −0.36 and −0.28 electrons, respectively, as opposed to −0.14 electrons in isomer 56D (56C has no axial hexagon nitrogens). The larger nitrogen charges stabilize the cations in isomers 56A and 56B and lower their energies relative to 56C and 56D.

For the $N_{10}C_{10}H_{10}$ cages, the effects of oblate structure and axial vs equatorial nitrogens are a nonissue, because, in the dodecahedron, all 20 cage positions are about the same distance from the center. Therefore, placement of the nitrogen atoms is irrelevant to the interaction between the cage and the cations. The data in Tables 1 and 2 bear this out. For the empty cages, isomer B is more stable than isomer C by 10 kcal/mol. For the endohedral cations, isomer B is more stable than isomer C by 4−12 kcal/mol in all cases, so the isomer energy reversals and larger swings in energy shown for $N_{12}C_{12}H_{12}$ (especially isomers 56C and 56D) do not occur for $N_{10}C_{10}H_{10}$.

Relative Accuracy of the DFT Methods. For selected cages and interiors, MP2/cc-pVDZ geometries have been optimized, and MP4(SDQ)/cc-pVDZ energies have been calculated at the MP2 geometries. These results are used as a benchmark for determining the relative accuracy of the B3LYP and PBE1PBE methods. B3LYP is a long-standing functional with a long track record, and the PBE1PBE functional is representative of a more recent approach that has been shown[38−41] to give good results for molecules and solids. The comparison is this study is intended to test the two functionals for large molecules. The results are shown in Table 5. In nine of the 11 trials, the PBE1PBE outperformed B3LYP. In those nine trials, the PBE1PBE method recovered, on average, 51% of the energy difference between B3LYP and MP4(SDQ). In the four trials involving isomer 56B, the isomer most structurally similar to reference molecule 56A, PBE1PBE, was more successful across the board, recovering an average of 61% of the energy difference between B3LYP and MP4(SDQ).

Energy Release upon Decomposition. Table 6 shows the results of energy calculations on the reaction $X@N_{12}C_{12}H_{12} \rightarrow 6N_2 + 2C_6H_6 + X$, where X is the endohedral atom or ion. Isomer 56A is chosen for study because it is the most stable isomer of $N_{12}C_{12}H_{12}$. Since 56A has very few of the relatively weak N−N bonds and none of the sterically strained four-membered rings, it is the best candidate for a kinetically stable high-energy material. The data in Table 6, calculated at the PBE1PBE/cc-pVDZ level of theory, show the energy properties of isomer 56A of $N_{12}C_{12}H_{12}$ and the influence of the endohedral atoms/ions on the energy properties. In general, the noble gases increase the energy release of the cage because steric repulsions between the noble gases and the cage raise the energy of the cage, an effect that increases with increasing size of the noble gas atom. The first-row ions decrease the energy release of the cage, because the cage is stabilized by ion-dipole interactions between the ion and the cage. The second-row ions are most likely causing both steric interactions (destabilizing) and ion-dipole interactions (stabilizing), and therefore the influence of second-row atoms on the energy release properties is more erratic.

Conclusions

The following conclusions arise from this study: (1) Four-membered rings and homonuclear bonds are the primary destabilizing factors for $N_{10}C_{10}H_{10}$ and $N_{12}C_{12}H_{12}$ cages. (2) The ability of the cages to accommodate noble gases depends primarily on the size and shape of the interior cavity of each cage but less dependent on the precise placement of individual atoms. (3) The ability of the cages to accommodate cations is very dependent on the precise placement of atoms on the framework, especially regarding the proximity of nitrogen atoms to the molecule's center. (4) The PBE1PBE density functional method consistently outperforms B3LYP for these systems.

Acknowledgment. The Alabama Supercomputer Authority is gratefully acknowledged for a grant of computer time on the SGI Altix in Huntsville, AL. This work was partially supported by the National Computational Science Alliance under grant number CHE050022N and utilized the IBM p690 cluster in Champaign, IL. J.J., A.G., and D.S. are undergraduate scholars supported by the Minority Access to Research Careers (MARC) program administered by the National Institute of General Medical Sciences (NIH/NIGMS 5T34GM08167-20). D.L.S. is supported by MARC as a faculty research mentor. This work was supported by the National Science Foundation (NSF/HBCU-UP grant 0505872). D.M., R.L., and K.C. are supported by NSF/HBCU-UP as undergraduate scholars. This work was also supported by the National Institutes of Health (NIH/NCMHD 1P20MD-000547-01) and the Petroleum Research Fund (ACS/PRF 43798-B6). The taxpayers of the state of Alabama in particular and the United States in general are gratefully acknowledged.

References

(1) Fau, S.; Bartlett, R. J. *J. Phys. Chem. A* **2001**, *105*, 4096.

(2) Tian, A.; Ding, F.; Zhang, L.; Xie, Y.; Schaefer, H. F., III *J. Phys. Chem. A* **1997**, *101*, 1946.

(3) Chung, G.; Schmidt, M. W.; Gordon, M. S. *J. Phys. Chem. A* **2000**, *104*, 5647.

(4) Strout, D. L. *J. Phys. Chem. A* **2002**, *106*, 816.

(5) Thompson, M. D.; Bledson, T. M.; Strout, D. L. *J. Phys. Chem. A* **2002**, *106*, 6880.

(6) Li, Q. S.; Liu, Y. D. *Chem. Phys. Lett.* **2002**, *353*, 204. Li, Q. S.; Qu, H.; Zhu, H.S. *Chin. Sci. Bull*. **1996**, *41*, 1184.

(7) Li, Q. S.; Zhao, J. F. *J. Phys. Chem. A* **2002**, *106*, 5367. Qu, H.; Li, Q. S.; Zhu, H. S. *Chin. Sci. Bull*. **1997**, *42*, 462.

(8) Gagliardi, L.; Evangelisti, S.; Widmark, P. O.; Roos, B. O. *Theor. Chem. Acc.* **1997**, *97*, 136.

(9) Gagliardi, L.; Evangelisti, S.; Bernhardsson, A.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2000**, *77*, 311.

(10) Schmidt, M. W.; Gordon, M. S.; Boatz, J. A. *Int. J. Quantum Chem.* **2000**, *76*, 434.

(11) Christe, K. O.; Wilson, W. W.; Sheehy, J. A.; Boatz, J. A. *Angew. Chem., Int. Ed.* **1999**, *38*, 2004.

(12) Vij, A.; Pavlovich, J. G.; Wilson, W. W.; Vij, V.; Christe, K. O. *Angew. Chem., Int. Ed.* **2002**, *41*, 3051. Butler, R. N.; Stephens, J. C.; Burke, L. A. *Chem. Commun.* **2003**, *8*, 1016.

(13) Eremets, M. I.; Gavriliuk, A. G.; Trojan, I. A.; Dzivenko, D. A.; Boehler, R. *Nat. Mater.* **2004**, *3*, 558.

(14) Fau, S.; Wilson, K. J.; Bartlett, R. J. *J. Phys. Chem. A* **2002**, *106*, 4639.

(15) Dixon, D. A.; Feller, D.; Christe, K. O.; Wilson, W. W.; Vij, A.; Vij, V.; Jenkins, H. D. B.; Olson, R. M.; Gordon, M. S. *J. Am. Chem. Soc*. **2004**, *126*, 834.

(16) Knapp, C.; Passmore, J. *Angew. Chem., Int. Ed.* **2004**, *43*, 4834.

(17) Haiges, R.; Schneider, S.; Schroer, T.; Christe, K. O. *Angew. Chem., Int. Ed.* **2004**, *43*, 4919.

(18) Huynh, M. V.; Hiskey, M. A.; Hartline, E. L.; Montoya, D. P.; Gilardi, R. *Angew. Chem., Int. Ed.* **2004**, *43*, 4924.

(19) Klapotke, T. M.; Schulz, A.; McNamara, J. *J. Chem. Soc., Dalton Trans.* **1996**, 2985. Klapotke, T. M.; Noth, H.; Schutt, T.; Warchhold, M. *Angew. Chem., Int. Ed.* **2000**, *39*, 2108. Klapotke, T. M.; Krumm, R.; Mayer, P.; Schwab, I. *Angew. Chem., Int. Ed.* **2003**, *42*, 5843.

(20) Glukhovtsev, M. N.; Jiao, H.; Schleyer, P. v. R. *Inorg. Chem.* **1996**, *35*, 7124.

(21) Bruney, L. Y.; Bledson, T. M.; Strout, D. L. *Inorg. Chem.* **2003**, *42*, 8117.

(22) Strout, D. L. *J. Phys. Chem. A* **2004**, *108*, 2555.

(23) Zhou, H.; Wong, N.-B.; Zhou, G.; Tian, A. *J. Phys. Chem. A* **2006**, *110*, 7441.

(24) Sturdivant, S. E.; Nelson, F. A.; Strout, D. L. *J. Phys. Chem. A* **2004**, *108*, 7087.

(25) Strout, D. L. *J. Phys. Chem. A* **2004**, *108*, 10911.

(26) Strout, D. L. *J. Phys. Chem. A* **2003**, *107*, 1647.

(27) Sturdivant, S. E.; Strout, D. L. *J. Phys. Chem. A* **2004**, *108*, 4773.

(28) Strout, D. L. *J. Chem. Theory Comput.* **2005**, *1*, 561.

(29) Colvin, K. D.; Cottrell, R.; Strout, D. L. *J. Chem. Theory Comput.* **2006**, *2*, 25.

(30) Strout, D. L. *J. Phys. Chem. A* **2006**, *110*, 7228.

(31) Cottrell, R.; McAdory, D.; Jones, J.; Gilchrist, A.; Shields, D.; Strout, D. L. *J. Phys. Chem. A* **2006**, *110*, 13889.

(32) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(33) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(34) Perdew, J. P.; Ernzerhof, M. *J. Chem. Phys.* **1996**, *105*, 9982. Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029. Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(35) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(36) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(37) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision B.01*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(38) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

(39) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *Phys. Rev. B* **2004**, *69*, 075102.

(40) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274.

(41) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *121*, 1187.

CT700090H

JCTC Journal of Chemical Theory and Computation

Relativistic Effects on the Topology of the Electron Density

Georg Eickerling,[†] Remigius Mastalerz,[†] Verena Herz,[‡] Wolfgang Scherer,*[,‡]
Hans-Jörg Himmel,*[,§] and Markus Reiher*[,†]

Laboratorium für Physikalische Chemie, ETH Zurich, Hönggerberg Campus, Wolfgang-Pauli-Strasse 10, CH-8093 Zurich, Switzerland, Institut für Physik, Universität Augsburg, Universitätsstrasse 1, D-86159 Augsburg, Germany, and Institut für Anorganische Chemie, Ruprechts-Karls-Universität Heidelberg, Im Neuenheimer Feld 270, D-69120 Heidelberg, Germany

Received July 16, 2007

Abstract: The topological analysis of electron densities obtained either from X-ray diffraction experiments or from quantum chemical calculations provides detailed insight into the electronic structure of atoms and molecules. Of particular interest is the study of compounds containing (heavy) transition-metal elements, which is still a challenge for experiment as well as from a quantum-chemical point of view. Accurate calculations need to take relativistic effects into account explicitly. Regarding the valence electron density distribution, these effects are often only included indirectly through relativistic effective core potentials. But as different variants of relativistic Hamiltonians have been developed all-electron calculations of heavy elements in combination with various electronic structure methods are feasible. Yet, there exists no systematic study of the topology of the total electron density distribution calculated in different relativistic approximations. In this work we therefore compare relativistic Hamiltonians with respect to their effect on the electron density in terms of a topological analysis. The Hamiltonians chosen are the four-component Dirac−Coulomb, the quasi-relativistic two-component zeroth-order regular approximation, and the scalar-relativistic Douglas−Kroll−Hess operators.

1. Introduction

To base chemical concepts on elements of a quantum mechanical many-electron theory for molecules has a long history. For example, Hinze and Jaffe[1−3] explicated Mulliken's definition of electronegativity[4] in terms of 'orbital electronegativities' and the 'valence state of an atom in a molecule'. For historical reasons, these very successful early conceptual developments were deeply rooted in some sort of molecular-orbital-based picture. In recent years, however, complementary approaches refer to electron density distributions as a central quantity for interpretive studies.[5−8] The study of the topology of the total electron density $\rho(\boldsymbol{r})$ allows a detailed characterization of electronic densities and, within Bader's theory of atoms in molecules,[6] of interatomic interactions. One major advantage of the density-based approaches is that $\rho(\boldsymbol{r})$ is an observable and, hence, available, for instance, from quantum chemical calculations and X-ray or electron diffraction experiments.[9] Owing to the advances in experimental techniques such as low-temperature devices and fast and highly accurate area detectors, high-resolution X-ray diffraction experiments with a subsequent multipolar refinement[10] based on precise, high-resolution X-ray diffraction data became the most convenient experimental technique to analyze the charge density distribution of molecules and solids.[11−15]

The *static* electron density distribution as a physical observable provides a direct linkage between theory and

* Corresponding author e-mail: markus.reiher@phys.chem.ethz.ch (M.R.), wolfgang.scherer@physik.uni-augsburg.de (W.S.), hans-jorg.himmel@aci.uni-heidelberg.de (H.-J.H.).

[†] ETH Zurich.

[‡] Universität Augsburg.

[§] Ruprechts-Karls-Universität Heidelberg.

10.1021/ct7001573 CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/20/2007

experiment. This has recently been demonstrated in the case of transition-metal complexes displaying highly unusual structures (e.g., non-VSEPR complexes) or activated bonds as in agostic complexes. Combined studies showed a very good agreement between the charge density distributions obtained from sophisticated quantum-chemical calculations and advanced X-ray studies.[16] Recent studies even allowed the experimental verification of so-called ligand-induced charge concentrations (LICCs) in the valence shell charge concentration of a transition-metal atom[17]—a phenomenon predicted by theory already in 1995.[18] However, the experimental determination of reliable charge density distributions in the case of compounds containing heavy elements (with, say, nuclear charge numbers $Z > 36$) is still a challenge for both theory and experiment. From an experimental point of view the treatment of heavy elements in standard X-ray diffraction techniques is complicated because of the presence of severe absorption in addition to problems arising from extinction, thermal diffuse scattering, Umweganregung (i.e., the Renninger effect which may cause symmetry forbidden reflections to appear in the diffraction pattern due to multiple diffraction within the crystal), thermal motion, and partial structural disorder.[19] The compensation of these experimental error sources requires sophisticated data reduction and correction techniques. Furthermore, experimental studies are difficult because of the small number of valence electrons compared to the large total number of electrons which makes it difficult to describe the small fraction of nonspherically distributed electrons in the valence region within the multipolar model. All of this can be summarized in the suitability factor S which is defined as the ratio of the unit cell volume and the sum of the square of the number of core electrons treated spherically symmetric in the multipolar refinements.[20] For crystals of organic molecules, S varies typically from 3 to 5, while for first-row transition-metal complexes it is typically lower than 0.3. Accordingly, only a rather small number of experimental studies were carried out on compounds containing transition-metal elements.

On the other hand, quantum chemical calculations, which are used for comparison in many of the experimental studies, often employ effective core potentials to replace the core electrons of the heavy elements. For a direct comparison with experimental results, however, relativistic all-electron calculations [21−23] are needed as a reference—especially for heavy elements. The need of a detailed and thorough *all-electron* analysis employing relativistic Hamiltonians was recently exemplified for the series of $M(C_2H_4)_3$ complexes with M = Ni, Pd, Pt.[24] For this series of complexes we could demonstrate that the shell structure as well as the polarization pattern displaying zones where the charge density is locally concentrated or depleted are quantitatively and even qualitatively different in effective core potential calculations when compared to relativistic all-electron calculations. A careful all-electron analysis of the charge density distribution of these complexes revealed in agreement with previous findings of Kohout, Savin, and Preuss for isolated atoms[25] as well as with the work of Sagar et al.[43] that the negative Laplacian of the charge density distribution fails to recover the complete shell structure. The two outermost shells, i.e., the sixth and fifth shell of Pt, are not resolved while in the case of the 3d and 4d metals, Ni and Pd, respectively, solely the outermost shell is missing. Furthermore, in contrast to the calculations using scalar-relativistic effective core potentials our all-electron calculations employing the scalar-relativistic zeroth-order regular approximation (ZORA) Hamiltonian did also not recover any local zones of charge concentrations and charge depletion in the valence shell of charge concentrations at the Pt center in $Pt(C_2H_4)_3$.[24]

Various relativistic Hamiltonians are nowadays available to include relativistic effects in first-principles calculations. But no systematic study of the effects of these Hamiltonians on the topology of the resulting total electron density exists in the literature. In this work we therefore present the results of a comparative study of calculations on $M(C_2H_2)$ (M = Ni **1**, Pd **2**, Pt **3**) employing several relativistic and quasi-relativistic Hamiltonians, namely the four-component Dirac−Coulomb, the two-component ZORA, and the scalar-relativistic Douglas−Kroll−Hess operators, which are briefly introduced in section 2. Section 3 introduces the computational methodology. We chose a series of homologous complexes as it is well known that relativistic effects increase with the nuclear charge number Z.[21−23] For comparison, we also included results calculated with the standard *nonrelativistic* many-electron Hamiltonian which allows us to assess the general magnitude of relativistic effects on the electron density and its topology. After briefly comparing the molecular geometries in section 4, we first discuss the effect of the choice of the Hamiltonian on the topology of the electron density. Although it is sufficient to investigate the role of the Hamiltonian for the most simple approximation to the many-electron wave function, namely for a single Slater determinant in the framework of Hartree−Fock theory, we consider the relative magnitude of electron-correlation effects in section 6 by comparison with density functional theory (DFT) calculations. Finally, in section 7 we discuss the effect of the relativistic approximations on the Laplacian of the electron density. Here we start from the radial Laplacian of isolated atoms and compare its properties to metal atoms bound in the complexes under consideration.

2. Theoretical Background

Within the Hartree−Fock approximation, the N-electron wave function is represented by the antisymmetrized product of N single-particle functions, the molecular orbitals ϕ_i, which can be written in form of a Slater determinant (N is the total number of electrons). This single-determinant wave function will serve as a standard for our comparative study. We thus exclude effects of electron correlation. In section 6, however, we will compare the results of the Hartree−Fock-calculations to a simple model which takes electron correlation effects into account—namely to density functional theory. Although electron correlation is only treated approximately within present-day DFT, this is sufficient for our purpose as we are solely interested in assessing the approximate magnitude of electron correlation compared to relativistic effects on the electron density. The various Hamiltonians used within this study will be briefly introduced in the following subsections.

2.1. Four-Component Methods. In order to systematically study the effect of approximate relativistic Hamiltonians a well-defined reference is of particular importance. The well-established and often called "fully relativistic" reference theory for the description of atoms and molecules in quantum chemistry is based on Dirac's theory of the electron (see ref 26 for a review of these so-called four-component methods). Accordingly, the most appropriate reference Hamiltonian is the Dirac−Coulomb Hamiltonian

$$H_{\mathrm{DC}} = \sum_{i=1}^{N} h_{\mathrm{D}}(i) + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{1}$$

where r_{ij} is the distance of two electrons i and j. Note that we use Hartree atomic units throughout and that we have omitted the nucleus−nucleus repulsion terms for the sake of brevity. The four-component one-electron Dirac operator $h_{\mathrm{D}}(i)$ is given in standard notation as

$$h_{\mathrm{D}}(i) = c\boldsymbol{\alpha}\cdot\boldsymbol{p}_i + (\beta - 1)c^2 - \sum_{A=1}^{M}\frac{Z_A}{R_{iA}} \tag{2}$$

Here, c denotes the speed of light, $\boldsymbol{\alpha}$ represents a 3-vector whose components are (4 × 4) matrices built from Pauli spin matrices $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ on the off-diagonal, and $\boldsymbol{p}_i$ is the standard linear momentum operator. The second term on the right-hand side of eq 2 contains a shift in energy by the rest energy c^2 (in Hartree atomic units) in order to match the nonrelativistic energy scale. Finally, β is a diagonal (4 × 4) matrix with (1, 1, −1, −1) entries on the diagonal. The last sum in eq 2 describes the attractive Coulomb interaction between electron i and all nuclei A in the molecule.

Due to the structure of the Dirac operator the one-particle functions in the Slater determinant become four-component molecular spinors

$$\phi_i = \begin{pmatrix}\phi_i^1\\ \phi_i^2\\ \phi_i^3\\ \phi_i^4\end{pmatrix} \equiv \begin{pmatrix}\phi_i^L\\ \phi_i^S\end{pmatrix} \tag{3}$$

for which we introduced the large and small two-component spinors ϕ_i^L and ϕ_i^S. A one-particle eigenvalue equation for h_{D} can now be written in split notation as

$$(V - \epsilon_i)\phi_i^L + c\boldsymbol{\sigma}\cdot\boldsymbol{p}\phi_i^S = 0 \tag{4}$$

$$c\boldsymbol{\sigma}\cdot p\phi_i^L + (V - 2c^2 - \epsilon_i)\phi_i^S = 0 \tag{5}$$

Within the Hartree−Fock approach chosen here, the N-particle wave function approximated by a Slater determinant Φ provides a total charge density $\rho_{4\mathrm{comp}}(\boldsymbol{r})$ that is a sum over all (occupied) four-component molecular spinors

$$\begin{aligned}\rho_{4\mathrm{comp}}(\boldsymbol{r}_1) &= \int ds_1 \int d\tau_2 \cdots \int d\tau_N \Phi^*(\tau_1,\tau_2,...,\tau_N)\Phi(\tau_1,\tau_2,...,\tau_N)\\ &= \sum_{i=1}^{N}\phi_i^{\dagger}(\boldsymbol{r}_1)\phi_i(\boldsymbol{r}_1) = \sum_{i=1}^{N}\sum_{k=1}^{4}\phi_i^{\mathrm{k}*}(\boldsymbol{r}_1)\phi_i^{\mathrm{k}}(\boldsymbol{r}_1)\end{aligned} \tag{6}$$

where τ_i denotes the set of spatial and spin coordinates $\boldsymbol{r}_i$ and s_i, respectively. In our study, we use this density $\rho_{4\mathrm{comp}}(\boldsymbol{r})$ as the reference density.

2.2. Elimination Techniques. Due to the fact that four-component methods are computationally very demanding, elimination and transformation techniques have been devised in order to decouple the positive- and negative-energy parts of the spectrum of the Dirac Hamiltonian (see ref 27 for a most recent review). These methods aim at a reduction of the four-component Dirac equation to an effective two-component form. In this study we consider two quasi-relativistic Hamiltonians to investigate the effect of such approximations on the topology of the electron density.

One efficient and widely used method to achieve the decoupling of the large and the small components is ZORA.[28−30] Within this approximation one solves eq 5 for ϕ_i^S

$$\phi_i^S = X(\epsilon_i)\phi_i^L \tag{7}$$

where the energy-dependent X-operator reads

$$X(\epsilon_i) = \frac{c\boldsymbol{\sigma}\cdot\boldsymbol{p}}{(\epsilon_i - V + 2c^2)} \tag{8}$$

Inserting $X(\epsilon_i)$ into the upper part of the Dirac equation, namely into eq 4, we obtain

$$(V - \epsilon_i)\phi_i^L + \frac{1}{2c^2}(c\boldsymbol{\sigma}\cdot\boldsymbol{p})\left[\frac{2c^2}{\epsilon_i - V + 2c^2}\right](c\boldsymbol{\sigma}\cdot\boldsymbol{p})\phi_i^L = 0 \tag{9}$$

The resulting expression for the Hamiltonian is energy dependent and can be rewritten and expanded in terms of a Taylor series to finally yield the two-component ZORA Hamiltonian

$$h_{\mathrm{ZORA}} = \boldsymbol{\sigma}\cdot\boldsymbol{p}\frac{2c^2}{2c^2 - V}\boldsymbol{\sigma}\cdot\boldsymbol{p} + V \tag{10}$$

The ZORA Hamiltonian h_{ZORA} is widely used instead of h_{D} of eq 2 for calculations including scalar-relativistic effects as well as spin−orbit coupling.

2.3. Transformation Techniques. The generalized Douglas−Kroll−Hess (DKH) unitary transformation technique[31−33] (see ref 34 for a recent review of conceptual aspects of this theory) aims at a block-diagonalization of the Dirac Hamiltonian resulting in two independent (2 × 2) matrix operators h_+ and h_- which describe the electronic and the so-called positronic eigenstates, respectively

$$h_{\mathrm{bd}} = Uh_{\mathrm{D}}U^{\dagger} = \begin{pmatrix}h_+ & 0\\ 0 & h_-\end{pmatrix} \tag{11}$$

This block-diagonalization can be accomplished by a sequence of unitary transformations

$$h_{\mathrm{bd}} = ...U_3U_2U_1U_0h_{\mathrm{D}}U_0^{\dagger}\,U_1^{\dagger}\,U_2^{\dagger}\,U_3^{\dagger}... \tag{12}$$

Every U_n in this sequence of unitary transformations is parametrized in terms of a power series expansion of an antihermitian operator W_n,[33] which is chosen to diminish the off-diagonal contributions order by order in the external potential. According to eq 11 a complete decoupling of the Dirac operator leads to a two-component formulation based on h_+. However, the operator h_+ can be further separated into a one-component spin-free and a spin-dependent part. The DKH approach is most efficient in its scalar-relativistic, spin-free variant, which we chose for this work. We should emphasize that for all non-*p*-block elements of the periodic table of the elements with not too large nuclear charges, $Z \lesssim 100$, spin−orbit coupling does not play a decisive role. This is the reason why we challenge the four-component reference results with scalar-relativistic high-order DKH calculations.

In principle, exact decoupling of the Dirac Hamiltonian would require an infinite number of unitary transformations. However, given a certain accuracy determined by the computational set up (mainly by the quality of the basis set) the expansion may be truncated at a certain order *m*, defining the DKH*m* method.[35] We consider the tenth-order DKH10 Hamiltonian as sufficient for exact decoupling in the scalar-relativistic regime. The order of the DKH operator is related to the *n*th unitary matrix by the so-called $(2n + 1)$-rule. According to this rule, for instance, the tenth-order DKH Hamiltonian is completely defined by the first six unitary matrices U_0, U_1, U_2, U_3, U_4, and U_5. Transformation techniques closely related to the DKH theory which aim at the exact decoupling of the Dirac Hamiltonian such as the infinite-order two-component (IOTC) theory have recently been developed and implemented.[36,37]

2.4. Topological Analysis of the Electron Density. The differences of the densities obtained from calculations employing the Hamiltonians introduced in sections 2.1−2.3 are qualitatively discussed in terms of difference density plots, which directly reveal the change in the spatial distribution of the density. A quantitative measure of the differences is given by the values of $\rho(\boldsymbol{r})$ at a set of characteristic points within the molecule. These points were chosen as the stationary points of the three-dimensional electron density distribution, which are given a special meaning within the theory of atoms in molecules.[6] In particular we analyzed the bond critical points (BCPs) and the ring critical points (RCPs). Within the atoms-in-molecules theory these critical points are classified according to the signs of the eigenvalues of the Hessian matrix which contains the nine second derivatives of $\rho(\boldsymbol{r})$ with respect to the spatial coordinates $\boldsymbol{r} = (x, y, z)$. For example, the atomic positions at which $\rho(\boldsymbol{r})$ adopts maximum values are classified as $(3, -3)$ critical points. Here, (ω, σ) denotes the rank ω, i.e. the number of nonzero eigenvalues and the signature σ which is the sum of the signs of the eigenvalues of the Hessian matrix. Thus, a BCP is classified as a $(3,-1)$ critical point and a RCP as a $(3,+1)$ critical point. Besides this qualitative classification of the critical points the curvature of $\rho(\boldsymbol{r})$ at a bond critical point can be quantitatively characterized utilizing the three eigenvalues of the diagonalized Hessian matrix of $\rho(\boldsymbol{r})$ λ_1, λ_2, and λ_3. The eigenvector belonging to the positive eigenvalue λ_3 points along the bond axis and one can define the bond ellipticity ϵ as[38]

$$\epsilon = \frac{\lambda_1}{\lambda_2} - 1 \tag{13}$$

Here, λ_1 and λ_2 denote the two eigenvalues belonging to the two eigenvectors which span a plane perpendicular to the direction of the bond path and for which $|\lambda_1| \geq |\lambda_2|$ holds. According to this definition, ϵ is always positive. For a rotationally symmetric σ-bond λ_1 and λ_2 are equal and thus $\epsilon = 0$.

Another property sensitive to changes in the topology of the electron density is the topology of the bond path. The bond path is defined for molecules and solids at equilibrium geometry as a set of two gradient lines which originate at the BCP and terminate each at one of the nuclei of a bonded atom pair. Thereby the bond path follows the maximum slope of $\rho(\boldsymbol{r})$ and thus needs not be a straight line but can exhibit a complicated curvy-linear behavior which can be used to characterize the type of a chemical interaction.[39] At this point we should stress that the foundations of the atoms-in-molecules theory have not been rigorously defined in a relativistic four-component theory compared to Schrödinger quantum mechanics.[40] This is, however, no obstacle for our study, because we are interested in the shape of the total electron density, which we simply study in terms of the atoms-in-molecules notation.

In addition to the topology of $\rho(\boldsymbol{r})$, we also analyze the negative Laplacian of the electron density.

$$L(\boldsymbol{r}) = -\nabla^2\rho(\boldsymbol{r}) \tag{14}$$

The Laplacian is the trace of the (3×3) Hessian matrix of $\rho(\boldsymbol{r})$, which is invariant under basis transformations. This definition is convenient, because a *positive* value of $L(\boldsymbol{r})$ corresponds to a region where charge is locally *concentrated* whereas a *negative* sign of $L(\boldsymbol{r})$ corresponds to a region suffering of local charge *depletion*.[6] Analyzing the topology of $L(\boldsymbol{r})$ in the same way as described above for $\rho(\boldsymbol{r})$ provides another set of characteristic (i.e., stationary) points. The local maxima in the negative Laplacian distribution indicate the positions of locally enhanced charge concentration (CC), found within the valence shell of charge concentration ($L(\boldsymbol{r}) > 0$) in the valence region of atoms in molecules. Bader et al. suggested that the outermost shell of CC of an atom (second shell of CC of the carbon atoms and third shell of CC of the nickel atom) represents its (effective) valence shell charge concentration (VSCC).[25,41−46,82]

3. Computational Methodology

The molecular geometries of $M(C_2H_2)$ (M = Ni, Pd, Pt) were taken from a structure optimization with the Turbomole program package[47] employing the BP86 density functional,[48,49] effective core potentials from the Stuttgart group (ecp-10-mdf,[50] ecp-28-mwb,[51] and ecp-60-mwb[51] for Ni, Pd, and Pt, respectively), and basis sets of Gaussian-type functions (GTFs) of triple-ζ plus polarization quality (TZVP) as implemented in Turbomole. The molecular structures obtained from these calculations will be used as default if not explicitly mentioned otherwise, and we will refer to them as **A**. For comparison, we also performed structure optimizations employing all electron TZ2P basis sets of Slater-type functions (STFs), the BP86 density functional,[48,49] and the scalar-relativistic ZORA approximation using the ADF program package.[52,53] In the following this level of approximation will be denoted **B**.

All four-component Hartree–Fock and DFT as well as the two-component ZORA calculations were performed with the Dirac electronic structure program[54–56] employing completely decontracted basis sets. The density functionals LDA, BLYP, BP86, and B3LYP were chosen as implemented in Dirac.[48,49,57,58] In the cases of Pt and Pd we applied the relativistic quadruple-ζ basis sets devised by Dyall[59,60] in a completely decontracted way, i.e., with all exponents taken as primitive basis functions. This results in the following basis set sizes Pt: (34*s*, 30*p*, 19*d*, 12*f*, 7*g*, 4*h*, 1*i*), Pd: (33*s*, 25*p*, 17*d*, 9*f*, 6*g*, 3*h*). For Ni we supplemented the exponents of the basis set by Pou-Amerigo et al.,[61] which constitutes an expansion of the original basis set by Partridge,[62] by two additional diffuse *h*-type functions with exponents 0.1 and 0.01, yielding a final size of (21*s*, 15*p*, 10*d*, 6*f*, 4*g*, 2*h*). The exponents of Dunning's[63] cc-pVQZ basis set provided the exponents for the lighter elements C and H and result in the following basis set sizes C: (12*s*, 6*p*, 3*d*, 2*f*, 1*g*), H: (6*s*, 3*p*, 2*d*, 1*f*). In view of the overall size of the basis sets used in this study they may well be considered to be close to the basis set limit.

The scalar-relativistic Douglas–Kroll–Hess and the nonrelativistic Hartree–Fock calculations were performed with the Molpro2006.2 program package[64] using the same large primitive basis sets as employed in the four-component calculations. The DKH10 calculations were possible owing to our recent implementation of the arbitrary-order DKH Hamiltonian[65] into the Molpro package.

In each case, the total electron density was then calculated on a cubic grid of points (200 × 200 × 200 points, step size 3 pm). We ensured that the cubic grids were identical in Dirac and Molpro. The search for bond and ring critical points and the determination of the topological parameters was performed using the Integrity program written by P. Rabillier.[66] For unit conversion to eÅ^{-3} and eÅ^{-5}, the results in Hartree atomic units were multiplied with a conversion factor of 6.748315 Å^{-3} and 24.098731 Å^{-5}, respectively. We should note that the explicit multiplication with the elementary charge *e* in the unit would convert the electron density, which is a particle density distribution, into a (positive) electron density distribution of *N* elementary charges. However, this multiplication is hardly made explicit, instead one refers to "eÅ^{-3}" as a fraction of electrons per cubic Ångstrøm. The (negative) *charge* density can be obtained by multiplication of the electron density by −1.

The Laplacian of the total charge density was calculated numerically using a Mathematica[67] routine written by M. Presnitz.[68] The Laplacian, which was then also obtained on a grid of points, was again analyzed using the Integrity program to locate the stationary points.

To assess the error due to the numerical determination of the topological parameters on a grid of points we compared the results of grids with 3 pm and 0.015 pm grid spacing. The values of the electron density at the critical points are the same for both grids. However, the values of the Laplacian at the critical points show some deviations. For example for the C–C bond critical point in complex **1** we find for $L(\boldsymbol{r})$ −26.7 and −26.6 eÅ^{-5} for the grid with 0.03 and 0.015 pm spacing, respectively, which corresponds to a difference of only 0.4%. Only for smaller values of $L(\boldsymbol{r})$ as they are for instance found at the M–C bond critical points, the deviations are somewhat larger. Here we find for compound **1** values of 5.1 and 5.4 eÅ^{-5} using a 0.03 and 0.015 pm grid, respectively. However, these deviations do not affect the result of our study, as we will point out below.

Table 1. BP86 Bond Distances and Angles of M(C_2H_2) (M = Ni, Pd, Pt) in pm and deg as Obtained from the GTF-TZVP Nonrelativistic Calculations (**A**) and the STF-TZ2P ZORA Calculations (**B**)

A	M–C	C–C	C–H	CMC	MCC	MCH
Ni	184.1	128.2	108.5	40.8	69.6	139.6
Pd	204.1	126.5	108.0	36.1	71.9	133.3
Pt	200.5	128.5	108.2	37.4	71.3	135.2
B	M–C	C–C	C–H	CMC	MCC	MCH
Ni	181.1	128.8	108.5	41.7	69.2	139.9
Pd	203.9	126.5	108.0	36.1	71.9	133.5
Pt	198.3	128.7	108.1	37.9	71.1	136.3

Atomic Hartree–Fock calculations on the metal atoms were carried out using a fully numerical four-component (MC)SCF program,[69] in which all angular degrees of freedom are treated analytically, while the two radial functions $F_i(r) = P_i(r)/r$ and $G_i(r) = Q_i(r)/r$ of the 4-spinor are represented on an equidistant (logarithmic) grid of points in the new variable *s*, which is calculated from the radial variable *r* (see ref 70 for details on this type of radial grid).

4. Structural Comparison

The final bond distances and angles obtained from the structure optimizations of **1**–**3** are summarized in Table 1. Comparing the molecular geometries obtained from the GTF-TZVP/nonrel and the STF-TZ2P/ZORA calculations (methods **A** and **B**, respectively), the only significant deviations are found for the M–C bond distances of **1** and **3**, where the differences amount to 3 and 2.2 pm, respectively.

We may compare these generic systems to experimentally known acetylene and ethylene complexes. First of all, the structure of Ni(C_2H_2) agrees well with the one obtained by X-ray diffraction for Ni(C_2H_2)(PPh_3)$_2$ **4**.[71] Due to the C_{2v} symmetry, the two Ni–C distances are equal in **1** (184.1 pm), while for **4** two slightly different bond lengths were found (187.1 and 188.1 pm). The same holds true for the two CCH angles, which are again the same in **1** (150.8°) but differ in **4** (146.8° and 149.4°). The bond distances and angles are also in agreement with those reported earlier on the basis of B3LYP/6-311+G(2d,p) calculations for Ni(C_2H_2) (C–C and C–H distances of 127.6 and 107.8 pm and HCC angles of 148.5°).[72]

The qualitative comparison of the structures for the three generic systems **1**–**3** optimized in both schemes **A** and **B** shows that the C–C bond is less elongated—compared to free acetylene (120 pm)—for **2** (126.5 pm) than for **1** (128.2

Table 2. Values of $\rho(\boldsymbol{r})$ in eÅ^{-3} at the M−C BCPs and at the RCP in Complexes **1**−**3** Using Various Many-Electron Hamiltonians within the Hartree−Fock Approximation for the Total Wave Function[a]

	$\rho(\boldsymbol{r}_{\mathrm{BCP}})$	dev	$\rho(\boldsymbol{r}_{\mathrm{RCP}})$	dev
		M = Ni		
nonrel	0.97	0.0	0.85	0.0
DKH2	0.97	0.0	0.85	0.0
DKH10	0.97	0.0	0.85	0.0
ZORA	0.97	0.0	0.85	0.0
four-comp	0.97	−	0.85	−
		M = Pd		
nonrel	0.79	1.3	0.76	1.3
DKH2	0.80	0.0	0.78	4.0
DKH10	0.80	0.0	0.78	4.0
ZORA	0.80	0.0	0.78	4.0
four-comp	0.80	−	0.75	−
		M = Pt		
nonrel	0.94	6.0	0.89	4.3
DKH2	1.00	0.0	0.93	0.0
DKH10	1.00	0.0	0.93	0.0
ZORA	1.00	0.0	0.93	0.0
four-comp	1.00	−	0.93	−

[a] In addition the deviations of the values relative to the four-component calculations (dev) are given in %.

Table 3. Values of $L(\boldsymbol{r})$ in eÅ^{-5} and ϵ at the M−C BCPs and $L(\boldsymbol{r})$ at the RCP in Complexes **1**−**3** Using Various Many-Electron Hamiltonians within the Hartree−Fock Approximation for the Total Wave Function[a]

	$L(\boldsymbol{r}_{\mathrm{BCP}})$	dev	$\epsilon(\boldsymbol{r}_{\mathrm{BCP}})$	dev	$L(\boldsymbol{r}_{\mathrm{RCP}})$	dev
			M = Ni			
nonrel	−5.61	7.9	0.28	17.7	−14.28	4.8
DKH2	−4.83	7.1	0.36	5.9	−13.58	0.3
DKH10	−5.09	2.1	0.32	5.9	−13.67	0.4
ZORA	−5.33	2.5	0.29	14.7	−14.11	3.6
four-comp	−5.20	−	0.34	−	−13.62	−
			M = Pd			
nonrel	−6.94	8.4	1.34	22.9	−10.69	2.9
DKH2	−5.91	7.7	1.17	7.3	−9.79	5.8
DKH10	−6.09	4.8	1.20	10.1	−10.42	0.3
ZORA	−6.20	3.1	1.15	5.5	−10.40	0.1
four-comp	−6.40	−	1.09	−	−10.39	−
			M = Pt			
nonrel	−5.88	90.3	0.86	43.3	−11.53	19.5
DKH2	−3.19	3.1	0.61	1.7	−9.54	1.1
DKH10	−3.42	10.7	0.58	3.3	−9.09	5.8
ZORA	−3.29	6.5	0.57	5.0	−9.31	3.5
four-comp	−3.09	−	0.60	−	−9.65	−

[a] In addition the deviations of the values relative to the four-component calculations (dev) are given in %.

pm) and **3** (128.5 pm); only values for scheme A are specified. This is in agreement with experimental results for (d^{i}ppe)M(C_2H_2) (M = Ni **5**, Pt **6**) and (d^{i}ppe)-Pd(C_2PhH) **7** (d^{i}ppe = ^{i}Pr$_2$PCH$_2$CH$_2$P^{i}Pr$_2$) where the shortest C−C bond is also found for **7** (124.6(7) pm, compared to 128.7(7) and 137(3) pm for **5** and **6**, respectively).[73] In previous BP86 calculations on complexes (dpe)M(C_2H_2) (dpe = diphosphinoethane) employing triple-ζ Slater-type basis sets, C−C distances of 127.8, 126.9, and 129.0 pm and MCH angles of 141.0°, 137.4°, and 142.9° for M = Ni, Pd, and Pt, respectively, were found.[74] We note that the C−C bond length as obtained from a standard X-ray diffraction study employing the model of independent atoms for the structure factor refinement turns out to be 6 pm too short compared to results obtained from a multipolar refinement which includes aspherical density contribution due to bond formation.[19] With 49.3 and 46.7 kcal mol^{-1} the ligand dissociation energies for acetylene are similar for M = Ni and Pt. For M = Pd, a smaller value was reported (32.3 kcal mol^{-1}).[74]

5. Choice of the Hamiltonian and Topology of the Electron Density

To analyze the importance of the different Hamiltonians for the resulting electron density $\rho(\boldsymbol{r})$ we employ the Hartree−Fock approximation for the total electronic wave function from which $\rho(\boldsymbol{r})$ is calculated. The results concerning the critical points of the total electron density are summarized in Tables 2 and 3.

Comparing the two extreme cases of densities obtained from calculations with the nonrelativistic and the four-component Dirac−Coulomb Hamiltonian for complexes **1**−**3** one finds that for M = Ni there is no difference in total density at the bond critical point, while for M = Pt it amounts to $\Delta\rho(\boldsymbol{r}_{\mathrm{BCP}}) = 0.06$ eÅ^{-3}. Also at the ring critical point the density difference is zero for **1**, while for **3** it is $\Delta\rho(\boldsymbol{r}_{\mathrm{RCP}}) = 0.04$ eÅ^{-3}. In the case of M = Pd a small deviation of 0.01 eÅ^{-3} is observed for $\Delta\rho(\boldsymbol{r}_{\mathrm{BCP}})$ as well as for $\Delta\rho(\boldsymbol{r}_{\mathrm{RCP}})$. These results confirm the expected trend of increasing importance of relativistic corrections when moving from the lighter to the heavier elements in the group.

In order to assess the significance of the deviations found for $\Delta\rho(\boldsymbol{r}_{\mathrm{BCP}})$ and $\Delta\rho(\boldsymbol{r}_{\mathrm{RCP}})$ in the calculations one can use as reference the estimated standard deviations obtained for the topological parameters from *experimental* charge density studies. As the result of a detailed study concerning the reproducibility of the electron density obtained from high-resolution X-ray diffraction experiments by the International Union of Crystallography in 1984 a mean error in the electron density maps of 0.15 eÅ^{-3} was reported.[75] In recent experimental studies the *estimated standard deviations* for the electron density at the critical points range from 0.01 to 0.05 eÅ^{-3}. These standard deviations are, however, only estimates and are not determined directly from the experimental errors. In the current literature one finds only a few studies concerned with the reproducibility of topological parameters obtained from X-ray experiments.[14] One of these is the comparison of different measurements on glycyl-L-threonine by Lecomte et al.[76] and by Luger et al.[77] which showed that the topological parameters at the critical points agree to 99%, 95%, and 88% for $\rho(\boldsymbol{r}_{\mathrm{BCP}})$, $\rho(\boldsymbol{r}_{\mathrm{RCP}})$, and $L(\boldsymbol{r}_{\mathrm{BCP}})$, respectively.[14] Applying this to the topological parameters of a peptide bond leads to deviations of ~0.02 eÅ^{-3} and ~3.0 eÅ^{-5} for the density and the Laplacian at the bond critical points, respectively.[14] Another important matter in this context is the general agreement between experimental and theoretical topological parameters. The various factors which have to be taken into account for such a comparison were, for instance, recently summarized by Coppens and Volkov.[15] Studies on transition-metal compounds show that the general agreement between experimental and some theoretical values of $\rho(\boldsymbol{r}_{\mathrm{BCP}})$ and $L(\boldsymbol{r}_{\mathrm{BCP}})$ typically lie in the range of 0.01−0.03 eÅ^{-3} and 0.77−1.13 eÅ^{-5}.[78,79] These values suggest that a deviation of 0.06 eÅ^{-3} found for the $\rho(\boldsymbol{r}_{\mathrm{BCP}})$ when including relativistic corrections in complex **3** is indeed significant when comparing results from experiment and from quantum chemical calculations.

Finally the question arises whether the changes of the topological parameters are due to a shift of the positions of the critical points or if they are truly effects of a change in the topology of $\rho(\boldsymbol{r})$. One way to answer this question is to compare the distances of the critical points from the atomic positions for the different Hamiltonians employed. The distances between the M−C bond critical point and M lie in a range between 94.1 and 94.4 pm for M = Ni, 111.9 and 112.0 pm for M = Pd, and 113.0 and 114.0 pm for M = Pt. The corresponding distances between the metal atom and the ring critical point are in a range of 96.1 and 96.2 pm for M = Ni, 111.1 and 111.3 pm for M = Pd, and 113.0 and 113.5 pm for M = Pt. Thus, the changes in the positions of the critical points are small, and the changes in the topological parameters discussed above can clearly be attributed to a change in the topology of $\rho(\boldsymbol{r})$ rather than to a shift in the positions of the critical points.

Comparing densities from all Hamiltonians employed, one finds no effect of the different levels of approximation used for the calculation of the wave function on the values of $\rho(\boldsymbol{r}_{\mathrm{BCP}})$ for complex **1** (Table 2). For complex **2** already the DKH2 Hamiltonian gives the same value for $\rho(\boldsymbol{r}_{\mathrm{BCP}})$ as is found for the four-component Hamiltonian. Yet, this is not the case for $\rho(\boldsymbol{r}_{\mathrm{RCP}})$. While a small deviation of only 1.3% occurs when comparing the nonrelativistic to the four-component result, it increases to 4% for the three other approximate Hamiltonians we included in our study. The scalar-relativistic DKH Hamiltonian as well as the ZORA Hamiltonian including spin−orbit coupling terms overestimate $\rho(\boldsymbol{r}_{\mathrm{RCP}})$ by 0.03 eÅ^{-3}. This discrepancy is almost as large as the deviation found between the nonrelativistic and the four-component calculation of complex **3**, where $\Delta\rho(\boldsymbol{r}_{\mathrm{RCP}}) = 0.04$ eÅ^{-3}, corresponding to a relative error of 4.3%. In contrast to complex **2**, for M = Pt already the DKH2 Hamiltonian exactly reproduces the value for $\rho(\boldsymbol{r}_{\mathrm{RCP}})$ found in the four-component calculation.

Considering the Laplacian of the electron density at the bond and ring critical points as a very sensitive quantity to detect changes in a density distribution, one finds significantly larger relative deviations (see Table 3). Thus, nonrelativistic calculations for complex **1** result in a difference $\Delta L(\boldsymbol{r}_{\mathrm{BCP}})$ of 7.9% compared to results from the four-component Hamiltonian. This error is reduced to 7.1% for the DKH2 level of approximation. Only when employing the DKH10 or ZORA Hamiltonian the error is substantially reduced to 2.1% and 2.5%, respectively. However, for $L(\boldsymbol{r}_{\mathrm{RCP}})$ the largest deviation is still found for the nonrelativistic Hamiltonian, but for scalar-relativistic DKH2 calculations it is already as low as 0.3%. DKH10 calculations do not reduce the error further, and including spin−orbit effects by applying the ZORA Hamiltonian leads to an increase to 3.6%. Thus, the trend suggested by the values found for $L(\boldsymbol{r}_{\mathrm{BCP}})$ cannot be confirmed for $L(\boldsymbol{r}_{\mathrm{RCP}})$. A similar situation emerges for the Laplacian at the critical points in the case of complex **2**. Comparatively large deviations of $\Delta L(\boldsymbol{r}_{\mathrm{BCP}})$ are found for densities from calculations with the nonrelativistic and the DKH2 Hamiltonian (8.4 and 7.7%, respectively), which are reduced to 4.8 and 3.1% by applying the DKH10 and the ZORA approximations. In this case, $\Delta L(\boldsymbol{r}_{\mathrm{RCP}})$ appears to follow the same trend, resulting in remarkably low deviations for the DKH10 and the ZORA Laplacians of only 0.3 and 0.1%.

As expected, the largest error is found for the nonrelativistic calculation for complex **3**, which overestimates the absolute values of $L(\boldsymbol{r}_{\mathrm{BCP}})$ by over 90% and $L(\boldsymbol{r}_{\mathrm{RCP}})$ by almost 20% relative to the four-component result. These deviations are significantly reduced by the use of any of the relativistic Hamiltonians. The fact that it is the DKH2 Hamiltonian which almost reproduces the values derived from the four-component calculation should not lead to the interpretation that this Hamiltonian is best suited for the description of complex **3** but rather as error compensation between the approximate treatment of scalar- and spin−orbit effects. This can be explained after considering that the DKH10 approach, which includes a more accurate treatment of scalar-relativistic effects, leads to larger deviations compared to the four-component results.

Summarizing these results we conclude that relativistic effects on the topological parameters at the critical points can be observed in complexes **2** and **3**. The deviations of the density at the critical points amount to 1.3% for M = Pd and 6% in the case of M = Pt. Due to the larger variance of the values of the Laplacian of the electron density the relativistic effects are more difficult to quantify based on the data presented in this work. Still it is clear that for complex **3** the values of $L(\boldsymbol{r}_{\mathrm{BCP}})$ and $L(\boldsymbol{r}_{\mathrm{RCP}})$ are substantially biased if relativistic effects are neglected and that—as it was the case for the density itself—any of the three relativistic Hamiltonians significantly improve the results with respect to the four-component reference calculation.

We note that the magnitude of some of the deviations discussed above are within the error introduced by the use of the finite grid of points for the analysis of the electron density. However, we compare results obtained from the same set of points so that the relative error introduced by the use of a numerical analysis should be small and not relevant for our discussion. As pointed out in section 3 the values of $\rho(\boldsymbol{r})$ do not change with the step size of the grid, and only for small absolute values of $L(\boldsymbol{r})$ a notable deviation is found. Still, this does not affect the main result of our study as especially for complex **3** the errors introduced by neglecting the relativistic effects are much larger than the error due to the use of the finite grid.

The topological parameters discussed so far only provide a very local measure of the relativistic effects on the electron density. Moreover, the largest effect due to relativity is expected within the inner shells of the heavy atoms. The critical points, however, are located in the valence region of the atoms so that one might expect the deviations due to

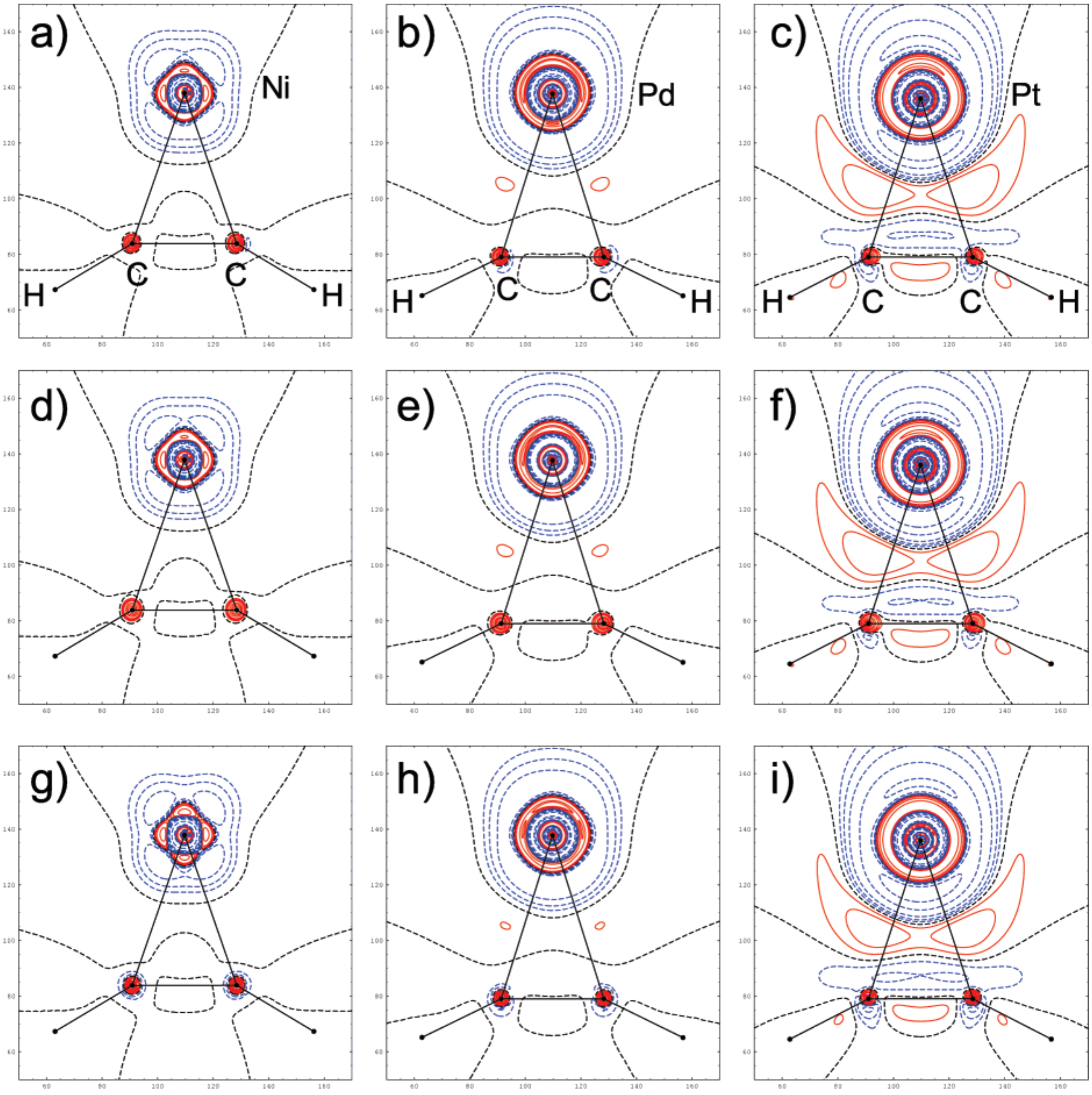


Figure 1. Difference densities in the molecular plane, $\rho_{4comp}(\boldsymbol{r}) - \rho_{nonrel}(\boldsymbol{r})$ for **1** (a), **2** (b) and **3** (c), $\rho_{ZORA}(\boldsymbol{r}) - \rho_{nonrel}(\boldsymbol{r})$ for **1** (d), **2** (e), and **3** (f), $\rho_{DKH10}(\boldsymbol{r}) - \rho_{nonrel}(\boldsymbol{r})$ for **1** (g), **2** (h), and **3** (i). Values of positive and negative difference densities are indicated by solid and dashed lines, respectively. Contour lines are drawn at ±2, ±4, ±8 × 10^n eÅ^{-3} with $n = 0, 1, 2$. Note that the axes labels denote grid points.

relativistic effects to be small at these points. To globally assess the differences due to the various levels of approximation used within our study we will now compare difference densities in the molecular plane obtained by subtracting the nonrelativistic densities from the densities obtained from the various relativistic Hamiltonians. Figure 1a),d),g) depicts the differences in the electron density obtained from four-component, ZORA, and DKH10 calculations, respectively, with respect to the nonrelativistic density for complex **1**. Even in the case of the light first transition row element nickel a significant difference is observed. For all three relativistic Hamiltonians four local maxima are found in the difference maps at the nickel atom of which especially the one facing the acetylene ligand is less pronounced in the case of the four-component and ZORA compared to the DKH10 difference map [in the former cases it is smaller than 0.2 eÅ^{-3}, as can be seen by the missing contour line in Figure 1a),d)]. As we will discuss below, the positions of these maxima resemble the positions of the local charge concentrations found in the valence shell of the metal atom. Thus, with relativistic Hamiltonians these regions of local charge concentration should be more pronounced, relative to the nonrelativistic case. In general, the maxima are more pronounced when using the scalar-relativistic DKH10 Hamiltonian compared to the Hamiltonians which include spin−orbit effects. In the case of M = Pd [Figure 1b),e),h)], the difference density maps show a similar scenario compared to **1**. Although weaker, also for complex **2** four maxima can be found in the outer most circular region of positive difference density around the metal atomic nucleus. While the circular maxima and minima can be clearly attributed to the changes in the radial extension of the atomic subshells

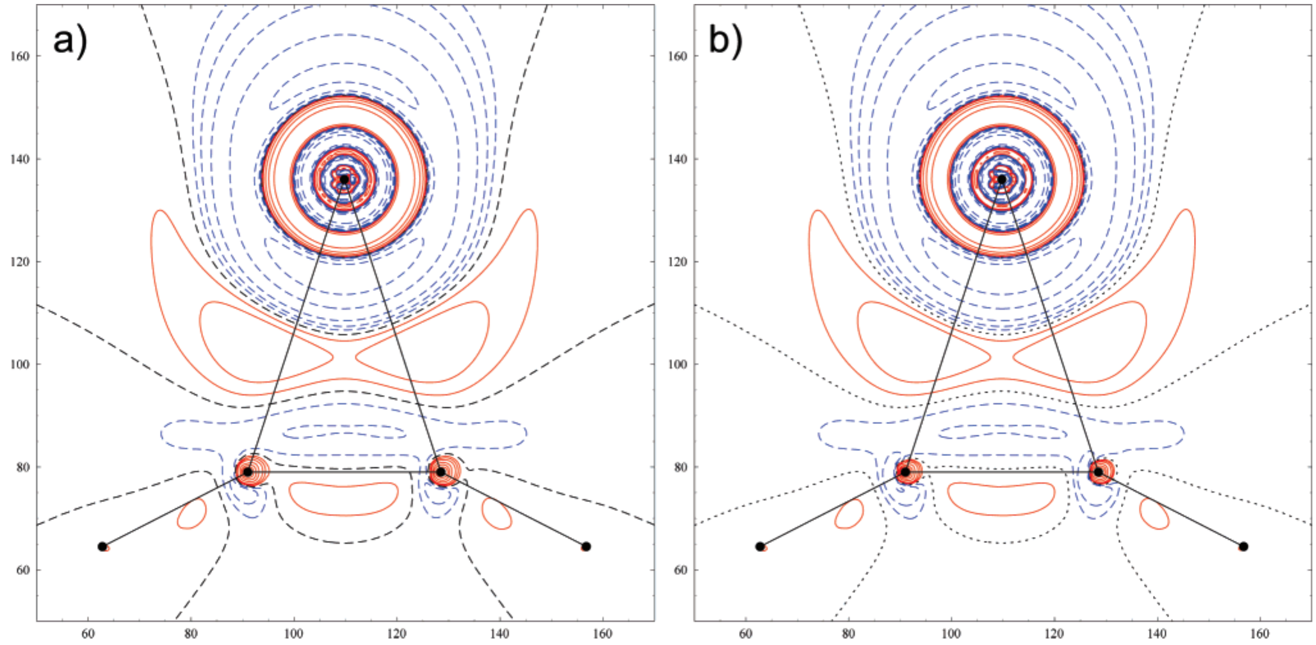


Figure 2. Difference densities in the molecular plane, $\rho_{4\mathrm{comp}}(\boldsymbol{r}) - \rho_{\mathrm{nonrel}}(\boldsymbol{r})$ for **3** (a), and $\rho^{L}_{4\mathrm{comp}}(\boldsymbol{r}) - \rho_{\mathrm{nonrel}}(\boldsymbol{r})$ (b). Values of postitive and negative difference densities are indicated by solid and dashed lines, respectively. For the specification of the contour levels see Figure 1. Note that the axes labels denote grid points.

due to scalar relativistic effects, the four local maxima indicate again a change in the electron density in the regions of the local charge concentrations although this is less pronounced in **2**, see below. An additional feature in the difference density maps for complex **2** emerges in the bonding region between the metal atom and the ligand. Two weak maxima appear at positions close to the M−C bond critical points, which are indicative for the small but noticeable relativistic effect on the electron density in the bonding region. This was already noticed above (Table 2) as a difference of 0.01 eÅ^{-3} at the M−C bond critical points.

As expected, the difference density maps for complex **3** [Figure 1c),f),i)] reveal the largest differences for the three metals under consideration in our study. The changes in the radial extension of the subshells again lead to circular maxima and minima around the Pt nucleus position, but here only the ZORA and the four-component difference densities feature one local maximum in the *trans* position to the ligand in the outer most of these positive regions of difference density. The value of 1.4 eÅ^{-3} is significantly larger than the values found for the corresponding local maxima in complexes **1** (0.2 eÅ^{-3}) and **2** (0.4 eÅ^{-3}). The maxima in the metal−ligand bonding region are also more pronounced compared to complex **2** and extend to a much larger region between the Pt atom and the ligand.

Taking into account the result that already the ZORA and the DKH methods are able to reproduce the density obtained from the four-component calculations we finally analyzed the topological changes of the electron density with respect to the contribution of the small components ϕ_i^S of the four-component wave function. As an example Figure 2 depicts the difference densities between the nonrelativistic and the total electron density as obtained from the four-component calculation a) and the corresponding difference considering only the electron density of the large component $\rho^L(\boldsymbol{r})$ b). Due to the local nature of the small component, the differences between both maps are small and only detectable in the close vicinity of the atomic nuclei. Thus the values of $L(\boldsymbol{r})$ at the M−C bond critical point and at the ring critical point differ only by 3.9 and 4.1%, while the values of $\rho(\boldsymbol{r})$ remain unchanged. This is the reason why the two-component methods which do not completely eliminate the small component are still able to account for most of the relativistic effects on the topology of the electron density.

In summary, comparing the overall topology of the difference density maps presented in Figure 1 we note several points: (i) The differences between the nonrelativistic and the relativistic densities increase with the nuclear charge Z of the metal atom from the 10th group of the periodic table. (ii) The differences in the absolute values of $\rho(\boldsymbol{r})$ reach values of more than 1 eÅ^{-3} in the valence region of the platinum atom. (iii) The overall changes in the electron density due to relativistic effects in compounds **1**−**3** are already well accounted for by the scalar-relativistic DKH10 Hamiltonian. (iv) Inclusion of spin−orbit effects by the two-component ZORA approximation does not improve on the DKH10 results when compared to the four-component reference.

6. Effect of Electron Correlation on the Topology of the Electron Density

In order to assess the importance of relativistic effects on the total electron density in relation to electron correlation effects, we compare results from the Dirac−Coulomb Hamiltonian in Hartree−Fock calculations (abbreviated Dirac−Hartree−Fock in the following) with those obtained by four-component Kohn−Sham DFT calculations using various density functionals. The results are summarized in Table 4.

For all three different BCPs present in complex **1** the Dirac−Hartree−Fock approximation overestimates the value

Table 4. Values of $\rho(\boldsymbol{r})$ in eÅ^{-3} at the Bond and Ring Critical Points of **1**−**3** Obtained from Four-Component Calculations Using Various Different Density Functionals[a]

	Ni		Pd		Pt	
	$\rho(\boldsymbol{r})$	dev	$\rho(\boldsymbol{r})$	dev	$\rho(\boldsymbol{r})$	dev
			M−C			
DHF	0.97	−	0.80	−	1.00	−
LDA	0.95	2.1	0.80	0.0	0.98	2.0
BLYP	0.94	3.1	0.79	1.3	0.97	3.0
BP86	0.94	3.1	0.79	1.3	0.97	3.0
B3LYP	0.94	3.1	0.79	1.3	0.98	2.0
			C−C			
DHF	2.56	−	2.64	−	2.56	−
LDA	2.50	2.3	2.59	1.9	2.50	2.3
BLYP	2.53	1.2	2.62	0.8	2.53	1.2
BP86	2.52	1.6	2.61	1.1	2.52	1.6
B3LYP	2.53	1.2	2.62	0.8	2.54	0.8
			C−H			
DHF	1.96	−	1.99	−	1.99	−
LDA	1.87	4.6	1.89	5.0	1.90	4.5
BLYP	1.91	2.6	1.94	2.5	1.94	2.5
BP86	1.92	2.0	1.94	2.5	1.95	2.0
B3LYP	1.92	2.0	1.95	2.0	1.95	2.0
			RCP			
DHF	0.85	−	0.78	−	0.93	−
LDA	0.91	7.1	0.78	0.0	0.93	0.0
BLYP	0.89	4.7	0.76	2.6	0.91	2.2
BP86	0.90	5.9	0.77	1.3	0.92	1.1
B3LYP	0.89	4.7	0.77	1.3	0.92	1.1

[a] The relative deviation (dev) in % is given with respect to the results of the Dirac−Hartree−Fock (DHF) calculation.

of $\rho(\boldsymbol{r})$ and at the ring critical point it underestimates $\rho(\boldsymbol{r})$ compared to the DFT results. The largest deviations, 7.1% for the local density approximation (LDA) and 5.1% for the generalized gradient approximation (GGA) functionals (like BLYP, BP86, B3LYP), are found for $\rho(\boldsymbol{r})$ at the ring critical point (deviations given relative to the DHF results). The bond path profile is found to be V-shaped for all cases, but the path is found to be more exocyclic in the case of the Dirac−Hartree−Fock calculations, indicated by the distance between the M−C BCPs of 81.8 pm for Dirac−Hartree−Fock compared to 62.1 pm for the LDA and 66.0 pm for the GGA functionals.

In the case of complex **2** there is no clear trend found for the effect of the various different functionals used as it was the case for M = Ni. The values for $\rho(\boldsymbol{r})$ at the M−C bond critical points are the same comparing the Dirac−Hartree−Fock and the LDA results and only 0.01 eÅ^{-3} higher than those obtained with the GGA functionals. The same holds true for $\rho(\boldsymbol{r})$ at the ring critical points, where the deviations are also remarkably small (same values for Dirac−Hartree−Fock and LDA, 1.7% difference (averaged values) between Dirac−Hartree−Fock and GGA). Only for the C−C and the C−H bonds $\rho(\boldsymbol{r})$ at the bond critical points is again overestimated by the Dirac−Hartree−Fock calculation (LDA: 1.9%, 5.0%; GGA: 0.9%, 2.3% for the C−C and the C−H bond, respectively). The profile of the bond path is again V-shaped in all cases, as it was the case for M = Ni. It is found to be more endocyclic in the DFT calculations (d(M−C BCP) = 67.2, 60.5, and 64.1 pm for Dirac−Hartree−Fock, LDA, and GGA, respectively). However, the difference between the Dirac−Hartree−Fock and the GGA results is significantly smaller for M = Pd (3.1 pm) compared to M = Ni (15.8 pm).

For **3** the same trend for $\rho(\boldsymbol{r})$ at the bond critical points is found as in the case of M = Ni. Dirac−Hartree−Fock overestimates the values compared to the four-component DFT results. But for M = Pt, the values of $\rho(\boldsymbol{r})$ at the ring critical point are very similar for all cases. As for Ni and Pd, the bond path profile is always V-shaped and more endocyclic for the DFT calculations (distances between the M−C bond critical points d(M−C BCP) = 78.2, 71.1, and 73.8 pm for Dirac−Hartree−Fock, LDA, and GGA, respectively). The difference between the Dirac−Hartree−Fock and the GGA results is again small (4.4 pm).

To conclude, the inclusion of electron correlation within DFT leads to reduced values of $\rho(\boldsymbol{r})$ at the bond critical points of the C−C and C−H bonds for all complexes. The same holds true for the M−C bond in the case of M = Ni and Pt. This indicates that the very good agreement of $\rho(\boldsymbol{r})$ at the critical points between the Dirac−Hartree−Fock and the DFT calculations for M = Pd is due to a cancellation of two complementary effects. The qualitative nature of the topology (i.e., V-shaped bond path profile) is not affected by the inclusion of electron correlation, though it still leads to a change in the curvature of the bond path profile. These results are in agreement with earlier studies that investigated the effects of electron correlation on the topology of the electron density.[80] If we assume that no artifacts are introduced through the approximate exchange functional, then the effect on the topology of $\rho(\boldsymbol{r})$ as exerted by electron correlation is comparable in magnitude to the effects found for the different relativistic Hamiltonians discussed above.

7. The Laplacian of the Electron Density

Up to now only the topology of $\rho(\boldsymbol{r})$ has been discussed in detail. The analysis of the Laplacian of the total electron density and − as already mentioned in section 5 − of the local charge concentrations found in the valence shell of charge concentrations of the metal atoms should provide detailed insight into the relativistic effect on $\rho(\boldsymbol{r})$. As was shown by Shi and Boyd[81] and by Sagar et al.[43] for nonrelativistic wave functions and later by Kohout, Savin, and Preuss[25] by relativistic calculations on isolated third-row transition-metal atoms the shell structure of *isolated* atoms is not completely resolved by the Laplacian. For light main group elements regions of positive and negative values of $L(\boldsymbol{r})$ found for each subshell are clearly distinguishable. Starting with the transargonic elements the fourth shell of charge concentration is already so weakly visible in $L(\boldsymbol{r})$ that it might only appear as a small shoulder in the *negative* region of $L(\boldsymbol{r})$.[82] For third-row transition-metals, no maximum for the $n = 6$ shell could be found at all (with n being the principle quantum number).

In the following we will discuss to what extent these earlier findings on isolated atoms are transferable to the transition-metal complexes **1**−**3**. As in the case of the M(C_2H_2) molecules, we solely rely on a decomposition of the total

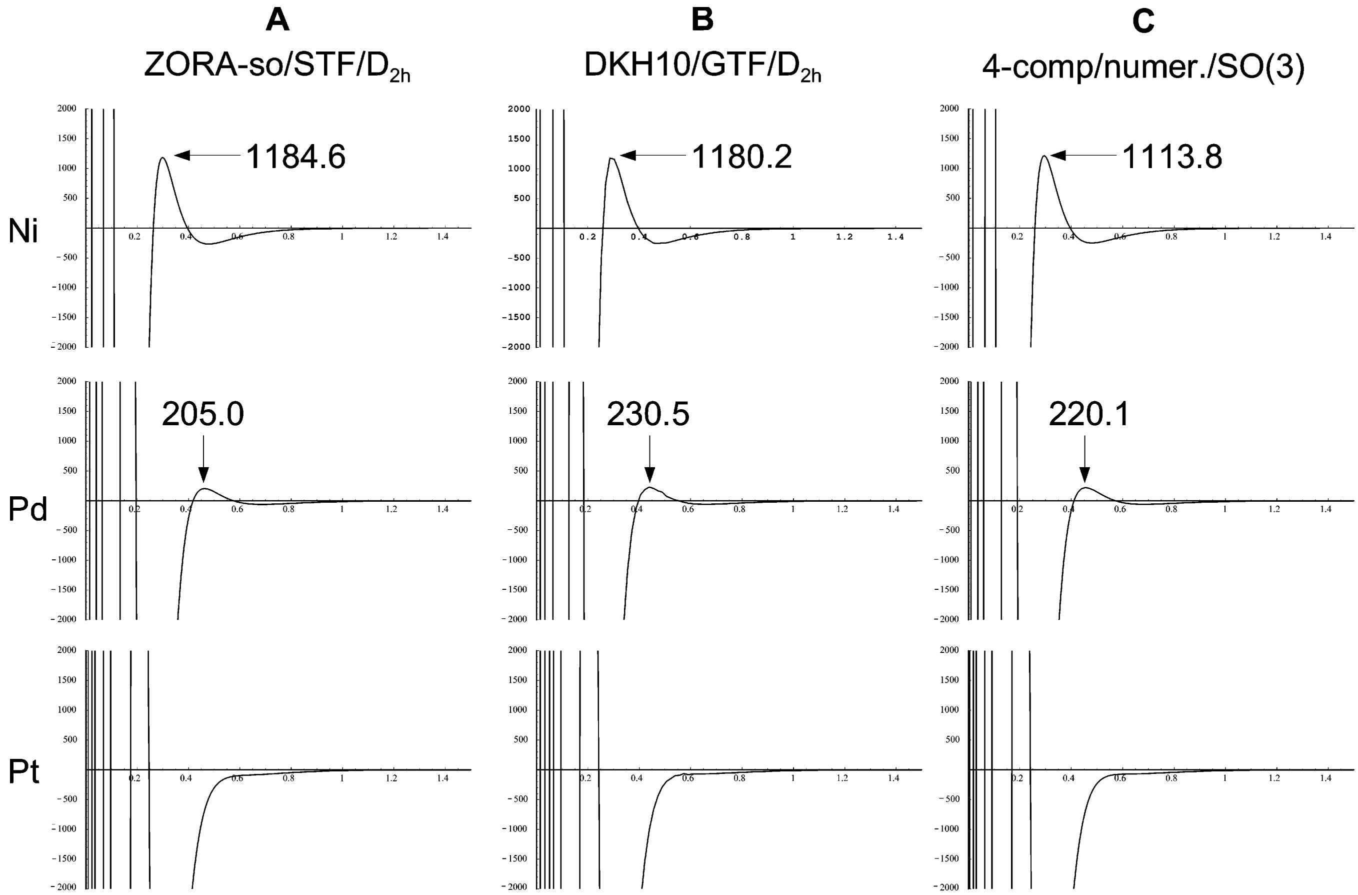


Figure 3. Comparison of $L(\boldsymbol{r})$ (in eÅ^{-5}) for an isolated Ni, Pd, and Pt atom as obtained from a quasi two-component ZORA calculation (D_{2h} symmetry) using a triple-ζ basis set of Slater functions (STF) (A), a scalar relativistic DKH10 calculation (D_{2h} symmetry) using a GTF basis set (B) and a fully numerical four-component calculation of a pure $ns^0(n-1)d^{10}$ configuration (in radial SO(3) symmetry) (C).

electron density of an isolated atom in terms of the orbitals of a single configuration. Consequently, we again adopt the Hartree−Fock model also for the atomic calculations and hence consider only the singlet configuration $ns^0(n-1)d^{10}$. We first analyze the spherically averaged density obtained from a four-component Hartree−Fock calculation using a fully numerical code where the atomic 4-spinors in the Slater determinant are given by

$$\phi_{n_i,\kappa_i,m_{j(i)}}(r,\vartheta,\varphi)=\begin{pmatrix} F_i(r) & \chi_{\kappa_i,m_{j(i)}}(\vartheta,\varphi) \\ \mathrm{i}G_i(r) & \chi_{-\kappa_i,m_{j(i)}}(\vartheta,\varphi)\end{pmatrix} \quad (15)$$

The radial averaged density $\bar{\rho}(\boldsymbol{r})$ is then subjected to the action of the Laplacian operator transformed from Cartesian coordinates to spherical coordinates which then reads[43]

$$\Delta(x,y,z)\rightarrow\Delta(r,\vartheta,\varphi)\rightarrow\Delta(r)=\frac{\partial^2}{\partial r^2}+\frac{2}{r}\frac{\partial}{\partial r} \quad (16)$$

The $L(\boldsymbol{r})$ maps obtained for a nickel, palladium, and a platinum atom are shown in Figure 3C. Starting at $r = 0$ $L(\boldsymbol{r})$ is positive infinite.[43] According to Bader[42] we count the first zero-crossing as the first shell. The total of five zero-crossings can be attributed to three clearly distinguishable shells in the case of the nickel atom. Moving to the heavier palladium atom, an analogous scenario emerges. Here, the seven zero-crossings indicate four resolved subshells in $L(\boldsymbol{r})$. Yet, the maximum in the Laplacian indicative for the outermost shell is significantly weakened compared to the lighter nickel atom (220.1 and 1113.8 eÅ^{-5}, respectively). For platinum the fifth shell is even less pronounced, and thus only a very weak maximum is found (6.8 eÅ^{-5}). This trend is also observed when instead of the four-component Hamiltonian the ZORA or the DKH10 approximation is applied (Figure 3, parts A and B, respectively). The absolute values found for the maxima in $L(\boldsymbol{r})$ corresponding to the outermost shell are similar for all three model Hamiltonians in the case of nickel and palladium. In the case of platinum no maximum to be attributed to the fifth shell is found for any of the calculations employing either the ZORA, the DKH10, or the four-component Hamiltonian.

Turning from the Laplacian of the isolated metal atoms to the molecular systems **1**−**3** Figure 4 depicts the Laplacian along a line perpendicular to the C−C bonding axis moving from the metal atom toward the C−C BCP of the acetylene ligand. In Figure 4a,b the Laplacian as obtained from calculation using the nonrelativistic, the DKH10, the ZORA, and the four-component Hamiltonian are compared. As was already the case for the density and the Laplacian at the critical points no significant difference can be observed for complex **1**. Only in complex **3** a significant difference is found. Here, the nonrelativistic calculations reveal two weak maxima in the negative region of $L(\boldsymbol{r})$ at approximately 0.75

Å from the nucleus (−3.36 and −3.55 eÅ^{-5} for the maximum facing and opposite to the ligand, respectively). These maxima are not observed when any of the three relativistic Hamiltonians considered within our study is used. In addition, in the relativistic calculation the electron density in the region of the valence shell cis to the ligand appears to be less concentrated than in the region trans to the ligand, which is not the case for the nonrelativistic Laplacians. Thus, in general the result found for the isolated atoms, namely the diminishing of the valence shell charge concentration starting from the second transition-metal period, is also found for metal atoms bound in molecules. The region of charge concentration which can be attributed to the valence shell of a first transition row atom is less pronounced in second-row transition metals and finally reduced to a weak maximum in the *negative* region of $L(\mathbf{r})$ in the case of nonrelativistic calculations. For isolated atoms relativistic effects can change the situation qualitatively and cause a valence shell still to be observed as was shown by the *positive* value of $L(\mathbf{r})$ in the valence region of the platinum atom. Yet, for the Pt atom bound to a ligand in **3** the opposite effect is observed, and the subtle maximum in the negative region of $L(\mathbf{r})$ which is found in the nonrelativistic calculation vanishes in the relativistic calculations.

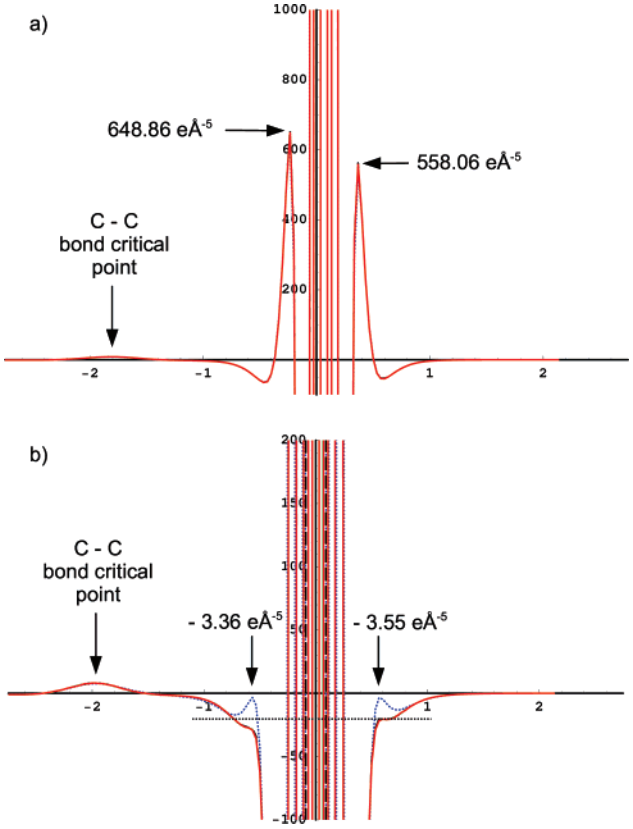


Figure 4. Comparison of $L(\mathbf{r})$ (in eÅ^{-5}) along a line through the metal atom position and the C−C bond critical point of the acetylene ligand in complexes a) **1** and b) **3**. The dotted and solid lines represent the nonrelativistic and the relativistic Hamiltonians, respectively. Note the different scales for the *y*-axis in a) ($L(\mathbf{r})_{\max}$ = 1000 eÅ^{-5}) and b) ($L(\mathbf{r})_{\max}$ = 200 eÅ^{-5}). The values marked in a) are referring to the four-component calculation, while in b) they refer to the nonrelativistic calculation.

Table 5. Values of $L(\mathbf{r})$ in eÅ^{-5} and $\rho(\mathbf{r})$ in eÅ^{-3} at the Positions of the Ligand Induced Charge Concentrations (LICC) in Complexes **1** and **2**, i.e. for $M(C_2H_2)$ with M = Ni and M = Pd, Respectively

	$L(\mathbf{r}_{\mathrm{LICC}})$			$\rho(\mathbf{r}_{\mathrm{LICC}})$		
	LICC1	LICC2	LICC3	LICC1	LICC2	LICC3
			M = Ni			
nonrel	642	612	544	39.76	38.25	36.77
DKH10	672	639	567	40.28	38.79	37.37
ZORA	667	637	565	40.04	39.66	38.36
four-comp	666	636	563	40.16	38.76	37.28
			M = Pd			
nonrel	85	80	74	13.11	12.54	12.41
DKH10	81	77	74	13.29	12.79	12.73
ZORA	81	76	73	13.32	12.82	12.71
four-comp	82	76	73	13.32	12.81	12.71

Finally we will now compare the overall topology of $L(\mathbf{r})$ in the molecular plane of complexes **1**−**3** as obtained from the calculations using the different relativistic Hamiltonians (see Figure 5). Beginning again with **1** for which $L(\mathbf{r})$ is shown in Figure 5a),d),g),j) for the four-component, the ZORA, the DKH10, and the nonrelativistic Hamiltonian, respectively, no differences are observed at first sight. For all cases, four ligand induced charge concentrations[82] are found in the valence region of the nickel atom, LICC1 facing the acetylene ligand, LICC2 (and the symmetry equivalent LICC2′) on a line parallel to the C−C bond axis, and LICC3 opposite to the ligand. The origin of this polarization was recently investigated within an experimental study on the nickel complex [Ni(C_2H_4)dbpe] (dbpe = $Bu^t_2PCH_2CH_2PBu^t_2$).[39] There it was shown that the occurrence of the four ligand induced charge concentrations in the MCC plane of the valence shell of the metal atom can directly be attributed to the π back-donation of electron density from the occupied metal *d* orbitals to the empty π^* orbitals of the ligand. The positions of LICC1-4 resemble that of the maxima found in the difference densities discussed in section 5. The values of $L(\mathbf{r})$ and $\rho(\mathbf{r})$ at the positions of the ligand induced charge concentrations are given in Table 5. The first thing we note is that the values of $L(\mathbf{r})$ are generally reduced by about a factor of 2 compared to the values found for the valence shell of the isolated nickel atom (600 compared to 1100 eÅ^{-5}). Closer inspection of the values for the three different local charge concentrations reveals an increase relative to the nonrelativistic calculation in the value of $L(\mathbf{r})$ for LICC1 by 30, 25, and 24 eÅ^{-5} using the DKH10, the ZORA, and the four-component Hamiltonian, respectively. A similar increase of 25 and 21 eÅ^{-5} (averaged values) is observed for the two charge concentrations denoted as LICC2 and LICC3, respectively. Thus, even in the case of the lightest of the metal atoms considered within our study, relativistic effects on the magnitude of the local charge concentrations are clearly observable. As was already discussed in section 5 by means of the difference density maps, the increase of charge concentration in these regions is at the same time accompanied by an increase in the values of $\rho(\mathbf{r})$ (Table 5). For complex **2**, the inspection of the

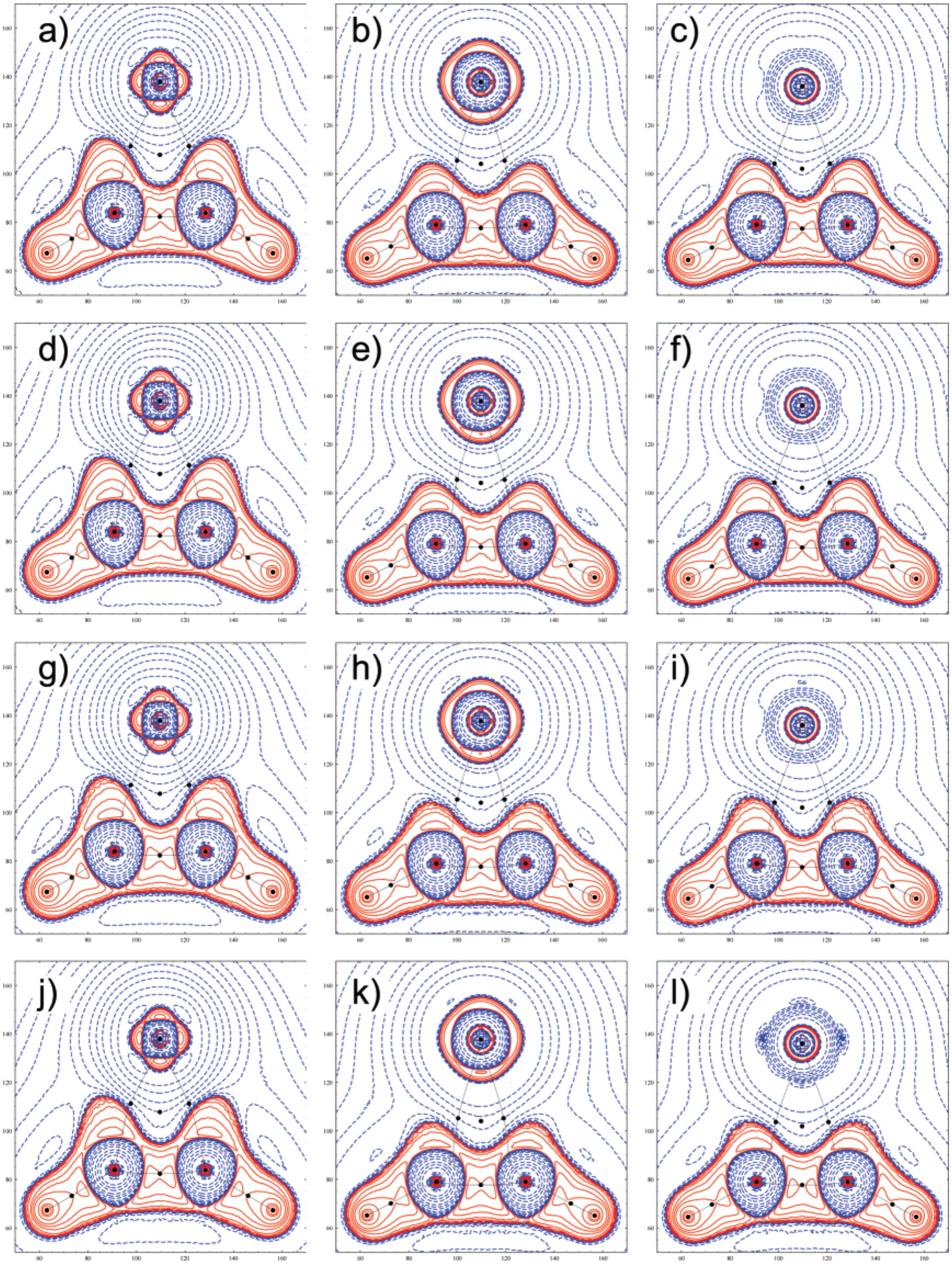


Figure 5. *L*(***r***) in the molecular plane of complexes **1**−**3** as obtained from four-component (a−c), ZORA (d−f), DKH10 (g−i), and nonrelativistic (j−l) calculations. Positive and negative values of *L*(***r***) are denoted by solid and dashed lines, respectively, and the bond paths are drawn as black solid lines. The positions of the critical points are indicated by filled black circles. The contour lines are drawn at the default values specified in Figure 1.

Laplacian maps in Figure 5b),e),h),k) shows no qualitative differences similar to the situation for **1**. In the case of M = Pd there is (relative to the absolute values) a similar change in the values of $L(\boldsymbol{r})$ at the positions of the ligand induced charge concentrations (Table 5) as was found for complex **1** (approximately 5%). The absolute values of $L(\boldsymbol{r})$ at the positions of the local concentrations are significantly smaller, as already discussed above, and only for LICC1 and LICC2 a decrease by approximately 4 eÅ^{-5} is observed when comparing the different relativistic Hamiltonians to the nonrelativistic case. In addition, there is a slight increase in the values of $\rho(\boldsymbol{r})$ at the positions of the local charge concentrations, as was the case for complex **1**. The significantly smaller values of $\rho(\boldsymbol{r})$ at the positions of the ligand induced charge concentrations in **2** compared to **1** are due to the fact that the distance between the nucleus and the local maxima in $L(\boldsymbol{r})$ is larger for **2** (0.46 Å) than for **1** (0.28 Å). At these distances to the nucleus the values of $\rho(\boldsymbol{r})$ as obtained by numerical four-component calculations on the free atoms are 36.07 eÅ^{-3} for the nickel and 12.28 eÅ^{-3} for the palladium atom, which in both cases is close to the values found for the positions of the local charge concentrations in **1** and **2**. Comparing finally $L(\boldsymbol{r})$ in the molecular plane of complex **3** [Figure 5c),f),i),l)] one finds a significant difference between the nonrelativistic and the relativistic calculations but again no change between the three relativistic Hamiltonians. As already indicated by the radial plots of $L(\boldsymbol{r})$ in Figure 4, local maxima in the negative region around the platinum atom are found when using the nonrelativistic Hamiltonian. The positions and distances to the nucleus closely resemble the positions of the ligand induced charge concentrations found for **2**. These maxima are not found with the relativistic Hamiltonians. Apart from this there are no significant differences in the topology of $L(\boldsymbol{r})$.

8. Summary and Conclusion

In this work we conducted a systematic study of relativistic effects on the total electron density. We also compared these effects to those due to electron correlation. In this way, we were able to assess theory-inherent deficiencies in electron density studies. This is important because so far experimental and calculated densities have been compared directly neglecting the fact that both are affected by measurement errors and by a method-inherent error, respectively. Hence, hardly any reliable error estimates are available for either experiment or theory. One aim of the present study was to close this gap for the theoretical approaches. We should, however, note that we did not investigate the magnitude of method-inherent errors as introduced by a small-sized basis set (since our results were obtained close to the basis set limit; compare also refs 14, 15, 83, and 84 in this context) or by the fact that the study of an isolated molecule does not necessarily represent a true benchmark for X-ray diffraction studies of molecular crystals.

While scalar-relativistic effects were included through DKH Hamiltonians, spin−orbit effects were included in the ZORA framework. Results from these calculations were compared to the limiting reference cases, namely to nonrelativistic and four-component results. We could show that especially for the platinum complex **3** the differences in the topological parameters at the critical points and thus even in the bonding region due to relativistic effects are of significant magnitude when comparing results obtained from experimental and theoretical electron densities. This is best illustrated by the difference in $\rho(\boldsymbol{r})$ at the M−C bond critical point in **3** which is underestimated by 0.06 eÅ^{-3} (corresponding to a relative deviation of 6%) when a nonrelativistic Hamiltonian is employed compared to the four-component result.

The comparison of the electron densities obtained from calculations employing the DKH, ZORA, and the Dirac−Hartree−Fock Hamiltonian suggests, however, that the relativistic effects in complexes **1**−**3** are already accounted for by a scalar-relativistic approximation, so that computationally more demanding two-component calculations including spin−orbit effects or even four-component calculations are not necessarily required. The corresponding deviations in the Laplacian can be much larger (up to 90% for $L(\boldsymbol{r}_{\mathrm{BCP}})$ in complex **3**). This fact again demonstrates how the Laplacian can be employed to detect subtle changes in an electron density distribution. A detailed analysis of the Laplacian and especially of the local charge concentrations of complexes **1** and **2** showed that the scalar relativistic contraction of the electronic core shells of the metal atoms leads to an increase in $L(\boldsymbol{r})$ as well as in $\rho(\boldsymbol{r})$ at the positions of the local charge concentrations. Yet, as we demonstrated by a detailed analysis of the electronic shell structure of the isolated atoms and the metal centers in the complexes **1**−**3** the vanishing outer most shell as revealed by the Laplacian plays the far greater role when comparing the topology of the Laplacian within the 10th group of the periodic table.

Finally, comparing results obtained within the Hartree−Fock approximation to results obtained from DFT calculations for complexes **1**−**3** we could show that the effect of electron correlation on the topology of the electron density as accounted for within present-day density functional theory (up to 5.9% change in $\rho(\boldsymbol{r})$ at the M−C bond critical point in complex **1**) can be of the same order of magnitude as the relativistic effects.

Acknowledgment. This project has been generously supported by the German Science Foundation DFG (SPP 1178 “Experimental Electron Density as a Key to the Understanding of Chemical Interactions”). It is a pleasure for M.R. to gratefully acknowledge that it was made possible by the German Science Foundation DFG to continue funding (through Re1703/2-1) in Switzerland after the relocation of M.R. from Jena to Zurich. V.H. thanks NanoCat, an International Graduate Program within the Elitenetzwerk Bayern for financial support. This paper is dedicated to Professor Juergen Hinze, Ph.D., on the occasion of his 70th birthday.

References

(1) Hinze, J.; Jaffé, H. H. *J. Am. Chem. Soc.* **1962**, *84*, 540−546.

(2) Hinze, J.; Jaffé, H. H. *Can. J. Chem.* **1963**, *41*, 1315−1328.

(3) Hinze, J.; Jaffé, H. H. *J. Phys. Chem.* **1963**, *67*, 1501−1506.

(4) Mulliken, R. S. *J. Chem. Phys* **1934**, *2*, 782−793.

(5) Parr, R. G. *J. Chem. Phys* **1978**, *68*, 3801−3807.

(6) Bader, R. *Atoms in Molecules;* Clarendon Press: Oxford, 1990.

(7) Geerlings, P.; DeProft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793−1873.

(8) Ayers, P. *Faraday Discuss.* **2007**, *135*, 161−190.

(9) Zuo, J. M.; Kim, M.; O’Keeffe, M.; Spence, J. C. H. *Nature* **1999**, *401*, 49−52.

(10) Hansen, N. K.; Coppens, P. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1978**, *34*, 909−921.

(11) Coppens, P. *X-Ray Charge Densities and Chemical Bonding;* Oxford University Press: Oxford, New York, 1997.

(12) Gatti, C. *Z. Kristallogr.* **2005**, *220*, 399−457.

(13) Tsirelson, V. G.; Ozerov, R. P. *Electron Density and Bonding in Crystals;* Institute of Physics Publishing: Bristol, 1996.

(14) Koritsanszky, T. S.; Coppens, P. *Chem. Rev.* **2001**, *101*, 1583−1627.

(15) Coppens, P.; Volkov, A. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2004**, *A60*, 357−364.

(16) Scherer, W.; McGrady, G. S. *Angew. Chem., Int. Ed.* **2004**, *43*, 1782−1806.

(17) Scherer, W.; Sirsch, P.; Shorokhov, D.; Tafipolsky, M.; McGrady, G. S.; Gullo, E. *Chem. Eur. J.* **2003**, *9*, 6057−6070.

(18) Byetheway, I.; Gillespie, R. J.; Tang, T. H.; Bader, R. F. W. *Inorg. Chem.* **1995**, *34*, 2407−2414.

(19) Reisinger, A.; Trapp, N.; Krossing, I.; Altmannshofer, S.; Herz, V.; Presnitz, M.; Scherer, W. *Angew. Chem.* **2007**, in press.

(20) Stevens, E. D.; Coppens, P. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1976**, *32*, 915−917.

(21) Schwerdtfeger, P. *Relativistic Electronic Structure Theory. Part I. Fundamentals;* Elsevier: Amsterdam, 2002.

(22) Schwerdtfeger, P. *Relativistic Electronic Structure Theory. Part II. Applications;* Elsevier: Amsterdam, 2004.

(23) Hess, B. A. *Relativistic Effects in Heavy-Element Chemistry and Physics;* Wiley: Chichester, 2003.

(24) Hebben, N.; Himmel, H.-J.; Eickerling, G.; Herrmann, C.; Reiher, M.; Herz, V.; Presnitz, M.; Scherer, W. *Chem. Eur. J.* **2007**, DOI: 10.1002/chem.200700885.

(25) Kohout, M.; Savin, A.; Preuss, H. *J. Chem. Phys.* **1991**, *95*, 1928−1942.

(26) Reiher, M.; Hinze, J. Four-component ab initio Methods for Electronic Structure Calculations of Atoms, Molecules and Solids. In *Relativistic Effects in Heavy-Element Chemistry and Physics;* Wiley-VCH: Weinheim, 2003; pp 61−88.

(27) Reiher, M.; Wolf, A.; Hess, B. A. Relativistic Quantum Chemistry: From quantum electrodynamics to quasi-relativistic methods. In *Handbook of Theoretical and Computational Nanotechnology*; Rieth, M., Schommers, W., Eds.; 2006; Vol. 1, pp 401−444.

(28) Chang, C.; Pelissier, M.; Durand, P. *Phys. Scr.* **1986**, *34*, 394−404.

(29) van Lenthe, E.; Baerends, E.-J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597−4610.

(30) van Lenthe, E.; Baerends, E.-J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783−9792.

(31) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89−155.

(32) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742−3748.

(33) Wolf, A.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 9215−9226.

(34) Reiher, M. *Theor. Chem. Acc.* **2006**, *116*, 241−252.

(35) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 2037−2047.

(36) Barysz, M.; Sadlej, A. J. *J. Chem. Phys.* **2002**, *116*, 2696−2704.

(37) Ilias, M.; Saue, T. *J. Chem. Phys.* **2007**, *126*, 064102.

(38) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. *J. Am. Chem. Soc.* **1983**, *105*, 5061−5068.

(39) Scherer, W.; Eickerling, G.; Shorokhov, D.; Gullo, E.; McGrady, G. S.; Sirsch, P. *New J. Chem.* **2006**, *30*, 309−312.

(40) Cioslowski, J.; Karwowski, J. *Fundamentals of Molecular Similarity;* Kluwer Academic: New York, 2001.

(41) Bader, R. F. W.; MacDougall, P. J.; Lau, C. D. H. *J. Am. Chem. Soc.* **1984**, *106*, 1594−1605.

(42) Bader, R. F.; Essén, H. *J. Chem. Phys.* **1984**, *80*, 1943−1960.

(43) Sagar, R. P.; Ku, A. C. T.; Smith, V. H., Jr. *J. Chem. Phys.* **1988**, *88*, 4367−4374.

(44) Shi, Z.; Boyd, J. R. *J. Chem. Phys.* **1988**, *88*, 4375−4377.

(45) Chan, W.-T.; Hamilton, I. P. *J. Chem. Phys.* **1998**, *108*, 2473−2485.

(46) Bader, R. F. W.; Gillespie, R. J.; Martín, F. *Chem. Phys. Lett.* **1998**, *290*, 488−494.

(47) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Klömel, C. *Chem. Phys. Lett.* **1989**, *162*, 165−169.

(48) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(49) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822−8824.

(50) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866−872.

(51) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123−141.

(52) ADF2006.01, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands. http://www.scm.com (accessed November 2006).

(53) te Velde, G.; Bickelhaupt, F.; van Gisbergen, S.; Guerra, C. F.; Baerends, E.; Snijders, J.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931−967.

(54) Jensen, H. J. Aa; Saue, T.; Visscher, L. with contributions from Bakken, V.; Eliav, E.; Enevoldsen, T.; Fleig, T.; Fossgaard, O.; Helgaker, T.; Laerdahl, J.; Larsen, C. V.; Norman, P.; Olsen, J.; Pernpointner, M.; Pedersen, J. K.; Ruud, K.; Salek, P.; van Stralen, J. N. P.; Thyssen, J.; Visser, O.; Winther, T. *Dirac, a relativistic ab initio electronic structure program, Release DIRAC04.0 (2004)*; http://dirac-.chem.sdu.dk (accessed September 2005).

(55) Saue, T.; Faegri, K.; Helgaker, T.; Gropen, O. *Mol. Phys.* **1997**, *91*, 937−950.

(56) Saue, T.; Helgaker, T. *J. Comput. Chem.* **2002**, *23*, 814−823.

(57) Dirac, P. A. M. *Proc. Camb. Phil. Soc.* **1930**, *26*, 376−385.

(58) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200−1211.

(59) Dyall, K. G. *Theor. Chem. Acc.* **2007**, *117*, 483−489.

(60) Dyall, K. G. *Theor. Chem. Acc.* **2004**, *112*, 403−409.

(61) Pou-Amerigo, R.; Merchan, M.; Nebotgil, I.; Widmark, P. O.; Roos, B. O. *Theor. Chim. Acta* **1995**, *92*, 149−181.

(62) Partridge, H. *J. Chem. Phys.* **1989**, *90*, 1043−1047.

(63) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(64) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Schütz, M. et al. Molpro 2006.2, a package of ab initio programs.

(65) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945−10956.

(66) Katan, C.; Rabiller, P.; Lecomte, C.; Guezo, M.; Oison, V.; Souhassou, M. *J. Appl. Crystallogr.* **2003**, *36*, 65−73.

(67) Wolfram Research, Inc. *Mathematica Version 5.2;* Wolfram Research, Inc.: Champaign, IL, 2005.

(68) Presnitz, M.; Mayer, F.; Herz, V.; Eickerling, G.; Scherer, W. “calc.lap.nb”, Universität Augsburg (Lehrstuhl CPM), 2007 Mathematica Script for the Processing of Volume Data (Calculation of the Laplacian Field).

(69) Reiher, M. Ph.D. Thesis, University of Bielefeld, 1998.

(70) Andrae, D.; Reiher, M.; Hinze, J. *Int. J. Quantum Chem.* **2000**, *76*, 473−499.

(71) Pörschke, K. R.; Yi-Hung, T.; Krüger, C. *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 323−324.

(72) Pilme, J.; Silvi, B.; Alikhani, M. E. *J. Phys. Chem. A* **2005**, *109*, 10028−10037.

(73) Schager, F.; Bonrath, W.; Pörschke, K. R.; Kessler, M.; Krüger, C.; Seevogel, K. *Organometallics* **1997**, *16*, 4276−4286.

(74) Massera, C.; Frenking, G. *Organometallics* **2003**, *22*, 2758−2765.

(75) Coppens, P. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1984**, *A40*, 184−195.

(76) Benabicha, F.; Pichon-Pesme, V.; Jelsch, C.; Lecomte, C.; Khmou, A. *Acta Crystallogr., Sect. B: Struct. Sci*. **2000**, *B56*, 155−165.

(77) Dittrich, B.; Flaig, R.; Koritsanszky, T.; Krane, H.-G.; Morgenroth, W.; Luger, P. *Chem. Eur. J.* **2000**, *6*, 2582−2589.

(78) Scherer, W.; Eickerling, G.; Tafipolsky, M.; McGrady, G. S.; Sirsch, P.; Chatterton, N. P. *Chem. Commun. (Cambridge)* **2006**, 2986−2988.

(79) Rohrmoser, B.; Eickerling, G.; Presnitz, M.; Scherer, W.; Eyert, V.; Hoffmann, R.-D.; Rodewald, U. C.; Vogt, C.; Pöttgen, R. *J. Am. Chem. Soc.* **2007**, *129*, 9356−9365.

(80) Gatti, C.; McDougall, P. J.; Bader, R. W. F. *J. Chem. Phys.* **1987**, *88*, 3792−3804.

(81) Shi, Z.; Boyd, J. *J. Chem. Phys.* **1988**, *88*, 4375−4377.

(82) McGrady, G. S.; Haaland, A.; Verne, H. P.; Volden, H. V.; Downs, A. J.; Shorokhov, D.; Eickerling, G.; Scherer, W. *Chem. Eur. J.* **2005**, *11*, 4921−4934.

(83) Volkov, A.; Abramov, Y.; Coppens, P.; Gatti, C. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2000**, *A56*, 332−339.

(84) Henn, J.; Ilge, D.; Leusser, D.; Stalke, D.; Engels, B. *J. Phys. Chem. A* **2004**, *108*, 9442−9452.

CT7001573

JCTC Journal of Chemical Theory and Computation

Theoretical Characterization of a Tridentate Photochromic Pt(II) Complex Using Density Functional Theory Methods

Jay C. Amicangelo*

School of Science, Penn State Erie, The Behrend College, 4205 College Drive, Erie, Pennsylvania 16563-0203

Received June 26, 2007

Abstract: Density functional theory methods have been used to characterize a tridentate photochromic Pt(II) complex [Pt(AAA)Cl], its acetonitrile complex [Pt(AAA)Cl•CH_3CN], and the transition state in the complexation reaction. B3LYP/6-31G* (effective core potential for Pt) optimized geometries of Pt(AAA)Cl and Pt(AAA)Cl•CH_3CN are found to be in reasonably good agreement with most of the applicable parameters for the available experimental crystal structures of Pt(AAA)Cl and a Pt(AAA)Cl-triphenylphoshine complex, with the exception of one of the dihedral angles, the deviation of which is determined to be due to a steric cis versus trans effect. Vibrational frequencies are calculated for Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN, and the predicted shift in the benzaldehyde carbonyl frequency is found to be in line with that observed experimentally. Singlet vertical excitation energies are calculated for Pt(AAA)Cl and *cis*-Pt(AAA)-Cl•CH_3CN using time-dependent density-functional theory and are found to be in good agreement with the experimental transition energies, although for *cis*-Pt(AAA)Cl•CH_3CN, the calculations suggest a reassignment of the experimental S_1 and S_2 transitions. Single point energies are calculated at the B3LYP/6-311+G(2d,2p) level (effective core potential for Pt) and the calculations predict the complexation reaction (dark reaction) to be exothermic and, after a correction to the entropy, to be exoergic at 298 K and to proceed with a reasonable activation energy. Based on singlet and triplet vertical excitation energies, it is speculated that the photoreaction occurs via an intersystem crossing from S_1 to T_1 for *cis*-Pt(AAA)Cl•CH_3CN followed by an adiabatic reaction along the T_1 surface and then nonradiative intersystem crossing to the S_0 state of Pt(AAA)Cl.

Introduction

Photochromism is generally defined as a reversible photoinduced transformation of a chemical species between two forms having distinct absorption spectra.[1] The majority of the studies on photochromic compounds have been for organic systems,[2] with a smaller amount of work being focused on inorganic compounds.[3] Within the inorganic systems, an even smaller amount has been concerned with transition-metal compounds and complexes.[45]

One particularly interesting example of a photochromic transition-metal compound is a Pt(II) complex, known as *cis*-[N-(*o*-aminobenzylidene)anthranilaldehydato-*O*,*N*,*N*′]-chloroplatinum or Pt(AAA)Cl, that was synthesized and characterized by Mertes and co-workers.[6,7] Structurally, this compound is a complex between Pt(II) and a tridentate ligand that is a Schiff base condensate of *o*-aminobenzaldehyde (Figure 1). Shortly after the initial report about the structural characterization of Pt(AAA)Cl appeared in the literature,[6] Mertes and co-workers reported that this complex underwent a reversible, photochromic solvolysis reaction, shown in Figure 1, in coordinating solvents such as acetonitrile and dimethylsulfoxide.[7] The unusual feature about this photochromic reaction was that it was appeared to operate in a reverse fashion to that observed for most other photochemical and photochromic transition-metal solvolysis reactions,[8] in that the solvolysis reaction was the reaction occurring in the

* Corresponding author e-mail: jca11@psu.edu.

10.1021/ct700158g CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/04/2007

dark and the recoordination of the ligand (aldehyde group) was the photoactive reaction. The primary physical evidence for the proposed photochromic solvolysis reaction (Figure 1) was a red-shift of the aldehyde carbonyl stretching frequency upon uncoordination of the aldehyde and a blue-shift in the visible absorption spectrum upon solvent coordination. Due to the instability of the solvated complex [Pt(AAA)Cl•S] upon evaporation of the solvent, Mertes and co-workers were unable to obtain a crystal structure for the acetonitrile or dimethylsulfoxide complexes.[9] However, these researchers were able to obtain a crystal structure of a related complex with triphenylphoshine, Pt(AAA)Cl•PPh_3,[10] whose infrared and visible spectral characteristics agreed with the acetonitrile and dimethylsulfoxide systems, supporting the proposed reaction.

With the ease of use and availability of computational chemistry software packages and the fast computational speed afforded by modern computers, additional support of a reaction mechanism or molecular structures can often now be obtained by high level theoretical calculations.[11] Particularly relevant to transition-metal systems was the development of reliable density functional theory methods (DFT).[12,13] In this paper, the photochromic solvolysis reaction of Pt(AAA)Cl with acetonitrile is theoretically characterized (geometries, vibrational frequencies, singlet and triplet vertical excitations, and single point energies) with density functional theory methods using the B3LYP hybrid functional.[14,15] The primary motivation for this work is the preliminary, unpublished results of Jircitano and co-workers,[16] in which the rate constants for the acetonitrile dark reaction of F, Cl, and CH_3 substituted Pt(AAA)Cl derivatives (substituted on the aromatic rings) have been measured and found to display the trend that electron withdrawing groups increased the rate and electron donating groups decreased the rate relative to Pt(AAA)Cl. It was thought that this might be due to changes in the activation energy with the substituents for the solvolysis reaction and that density functional theory calculations may be able to lend theoretical support to this idea. However, prior to embarking on density functional theory calculations for a whole series of Pt(AAA)Cl derivatives, a full theoretical characterization of Pt(AAA)Cl and its photochromic solvolysis reaction was undertaken, and this comprises the material described in this paper.

Computational Details

All theoretical calculations were carried out using the Gaussian 98[17] or Gaussian 03[18] suite of programs. Because the standard basis sets available in Gaussian 98 or Gaussian 03 are not developed for use with Pt, the Hay-Wadt (HW) relativistic effective core potential (ECP) and valence basis set,[19] modified for use with cations,[20] was used for Pt in all calculations. For all other atoms, the standard Gaussian basis sets were used. In the following descriptions and throughout the paper, only the method and standard basis sets will generally be listed, with the implicit understanding that the modified HW ECP was used for Pt in all calculations.

Ground-state geometry optimizations were performed in a series of steps with basis sets of increasing size for all atoms except Pt, involving the following general sequence: HF/STO-3G → HF/3-21G → HF/6-31G* → B3LYP[14,15]/6-31G*. For Pt(AAA)Cl, the input geometry was that of the crystal structure reported by Mertes and co-workers[6] for the heavy atoms, with the hydrogen atoms added at standard positions using the GaussView[21] program. For the Pt(AAA)Cl•S complexes (S = CH_3CN or PH_3), most of the input geometry for the heavy atoms was taken from the crystal structure of a triphenylphosphine complex of Pt(AAA)Cl reported by Jircitano, Rohly, and Mertes,[10] except that the triphenylphosphine was replaced with a CH_3CN or a PH_3 ligand at the appropriate position (cis or trans) and again the hydrogens were added at standard positions. For the *cis*-Pt(AAA)Cl•CH_3CN transition state, the optimized *cis*-Pt(AAA)Cl•CH_3CN geometry was manipulated until an initial geometry was found that had a large negative frequency that appeared to correspond to the desired reaction coordinate. This geometry was then used as the input geometry for the transition-state optimization using the Gaussian opt=TS keyword. The NoEigenTest option to the opt keyword was used for the first step in the optimization (HF/STO-3G level), due to the presence of several other small negative frequencies for the input structure. Once the HF/STO-3G optimization was complete, a frequency analysis was performed, and this confirmed that the structure was indeed a transition state with only one negative frequency. The HF/STO-3G optimized transition-state geometry was then used as the input for the remaining series of optimizations at successively higher levels of theory, with a frequency analysis being performed at each level to confirm the optimized structure was a transition state. Single point energies were calculated at the B3LYP/6-311+G(2d,2p) level using the B3LYP/6-31G* optimized geometries and were corrected for zero-point and thermal energies using scaled B3LYP/6-31G* vibrational frequencies (scaling factor = 0.9804[22]). Singlet and triplet vertical excitation energies were calculated for the B3LYP/6-31G* optimized geometries using time-dependent density-functional theory (TDDFT), as implemented in Gaussian 98 and Gaussian 03,[23] primarily with the B3LYP functional and the 6-31+G*, 6-311+G*, and the 6-311+G(2d,2p) basis sets.

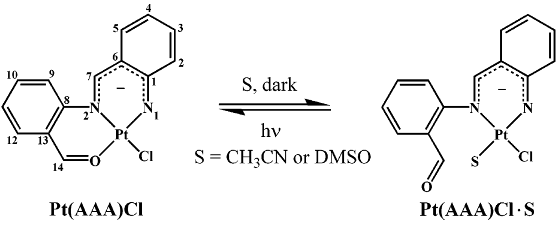


Figure 1. Structure of Pt(AAA)Cl and a schematic representation of its reversible, photochromic solvolysis reaction. The numbering scheme for Pt(AAA)Cl is the same as that used by Mertes and co-workers for the Pt(AAA)Cl crystal structure[6] and is as follows: the nitrogens are labeled as N(1) and N(2) and the carbons are labeled as C(1)−C(14).

Results and Discussion

Optimized Geometries. B3LYP/6-31G* optimized geometries were obtained for the Pt(AAA)Cl and Pt(AAA)Cl•CH_3-

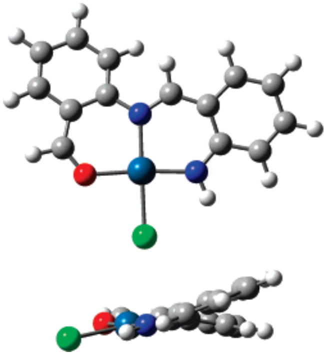


Figure 2. Optimized structure of Pt(AAA)Cl calculated using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom. The view in the lower panel is from the right side of the upper panel to emphasize the canting angle between the two ring systems. Color scheme: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is light green, and platinum is teal.

CN complexes and for the *cis*-Pt(AAA)Cl•CH_3CN transition state. In the case of the Pt(AAA)Cl•CH_3CN complex, an optimized geometry was obtained for the configuration in which the CH_3CN group was cis to the benzaldehyde as well as for the configuration in which the CH_3CN group was trans to the benzaldehyde. As will be discussed below, this was primarily done to assess the effect of the CH_3CN orientation on the dihedral angle of the benzaldehyde group relative to the platinum square plane, due to a significant deviation in this dihedral angle between the calculated *cis*-Pt(AAA)Cl•CH_3CN geometry and that of the X-ray crystal structure of a *trans*-triphenylphosphine adduct of Pt(AAA)Cl (abbreviated as Pt(AAA)Cl•PPh_3).[10] To further examine this effect and examine the effect of the identity of the ligand itself, cis and trans optimized geometries were also obtained for a model compound of Pt(AAA)Cl•PPh_3 in which a PH_3 group was used in place of the triphenylphosphine group, Pt(AAA)-Cl•PH_3. For all of the complexes theoretically examined in this study, the calculated bond lengths and bond angles of the tridentate *N*-(*o*-aminobenzylidene)anthranilaldehydato ligand itself were found to be in very good agreement with the values from the Pt(AAA)Cl and Pt(AAA)Cl•PPh_3 crystal structures.[6,10] Therefore, the discussion that follows will be focused primarily on the geometry around the platinum.

The B3LYP/6-31G* optimized geometry of the Pt(AAA)-Cl complex is shown in Figure 2, and selected values of the bond lengths, bond angles, and dihedral angles are listed in Table 1, along with the values determined by X-ray crystallography by Mertes and co-workers.[6] As can be seen in Figure 2 and Table 1, the approximate square planar coordination around the platinum in the Pt(AAA)Cl complex is reproduced fairly well by the optimized geometry as compared to the crystal structure. This is evidenced by a mean absolute deviation of 0.05 ± 0.02 Å between the experimental and theoretical bond lengths and a mean absolute deviation of 1.7 ± 1.1° between the experimental and theoretical bond angles. These can be reasonably compared to the average uncertainties for the bond length and bond angle parameters of the crystal structure listed in Table 1, which are 0.01 Å and 0.6°, respectively. Another feature of the complex that is well reproduced by the calculations is the significant nonplanarity (canting) between the two chelate ring systems, as can be seen in the lower panel of Figure 2. The canting angle between the chelate rings can be approximated by 180° minus the C(7)−N(2)−C(8)−C(13) dihedral angle, and this is found to be 24° for the crystal structure and 23.0° for the optimized geometry.

The B3LYP/6-31G* optimized geometry of the *cis*-Pt-(AAA)Cl•CH_3CN complex is shown in Figure 3, and selected values of the bond lengths, bond angles, and dihedral angles are listed in Table 1. A crystal structure has not been obtained for the Pt(AAA)Cl•CH_3CN complex; however, as mentioned above, Mertes and co-workers were able to determine a crystal structure for a triphenylphosphine adduct of Pt(AAA)-Cl,[10] and selected values of the experimental bond lengths, bond angles, and dihedral angles for the Pt(AAA)Cl•PPh_3 complex are listed in Table 1. It should be noted that the experimentally observed configuration of the PPh_3 is trans to the benzaldehyde group in the Pt(AAA)Cl•PPh_3 complex. The reason for this was thought to be due to unfavorable steric interactions between the PPh_3 and the benzaldehyde group.[10] Therefore, a direct comparison of all of the experimental geometric parameters of the Pt(AAA)Cl•PPh_3 complex with the calculated parameters for the *cis*-Pt(AAA)-Cl•CH_3CN complex is not possible, and only the most appropriate values are compared in Table 1.

As can be seen from Figure 3 and Table 1, the approximate square planar coordination around the platinum for the *cis*-Pt(AAA)Cl•CH_3CN complex is retained, with the acetonitrile group is acting as the fourth ligand. The calculated Pt−Cl, Pt−N(1), and Pt−N(2) bond lengths are in reasonable agreement with those for the Pt(AAA)Cl•PPh_3 complex, with deviations of 0.035, 0.02, and 0.01 Å, respectively. In terms of bond angles about the platinum, only the N(1)−Pt−N(2) angle can be compared to an experimental value, and it is in reasonable agreement with the experimental value, with a deviation of 1.2°. The largest discrepancy with the experimental Pt(AAA)Cl•PPh_3 structure appears to be with the angle between the two chelate rings, as judged by comparing the C(7)−N(2)−C(8)−C(13) dihedral angles of 93.3° and 69.1° for the experimental and calculated structures, respectively. With respect to the platinum square plane, these dihedral angles put the benzaldehyde ring at approximate angles of 87° and 111° for the experimental and calculated structures, respectively. Since the *cis*-Pt(AAA)Cl•CH_3CN optimization started with the C(7)−N(2)−C(8)−C(13) dihedral angle at the Pt(AAA)Cl•PPh_3 crystal structure value and optimized to the smaller value, it is clear that a dihedral angle near 90° is not a geometric minimum for the cis structure. This large deviation in the C(7)−N(2)−C(8)−C(13) is also the reason for the large deviation between the experimental and calculated Pt−O distance, 0.47 Å. One possibility for this large deviation could be that the cis orientation of the CH_3CN ligand causes more steric repulsions with the benzaldehyde ring, resulting in an increased dihedral angle to relieve these repulsive interactions, as opposed to the trans configuration of the PPh_3 ligand, which would not have these repulsive interactions.

Table 1. Selected Bond Lengths (Å), Bond Angles (deg), and Dihedral Angles (deg) for Pt(AAA)Cl, *cis*- and *trans*-Pt(AAA)Cl·CH_3CN, and *cis*- and *trans*-Pt(AAA)Cl·PH_3 Optimized Using Density Functional Theory and for Pt(AAA)Cl and *trans*-Pt(AAA)ClPPh_3 Determined by X-ray Crystallography

	Pt(AAA)Cl		Pt(AAA)Cl·L (L = CH_3CN or PH_3)				
parameter[a]	crystal[b]	theory[c]	*trans*-PPh_3 crystal[d]	*cis*-CH_3CN theory[c]	*trans*-CH_3CN theory[c]	*cis*-PH_3 theory[c]	*trans*-PH_3 theory[c]
			Bond Lengths				
Pt−Cl	2.309 (5)	2.37	2.355 (6)	2.39	2.41	2.40	2.42
Pt−N(1)	1.93 (2)	1.96	2.00 (1)	1.98	2.01	2.01	2.01
Pt−N(2)	1.99 (1)	2.06	2.07 (1)	2.06	2.03	2.06	2.08
Pt−N_{Ac}				2.02	2.00		
Pt−P			2.274 (6)			2.31	2.27
Pt−O[e]	2.01 (1)	2.04	3.72 (1)	4.19	3.67	4.19	3.60
			Bond Angles				
Cl−Pt−N(1)	87.0 (5)	86.6		87.0		87.0	
N_{Ac}−Pt−N(1)					90.2		
P−Pt−N(1)			92.5 (3)				93.1
N(1)−Pt−N(2)	93.6 (7)	92.5	89.2 (4)	90.4	90.4	89.5	89.7
N(2)−Pt−O	94.4 (6)	92.5					
N(2)−Pt−N_{Ac}				95.2			
N(2)−Pt−P						99.7	
N(2)−Pt−Cl			91.8 (4)		93.8		94.4
O−Pt−Cl	85.1 (4)	88.4					
N_{Ac}−Pt−Cl				87.5	85.5		
P−Pt−Cl			86.7 (2)			83.9	81.7
Pt−N_{Ac}−C_{Ac}				171.6	173.7		
			Dihedral Angles				
C(6)−C(7)−N(2)−C(8)	172 (2)	168.8	176[f]	176.4	179.3	170.7	179.6
C(7)−N(2)−C(8)−C(13)	156 (2)	157.0	93[f]	69.1	90.0	70.1	93.6

[a] Numbering scheme corresponds to that shown in Figure 1; N_{Ac} and C_{Ac} refer to the nitrogen and first carbon of the acetonitrile group, respectively. [b] Reference 6; the number in parentheses is the uncertainty in the last digit. [c] Calculations performed using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, P, and H atoms and a modified Hay-Wadt ECP on the Pt atom. [d] Reference 10; the number in parentheses is the uncertainty in the last digit. [e] Distance between the oxygen and platinum atoms. [f] Not reported by Mertes and co-workers in ref 10; determined by examining the Pt(AAA)Cl PPh_3 crystal structure in GaussView.

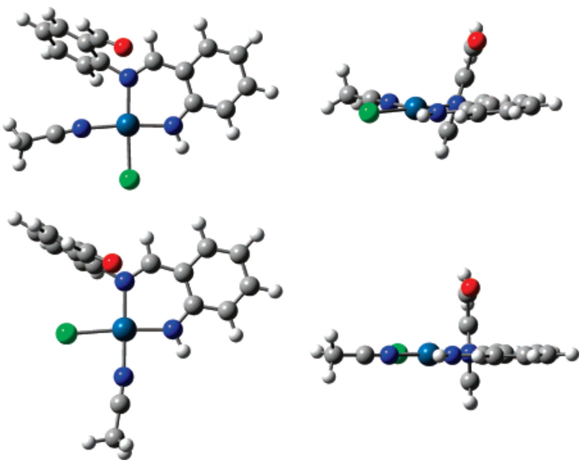


Figure 3. Optimized structure of *cis*-Pt(AAA)Cl·CH_3CN (upper) and *trans*-Pt(AAA)Cl·CH_3CN (lower) calculated using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom. The right panel in both upper and lower is the view along the N(2)−C(8) bond to emphasize the C(7)−N(2)−C(8)−C(13) dihedral angle. The color scheme is the same as that in Figure 2.

To assess if the large deviation of the C(7)−N(2)−C(8)−C(13) dihedral angle is due to the inverted configuration of the Pt(AAA)Cl·PPh_3 crystal structure as compared to that for the calculated structure of *cis*-Pt(AAA)Cl·CH_3CN, an optimization was performed for the Pt(AAA)Cl·CH_3CN complex with the CH_3CN trans to the benzaldehyde group. The B3LYP/6-31G* optimized geometry of the *trans*-Pt(AAA)Cl·CH_3CN complex is shown in Figure 3, and selected values of the bond lengths, bond angles, and dihedral angles are listed in Table 1. As can be seen in Table 1, the approximate square planar coordination around the platinum for this complex is still retained in this configuration. The calculated Pt−Cl, Pt−N(1), and Pt−N(2) bond lengths are again in reasonable agreement with those for the Pt(AAA)Cl·PPh_3 complex, with deviations of 0.055, 0.01, and 0.04 Å, respectively. Two of the bond angles about the platinum, N(1)−Pt−N(2) and N(2)−Pt−Cl, can be compared to experimental values, and the calculated angles are both in reasonable agreement with the experimental values, with deviations of 1.2° and 2.0°, respectively. The C(7)−N(2)−C(8)−C(13) dihedral angle in the trans CH_3CN complex is calculated to be 90.0°, which is in better agreement with the Pt(AAA)Cl·PPh_3 crystal structure value of 93.3°, although still off by 3.3°. This calculated dihedral angle now puts the benzaldehyde ring at an approximate angle of 90° with respect to the platinum square plane, as compared to 87° for the Pt(AAA)Cl·PPh_3 structure. The increase in the calculated C(7)−N(2)−C(8)−C(13) dihedral angle has also

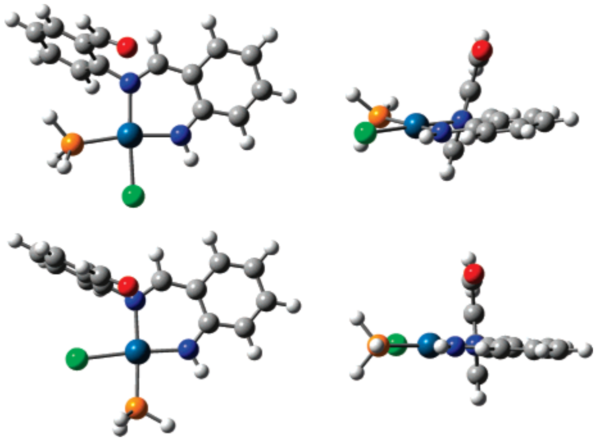


Figure 4. Optimized structure of *cis*-Pt(AAA)Cl•PH_3 (upper) and *trans*-Pt(AAA)Cl•PH_3 (lower) calculated using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, H, and P atoms and a modified Hay-Wadt ECP on the Pt atom. The right panel in both the upper and lower is the view along the N(2)−C(8) bond to emphasize the C(7)−N(2)−C(8)−C(13) dihedral angle. The color scheme is the same as that in Figure 2, with the addition that phosphorus is orange.

improved the agreement between the experimental and calculated Pt−O distance, with a deviation of 0.05 Å.

In order to further investigate the relative benzaldehyde dihedral angle for cis versus trans geometries, two more calculations were performed to determine if the identity of the ligand (CH_3CN versus PPh_3) has any effect on the calculated dihedral angle. Due to the computational expense of using the triphenylphosphine ligand itself, calculations were performed on cis and trans configurations of a model compound of Pt(AAA)Cl•PPh_3 in which a PH_3 group was used in place of the triphenylphosphine group, Pt(AAA)Cl•PH_3. The B3LYP/6-31G* optimized geometries for the cis and trans configurations of the Pt(AAA)Cl•PH_3 complexes are shown in Figure 4, and selected values of the bond lengths, bond angles, and dihedral angles are listed in Table 1. Similar to the CH_3CN complexes, the cis and trans PH_3 complexes both reproduce the approximate square planar geometry around the platinum fairly well as compared to the Pt(AAA)Cl•PPh_3 structure. The mean absolute deviation of the bond lengths is found to be 0.03 ± 0.02 Å (4 values) and 0.02 ± 0.03 Å (4 values), respectively, for the cis and trans complexes, and the mean absolute deviation of the bond angles is found to be 1.5 ± 1.8° (2 values) and 2.2 ± 2.1° (4 values), respectively. The most significant difference between the optimized *cis*- and *trans*-Pt(AAA)Cl•PH_3 geometries is with the C(7)−N(2)−C(8)−C(13) dihedral angles, which are 70.1 and 93.6°, respectively. These values deviate from the Pt(AAA)Cl•PPh_3 crystal structure value (93.3°) by 23.2 and 0.3°, respectively. Similar to the calculated dihedral angles for the *cis*- and *trans*-Pt(AAA)-Cl•CH_3CN complexes, the dihedral angle for the cis configuration deviates significantly from that of the Pt(AAA)Cl•PPh_3 crystal structure, while the dihedral angle for the trans configuration is much closer to the experimental value and, in fact, in this case is in excellent agreement with the experimental value.

The results of the cis and trans optimizations with both ligands, when taken together, clearly suggest that the calculated benzaldehyde dihedral angle is affected by two influences, one large and one small. The larger influence is the cis versus trans orientation of the ligand, with the cis orientation resulting in a smaller dihedral angle, presumably due to repulsive steric interactions between the ligand and the benzaldehyde ring. Given the similar dihedral angles for the cis structures of both ligands, the magnitude of this repulsive cis interaction appears to be approximately the same for both ligands, although the smaller PH_3 ligand does seem to have slightly less repulsive interactions since it gives rise to the larger dihedral angle. The smaller, more subtle influence is with the identity of the ligand when it is in the trans orientation, with the PH_3 ligand resulting in a slightly larger dihedral angle. This appears to be a classic "trans effect",[24] in that the nature and bonding strength of the ligand has an effect on the strength and therefore the length of the bond trans to itself, which is the Pt−N(2) bond in these trans complexes. As can be seen from Table 1, the Pt−N(2) bond length is larger in the PH_3 complex (2.08 Å) as compared to the CH_3CN complex (2.03 Å), and this most likely results in slightly smaller repulsive interactions between the benzaldehyde ring and the Cl atom and, therefore, a larger dihedral angle for the PH_3 complex.

Even though optimizations were performed on both cis and trans configurations of the Pt(AAA)Cl•CH_3CN and Pt-(AAA)Cl•PH_3 complexes in order to determine the reason for the large difference between the calculated C(7)−N(2)−C(8)−C(13) dihedral angle for the *cis*-Pt(AAA)Cl•CH_3CN complex and the experimental value for the Pt(AAA)Cl•PPh_3 crystal structure, the discussion of the optimized Pt(AAA)-Cl•CH_3CN transition-state geometry as well as the remaining portions of this paper (vibrational frequencies, excited-state calculations, and reaction energetics) will, in general, only be concerned with the cis configuration. The reasoning for this is 2-fold. The first is that the energy of the *trans*-Pt-(AAA)Cl•CH_3CN complex is found to be approximately 15 kJ/mol higher than that of the *cis*-Pt(AAA)Cl•CH_3CN complex at the B3LYP/6-31G* level, and the second is that there is a large amount of experimental evidence regarding the retention of configuration in square planar substitution reactions,[24,25] of which this reaction could be classified as.

The optimized geometry of the *cis*-Pt(AAA)Cl•CH_3CN transition state is shown in Figure 5, and selected values of the bond lengths, bond angles, and dihedral angles are listed in Table 2. It was verified that this structure was a transition state by a vibrational frequency analysis that indicated one imaginary frequency (-163 cm^{-1}) corresponding primarily to the motion of the benzaldehyde and acetonitrile groups. As can be seen from Figure 5, the transition-state geometry around the platinum is a five-coordinate distorted trigonal bipyramid. The bond lengths of the Pt−Cl, Pt−N(1), and Pt−N(2) bond lengths are similar to those calculated in the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN complexes (Table 1), while the Pt−O and Pt−N_{Ac} distances are significantly larger at 2.57 and 2.49 Å, respectively, which is to be expected since these are the groups undergoing the primary changes in the transition state. In terms of bond angles, the

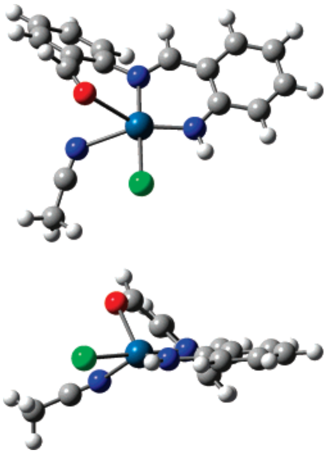


Figure 5. Optimized structure of the *cis*-Pt(AAA)Cl·CH_3CN transition state calculated using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom. The view in the lower panel is from the right side of the upper panel. The color scheme is the same as that in Figure 2. Note that the lines drawn between the platinum atom and the benzaldehyde oxygen and between the platinum atom and the acetonitrile nitrogen are not intended to indicate formal bonds but rather to indicate the geometry around the platinum.

Table 2. Selected Bond Lengths (Å), Bond Angles (deg), and Dihedral Angles (deg) for the *cis*-Pt(AAA)Cl·CH_3CN Transition State Optimized Using Density Functional Theory[a]

parameter[b]	value	parameter[b]	value
	Bond Lengths		
Pt−Cl	2.39	Pt−N_{Ac}	2.49
Pt−N(1)	1.96	Pt−O	2.57
Pt−N(2)	2.04		
	Bond Angles		
Cl−Pt−N(1)	86.6	O−Pt−Cl	98.8
N(1)−Pt−N(2)	92.7	N(1)−Pt−O	140.7
N(2)−Pt−N_{Ac}	94.9	N(1)−Pt−N_{Ac}	150.9
N(2)−Pt−O	81.5	O−Pt−N_{Ac}	68.3
N_{Ac}−Pt−Cl	85.8	Pt−N_{Ac}−C_{Ac}	124.6
	Dihedral Angles		
C(6)−C(7)−N(2)−C(8)	173.8	C(7)−N(2)−Pt−N_{Ac}	155.3
C(7)−N(2)−C(8)−C(13)	133.3	C(7)−N(2)−Pt−O	137.6

[a] Calculations performed using the B3LYP functional and the 6-31G* basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom. [b] Numbering scheme corresponds to that shown in Figure 1; N_{Ac} and C_{Ac} refer to the nitrogen and first carbon of the acetonitrile group, respectively.

Cl−Pt−N(1), N(1)−Pt−N(2), N(2)−Pt−N_{Ac}, and N_{Ac}−Pt−Cl angles are fairly similar to those calculated in the Pt(AAA)Cl and *cis*-Pt(AAA)Cl·CH_3CN complexes; however, the angles involving the carbonyl oxygen (N(2)−Pt−O and O−Pt−Cl) are different. Again, this is due to the fact that the benzaldehyde is one of the groups undergoing the primary changes in the transition state. By comparing the C(7)−N(2)−C(8)−C(13) dihedral angles calculated for the Pt(AAA)Cl and *cis*-Pt(AAA)Cl·CH_3CN complexes with that of the transition state, one can also see that the transition-state dihedral angle is intermediate to that of the Pt(AAA)Cl and *cis*-Pt(AAA)Cl·CH_3CN complexes, which is again as expected. As mentioned above, the overall structure about the platinum is described as a distorted trigonal bipyramid, mostly because of the enlarged N(1)−Pt−O and N(1)−Pt−N_{Ac} angles at 140.7° and 150.9°, respectively, and the squeezed O−Pt−N_{Ac} angle at 68.3°. It is also interesting to note that the acetonitrile group is considerabley off-axis in terms of its approach to the platinum, as judged by the low Pt−N_{Ac}−C_{Ac} angle; however, it is unclear what the cause of this is.

Table 3. Calculated and Experimental Benzaldehyde C═O Vibrational Frequencies (cm^{-1}) for Pt(AAA)Cl, the *cis*-Pt(AAA)Cl·CH_3CN Transition State, and *cis*-Pt(AAA)Cl·CH_3CN[a]

species	HF/ STO-3G	HF/ 3-21G	HF/ 6-31G*	B3LYP/ 6-31G*	expt
Pt(AAA)Cl	1968	1814	1919	1656	1621[b]
Pt(AAA)Cl·CH_3CN transition state	2028	1876	1994	1765	
Pt(AAA)Cl·CH_3CN	2039	1912	2014	1805	1690[c]

[a] The theoretical method and the basis set used for the C, O, N, Cl, and H atoms is indicated in the column headings; a modified Hay-Wadt ECP was used on the Pt atom for all calculations. [b] Reference 6. [c] Reference 7.

Calculated C═O Vibrational Frequencies. Vibrational analysis was performed for the optimized structures of the Pt(AAA)Cl and *cis*-Pt(AAA)Cl·CH_3CN complexes as well as the *cis*-Pt(AAA)Cl·CH_3CN transition state at several levels of theory, and the calculated frequencies for the benzaldehyde C═O stretching modes are listed in Table 3. Experimental vibrational frequencies for the benzaldehyde C═O stretching mode of the Pt(AAA)Cl and Pt(AAA)Cl·CH_3CN complexes are also listed in Table 3.[6,7] As can be seen in Table 3, the C═O stretching frequency is predicted to increase on going from being fully coordinated in the Pt(AAA)Cl complex, partially coordinated in the *cis*-Pt(AAA)Cl·CH_3CN transition state, and uncoordinated in *cis*-Pt(AAA)Cl·CH_3CN complex at each level of theory examined. These results are in qualitative agreement with the experimentally observed behavior for the Pt(AAA)Cl and *cis*-Pt(AAA)Cl·CH_3CN complexes, and the calculated frequencies lend support to the experimental assignments. A decrease in the C═O frequency upon complexation has also been experimentally observed with other metal complexes of ketones and aldehydes.[26]

Since the experimental C═O stretching frequencies have been reported for the Pt(AAA)Cl and Pt(AAA)Cl·CH_3CN complexes, a quantitative comparison can be made with the theoretically calculated values. At all of the levels of theory examined, the theoretical C═O stretching frequencies are predicted to be larger than the experimental values, which is typical and due to known systematic factors such as neglect of anharmonicity, complete or incomplete neglect of electron correlation, and the use of finite basis sets.[11,27] The magnitude of the deviations are found to change with the method and the basis set, with the largest average deviation being at the HF/STO-3G level (∼17.4%) and the smallest average deviation being at the B3LYP/6-31G* level (∼4.2%), consistent with both an increase in the size of the basis set and the inclusion of electron correlation.[11,27,28] There does, however, appear to be a discontinuity in the deviations, with the average deviation at the HF/3-21G level (∼11.1%) being lower than the average deviation at the HF/6-31G* level (∼15.8%). This seemingly anomalous trend has been observed and reported previously in a systematic study comparing the theoretical versus experimental vibrational frequencies of over 1000 molecules at various levels of theory.[28] In this study, it was found that the average deviation at the HF/3-21G level was approximately 9.2%, and at the HF/6-31G* level the average deviation was 10.5%.

In addition to comparing the absolute magnitudes of the experimental and calculated C═O stretching frequencies for the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN complexes, it is also of interest to quantitatively compare the experimental versus the theoretical shift in this frequency between the two complexes. As shown in Table 3, the experimental C═O stretching frequencies for the Pt(AAA)Cl and Pt(AAA)Cl•CH_3CN complexes were determined to be 1621 and 1690 cm^{-1}, respectively, corresponding to a shift of 69 cm^{-1}. The theoretical shift for this vibrational mode in the two complexes is predicted to be 71, 98, 95, and 149 cm^{-1} at the HF/STO-3G, HF/3-21G, HF/6-31G*, and B3LYP/6-31G* levels, respectively. Overall, the magnitudes of the predicted shifts are in reasonable agreement with the experimentally observed value, again lending support to the experimental assignment of these bands to the C═O stretching vibrations in the two complexes. It is worth noting that the best agreement appears to be at the HF/STO-3G level; however, this is most likely fortuitous.

Singlet Excitation Energies. Singlet vertical excitation energies were calculated for the B3LYP/6-31G* optimized Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN geometries using time-dependent density-functional theory (TDDFT) with the B3LYP hybrid functional and the 6-31+G*, 6-311+G*, and the 6-311+G(2d,2p) basis sets. The calculated excitation energies, transition wavelengths, and oscillator strengths for the first three excited states of the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN complexes are listed in Table 4. As can be seen from Table 4, the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN energies calculated for a given transition using the three different basis sets are very close to one another, with deviations between 0.02 and 0.04 eV, which suggests that basis set size does not have a large effect on the calculated excitation energies. However, it is worth noting that the transition energies do display a subtle basis set effect, in that a small decrease in the state energy is generally observed with an increasing basis set size. For simplicity in the following discussion, generally only the excited-state results using the largest basis set, 6-311+G(2d,2p), will be described in detail below.

Table 4. Low-Lying Singlet Excited States for Pt(AAA)Cl and for *cis*- and *trans*-Pt(AAA)Cl•CH_3CN Calculated with Time-Dependent Density Functional Theory Using the B3LYP Functional (Unless Noted) and Various Basis Sets[a]

	6-31+G*			6-311+G*			6-311+G(2d,2p)		
state	E^b (eV)	λ^c (nm)	$f^{\,d}$	$E^{\,b}$ (eV)	λ^c (nm)	$f^{\,d}$	E^b (eV)	λ^c (nm)	$f^{\,d}$
				Pt(AAA)Cl					
S_1	2.21	560	0.137	2.20	565	0.131	2.18	568	0.130
S_2	2.56	483	0.033	2.54	487	0.037	2.53	490	0.037
S_3	2.68	463	0.016	2.66	466	0.015	2.64	469	0.015
S_1^e	2.22	559	0.133						
S_2^e	2.57	482	0.038						
S_3^e	2.69	461	0.015						
				cis-Pt(AAA)Cl•CH_3CN					
S_1	2.07	600	0.017	2.06	601	0.017	2.05	604	0.017
S_2	2.85	435	0.047	2.84	437	0.047	2.82	440	0.047
S_3	2.91	426	0.006	2.91	426	0.005	2.93	423	0.004
S_1^e	2.08	595	0.017						
S_2^e	2.86	433	0.048						
S_3^e	2.92	424	0.004						
				trans Pt(AAA)Cl•CH_3CN					
S_1	2.49	498	0.000						
S_2	2.83	437	0.046						
S_3	3.03	409	0.000						

[a] The basis set used for the C, O, N, Cl, and H atoms is indicated in the column headings; a modified Hay-Wadt ECP was used on the Pt atom for all calculations. [b] Calculated transition energy. [c] Calculated transition wavelength. [d] Calculated oscillator strength. [e] Calculated using the B3PW91 functional.

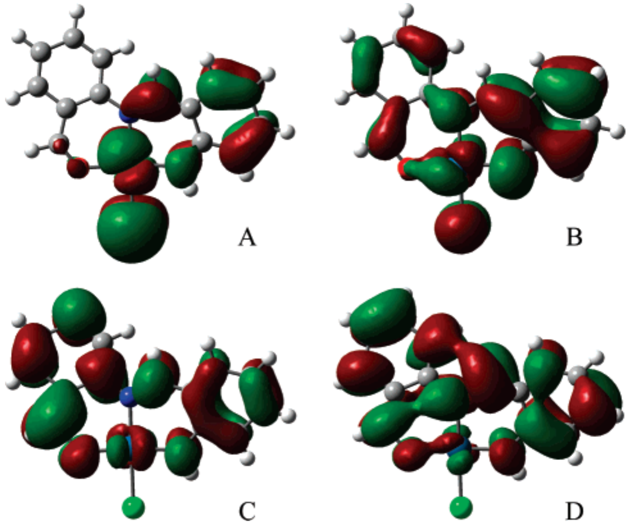


Figure 6. Contour plots of the (a) HOMO − 1, (b) HOMO, (c) LUMO, and (d) LUMO + 1 molecular orbitals of Pt(AAA)-Cl calculated using the B3LYP functional and the 6-311+G-(2d,2p) basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom.

For the Pt(AAA)Cl complex, the calculations predict the transition to the S_1 state to be the most intense of the first three excited states ($f = 0.130$), with a transition energy of 2.18 eV and a transition wavelength of 568 nm. The dominant orbital excitation for the S_1 state is the HOMO to LUMO transition, and both of these orbitals, calculated at the B3LYP/6-311+G(2d,2p) level, are displayed in Figure 6. As can be seen from Figure 6, both of these orbitals are primarily admixtures of Pt d orbitals and π orbitals on the AAA ligand. The most significant difference between them is that the HOMO orbital has significant Cl p orbital character, while the LUMO has zero Cl orbital character. Therefore, this transition can at least partially be classified as a Cl-to-Pt/AAA charge-transfer transition. The transitions to the S_2 and S_3 states are predicted to have energies of 2.53 and 2.64 eV, respectively, corresponding to wavelengths of 490 and 469 nm, respectively; however, their intensities are reduced by a factor of 3.5 ($f = 0.037$) and 8.7 ($f = 0.015$), respectively, as compared to the S_1 state. The dominant orbital excitations for the S_2 and S_3 states are the HOMO − 1 to LUMO and HOMO to LUMO + 1 transitions, respectively. The HOMO − 1 and LUMO + 1 orbitals, calculated at the B3LYP/6-311+G(2d,2p) level, are also displayed in Figure 6. The HOMO − 1 orbital is also primarily an admixture of a Pt d orbital, a Cl p orbital, and a π orbital on the AAA ligand; however, in this case the AAA ligand π orbital is almost exclusively centered on the o-aminobenzylidene (OAB) portion of the ring. Comparing the HOMO − 1 to the LUMO orbital, the S_2 transition can be classified as a combination of a Cl-to-Pt/AAA charge-transfer transition and an AAA intraligand charge-transfer transition (OAB ring to the benzaldehyde ring). Similar to the LUMO orbital, the LUMO + 1 orbital is an admixture of a Pt d orbital and a π orbital on the entire AAA ligand, with zero Cl orbital character. Comparing the HOMO to the LUMO + 1 orbital, the S_3 transition can be classified as a Cl-to-Pt/AAA charge-transfer transition.

Experimentally the visible absorption spectrum of the Pt(AAA)Cl complex has been reported by Mertes and co-workers in $CHCl_3$ and acetonitrile[6,7] and consists of a moderately intense, broad band centered at 578 ($\epsilon = 1.4 \times 10^4$ M^{-1} cm^{-1}) and 560 nm ($\epsilon = 1.0 \times 10^4$ M^{-1} cm^{-1}), respectively, which correspond to experimental transition energies of 2.14 and 2.21 eV, respectively. Comparing the experimental energies to the calculated energy for the S_1 state, it is found that the calculated energy of the S_1 state is in very good agreement with the two experimental values, with deviations of 0.04 and 0.03 eV, respectively.

Upon cis complexation of the CH_3CN, the calculations predict that the energy of the transition to the S_1 state is red-shifted compared to Pt(AAA)Cl, with a value of 2.05 eV (604 nm) and that its intensity is reduced by a factor of 7.6 ($f = 0.017$) as compared to Pt(AAA)Cl. The dominant orbital excitation for the S_1 state of *cis*-Pt(AAA)Cl•CH_3CN is the HOMO to LUMO transition, and both of these orbitals, calculated at the B3LYP/6-311+G(2d,2p) level, are displayed in Figure 7. As can be seen in Figure 7, the HOMO orbital is primarily an admixture of a Pt d orbital, a Cl p orbital, and a π orbital localized on the OAB portion of the AAA ligand. The LUMO orbital, on the other hand, is primarily a π orbital localized on the benzaldehyde portion of the AAA ligand, with small contributions from the Pt and Cl atoms. This transition can then be classified as a Pt/Cl/OAB-to-benzaldehyde charge-transfer type transition. The energies of the S_2 and S_3 states of the *cis*-Pt(AAA)Cl•CH_3CN complex are predicted to be blue-shifted with respect to those of Pt(AAA)Cl, with energies of 2.82 and 2.93 eV, respectively. The intensities of the transitions to the S_2 and S_3 states are predicted to increase by a factor of 1.3 ($f = 0.047$) and decrease by a factor of 3.8 ($f = 0.004$), respectively, relative to the intensities of these transitions for Pt(AAA)Cl. The dominant orbital excitations for the S_2 and S_3 states are the HOMO to LUMO + 1 and the HOMO to LUMO + 2 transitions, respectively, and the LUMO + 1 and LUMO + 2 orbitals calculated at the B3LYP/6-311+G(2d,2p) level are displayed in Figure 7. The LUMO + 1 orbital is primarily an admixture of a Pt d orbital and a π orbital that is delocalized over most of the AAA ligand and has zero Cl orbital character. Comparing the HOMO to the LUMO + 1 orbital, the S_2 transition can be classified as a combination of a Cl-to-Pt/AAA charge-transfer transition and an AAA intraligand charge-transfer transition (OAB ring to the benzaldehyde ring). The LUMO + 2 orbital is an admixture of a Pt d orbital, a Cl p orbital, and a fairly delocalized AAA π orbital. Comparing the HOMO to the LUMO + 2 orbital, the S_3 transition can be classified as a Cl/Pt/OAB-to-benzaldehyde charge-transfer transition.

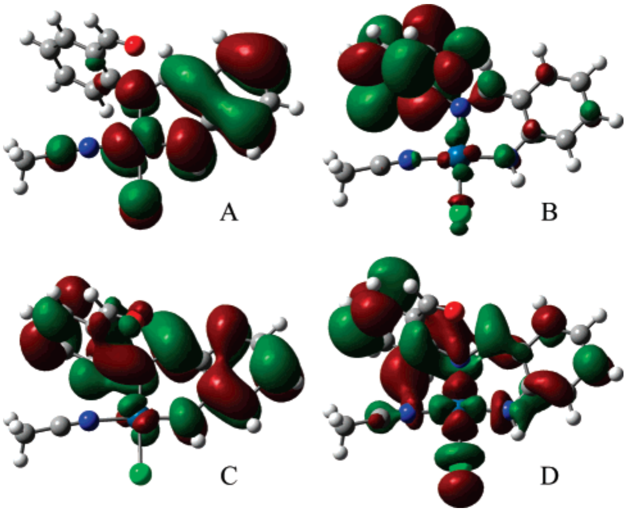


Figure 7. Contour plots of the (a) HOMO, (b) LUMO, (c) LUMO + 1, and (d) LUMO + 2 molecular orbitals of *cis*-Pt(AAA)Cl•CH_3CN calculated using the B3LYP functional and the 6-311+G(2d,2p) basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom.

Experimentally the absorption spectrum of the Pt(AAA)Cl•CH_3CN complex in the visible region has been reported by Mertes and co-workers in acetonitrile[7] and is comprised of a low intensity, broad band centered at 470 nm ($\epsilon = 4.0 \times 10^3$ M^{-1} cm^{-1}), which corresponds to an energy of 2.64 eV. Upon initial comparison of the energy of the experimentally reported λ_{max} and the energy of the calculated S_1 transition, it appears as if there is a large discrepancy of 0.59 eV. Given the good agreement between the calculated and experimental transition of the Pt(AAA)Cl complex, however, this was very surprising and seemed unlikely. Upon closer inspection of the published spectrum assigned to the Pt(AAA)Cl•CH_3CN complex by Mertes and co-workers, it is observed that the most intense feature in the visible region is indeed the band centered at 470 nm; however, there still remains a broad, low intensity ($\epsilon \approx 1.0 \times 10^3$ M^{-1} cm^{-1}) band between 550 and 600 nm in the spectrum. Since the calculated relative intensities of the S_1 and S_2 transitions for the *cis*-Pt(AAA)Cl•CH_3CN complex are similar to the experimental molar absorptivities of the 470 nm band and the region between 550 and 600 nm, this suggests that the energy of the 470 nm band (2.64 eV) should be quantitatively

Table 5. Zero-Point Corrected Total Energies, Enthalpies, Entropies, and Free Energies at 298 K for CH_3CN, Pt(AAA)Cl, *cis*-Pt(AAA)Cl•CH_3CN, and the *cis*-Pt(AAA)Cl•CH_3CN Transition State Calculated at the B3LYP/6-311+G(2d,2p) Level[a]

species	E^{298} (hartrees)	H^{298} (hartrees)	S^{298} (J/K•mol)	G^{298} (hartrees)
CH_3CN	−132.751806	−132.750862	242.6	−132.778421
Pt(AAA)Cl	−1304.292137	−1304.291193	539.8	−1304.352518
cis-Pt(AAA)Cl•CH_3CN	−1437.046106	−1437.045161	684.7	−1437.122948
Pt(AAA)Cl•CH_3CN TS[b]	−1437.014532	−1437.013588	660.0	−1437.088560

[a] Calculations performed using the 6-311+G(2d,2p) basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom. [b] *cis*-Pt(AAA)Cl•CH_3CN transition state.

compared to the calculated energy of the S_2 transition (2.82 eV). With this comparison, the agreement now seems reasonable, with a deviation of 0.18 eV. Another factor that supports this assignment is the good agreement of the relative molar absorptivities of the 560 nm band of Pt(AAA)Cl ($\epsilon = 1.0 \times 10^4$ M^{-1} cm^{-1}) to the 470 nm band of Pt(AAA)Cl•CH_3CN ($\epsilon = 4.0 \times 10^3$ M^{-1} cm^{-1}) with the relative calculated oscillator strengths of the S_1 transition of Pt(AAA)Cl ($f = 0.130$) to the S_2 transition of *cis*-Pt(AAA)Cl•CH_3CN ($f = 0.047$).

In order to further assess the validity of the calculated transitions for the *cis*-Pt(AAA)Cl•CH_3CN complex and the assignments made above, several other TDDFT excited-state calculations were performed. These additional TDDFT calculations were Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN at the B3PW91/6-31+G* level and *trans*-Pt(AAA)Cl•CH_3CN at the B3LYP/6-31+G* level, the results of which are given in Table 4. Comparing the results for Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN at the B3PW91/6-31+G* level to those at the B3LYP/6-31+G* level, it is clear that the B3PW91 functional predicts nearly the same transition energies and oscillator strengths for the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN complexes as the B3LYP calculations, supporting the validity of the B3LYP calculations. With the *trans*-Pt(AAA)Cl•CH_3CN complex at the B3LYP/6-31+G* level, the calculations predict the S_1 transition to be at a higher energy than for the *cis*-Pt(AAA)Cl•CH_3CN complex; however, the oscillator strength is predicted to be zero. The calculated energy and oscillator strength for the S_2 transition of the *trans*-Pt(AAA)Cl•CH_3CN complex, in contrast, are very close to the values for the S_2 transition of the *cis*-Pt(AAA)Cl•CH_3CN complex. Similar to the S_1 transition, the energy of the S_3 transition is predicted to be higher for the *trans*-Pt(AAA)Cl•CH_3CN complex than it is for the *cis*-Pt(AAA)Cl•CH_3CN complex, and the oscillator strength is calculated to be zero for the trans complex. Overall, the additional calculations support the hypothesis that once the CH_3CN is bound to the platinum, the S_2 transition becomes the most intense transition in the visible region and that it is blue-shifted and its intensity is decreased when compared to the most intense, S_1 transition for Pt(AAA)Cl.

Dark Reaction Thermochemistry and Activation Parameters. Single point energies were calculated at the B3LYP/6-311+G(2d,2p) level using the B3LYP/6-31G* optimized geometries for CH_3CN, the Pt(AAA)Cl complex, the *cis*-Pt(AAA)Cl•CH_3CN complex, and the *cis*-Pt(AAA)Cl•CH_3CN transition state. Using the scaled (0.9804[22]) B3LYP/6-31G* vibrational frequencies, zero-point and 298 K thermal corrections were determined, allowing for the calculation of 0 K energies and 298 K enthalpies, entropies, and free energies, which are listed in Table 5.

Utilizing the parameters for CH_3CN, Pt(AAA)Cl, and *cis*-Pt(AAA)Cl•CH_3CN in Table 5, the thermochemistry of the dark reaction, Pt(AAA)Cl + CH_3CN → Pt(AAA)Cl•CH_3CN, can be calculated. The 298 K reaction energy and reaction enthalpy are both calculated to be exothermic, with values of −5.7 and −8.2 kJ/mol, respectively, indicating that the dark reaction is slightly energetically favorable. The 298 K entropy change for the dark reaction is found to be negative, which is consistent with two molecules going to one molecule, and has a calculated value of −97.7 J/K•mol using the absolute entropies in Table 5. Somewhat surprisingly, since the reaction is experimentally known to occur spontaneously in solution,[7] the 298 K free energy of the dark reaction is predicted to be endergonic in the gas phase, with a value of +21.0 kJ/mol. The primary reason for the large positive calculated ΔG^{298} is due to the large negative entropy, which causes the magnitude of $T\Delta S^{298}$ to be larger than the ΔH^{298}. Experimentally, this reaction occurs in acetonitrile as the solvent, and, in order to calculate an approximate free energy for the reaction in the condensed phase, the liquid-phase absolute entropy of acetonitrile (149.6 J/K•mol[29]) can be used in the calculation of $\Delta S^{298.}$ Using the liquid-phase entropy for acetonitrile, the condensed-phase values of ΔS^{298} and ΔG^{298} are calculated to be −4.7 J/K•mol and −6.7 kJ/mol, respectively, and the reaction is now predicted to be spontaneous, as it is experimentally known to be. This seems to be an example of a reaction in which solvent/condensed-phase considerations have a large effect on the spontaneity of the reaction. Unfortunately, no experimental thermochemical parameters have been reported for the dark reaction of this complex, and so a comparison to experimental values is not possible. However, the thermochemistry of ligand substitution reactions for several square planar Pt(II) complexes has been studied both experimentally[30] and theoretically,[31] and the current values are found to be within the range of values reported previously.

Since single point and frequency calculations been performed on the *cis*-Pt(AAA)Cl•CH_3CN transition state, the activation parameters of the dark reaction can also be calculated using the values for CH_3CN, Pt(AAA)Cl, and the *cis*-Pt(AAA)Cl•CH_3CN transition state (Table 5). The 298 K activation energy is calculated to be 77.2 kJ/mol, and the 298 K activation enthalpy is calculated to be 74.7 kJ/mol. The gas-phase activation entropy and free energy are calculated to be −122.5 J/k•mol and 111.2 kJ/mol, respectively. However, in light of the discussion above concerning the condensed-phase reaction entropy and free energy, condensed-phase values are also calculated using the liquid-phase absolute entropy of acetonitrile and are found to be −29.5 J/k•mol and 83.5 kJ/mol for the condensed-phase activation entropy and the activation free energy, respectively. Unfortunately, experimental values of the activation energy and the activation enthalpy for the dark reaction of Pt(AAA)Cl have not been reported in the literature, and so a direct comparison to the calculated values is not possible. However, activation enthalpies have been measured for nucleophilic substitution reactions of several platinum(II) complexes, and these are found to vary from approximately 30−100 kJ/mol.[24] In light of these values, the calculated activation enthalpy for the dark reaction seems reasonable.

Energetics and Mechanism of the Photoreaction. In addition to the energetics of the dark reaction, it also of interest to attempt to characterize the energetics and mechanism of the photoreaction. In general, there are two typical mechanisms for photochemical or photoinduced reactions: adiabatic and diabatic reactions. In an adiabatic photoreaction, the reactants are excited and then traverse along the excited-state surface over an excited-state activation barrier to the excited state of the products. The ground-state products are then reached via radiative or nonradiative decay. In the diabatic reaction, the reactants are excited and traverse along the excited-state surface to a point at which the excited-state and ground-state surfaces closely approach one another, usually close to where the ground state is a maximum, i.e., the ground transition state.[32] This is then followed by a surface crossing from the excited-state surface to the ground-state surface and finally ground-state relaxation down to the products or back to the ground state of the reactants, in which case no overall reaction has occurred. Because the present reaction involves platinum(II) complexes and it has been long known that platinum(II) complexes often display very strong phosphorescence emission even at room temperature[33−38] and therefore have very large singlet to triplet intersystem crossing rates, the possibility that the photoreaction occurs via the lowest triplet-state surface needs to be also considered.

In order to determine along which surface the reaction proceeds and which of the two mechanisms is most probable, the location of the Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN minima and the *cis*-Pt(AAA)Cl•CH_3CN transition-state maxima along the S_1 and T_1 potential energy surfaces would be needed, which would require S_1 and T_1 geometry optimizations. Unfortunately, geometry optimizations coupled with TDDFT calculations are not currently available in the Gaussian suite of programs. However, using calculated S_1 and T_1 vertical excitations for the ground-state optimized Pt(AAA)Cl, *cis*-Pt(AAA)Cl•CH_3CN, and *cis*-Pt(AAA)Cl•CH_3CN transition state, some indications of the energetics and the most likely mechanism for the photoreaction can be inferred.

For consistency in comparison with the highest level ground-state energetics, only vertical excitation energies calculated at the B3LYP/6-311+G(2d,2p) level will be used in this discussion. For Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN, the S_1 excitation energies are taken from Table 4 and converted to kJ/mol, giving 210 and 198 kJ/mol, respectively. The S_1 vertical excitation energy for the *cis*-Pt(AAA)Cl•CH_3CN transition state was calculated to be 1.80 eV or 173 kJ/mol at the B3LYP/6-311+G(2d,2p) level. The T_1 vertical excitation energies for Pt(AAA)Cl, *cis*-Pt(AAA)Cl•CH_3CN, and the *cis*-Pt(AAA)Cl•CH_3CN transition state were calculated to be 1.42, 1.84, and 1.37 eV, respectively, at the B3LYP/6-311+G(2d,2p) level, which correspond to 137, 178, and 132 kJ/mol, respectively. By combining the S_1 and T_1 vertical excitation energies with the ground-state (S_0) energies of these species relative to the Pt(AAA)Cl and CH_3CN reactants, the relative energetics of these species along the S_1 and T_1 potential energy surfaces can be obtained, and these are represented in Figure 8.

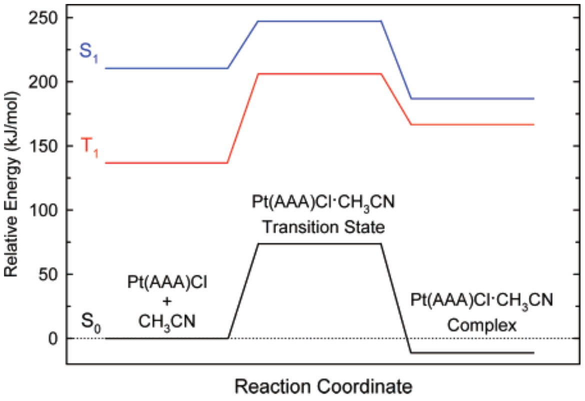


Figure 8. Relative energy diagram for the S_0, S_1, and T_1 states of the Pt(AAA)Cl + CH_3CN reactants, the *cis*-Pt(AAA)-Cl•CH_3CN transition state, and the *cis*-Pt(AAA)Cl•CH_3CN complex calculated using the B3LYP functional and the 6-311+G(2d,2p) basis set on the C, O, N, Cl, and H atoms and a modified Hay-Wadt ECP on the Pt atom.

From Figure 8, it can be seen that the reaction from *cis*-Pt(AAA)Cl•CH_3CN to Pt(AAA)Cl along the S_1 surface is predicted to be endothermic by 24 kJ/mol and to have an energy barrier of 62 kJ/mol. In contrast to the S_1 surface, the same reaction along the T_1 surface is predicted to be exothermic by 29 kJ/mol and to have an energy barrier of 42 kJ/mol. The S_1−T_1 energy gap for *cis*-Pt(AAA)Cl•CH_3CN is predicted to be 20.1 kJ/mol, which seems small enough that S_1 to T_1 intersystem crossing would probably occur with an appreciable rate. Therefore, based on these relative energies, it seems likely that upon absorption of a photon to reach the S_1 state or absorption to S_2 followed by fast internal conversion to S_1, rapid intersystem crossing to T_1 occurs, and the photoreaction then proceeds along the T_1 surface. The last question to address is whether the photoreaction follows an adiabatic mechanism along the T_1 surface or a diabatic mechanism involving a crossing from the T_1 surface to a maximum along the S_0 surface. Because the T_1 excitations for each of the molecules along the reaction involve fairly delocalized HOMO and LUMO orbitals (see Figures 6 and 7 for example) and are therefore not expected to cause very large changes in the bonding upon excitation, it seems reasonable to assume that the geometries of the minima and maxima along the T_1 surface would be similar to those along S_0. This then implies that a diabatic mechanism involving a surface crossing from a low-energy point along T_1 to a high-energy point along S_0 is unlikely and that the photoreaction proceeds adiabatically along T_1 from *cis*-Pt-

(AAA)Cl•CH_3CN over the T_1 energy barrier to the T_1 state of Pt(AAA)Cl. The final step in the photoreaction, which is the relaxation from the T_1 state to the S_0 state of Pt(AAA)-Cl, most likely occurs via nonradiative intersystem crossing, given that Mertes and co-workers[7] did not report observing the re-emission of light for the photoreaction of Pt(AAA)-Cl•CH_3CN back to Pt(AAA)Cl.

Summary and Conclusion

Density functional theory methods have been used to theoretically characterize a tridentate photochromic Pt(II) complex [Pt(AAA)Cl], its acetonitrile solvolyis product [Pt-(AAA)Cl•CH_3CN], and the transition state in the solvolysis reaction. The geometries were optimized at the B3LYP/6-31G* level of theory. The optimized geometry of Pt(AAA)-Cl was found to be in good agreement with the reported crystal structure.[6] The optimized geometry of *cis*-Pt(AAA)-Cl•CH_3CN was also found to be in good agreement with most of the applicable geometrical parameters for a crystal structure reported for a related complex with triphenylphosphine as the ligand, *trans*-Pt(AAA)Cl•PPh_3,[10] the exception being the C(7)−N(2)−C(8)−C(13) dihedral angle. Additional optimizations were performed for *trans*-Pt(AAA)-Cl•CH_3CN and for *cis*- and *trans*-Pt(AAA)Cl•PH_3. As a result, it was found that the deviation in the dihedral angle between the calculated *cis*-Pt(AAA)Cl•CH_3CN value and the experimental *trans*-Pt(AAA)Cl•PPh_3 value was primarily due to a steric cis versus trans effect. Vibrational frequencies were calculated for the optimized Pt(AAA)Cl and *cis*-Pt-(AAA)Cl•CH_3CN complexes at several levels of theory, and it was found that the predicted shift in the benzaldehyde carbonyl frequency for Pt(AAA)Cl to *cis*-Pt(AAA)Cl•CH_3-CN was in the same direction and close to that observed experimentally,[6,7] supporting the experimental assignments. Singlet vertical excitation energies were calculated for the B3LYP/6-31G* optimized Pt(AAA)Cl and *cis*-Pt(AAA)Cl•CH_3CN geometries using time-dependent density-functional theory (TDDFT). The most intense transition for Pt(AAA)-Cl was predicted to be to the S_1 state, and its energy was found to be in good agreement with the experimental value. The most intense transition for *cis*-Pt(AAA)Cl•CH_3CN, however, was predicted to be to the S_2 state rather than to the S_1 state, and the energy of this transition was found to be in reasonable agreement with the experimental value. Overall, the excited-state calculations support the experimental observation of a blue-shift and a decrease in intensity on going from Pt(AAA)Cl to *cis*-Pt(AAA)Cl•CH_3CN. Single point energies were calculated at the B3LYP/6-311+G(2d,-2p) level using the B3LYP/6-31G* optimized geometries for CH_3CN, the Pt(AAA)Cl complex, the *cis*-Pt(AAA)Cl•CH_3-CN complex, and the *cis*-Pt(AAA)Cl•CH_3CN transition state. The calculations predict the dark reaction to be slightly exothermic at 298 K and, after a correction to the entropy, to also be spontaneous at 298 K, and to proceed with a reasonable activation energy. For the photoreaction, approximate excited-state energies were obtained using the vertical S_1 and T_1 energies for *cis*-Pt(AAA)Cl•CH_3CN, the *cis*-Pt(AAA)Cl•CH_3CN transition state, and Pt(AAA)Cl, and based on these energies relative to the ground-state energies, it was speculated that the photoreaction occurs via an intersystem crossing from S_1 to T_1 for *cis*-Pt(AAA)Cl•CH_3-CN followed by an adiabatic reaction along the T_1 surface to the T_1 state of Pt(AAA)Cl and then nonradiative intersystem crossing to the S_0 state of Pt(AAA)Cl.

Acknowledgment. This work was supported by startup funds from Penn State Erie, The Behrend College. I thank Alan Jircitano and Tom Spudich for providing the initial motivation to engage in these calculations and for enlightening discussions during the course of this work. I additionally thank Alan Jircitano for supplying the initial ChemDraw structure that was used for Figure 1. I thank Peter B. Armentrout for providing a Gaussian input file containing the parameters of the modified Hay-Wadt ECP/valence that was used for the Pt in the present work.

Supporting Information Available: Gaussian input with route line, title, Cartesian coordinates, and Pt ECP/valence basis set parameters for all of the optimized geometries obtained in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) VerHoeven, J. W. *Pure Appl. Chem.* **1996**, *68*, 2223−2286.

(2) See the following general photochromism references. (a) *Organic Photochromic and Thermochromic Compounds*; Crano, J. C., Guglielmetti, R. J., Eds.; Plenum: New York, 1999; Vols. 1 and 2. (b) *Photochromism: Molecules and Systems*; Durr, H., Bouas-Laurent, H., Eds.; Elsevier: New York 1990; pp 1−1068. (c) See the following issue of *Chem. Rev.*: *Chem. Rev.* **2000**, *100* (5), 1683−1890. (d) See sections I and II of Exelby, R.; Grinter, R. *Chem. Rev.* **1965**, *65*, 247−260.

(3) See the following general reviews of photochromic inorganic compounds. (a) Faughan, B. W.; Staebler, D. L.; Zoltan, K. J. *Appl. Solid State Phys.* **1971**, *2*, 107−172. (b) Zoltan, K. J. *Physics Today* **1970**, *23*, 42−49. (c) Cohen, S. D.; Newman, G. A. *J. Photogr. Sci.* **1967**, *15*, 290−298. (d) See section III of Exelby, R.; Grinter, R. *Chem. Rev.* **1965**, *65*, 247−260.

(4) Particular sections of the following reviews of photochemistry of transition-metal compounds deal with photochromic transition-metal compounds. (a) Adamson, A. W. *Pure Appl. Chem.* **1969**, *20*, 25−52. (b) Adamson, A. W.; Waltz, W. L.; Zinato, E.; Watts, D. W.; Fleischauer, P. D.; D., L. R. *Chem. Rev.* **1968**, *68*, 541−585.

(5) See the following articles and references therein for recent examples of photochromic transition-metal compounds. (a) Miyamoto, Y.; Kikuchi, A.; Iwahori, F.; Abe, J. *J. Phys. Chem. A* **2005**, *109*, 10183−10188. (b) Matsuda, K.; Takayama, K.; Irie, M. *Inorg. Chem.* **2004**, *43*, 482−489. (c) Nishimura, H.; Matsushita, N. *Chem. Lett.* **2002**, *9*, 930−931. (d) Wakamatsu, K.; Nishimoto, K.; Shibahara, T. *Inorg. Chim. Acta* **1999**, *295*, 180−188.

(6) Timken, M. D.; Sheldon, R. I.; Rohly, W. G.; Mertes, K. B. *J. Am. Chem. Soc.* **1980**, *102*, 4716−4720.

(7) Rohly, W. G.; Mertes, K. B. *J. Am. Chem. Soc.* **1980**, *102*, 7939−7942.

(8) For examples, see the following and references therein. (a) Lutterman, D. A.; Fu, P. K.-L.; Turro, C. *J. Am. Chem. Soc.* **2006**, *128*, 738−739. (b) Martinez, M. S.; de Oliveira, E.

C.; Tfouni, E. *J. Photochem. Photobiol., A* **1999**, *122*, 103−108. (c) Kirk, A. D. *Comments Inorg. Chem.* **1993**, *14*, 89−121. (d) Moensted, L.; Moensted, O. *Acta Chem. Scand.* **1993**, *47*, 9−17. (e) Carlos, R. M.; Frink, M. E.; Tfouni, E.; Ford, P. C. *Inorg. Chim. Acta* **1992**, *193*, 159−165. (f) Pavanin, L. A.; Novais, da Rocha, Z.; Giesbrecht, E.; Tfouni, E. *Inorg. Chem.* **1991**, *30*, 2185−2190.

(9) Jircitano, A. J. Penn State Erie, The Behrend College, Erie, PA. Personal communication 2006.

(10) Jircitano, A. J.; Rohly, W. G.; Mertes, K. B. *J. Am. Chem. Soc.* **1981**, *103*, 4879−4883.

(11) Hehre, W. J.; Radom, L.; Schleyer, P. V. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley & Sons: New York, 1986; pp 1−548.

(12) Ziegler, T. *Chem. Rev.* **1991**, *91*, 651−667.

(13) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley-VCH: Weinheim, 2001; pp 1−300.

(14) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(15) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(16) Jircitano, A. J.; Spudich, T. M.; Ulrich, L.; Saxton, N. L. Penn State Erie, The Behrend College, Erie, PA. Personal communication 2004.

(17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kundin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A.11*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(18) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Lyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(19) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299−310.

(20) Ohanessian, G.; Brusich, M. J.; Goddard, W. A., III *J. Am. Chem. Soc.* **1990**, *112*, 7179−7189.

(21) Dennington, R. II.; Keith, T.; Millam, J.; Eppinnett, K.; Hovell, W. L.; Gilliland, R. *GaussView, Version 3.09*; Semichem, Inc.: Shawnee Mission, KS, 2003.

(22) Wong, M. W. *Chem. Phys. Lett.* **1996**, *256*, 391−399.

(23) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218−8224.

(24) Basolo, F.; Pearson, R. G. *Mechanisms of Inorganic Reactions: A Study of Metal Complexes in Solution*; Wiley & Sons: New York, 1967; pp 351−453.

(25) Wilkins, R. G. *The Study of Kinetics and Mechanism of Reactions of Transition Metal Complexes*; Allyn & Bacon: Boston, MA, 1974; pp 223−225.

(26) Nakamoto, K. *Infrared and Raman Spectra of Inorganic and Coordination Compounds. Part B: Applications in Coordination, Organometallic, and Bioinorganic Chemistry*, 5th ed.; Wiley & Sons: New York, 1997; pp 58−59.

(27) Pople, J. A.; Schlegel, H. B.; Raghavachari, K.; DeFrees, D. J.; Binkley, J. F.; Frisch, M. J.; Whitesides, R. F.; Hout, R. F.; Hehre, W. J. *Int. J. Quantum Chem Symp.* **1981**, *15*, 269−278.

(28) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502−16513.

(29) *CRC Handbook of Chemistry and Physics*, 70th ed.; Weast, R. C., Lide, D. R., Astle, M. J., Beyer, W. H., Eds.; CRC Press: Boca Raton, FL, 1989; p D-61.

(30) Coe, J. S. *MTP Int. Rev. Sci.: Inorg. Chem., Ser. 2* **1974**, 45−62.

(31) Cooper, J.; Ziegler, T. *Inorg. Chem.* **2002**, *41*, 6614−6622.

(32) Turro, N. J. *Modern Molecular Photochemistry*; University Science Books: Sausalito, CA, 1991; pp 72−74.

(33) Yersin, H.; Humbs, W.; Strasser, J. *Coord. Chem. Rev.* **1997**, *159*, 325−358.

(34) Balashev, K. P.; Puzyk, M. V.; Kotlyar, V. S.; Kulikova, M. V. *Coord. Chem. Rev.* **1997**, *159*, 109−120.

(35) Yersin, H.; Strasser, J. *Coord. Chem. Rev.* **2000**, *208*, 331−364.

(36) Hissler, M.; McGarrah, J. E.; Connick, W. B.; Geiger, D. K.; Cummings, S. D.; Eisenberg, R. *Coord. Chem. Rev.* **2000**, *208*, 115−137.

(37) Yersin, H.; Donges, D. *Top. Curr. Chem.* **2001**, *214*, 81−186.

(38) Castellano, F. N.; Pomestchenko, I. E.; Shikhova, E.; Hua, F.; Muro, M. L.; Rjapkse, N. *Coord. Chem. Rev.* **2006**, *250*, 1819−1828.

CT700158G

JCTC Journal of Chemical Theory and Computation

Influence of the Side Chain in the Structure and Fragmentation of Amino Acids Radical Cations

Adrià Gil,[†] Sílvia Simon,*[,‡] Luis Rodríguez-Santiago,[†] Juan Bertrán,[†] and Mariona Sodupe*[,†]

Departament de Química, Universitat Autonoma de Barcelona, Bellaterra 08193, Spain, and Institut de Química Computacional, Departament de Química, Universitat de Girona, Girona 17071, Spain

Received March 7, 2007

Abstract: The conformational properties of ionized amino acids (Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His) have been theoretically analyzed using the hybrid B3LYP and the hybrid-meta MPWB1K functionals as well as with the post-Hartree Fock CCSD(T) level of theory. As a general trend, ionization is mainly localized at the $-NH_2$ group, which becomes more planar and acidic, the intramolecular hydrogen bond in which $-NH_2$ acts as proton donor being strengthened upon ionization. For this reason, the so-called conformer IV(+) becomes the most stable for nonaromatic amino acid radical cations. Aromatic amino acids do not follow this trend because ionization takes place mainly at the side chain. For these amino acids for which ionization of the side chain prevails over the $-NH_2$ group, structures III(+) and II(+) become competitive. The $C_\alpha-X$ fragmentations of the ionized systems have also been studied. Among the different decompositions considered, the one that leads to the loss of $COOH^\bullet$ is the most favorable one. Nevertheless, for aromatic amino acids fragmentations leading to $R^\bullet$ or R^+ start being competitive. In fact, for His and Tyr, results indicate that the fragmentation leading to R^+ is the most favorable process.

Introduction

Protein, peptide, and amino acid radicals may play an important role in several biological processes. One of them is the oxidative damage of proteins, which is related to pathological disorders[1,2] and subsequent development of diseases such as Alzheimer[3−9] or glaucoma.[10] Since this effect is mainly due to reactions that take place in amino acids, the knowledge of their structure and reactivity upon ionization is of great importance. Moreover, their study is also important to understand the role of transient species involved in protein radical catalysis.[11] On the other hand, gas-phase studies have shown that radical cations of some oligopeptides can be produced by collision induced dissociation of $[Cu^{II}(dien)M]^{\bullet 2+}$ complex ions.[12] Their dissociation behavior is very rich and differs considerably from that of protonated peptides, which make them very attractive for peptide sequentiation. Because of that, in the past few years, the properties of different amino acid and derived radicals have attracted considerable attention, both from an experimental and theoretical point of view.[12−50]

Amino acids usually present intramolecular hydrogen bonds which are crucial to understand their structure and reactivity. However, these hydrogen bonds can be largely modified upon ionization. Previous studies have shown that removing an electron from such a system modifies both the acidity and the basicity of the groups involved in the hydrogen bond, in such a way that it is difficult to establish how this interaction would be affected by oxidation.[27,51−54] For glycine the observed changes in intramolecular hydrogen bonds have been related to the nature of the electron hole in different electronic states.[27] On the other hand, oxidized species can also lead to intermolecular spontaneous proton-

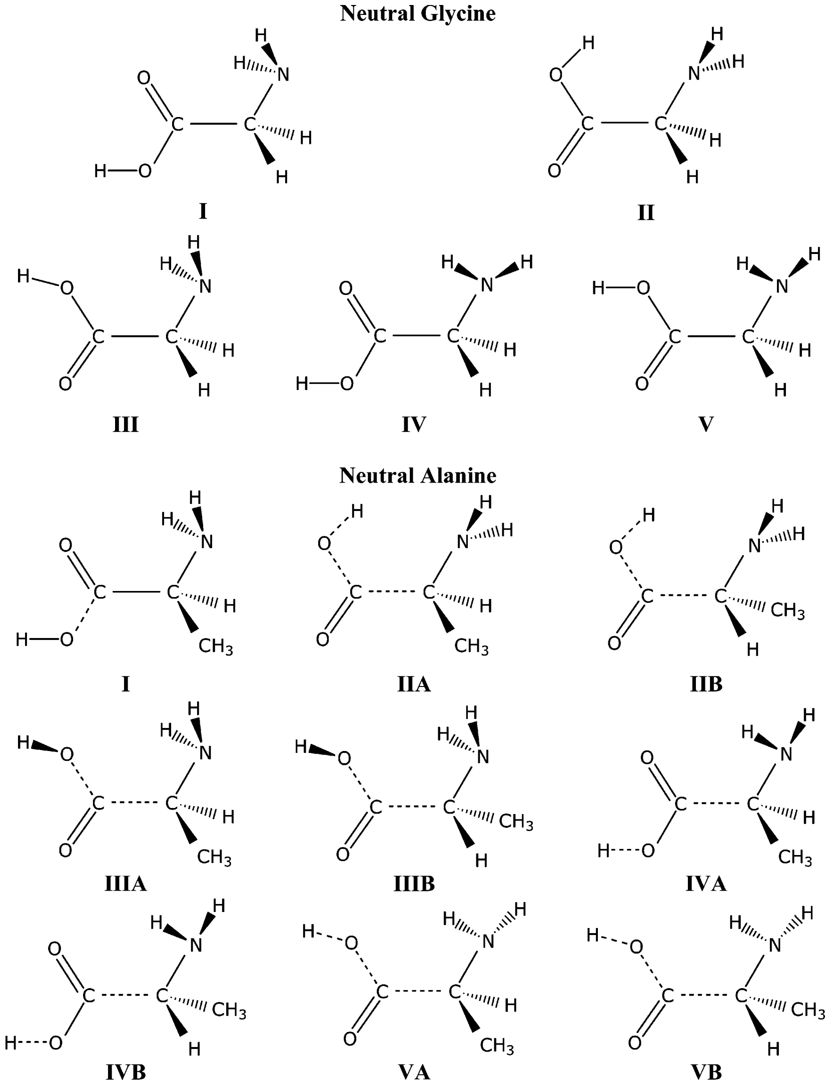
* Corresponding author e-mail: Mariona.Sodupe@uab.es (M.S.), silvia.simon@udg.edu (S.S.).

[†] Universitat Autonoma de Barcelona.

[‡] Universitat de Girona.

10.1021/ct700055p CCC: $37.00
Published on Web 10/17/2007

Scheme 1



transfer processes in solution. Rega et al.[26] have observed that the main product after glycine ionization in solution is the glycyl radical $[NH_2CHCOOH]^{\bullet}$, even at low pH, due to the large acidity of the $-CH_2$ group[55,56] in ionized species.

Glycine is the simplest amino acid and consequently an important model compound, which has been the subject of many experimental and theoretical investigations.[13−16,19,20,26−30,38,50] However, most of the studies have focused their attention on the structure and magnetic properties of the C-centered glycyl $[NH_2CHCOOH]^{\bullet}$ radical, one of the radiation products of glycine in solution. Glycyl radical has also been generated in the gas phase[57,58] by collisional neutralization of the stable glycyl cation $[NH_2CHCOOH]^+$, which is obtained by dissociative ionization of several amino acids such as phenylalanine or serine. Unimolecular decompositions are then studied by reionization mass spectrometry experiments. Moreover, photoion mass spectrometry studies of different amino acids in the 6−22 eV photon energy region have provided new information about their dissociative ionization products.[17,18] It has been shown that for the glycine radical cation, the most intense peak is due to the aminomethyl cation, $NH_2CH_2^+$, in complete agreement with a previous study,[28] where the loss of the COOH radical was calculated to be the lowest-energy ion fragmentation. This result was confirmed later on by Lu et al.[30]

Fewer conformational studies have been performed for the other amino acids[25,31−37,39−49] due to their higher conformational complexity. Their study, however, is interesting, because it introduces the influence of the side chain on the stability of the conformations as well as on the preference for any possible fragmentations upon ionization. This work reports an exhaustive gas-phase conformational study for 9 amino acids belonging to different groups (nonpolar, polar, acidic, basic, or aromatic) in their ionized forms. The studied amino acids are as follows: glycine, alanine, serine, cysteine, aspartic acid, glutamine, phenylalanine, tyrosine, and histidine. We expect that this exhaustive conformational study as well as the unimolecular decomposition analysis will help to explain the role of the side chain in oxidative processes of amino acids and to interpret mass spectrometry experiments.

Methods

It is well-known that amino acids can exist in a large number of conformations due to many single-bond rotamers. Given the conformational complexity introduced by the side chain,

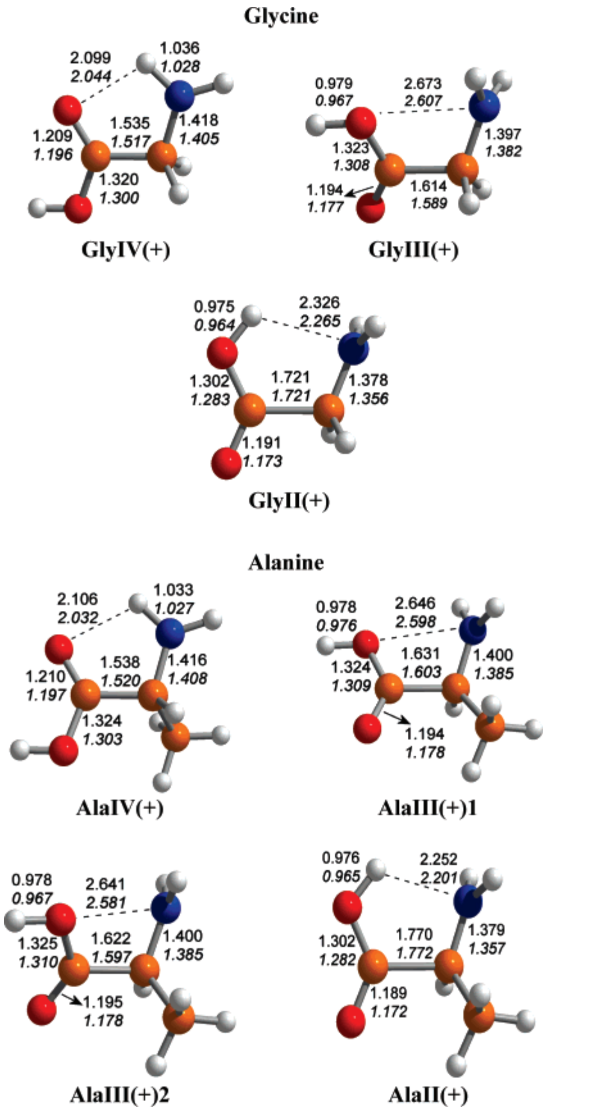


Figure 1. Optimized geometries for the lowest-energy conformer of Gly and Ala radical cations, at the B3LYP and MPWB1K/6-31++G(d,p) levels of theory. Distances are in angstroms.

the following strategy has been applied to find the lowest-energy conformations of each amino acid. First, we have performed a Monte Carlo Multiple Minimum (MCMM) conformational search[59,60] with the MMFF94s force field.[61,62] All plausible structures within an energy window of 50 kJ mol^{-1} were selected for subsequent quantum chemical optimizations of the neutral systems. Radical cation structures were then obtained by ionizing and reoptimizing each neutral conformation.

Optimized geometries and harmonic vibrational frequencies have been obtained using the hybrid B3LYP[63−65] and hybrid-meta MPWB1K[66] functionals with the 6-31++G(d,p) basis set. Geometry optimizations at the MP2/6-31++G(d,p) level of theory have been performed as well, and for all systems except tyrosine, which is very similar to phenylalanine, single-point calculations at the CCSD(T)/6-31++G(d,p) level have also been carried out. All valence electrons were correlated at the MP2 and CCSD(T) levels of theory. Mean average deviations of the used functionals as well as MP2 with respect to CCSD(T) with the 6-31++G(d,p) show that for these radical cation species MPWB1K tends to give results in much better agreement with CCSD(T) than B3LYP or MP2, the MPWB1K, B3LYP, and MP2 average deviations being 0.6, 1.6, and 2.1 kcal mol^{-1}, respectively (see the Supporting Information). On the other hand, the effect of further enlarging the basis set has been analyzed for glycine and alanine by performing calculations with the augmented aug-cc-pVDZ and aug-cc-pVTZ basis sets.[67]

Net atomic charges and spin densities have been obtained using the natural population analysis of Weinhold et al.[68] All DFT calculations and post Hartree−Fock MP2 and CCSD(T) with the small 6-31++G(d,p) basis set have been performed with the Gaussian 03 package,[69] and open-shell systems have been treated with an unrestricted formalism. CCSD(T) with the aug-cc-pVXZ (X = D and T) sets have been performed with the MOLPRO program and were based on a restricted Hartree−Fock reference wave function.[70] A Monte Carlo Multiple Minimum (MCMM) conformational search has been performed with the Macromodel 7.0 package.[71]

Results and Discussion

Removing an electron from neutral amino acids induces significant structural changes that vary depending on the starting conformation. In order to understand the influence of ionization in all amino acids, we will first analyze in detail the structural features of the radical cations of the two simplest amino acids: glycine (Gly) and alanine (Ala).

Table 1. Relative Energies (Δ*E*) in kcal mol^{-1} at Different Levels of Theory

			CCSD(T)// MPWB1K/6-31++G(d,p)		
structure	B3LYP 6-31++G(d,p)	MPWB1K 6-31++G(d,p)	6-31++G(d,p)	aug-pVDZ	aug-pVTZ
			Glycine		
GlyIV(+)	0.0	0.0	0.0	0.0	0.0
GlyIII(+)	−3.0	1.3	1.1	1.9	1.6
GlyII(+)	6.2	11.7	12.1	11.0	10.1
			Alanine		
AlaIV(+)	0.0	0.0	0.0	0.0	0.0
AlaIII(+)1	−3.1	0.4	0.9	1.0	0.8
AlaIII(+)2	−2.5	1.5	1.4	1.5	1.4
AlaII(+)	5.6	10.4	11.5	-	-

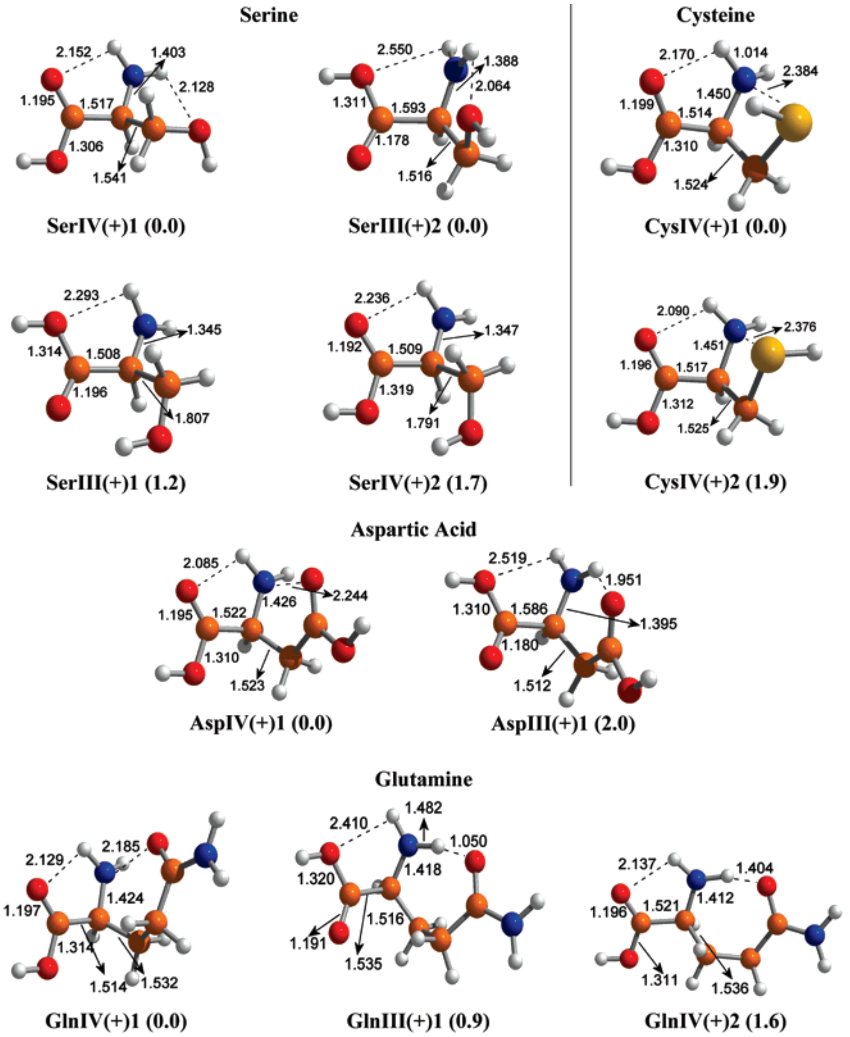


Figure 2. Optimized geometries and relative energies (Δ*E*) for the lowest-energy conformers of Ser, Cys, Asp, and Gln radical cations at the MPWB1K/6-31++G(d,p) level of theory. Distances are in angstroms and energies are in kcal mol^{-1}.

Second, we will consider the influence of ionizing serine (Ser), cysteine (Cys), aspartic acid (Asp), and glutamine (Gln) amino acids, which contain acidic and basic sites in their side chains that can be involved in intramolecular hydrogen bonds. Next, we will present the results corresponding to the aromatic amino acids phenylalanine (Phe), tyrosine (Tyr), and histidine (His), which have an easy ionizable side chain. Finally, all possible C$_\alpha$ cleavages of the ionized amino acids will be analyzed.

Structural Changes. *Gly and Ala*. Previous accurate theoretical studies have identified eight minimum energy conformers of neutral glycine.[50] Among them, five conformers present relative energies which are less than 1000 cm^{-1} (2.86 kcal mol^{-1}), the relative energies of the three remaining conformers being larger than 4.52 kcal mol^{-1} with respect to the ground-state structure. For alanine 13 conformers have been identified as minima on the potential energy surface.[32] However, only 9 present relative energies lower than 1000 cm^{-1}. The five major structures of neutral glycine and nine major structures of alanine are shown in Scheme 1. Notation used has been taken from refs 32 and 50. It can be observed that the increase in the number of stable conformers for Ala is due to the doubling of conformers for structures II, III, IV, and V because of the loss of symmetry plane. Nevertheless, upon ionization of these neutral conformers only three stable structures are found for glycine radical cation and four for alanine radical cation. Optimized geometries are shown in Figure 1, whereas relative energies Δ*E* at different levels of theory are given in Table 1. These conformers have been labeled as II(+), III(+), and IV(+), in analogy to the notation used for the neutral systems, which is related to the nature of intramolecular hydrogen bonds, but we have added a (+) symbol to indicate that it refers to the radical cation species.

Table 2. MPWB1K/6-31++G(d,p) Charge (Spin Density) from Natural Population Analysis for the Lowest-Energy Conformer of Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His Radical Cations

amino acid	NH$_2$	COOH	R	CH
GlyIV(+)	0.64 (0.90)	0.12 (0.00)	0.34 (0.06)	−0.10 (0.04)
AlaIV(+)	0.62 (0.88)	0.11 (0.00)	0.17 (0.08)	0.10 (0.04)
SerIV(+)1	0.63 (0.87)	0.12 (0.01)	0.16 (0.06)	0.10 (0.06)
CysIV(+)1	0.19 (0.40)	0.08 (0.00)	0.62 (0.60)	0.10 (0.00)
AspIV(+)1	0.51 (0.75)	0.10 (0.00)	0.27 (0.24)	0.11 (0.01)
GlnIV(+)1	0.46 (0.69)	0.08 (0.00)	0.36 (0.31)	0.10 (0.00)
PheII(+)1	−0.06 (0.03)	0.23 (0.20)	0.72 (0.72)	0.11 (0.03)
TyrII(+)1	−0.07 (0.02)	0.14 (0.11)	0.83 (0.84)	0.10 (0.03)
HisIII(+)1	−0.10 (0.00)	0.08 (0.03)	0.89 (0.96)	0.12 (0.01)

Moreover, an additional number has been included after (+) to distinguish between conformers with the same amino/carboxylic intramolecular hydrogen bond pattern.

First, it can be noted that structures I(+) and V(+) are not found to be a minima on the potential energy surface since ionization of I or V leads to structure III(+). On the other hand, structures IIA and IIB and IVA and IVB of alanine collapse to conformers II(+) and IV(+), respectively, which reduces significantly the number of stable conformers for the radical cation. This is not surprising considering that structures A and B differ on the relative orientation the carboxylic group with respect to the CH_3 side chain. For example, the OCCN dihedral angles for structures IIA and IIB are 169° and −167°, respectively. However, ionization introduces a positive charge that leads to a unique minimum with a dihedral angle of 179°.

As a general trend, it is observed that both for Gly and Ala ionization is localized at the $-NH_2$ group, and, thus, the hydrogen bonds that involve this group are modified. That is, the amino group becomes more planar, $-NH_2^+$ increases its acidity, and consequently the intramolecular hydrogen bonds in which $-NH_2$ acts as proton donor are strengthened. For this reason structure IV(+) becomes largely stabilized for glycine and alanine radical cations. In contrast, structure II(+) in which $-NH_2$ acts as proton acceptor becomes the most unstable one due to the decrease of basicity of $-NH_2$ upon ionization.

It can be observed in Figure 1 that the optimized geometries with the two functionals are quite similar. However, the computed relative energies largely depend on the functional used, the one that better compares to the CCSD(T) method being the hybrid-meta MPWB1K (see Table 1). That is, for both Gly and Ala at the MPWB1K level of theory, structure IV(+) is predicted to be the global minimum in agreement with the CCSD(T) calculations. However, at the B3LYP level a III(+)-like structure is determined to be the most stable one. It should be mentioned that the CCSD(T) values are almost the same regardless of whether we use the B3LYP or MPWB1K optimized geometries to perform the single-point CCSD(T) calculations. The discrepancy between both functionals is not surprising considering that structures III(+) present a two-center/three-electron bond between N and O, which has been shown to be overstabilized by the B3LYP functional, due to an overestimation of the self-interaction part of the exchange energy because of the delocalized nature of the electron hole.[72,73] These studies showed also that the admixture of exact exchange energy reduces the error, as found here with MPWB1K, which includes a 44% of exact exchange. On the other hand, it can be observed in Table 1 that the influence of further enlarging the basis set at the CCSD(T) level is much smaller, the values with the largest aug-cc-pVTZ basis set being in quite good agreement with the MPWB1K/6-31++G(d,p) results, which validates this latter level of theory as a cost-effective one for studying these systems. Because of that in the following sections, and in order to facilitate the discussion, only the MPWB1K results will be reported.

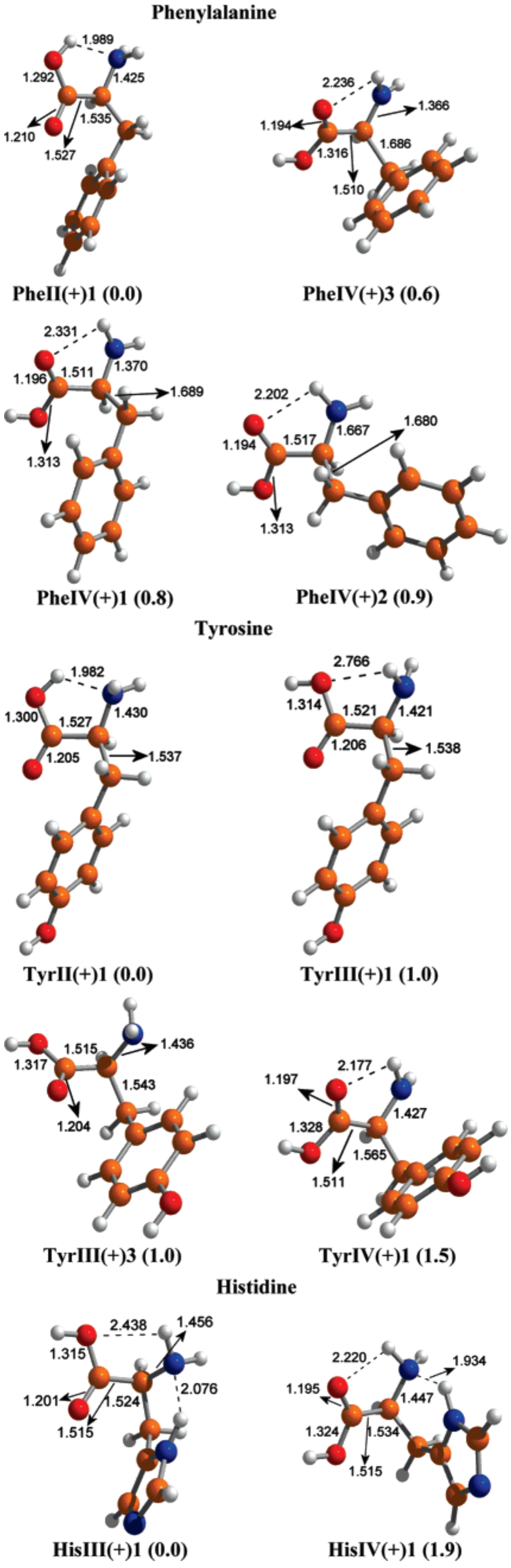


Figure 3. Optimized geometries and relative energies (Δ*E*) for the lower energy conformers of Phe, Tyr, and His at the MPWB1K/6-31++G(d,p) level of theory. Distances are in angstroms and energies are in kcal mol^{-1}.

Ser, Cys, Asp, and Gln. Let us now consider those amino acids that contain hydrocarbon side chains with acidic and

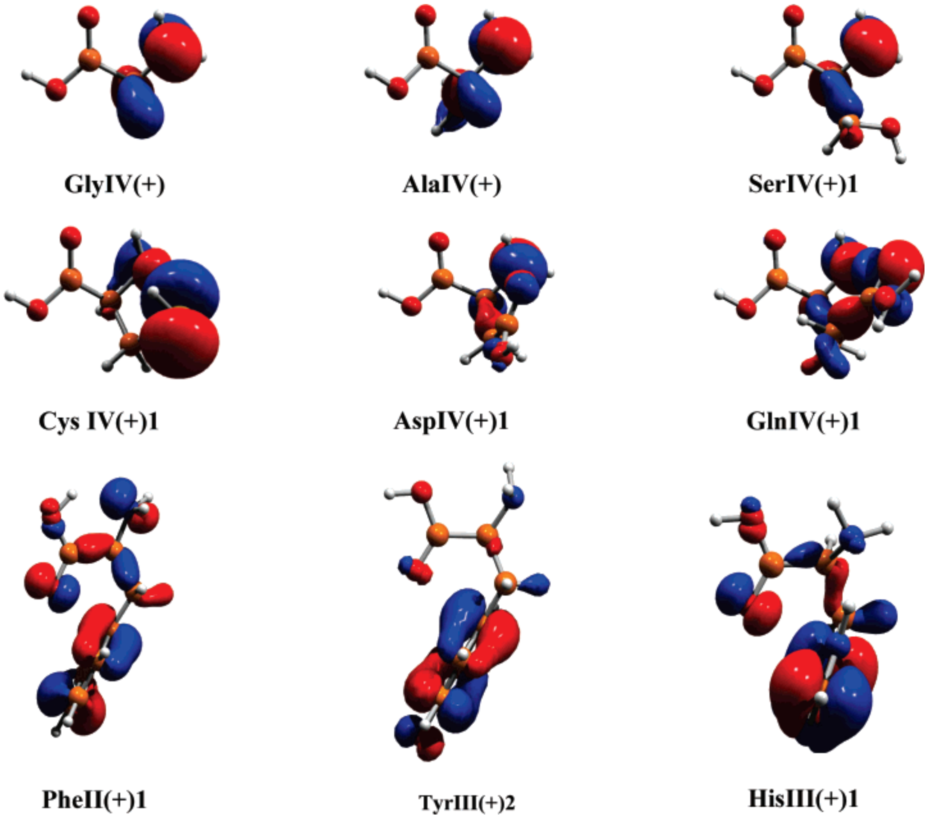


Figure 4. Single occupied molecular orbital of the lowest conformer of Gly, Ala, Ser, Cys Asp, Gln, Phe, Tyr, and His radical cations.

Table 3. MPWB1K/6-31++G(d,p) Internal Energies of Reaction (ΔU_{0K}) for the Unimolecular Decompositions of Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His Radical Cations (kcal mol^{-1})[a]

	$GX^{\cdot+} \rightarrow G^{+} + X^{\cdot}$	Gly	Ala	Ser	Cys	Asp	Gln	Phe	Tyr	His
(1)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CRCOOH]^{+} + H^{\cdot}$	33.0	21.8	23.5	31.8	25.8	27.5	26.5	39.5	50.3
(2)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [CHRCOOH]^{+} + [NH_2]^{\cdot}$	79.0	71.2	42.9	45.6	56.5	28.1	40.3	46.9	42.5
(3)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CHR]^{+} + [COOH]^{\cdot}$	23.5	12.5	22.6	22.4	16.7	18.6	20.5	27.3	37.4
(4)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CHCOOH]^{+} + R^{\cdot}$	33.0	28.2	28.2	29.9	28.3	47.0	26.4	37.2	38.4

	$GX^{\cdot+} \rightarrow G^{\cdot} + X^{+}$	Gly	Ala	Ser	Cys	Asp	Gln	Phe	Tyr	His
(5)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CRCOOH]^{\cdot} + H^{+}$	180.3	181.8	185.8	189.7	186.2	199.1	193.0	204.5	209.8
(6)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [CHRCOOH]^{\cdot} + [NH_2]^{+}$	163.7	163.3	169.9	172.1	172.0	185.1	175.7	186.7	191.1
(7)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CHR]^{\cdot} + [COOH]^{+}$	67.8	70.7	72.9	76.7	73.9	89.4	82.7	92.7	94.6
(8)	$[NH_2CHRCOOH]^{\cdot+} \rightarrow [NH_2CHCOOH]^{\cdot} + R^{+}$	180.3	88.8	37.1	40.8	68.1	66.1	27.2	25.8	34.4

[a] R = H, CH_3, CH_2OH, CH_2SH, CH_2COOH, $CH_2CH_2CONH_2$, $CH_2C_6H_5$, $CH_2C_6H_4OH$, $CH_2C_3N_2H_4$ for Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His, respectively.

basic groups such as Ser, Cys, Asp, and Gln for which R = $-CH_2OH$, $-CH_2SH$, $-CH_2COOH$, and $-CH_2CH_2CONH_2$, respectively. Because now the number of stable conformers is much larger due to the presence of many single-bond rotamers, the following strategy has been applied to find the lower conformers of each amino acid. Starting from the major structures of glycine (see Scheme 1) a Monte Carlo Multiple Minimum (MCMM) conformational search[59,60] with the MMFF94s force field[61,62] has been performed allowing only the internal rotations of the side chain. All plausible structures within an energy window of 50 kJ mol^{-1} were selected for subsequent quantum chemical optimizations of the neutral systems. Structures of radical cations were then obtained by reoptimizing these structures after removing one electron from the system. We expect that with this strategy the main conformers of the radical cations of these amino acids have been localized. Optimized geometries and relative energies of these low-lying conformers (up to 2 kcal mol^{-1} at the MPWB1K level) are shown in Figure 2. The remaining structures as well as their relative energies at different levels of theory are given in the Supporting Information.

It can be observed that, as found for Gly and Ala, the low-lying energy conformer of these amino acid radical cations are either of type III(+) or IV(+). In all cases the initially pyramidalized $-NH_2$ group becomes more planar in the radical cation species due to the fact that ionization mainly takes place at $-NH_2$. This is confirmed by natural population analysis which indicates that the spin density

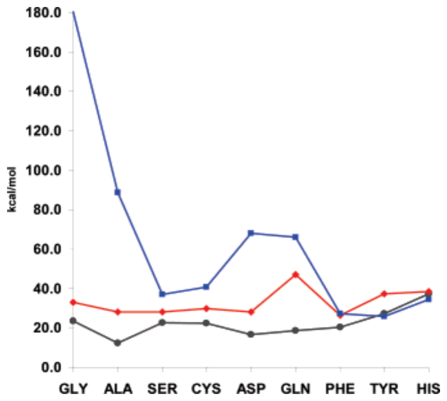


Figure 5. Reaction energies ΔU_{0K} for (●) $[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHR]^+ + [COOH]^\bullet$, (◆) $[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHCOOH]^+ + R^\bullet$, and (■) $[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHCOOH]^\bullet + R^+$ for each amino acid at the MPWB1K/6-31++G(d,p) level of theory.

mainly lies at this $-NH_2$ group (see Table 2). Thus, ionization increases the $-NH_2$ acidity, which favors intramolecular hydrogen bond interactions in which this group acts as proton donor. For this reason structure IV(+) becomes the lowest-energy conformer in all cases. Nevertheless, for Ser we have located an almost degenerate conformer of type III(+) in which the $-NH_2$ establishes a quite strong hydrogen bond interaction with the OH group of the side chain.

For Cys, the more stable conformers, CysIV(+)1 and CysIV(+)2, present a two-center/three-electron hemibond interaction between the $-NH_2$ and the $-SH$ group of the side chain. For Ser, however, we have not been able to locate a conformation with such an interaction with the side chain, in agreement with the fact that the $-OH$ group prefers to establish hydrogen bonds interaction than two-center/three-electron hemibonds.[51,53] In contrast, for Asp the most stable IV(+) structure presents a stabilizing hemibond interaction between the carbonyl oxygen of the side chain and the $-NH_2$ group, which shows that the existence or not of such hemibond interactions in radical cations results from a subtle balance between the stabilization gained by this hemibond interaction and the possibility of establishing intramolecular hydrogen bonds. In fact for glutamine we have located two structures IV(+), one in which the carbonyl oxygen of the side chain forms an intramolecular hydrogen bond with the NH_2 group and another one in which it establishes a two-center/three-electron interaction, the relative energies between them being 1.6 kcal mol^{-1}. It should be noted that for those structures that show a two-center/three-electron bond between the $-NH_2$ group and a basic site of the side chain (CysIV(+)1, AspIV(+)1, and GlnIV(+)1) natural population analysis indicates that the spin density is delocalized between the two interacting groups (see Table 2)

Glutamine is a particularly interesting amino acid since in many cases ionization leads to a spontaneous proton transfer from the C_α to the CO group of the side chain. In fact, the most stable structural isomer of glutamine radical cation corresponds to a diol $[NH_2C(CH_2CH_2CONH_2)C(OH)_2]^{\bullet+}$ species, which lies 27.0 kcal mol^{-1} below the most stable nonproton transferred structure shown in Figure 2. As found for glycine radical cation,[29] this diol structure is largely stabilized by captodative effects in the glycyl like species formed. However, in contrast to glycine, further studies[74] have confirmed that glutamine easily evolves to a diol structure due to its long side chain with basic groups which allow it to act as a proton-acceptor and also as a solvent assistant catalyst.

His, Phe, and Tyr. Optimized geometries and relative energies of the low-lying conformers of His, Phe, and Tyr radical cations are shown in Figure 3. For Tyr we have only included one of the two (almost degenerate) conformers associated with the rotation of the OH of the side chain. Aromatic amino acids do not follow the trends found for the previous amino acids because ionization mainly takes place at the side chain. This is in agreement with the spin density values and the nature of the open-shell orbital shown in Table 2 and Figure 4, respectively. Note that for Tyr and His the spin density at the side chain is 0.8−0.9 and that the open-shell orbital is mainly centered at the aromatic ring. For Phe the spin density is more delocalized although it still has its major contribution at the ring. For these amino acids for which ionization of the side chain prevails over ionization of the $-NH_2$ group, structures type III(+) and II(+) become competitive. In fact, the most stable structure for His is a distorted structure III(+) in which the NH group of the imidazole ring forms a hydrogen bond with the $-NH_2$ group. This imidazole NH group is more acidic due to the ionization of the side chain. On the other hand, for Phe and Tyr, structures derived from II(+) become the ground-state structures, which shows the importance of the side-chain nature in the effects of ionization.

Unimolecular Decompositions. In addition to the changes observed in intramolecular hydrogen bonds, other major geometry changes occur upon ionization, which can determine the fragmentations that will be observed in mass spectrometry experiments.

Table 3 shows the internal energy of reaction (ΔU_{0K}) corresponding to the different fragmentation processes of Gly, Ala, Ser, Cys, Gln, Asp, Phe, Tyr, and His radical cations. Four different $C_\alpha-R$ bond cleavages can be considered: $C_\alpha-COOH$, $C_\alpha-H$, $C_\alpha-NH_2$, and $C_\alpha-R$. Such cleavages can be produced in two different ways: that is, by losing a neutral radical ($COOH^\bullet$, $H^\bullet$, $NH_2^\bullet$,$R^\bullet$) or by losing a cation ($COOH^+$, H^+, NH_2^+, R^+). Thus, eight different reactions have been considered, the computed reaction energies being collected in Table 3.

Among the four decompositions that involve the loss of a neutral radical, the loss of $[COOH]^\bullet$ is the most favorable process for all amino acids. This fact, previously observed for Gly, Ala, Ser, and Cys,[46] is also true for Gln, Asp, and Phe and is in very good agreement with their mass spectra,[75] since the most intense peaks at $m/z = 30$, $m/z = 44$, $m/z = 60$, $m/z = 76$, and $m/z = 88$, respectively, can be assigned to the $[NH_2CH_2]^+$, $[NH_2CHCH_3]^+$, $[NH_2CHCH_2OH]^+$, $[NH_2CHCH_2SH]^+$, and $[NH_2CHCH_2COOH]^+$ ions formed by loss of the $[COOH]^\bullet$ radical. Phe mass spectra also present an intense peak at $m/z = 120$, corresponding to the decomposition: $[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHR]^+ + [COOH]^\bullet$.

As the side chain increases, the loss of $R^\bullet$ (eq 4) starts to be a competitive process. This fact, observed for Ser, Cys,

Table 4. MPWB1K/6-31++G(d,p) Adiabatic Ionization Energy of Each Amino Acid and the Different Fragments Formed (kcal mol^{-1})^a

	Gly	Ala	Ser	Cys	Asp	Gln	Phe	Tyr	His
$NH_2CHRCOOH$	207.1	202.6	201.1	196.0	199.4	185.0	192.5	181.6	181.8
$[NH_2CRCOOH]^{\cdot}$	164.7	152.0	149.8	154.1	151.6	140.4	145.5	147.0	152.5
$[CHRCOOH]^{\cdot}$	204.5	197.1	162.3	162.8	173.8	132.3	153.8	149.4	140.6
$[NH_2CHR]^{\cdot}$	143.8	129.8	137.7	133.7	130.8	117.2	125.8	122.6	130.8
$[NH_2CHCOOH]^{\cdot}$	164.7	164.7	164.7	164.7	164.7	164.7	164.7	164.7	164.7
$H^{\cdot}$	312.0	312.0	312.0	312.0	312.0	312.0	312.0	312.0	312.0
$[NH_2]^{\cdot}$	289.3	289.3	289.3	289.3	289.3	289.3	289.3	289.3	289.3
$[COOH]^{\cdot}$	188.0	188.0	188.0	188.0	188.0	188.0	188.0	188.0	188.0
$[R]^{\cdot}$	312.0	225.3	173.6	175.6	204.5	183.8	165.4	153.2	160.7

[a] R = H, CH_3, CH_2OH, CH_2SH, CH_2COOH, $CH_2CH_2CONH_2$, $CH_2C_6H_5$, $CH_2C_6H_4OH$ and $CH_2C_3N_2H_4$ for Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His, respectively.

Table 5. MPWB1K/6-31++G(d,p) Cα−X Dissociation Energies (D_0)^a for Neutral Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His (in kcal mol^{-1})

$GX \rightarrow G^{\cdot} + X^{\cdot}$	Gly	Ala	Ser	Cys	Asp	Gln	Phe	Tyr	His
$[NH_2CHRCOOH] \rightarrow [NH_2CRCOOH]^{\cdot} + H^{\cdot}$	75.4	72.4	74.9	73.7	73.6	72.1	73.5	74.1	79.6
$[NH_2CHRCOOH] \rightarrow [CHRCOOH]^{\cdot} + [NH_2]^{\cdot}$	81.5	76.7	81.7	78.8	82.1	80.8	78.9	79.1	83.6
$[NH_2CHRCOOH] \rightarrow [NH_2CHR]^{\cdot} + [COOH]^{\cdot}$	86.8	85.3	85.9	84.7	85.3	86.4	87.2	86.3	88.4
$[NH_2CHRCOOH] \rightarrow [NH_2CHCOOH]^{\cdot} + R^{\cdot\, b}$	75.4	66.2	64.6	61.2	62.9	67.3	54.2	54.1	55.5

[a] Zero point energy computed from harmonic vibrational frequencies. [b] R = H, CH_3, CH_2OH, CH_2SH, CH_2COOH, $CH_2CH_2CONH_2$, $CH_2C_6H_5$, $CH_2C_6H_4OH$, $CH_2C_3N_2H_4$ for Gly, Ala, Ser, Cys, Asp, Gln, Phe, Tyr, and His, respectively.

and aromatic Phe, Tyr, and His amino acids, is in very good agreement with the mass spectra of Ser, Cys, and Phe which show a very intense peak at $m/z = 74$ corresponding to the fragment $[NH_2CHCOOH]^+$. The third process that can compete with the loss of $[COOH]^\bullet$ and $[R]^\bullet$ (as the side chain increases) is the glycyl formation; that is, the loss of the cationic side chain. It can be observed in Table 3 that, among the four reactions (eqs 5−8) leading to the loss of a cation fragment, the C_α−R cleavage is the most favorable process for all amino acids except for Gly and Ala which prefer the loss of $COOH^+$. This is not surprising considering that the larger the side chain is, the better the positive charge is delocalized. Internal energy changes (ΔU_{0K}) of these three decompositions are shown in Figure 5. It can be observed how the loss of R^+ becomes very important for aromatic amino acids, this reaction becoming even the preferred unimolecular decomposition for Tyr and His. In fact, these two amino acids present a very intense peak in their mass spectra, $m/z = 81$ and $m/z = 107$, respectively, corresponding to the R^+ fragment.

As noted previously,[28,46] a simple thermodynamic cycle allows us to decompose the internal energy of reaction in

$$\Delta U_{0K}(-X^\bullet) = -IE(GX) + D_0(GX) + IE(G^\bullet)$$

or

$$\Delta U_{0K}(-X^+) = -IE(GX) + D_0(GX) + IE(X^\bullet)$$

depending on whether the radical cation loses a neutral or a cationic fragment.

D_0(GX) corresponds to the homolytic dissociation energy of the neutral amino acid, IE(GX) corresponds to the adiabatic ionization energy of the considered amino acid, and IE($G^\bullet$) and IE($X^\bullet$) correspond to the ionization energy of each fragment. That is, the internal energy of reaction (ΔU_{0K}) of an unimolecular decomposition depends on three different parameters: (i) the ionization energy of the corresponding amino acid, (ii) the dissociation energy of the neutral compound, and (iii) the ionization energy of each fragment. Ionization energies are given in Table 4, whereas D_0(GX) values are shown in Table 5. Since we are interested in analyzing which is the most favorable fragmentation within an amino acid, the first parameter, its ionization energy, remains constant and, thus, will not be discussed further in the text.

For each fragmentation let us start to analyze the preference in the loss of the neutral radical fragment $X^\bullet$ or cation fragment X^+. From the energy decomposition scheme it can be noted that for each amino acid this preference only depends on the $X^\bullet$ and $G^\bullet$ relative ionization energies. As the ionization energies of the fragment that contains C_α (IE($G^\bullet$)) is lower than IE($X^\bullet$), the loss of $X^\bullet$ is, in general, the most favorable process. For example, for Ala the ionization energy of $[NH_2CHCH_3]^\bullet$ (IE=129.8 kcal mol^{-1}) is lower than that of $[COOH]^\bullet$ (IE=188.0 kcal mol^{-1}), which makes the loss of neutral $[COOH]^\bullet$ the most favorable process. The same conclusion can be reached for each pair of fragmentations of all nine amino acids. The only exception comes when the decomposition process implies the loss of R^+ or $R^\bullet$ because for Phe, Tyr, and His the IE of $R^\bullet$ (165.4, 153.2, and 160.7 kcal mol^{-1}, respectively) is very similar to that of the glycyl radical, $[NH_2CHCOOH]^\bullet$, which is 164.7 kcal mol^{-1}. This fact explains why the loss of cationic side chain ($[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHCOOH]^\bullet + R^+$) is preferred over the loss of the neutral radical ($[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHCOOH]^+ + R^\bullet$**)** in the case of Tyr and His.

When different fragmentations are compared, the dissociation energy of the involved bond of the neutral amino acids (D_0(GX)) needs to be taken into account. That is, the

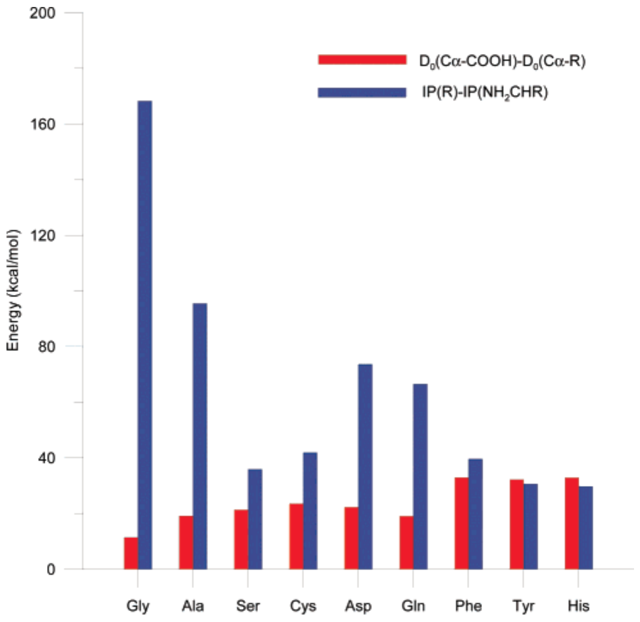


Figure 6. D_0(Cα−COOH)−D_0(Cα−R) and IE(R)−IE(NH_2-CHR) in kcal mol^{-1}.

preference for one reaction or another will depend not only on the ionization energy of the fragment that will finally support the positive charge but also on the energy required to break the corresponding bond. Since differences on the dissociation energies are rather small compared to the variations on ionization energies, usually the dominant term is the ionization energy in such a way that the preferred process is the one that leaves the positive charge in the fragment with the lower ionization energy; that is, reaction 3 $[NH_2CHRCOOH]^{\bullet+} \rightarrow [NH_2CHR]^+ + [COOH]^{\bullet}$ (see Table 3). Nevertheless, as the side chain becomes more voluminous, the C_α−R becomes weaker (see Table 5), in such a way that the C_α−R dissociation energy for aromatic amino acids (54−55 kcal mol^{-1}) becomes significantly smaller than the C_α−COOH (84−88 kcal mol^{-1}), C_α−NH_2 (78−83 kcal mol^{-1}), or the C_α−H (72−79 kcal mol^{-1}) dissociation energies. Therefore, the loss of the side chain, particularly R^+, $[NH_2\text{-}CHRCOOH]^{\bullet+} \rightarrow [NH_2CHCOOH]^{\bullet} + R^+$ reaction 8, becomes competitive for Phe, Tyr, and His. This can be clearly seen in Figure 6 where the difference on neutral dissociation energies (D_0(C_α−COOH)−D_0(C_α−R)) and ionization energies of the two cationic fragments ((IE($[R]^{\bullet}$)−IE($[NH_2\text{-}CHR]^{\bullet}$)) corresponding to these competing reactions are represented for each amino acid. These two quantities act in an opposite way; that is, the larger the first one is, the more favorable becomes reaction 8 (loss of R^+), whereas the larger the second term is, the more favorable becomes reaction 3 (loss of $[COOH]^{\bullet}$). It can be observed that for aromatic amino acids the two columns become almost equal so that the two fragmentation processes become energetically similar.

Summary

This work provides a theoretical study of the conformational behavior of nine ionized amino acids by means of the hybrid B3LYP and meta-hybrid MPWB1K functional as well as by means of post-Hartree Fock calculations at the CCSD(T) level of theory. Different kinds of amino acids have been chosen in order to study the effect of the side chain on the reorganization and fragmentation processes upon ionization. In almost all cases ionization of these amino acids takes place at the amino group, which becomes more planar and acidic. As a consequence NH···OC hydrogen bonds are strengthened, and conformer IV(+) is largely stabilized for the ionized species. In fact for all amino acids except the aromatic ones, a IV(+)-like conformer is the ground-state structure, the side chain being involved in additional intramolecular hydrogen bonds or in two-center/three-electron interactions with the ionized $-NH_2$ group. However, for Phe, Tyr, and His aromatic amino acids ionization takes place mainly at the aromatic ring. Because of that, for Phe and Tyr, structures II(+) become the most stable ones. In the case of His ionization increases the acidity of the imidazole −NH− group in such a way that it tends to form a hydrogen bond with the lone pair of the $-NH_2$ leading to a distorted structure III(+). Finally, among the different C_α−X fragmentation processes, the one that leads to the loss of $[COOH]^{\bullet}$ is the most favorable one. Nevertheless, for amino acids with an increasing size chain, fragmentations leading to R^+ or $R^{\bullet}$ start being competitive. In fact, for the aromatic amino acids Tyr and His, the fragmentation leading to R^+ is the most favorable process. This is important because it leads to the formation of glycyl radical, which is known to be involved in different protein radical processes.

Acknowledgment. Financial support from MCYT and DURSI, through the CTQ2005-08797-C02-02 and SGR2005-00244 projects, and the use of the Catalonia Supercomputer Centre (CESCA) are gratefully acknowledged.

Supporting Information Available: Low-lying energy conformers of Gly, Ala, Ser, Cys, Asp Gln, Phe, Tyr, and His radical cations and relative energies at different levels of theory. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Berlett, B. S.; Stadtman, E. R. *J. Biol. Chem.* **1997**, *272*, 20313.

(2) Stadtman, E. R. *Ann. Rev. Biochem.* **1993**, *62*, 797.

(3) Butterfield, D. A.; Boyd-Kimbal, D. *Biochim. Biophys. Acta* **2005**, *1703*, 149.

(4) Hou, L.; Shao, H.; Zhang, Y.; Li, H.; Menon, N. K.; Neuhaus, E. B.; Brewer, J. M.; Byeon, I.-J. L.; Ray, D. G.; Vitek, M. P.; Iwashita, T.; Makula, R. A.; Przybyla, A. B.; Zagorski, M. G. *J. Am. Chem. Soc.* **2004**, *126*, 1992.

(5) Clementi, M. E.; Martorana, G. E.; Pezzotti, M.; Giardina, B.; Misiti, F. *Int. J. Biochem. Cell Biol.* **2004**, *36*, 2066.

(6) Misiti, F.; Martorana, G. E.; Nocca, G.; Di Stasio, E.; Giardina, B; Clementi, M. E. *Neuroscience* **2004**, *126*, 297.

(7) Butterfield, D. A.; Bush, A. I. *Neurobiol. Aging* **2004**, *25*, 563.

(8) Ciccosto, G. D.; Barnham, K. J.; Cherny, R. A.; Masters, C. L.; Bush, A. I.; Curtain, C. C.; Cappai, R.; Tew, D. *Lett. Pept. Sci.* **2003**, *10*, 413.

(9) Barnham, K. J., et al. *J. Biol. Chem.* **2003**, *278*, 42959.

(10) Kantorow, M.; Hawse, J. R.; Cowell, T. L.; Benhamed, S.; Pizarro, G. O.; Reddy, V. N.; Hejtmanicik, J. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9654.

(11) Stubbe, J.; van der Donk, W. A. *Chem. Rev.* **1998**, *98*, 705.

(12) Bagheri-Majdi, E.; Ke, Y.; Orlova, G.; Chu, I. K.; Hopkinson, A. C.; Siu, K. W. M. *J. Phys. Chem. B* **2004**, *108*, 11170.

(13) Chis, V.; Brustolon, M.; Chis, V.; Morari, C.; Cozar, O.; David, L. *J. Mol. Struct.* **1999**, *482*, 283.

(14) Brustolon, M.; Chis, V.; Maniero, A. L.; Brunel, L. C. *J. Phys. Chem. A* **1997**, *101*, 4887.

(15) Sanderud, A.; Sagstuen, E. *J. Phys. Chem. B* **1998**, *102*, 9353.

(16) Bonifačic, M.; Štefanić, I.; Hug, G. L.; Armtrong, D. A.; Asmus, K. D. *J. Am. Chem. Soc.* **1998**, *120*, 9930.

(17) Jochims, H. W.; Schwell, M.; Chotin, J. L.; Clemino, M.; Dulieu, F.; Baumgärtel, H.; Leach, S. *Chem. Phys.* **2004**, *298*, 279.

(18) Lago, A. F.; Coutinho, L. H.; Marinho, R. R. T.; Naves de Brito, A.; De Souza, G. G. B. *Chem. Phys.* **2004**, *307*, 9.

(19) Messer, B. M.; Cappa, C. D.; Smith, J. D.; Wilson, K. R.; Gilles, M. K.; Cohen, R. C.; Saykally, R. J. *J. Phys. Chem. B* **2005**, *109*, 5375.

(20) Kumar, S.; Rai, A. K.; Singh, V. B.; Rai, S. B. *Spectrochim. Acta, Part A* **2005**, *61*, 2741.

(21) Snoek, L. C.; Robertson, E. G.; Kroemer, R. T.; Simons, J. P. *Chem. Phys. Lett.* **2000**, *321*, 49.

(22) Huang, Y.; Kenttämaa, H. *J. Am. Chem. Soc.* **2005**, *127*, 7952.

(23) Wilson, K. R.; Belau, L.; Nicolas, C.; Jimenez-Cruz, M.; Leone, S. R.; Ahmed, M. *Int. J. Mass. Spectrom.* **2006**, *249/250*, 155.

(24) Kovačevic, B.; Rožman, M.; Klansinc, L.; Srzić, D.; Maksić, Z. B.; Yáñez, M. *J. Phys. Chem. A* **2005**, *109*, 8329.

(25) Rauk, A.; Yu, D.; Armstrong, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 8848.

(26) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1998**, *120*, 5723,

(27) Rodríguez-Santiago, L.; Sodupe, M.; Oliva, A.; Bertran, J. *J. Phys. Chem. A* **2000**, *104*, 1256.

(28) Simon, S.; Sodupe, M.; Bertran, J. *J. Phys. Chem. A* **2002**, *106*, 5697.

(29) Simon, S.; Sodupe, M.; Bertran, J. *Theor. Chem. Acc.* **2003**, *111*, 217.

(30) Lu, H. F.; Li, F. Y.; Lin, S. H. *J. Phys. Chem. A* **2004**, *108*, 9233.

(31) Gronert, S.; O'Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, *117*, 2071.

(32) Császár, A. G. *J. Phys. Chem.* **1996**, *100*, 3541.

(33) Stepanian, S. G.; Reva, I. D.; Radchenko, E. D.; Adamowicz, L. *J. Phys. Chem. A* **1998**, *102*, 4623.

(34) Lambie, B.; Ramaekers, R.; Maes, G. *J. Phys. Chem. A* **2004**, *108*, 10426.

(35) Noguera, M.; Rodríguez-Santiago, L.; Sodupe, M.; Bertrán, J. *J. Mol. Struct. (THEOCHEM)* **2001**, *537*, 307.

(36) Blanco, S.; Lesarri, A.; López, J. C.; Alonso, J. L. *J. Am. Chem. Soc.* **2004**, *126*, 11675.

(37) Lakard, B. *J. Mol. Struct. (THEOCHEM)* **2004**, *681*, 183.

(38) Ai, H.; Bu, Y.; Li, P.; Li, Z. *J. Chem. Phys.* **2004**, *120*, 11600.

(39) Pecul, M.; Ruud, K.; Rizzo, A.; Helgaker, T. *J. Phys. Chem. A* **2004**, *108*, 4269.

(40) Jeon, I. S.; Ahn, D. S.; Park, S. W.; Lee, S.; Kim, B. *Int. J. Quantum Chem.* **2005**, *101*, 55.

(41) Miao, R.; Jin, C.; Yang, G.; Hong, J.; Zhao, C.; Zhu, L. *J. Phys. Chem. A* **2005**, *109*, 2340.

(42) Gong, X.; Zhou, Z.; Du, D.; Dong, X.; Liu, S. *Int. J. Quantum Chem.* **2005**, *103*, 105.

(43) Pecul, M. *Chem. Phys. Lett.* **2006**, *418*, 1.

(44) Kushwaha, P. S.; Mishra, P. C. *J. Photochem. Photobiol., A* **2000**, *137*, 79.

(45) Kushwaha, P. C.; Mishra, P. C. *J. Mol. Struct. (THEOCHEM)* **2001**, *549*, 229.

(46) Simon, S.; Gil, A.; Sodupe, S.; Bertran, J. *J. Mol. Struct. (THEOCHEM)* **2005**, *727*, 191.

(47) Dehareng, D.; Dive, G. *Int. J. Mol. Sci.* **2004**, *5*, 301.

(48) Zhang, M.; Huang, Z.; Lin, Z. *J. Chem. Phys.* **2005**, *122*, 134313.

(49) Huang, Z; Yu, W.; Lin, Z. *J. Mol. Struct. (THEOCHEM)* **2006**, *801*, 7.

(50) Császár, A. G. *J. Am. Chem. Soc.* **1992**, *114*, 9568.

(51) Sodupe, M.; Oliva, A.; Bertran, J. *J. Am. Chem. Soc.* **1994**, *116*, 8249.

(52) Sodupe, M.; Oliva, A.; Bertran, J. *J. Phys. Chem. A* **1997**, *101*, 9142.

(53) Sodupe, M.; Oliva, A.; Bertran, J. *J. Am. Chem. Soc.* **1995**, *117*, 8416.

(54) Rodríguez-Santiago, L.; Sodupe, M.; Oliva, A.; Bertran, J. *J. Am. Chem. Soc.* **1999**, *121*, 8882.

(55) Gil, A.; Bertran, J.; Sodupe, M. *J. Am. Chem. Soc.* **2003**, *125*, 7462.

(56) Gil, A.; Sodupe, M.; Bertran, J. *Chem. Phys. Lett.* **2004**, *27*, 395.

(57) Tureček, F.; Carpenter, F. H.; Polce, M. J.; Wesdemiotis, C. *J. Am. Chem. Soc.* **1999**, *121*, 7955.

(58) Turečec, F.; Carpenter, F. H. *J. Chem. Soc., Perkin Trans. 2* **1999**, 2315.

(59) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379.

(60) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc.* **1990**, *112*, 1419.

(61) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 720.

(62) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730.

(63) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(64) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(65) Stephens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(66) Zaho, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(67) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(68) Reed, A. E.; Curtiss, L. A.; Weinhold, *Chem. Rev. (Washington, D. C.)* **1988**, *88*, 899.

(69) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(70) MOLPRO is a package of ab initio programs written by H.-J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, M. Schütz, P. Celani, T. Korona, G. Rauhut, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, A. W. Lloyd, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, P. Palmieri, R. Pitzer, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni, and T. Thorsteinsson.

(71) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440.

(72) Sodupe, M.; Bertran, J.; Rodríguez-Santiago, L.; Baerends, E. J. *J. Phys. Chem. A* **1999**, *103*, 166.

(73) Braïda, B.; Hiberty, P. C.; Savin, A. *J. Phys. Chem. A* **1998**, *102*, 7872.

(74) Gil, A.; Simon, S.; Sodupe, M.; Bertran, J. *Theor. Chem. Acc.* **2007**, *118*, 589.

(75) Stein, S. E. NIST Mass Spectrometry Data Center, 2006.

CT700055P

JCTC Journal of Chemical Theory and Computation

Theoretical Study of Binding Site Preference in [2]Rotaxanes

Michael E. Foster and Karl Sohlberg*

Department of Chemistry, Drexel University, 3141 Chestnut Street,
Philadelphia, Pennsylvania 19104

Received June 28, 2007

Abstract: Rotaxanes that can be switched between co-conformations by some external stimulus are of interest because the switching mechanism might be used to create molecular devices capable of producing useful work. Probably the most common approach to create a switchable rotaxane is to start with a rotaxane where the ring interacts more strongly with one of two possible binding sites along the shaft and then apply an external stimulus that weakens the binding interaction between the ring and the shaft at this site, thereby changing the binding site preference. We have investigated binding site preference in two rotaxanes and two pseudorotaxanes with electronic structure calculations at several levels of theory. To gain insight into the origins of the intercomponent binding, empirical approximations were applied to estimate the electrostatic and dispersion contributions. Dispersion has been thought to make an important contribution to the intercomponent interaction in the presence of $\pi-\pi$ stacking interactions between the components, but the role of dispersion interaction has been a controversial issue because many computational methods neglect this interaction. For example, AM1 semiempirical calculations neglect dispersion but often predict correct co-conformational preferences. This suggests that inclusion of the dispersion interaction is required for correct quantitative, but not qualitative, description of the intercomponent binding, a result that is supported by the analytic partitioning of the binding interactions. The origins of this result are investigated.

1. Introduction

In the search for molecular systems that can serve as functional components of mechanical nanodevices, switchable rotaxanes have become a center of focus.[1−3] A rotaxane is an interlocked molecular complex wherein a ring molecule is threaded by a long chain molecule. The long chain molecule is terminated with bulky functional groups that have larger radii than the internal diameter of the ring, preventing spontaneous ring unthreading. In a [2]rotaxane therefore, the two component molecules are mechanically linked but chemically independent. A *switchable* rotaxane is a rotaxane that can be switched between two co-conformational isomers through the application of an external stimulus. Switchable rotaxanes are therefore of special interest because they contain the essential features of a molecular "machine" or "device," and mechanical action is accomplished through the application of an external stimulus.

To facilitate the design of nanodevices based on switchable rotaxanes, it would be of great value to identify a robust and efficient theoretical modeling technique. Numerous approaches have been explored.[4,5] One technique that has shown considerable success is AM1[6] semiempirical electronic structure methodology. Semiempirical electronic structure calculations have at least two desirable features. First, they are computationally efficient (relative to ab initio approaches) and therefore may be routinely used on systems of the size of interlocked macromolecular complexes. Second, semiempirical methods explicitly describe the electronic structure and therefore may be used on various charge and electronic states of a system without reparametrization. This second advantage is especially valuable for switchable rotaxanes, because many proposed switching mechanisms depend on a

* Corresponding author e-mail: sohlbergk@drexel.edu.

10.1021/ct7001623 CCC: $37.00
Published on Web 10/18/2007

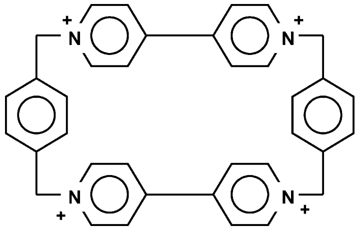


Figure 1. Tetracationic ring structure used in all rotaxane and pseudorotaxane systems considered here.

change in the charge or electronic state of the system. An alternative approach would be to employ molecular mechanics (MM) methods, which are also computationally efficient. MM methods, however, typically require reparametrization to treat different charge states, in particular the assignment of partial atomic charges. Such partial charges are often obtained by first carrying out semiempirical or ab initio electronic structure calculations on the same or a closely related system,[7] which partially negates the advantages of the MM methods. Since atomic partial charges vary with molecular conformation,[8] one must either carry out true electronic structure calculations for multiple conformations to establish this structure dependence or employ an empirical approach to assign them dynamically.[9] Since MM calculations do not yield explicit electronic structure information, they are also not generally applicable to obtain molecular orbital information or for modeling excited electronic states.

One of the key features that a modeling technique must predict reliably to be valuable as a tool for the design of switchable interlocked macromolecular systems is co-conformational selectivity. Selectivity is determined by the intercomponent nonbonding interactions.[10−13] Since the AM1 Hamiltonian produces qualitatively correct ordering of hydrogen-bonding interactions,[14] it may be confidently used to predict co-conformational selectivity in interlocked macromolecular complexes where hydrogen-bonding dominates the intercomponent interactions. Several rotaxanes and catenanes systems of this nature have been studied with success utilizing the AM1 method.[5,15,16] In many switchable rotaxanes, however, the dominant intercomponent interaction is $\pi-\pi$ stacking, which is governed by dispersion forces. Since AM1 is a HF-based technique, it neglects dispersion. One might expect, therefore, AM1 to be unreliable in application to such systems, but, remarkably, it has proven to be unexpectedly successful in this role, as shown in the present manuscript and elsewhere.[17] It is therefore of considerable importance to understand the limits of reliability of the semiempirical methods. We aim to better characterize the range of applicability of the semiempirical AM1 method, in general, and to understand in particular why AM1 is qualitatively successful in predicting binding site preference in $\pi-\pi$ stacked interlocked macromolecular complexes despite its neglect of dispersion.

In this manuscript we report electronic structure calculations at several levels of theory on several rotaxane and pseudorotaxane systems. All of these systems incorporate the same ring structure (shown in Figure 1). The different shaft structures are shown in Figures 2 and 3. To gain further insight into the interactions between the two components, empirical approximations to the electrostatic and dispersion interactions were also used. The results suggest an origin of the unexpected "success" of AM1 for modeling $\pi-\pi$ stacked interlocked macromolecular complexes.

2. Theoretical Methods

2.1. Electronic Structure Calculations. The interlocked macromolecular systems studied here contain from 92 to 242 atoms. The cyclobis(paraquat-*p*-phenylene) ring, used in all of the systems, contains 72 atoms. Determining the electronic structure of systems of this size presents a very considerable computational challenge. Because of their computational efficiency, semiempirical electronic structure methods hold promise in this application. We therefore seek to identify the limits of their applicability. Since it is widely accepted to be one of the most robust semiempirical methods, herein we explore the application of the AM1 Hamiltonian.[6]

To calibrate the accuracy of the AM1 calculations and to gain insight into the intercomponent interactions, Hartree−Fock self-consistent-field (HF-SCF) calculations and density functional theory (DFT) calculations, based on the B3LYP correlation-exchange functional, were carried out on a subset of the systems studied here. Various orbital basis sets were employed for the wave function expansion to test for convergence with respect to basis set completeness and to assess the sensitivity of the predicted intercomponent interactions to basis set size. In the HF-SCF and DFT calculations of the intercomponent interaction energies, the counterpoise (CP) correction to the basis set superposition error (BSSE) was determined as follows

$$\mathrm{CP} = (E_{\mathrm{s}}^{\mathrm{sp}'} - E_{\mathrm{s}}^{\mathrm{sp}}) + (E_{\mathrm{r}}^{\mathrm{sp}'} - E_{\mathrm{r}}^{\mathrm{sp}}) \qquad (1)$$

where $E_{\mathrm{r}}^{\mathrm{sp}}$ and $E_{\mathrm{s}}^{\mathrm{sp}}$ are the single-point energies of the ring and shaft in the rotaxane geometry, and $E_{\mathrm{r}}^{\mathrm{sp}'}$ and $E_{\mathrm{s}}^{\mathrm{sp}'}$ are the single-point energies of the ring and shaft in the rotaxane geometry using the rotaxane basis. It is important to note

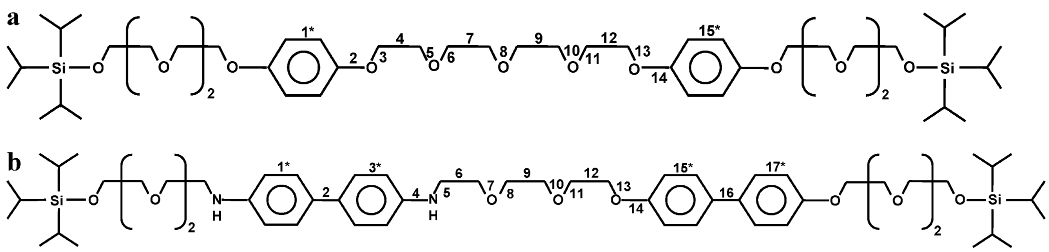


Figure 2. Top - rotaxane **1** shaft, bottom - rotaxane **2** shaft. The numbers on the shafts (a and b) correspond to the bond numbers in the appropriate figures.

that the presence of BSSE results in an overestimation of the interaction energy; therefore, the value obtained from eq 1 reduces the binding energy.

All electronic structure calculations were carried out in delocalized internal coordinates as implemented in the GAMESS suite of codes.[18]

2.2. Empirical Approximations. Empirical approximations were used to gain further insight into the electrostatic and dispersion contributions to the interactions between the components in the rotaxanes and pseudorotaxanes. The two different approximations are discussed and shown in algebraic form below.

2.2.1. Electrostatic Interactions. The electrostatic interactions were approximated by assuming a Coulomb interaction of atomic point charges by employing the following equation

$$E_{\text{electrostatic}} = \sum_{i_{\text{RING}}} \sum_{j_{\text{SHAFT}}} \frac{Q_i Q_j}{r_{ij}} \qquad (2)$$

where Q_i, Q_j, and r_{ij} are the net charge of an atom on the cyclophane ring, the net charge of an atom on the shaft, and the distance between these atoms, respectively. The net charges on each atom were taken to be the MOPAC charges from the AM1 calculations. The distances between the atoms were obtained from the corresponding AM1 optimized structure. It should be noted that since real atoms are not point charges, the equation is only approximate and works best in the long-range limit.

2.2.2. Dispersion Interactions. The dispersion interaction was estimated by employing an empirical approximation set forth by Grimme[19]

$$E_{\text{dispersion}} = -s \sum_{i_{\text{RING}}} \sum_{j_{\text{SHAFT}}} \frac{\sqrt{C^i C^j}}{{r_{ij}}^6} \left(\frac{1}{1 + \mathrm{e}^{-d(r_{ij}/R-1)}} \right) \qquad (3)$$

where C_i and C_j are dispersion coefficients corresponding to the element; r_{ij} and R are the distance between the two atoms and the sum of the atomic vdW radii, respectively; and S and d are scaling factors as reported by Grimme.[19] The equation was validated by applying it to several dimer systems with well-known dispersion interactions as presented in the Results section below.

2.3. Model Building. In order to generate reasonable starting structures for the AM1 calculations, partial optimizations of the components and the complexes were carried out using molecular mechanics (MM) methods with the AMBER force field.[20] It is important to note that MM was only used to generate starting structures not for final optimizations or energy calculations. All Graphical model building and preliminary partial optimizations were carried out with the Hyperchem software package.[21] Molekel advanced 3D-molecular graphics was used for viewing and to create the images provided herein.[22]

3. Computational Methods

The stating structures of the rotaxanes and pseudorotaxanes were obtained by first constructing the ring and shafts separately. The different components were built with a graphical-user-interface molecular editor and partially optimized using molecular mechanics (AMBER force field) to obtain reasonable component starting structures. The separate components were then each fully optimized at the AM1 level. The fully optimized components were then manually assembled into full complexes with a GUI molecular editor, centering the ring about the shaft slightly offset from a binding site. These interlocked structures were again partially optimized using MM to eliminate any close contacts. Finally, the full interlocked structures were fully optimized at the AM1 level. Difficulties in obtaining SCF convergence in the rotaxane structures were overcome by displacing the ring about the shaft in small steps until convergence could be achieved.

Once an optimized rotaxane was obtained, an effective potential energy curve (shuttling barrier) for translation of the ring along the shaft was obtained by mapping the potential energy at the AM1 level of theory with a series of restrained optimizations. Restraints were applied between the ring and the shaft so as to fix the position of the ring relative to specific atoms on the shaft. A set of optimized structures (with restraints) was generated for each position of the ring along the shaft between the two binding sites. At each position along the shaft, the ring was rotated by a random angle about the shaft and reoptimized. The new optimized coordinates were used for the subsequent rotation, and the process was repeated on average 10 times (not all optimizations converged). In some cases the SCF procedure failed to converge, typically because the starting structure that was autogenerated by our rotation procedure was unphysical. These unphysical structures were discarded. After performing constrained optimizations for each position of the ring along the shaft, the structures generated were regrouped by identifying the bond along the shaft to which the centroid of the ring was closest. (This was not always the bond to which the ring was constrained, due to the freedom of movement the remainder of the structure.) A Boltzmann-weighted average energy was obtained for each position of the ring along the shaft. A total of 165 structures was generated and optimized for rotaxane **1**, with a minimum of 5 and a maximum of 36 structures per position about the shaft. One hundred ninety-four structures were generated and optimized for rotaxane **2**, with a minimum of 4 and a maximum of 42 structures per position. The shaft of rotaxane **1** has symmetric connectivity about its midpoint; therefore, the symmetric bond positions were combined for analysis.

For pseudorotaxane **3** (pseudoshaft a) the PEC was mapped in a slightly different manner. The system was built and optimized in the same manner as the rotaxanes. Following Grabuleda and Jaime,[12] the ring was thrice shuttled back and forth along the shaft in a stepwise manner and restrained to specific atoms on the shaft at each step (no rotations of the ring about the shaft were preformed). The same approach was used to determine the closest bond along the shaft to the center of the ring as in the rotaxane systems. A total of 39 structures were generated by the shuttling process. Pseudoshaft a (Figure 3) also has symmetric connectivity about its midpoint allowing symmetric positions to be combined for analysis. As for the rotaxane systems, a

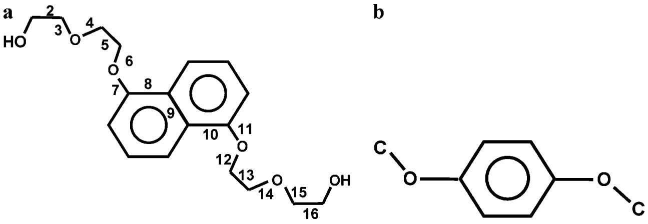


Figure 3. Left - pseudorotaxane **3** shaft, right - pseudorotaxane **4** shaft. The numbers on shaft (a) correspond to the bond numbers in the appropriate figures.

Boltzmann average was again taken of the set of energy values obtained at each position.

For pseudorotaxane **3**, the PEC for ring shuttling along the shaft was also mapped at the HF-SCF level. HF optimizations were performed on the lowest energy structures from each set as identified by the AM1 calculations, applying the same restraints to hold the ring in position. The HF calculations allowed for the BSSE to be determined by performing CP-correction calculations.

Pseudorotaxane **4** contains a very short shaft; therefore, the system was only studied in one relative position (associated). As in the other cases, the shaft was first optimized at the AM1 level and then manually associated with the optimized ring, and the entire complex was fully optimized. Full optimizations were also performed at the HF and DFT levels using the 6-31G(d) basis. In addition, single-point calculations at the HF-SCF level, using various basis sets, were performed on the different optimized structural geometries. CP-correction calculations were performed for all first-principles calculations.

To calibrate the reliability of the empirical dispersion expression (eq 3), four different dimer systems were studied: hydrogen, nitrogen, water, and nitromethane. The dispersion interaction values and reported geometries were found in the following references: hydrogen,[23] nitrogen,[24] water,[25] and nitromethane.[26,27] The reported geometries were used to determine the distances between the atoms of the two monomers as required for application of eq 3.

The water and nitromethane dimers were further studied to examine the consequences of performing single-point calculations on structures resulting from optimizations carried out at a different level of theory. These systems were fully optimized at the AMBER, AM1, HF/6-31G(d), B3LYP/6-31G(d), and B3LYP/6-311G(dp) levels; the water dimer was also optimized at the MP2/6-31G(d,p) level. As in the case of pseudorotaxane **4**, single-point calculations at the HF-SCF level, using various basis sets, were performed on the different optimized structural geometries. All of the calculations were CP-corrected. The results are presented and discussed below.

4. Results

4.1. Rotaxane 1. The shuttling barrier for movement of the ring between the two identical hydroquinone binding stations on the shaft of rotaxane **1** is shown in Figure 4. (A schematic of the shaft is shown in Figure 2a.) The AM1 PEC (Figure 4) shows that the ring preferentially lies at one of the equivalent binding stations. The lowest energy structure obtained from the AM1 method is shown in Figure 5. The energy required for the ring to move from one station to the other was determined to be 10.1 kcal/mol. This is in good agreement with computational results obtained by Grabuleda and Jaime,[12] where they determined the barrier to be 10.8 kcal/mol using molecular mechanics (MM3 force field) simulations. These results are in relatively good agreement with the experimental result of 13 kcal/mol obtained by Anelli.[28] It is important to note that the computational values from the literature and our own AM1 values differ from the corresponding experimental results by less than the standard error in calculations of their type.[29] In other words, the differences are not statistically significant.

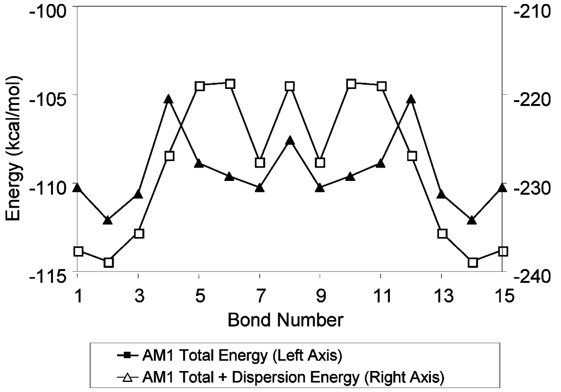


Figure 4. AM1 total energy and AM1 total energy plus dispersion energy for the shuttling process in rotaxane **1**. The bond numbers correspond to the number on the shaft as shown in Figure 2a.

Since van der Waals (dispersion) interactions are neglected in the AM1 method, eq 3 was used to estimate the magnitude of this interaction between the two components. The PEC with the empirical estimate of the dispersion interaction added to the AM1 binding energy is shown in Figure 4. The same preferred binding site is predicted as from the AM1 results alone. With the inclusion of the dispersion interaction, the shuttling barrier is 13.7 kcal/mol. Since the MM calculations of Grabuleda and Jaime[12] considered dispersion interactions, it is most appropriate to compare their result to our AM1 result with the inclusion of dispersion. In this case our result is indeed closer to the experimental value of 13 kcal/mol, although the difference is again not statistically meaningful.

The electrostatic interaction between the components as the ring is moved along the shaft as estimated using eq 2 is shown graphically in Figure 6. The electrostatic interaction also predicts the ring to lie preferentially at one of the equivalent binding sites. Note that the electrostatic energy as estimated with the Coulomb expression (eq 2) and the MOPAC point-charge approximation (obtained form the AM1 calculations) predicts qualitatively the same binding site preference as the AM1 calculation.

4.2. Rotaxane 2. The barrier to ring-shuttling in rotaxane **2** was mapped in a similar manner as for rotaxane **1**. The resulting PEC based on AM1 with/without inclusion of dispersion is shown in Figure 7. Rotaxane **2** contains benzidine and 4,4′-biphenol binding stations; a schematic of the shaft is shown in Figure 2b. The ring is predicted to reside

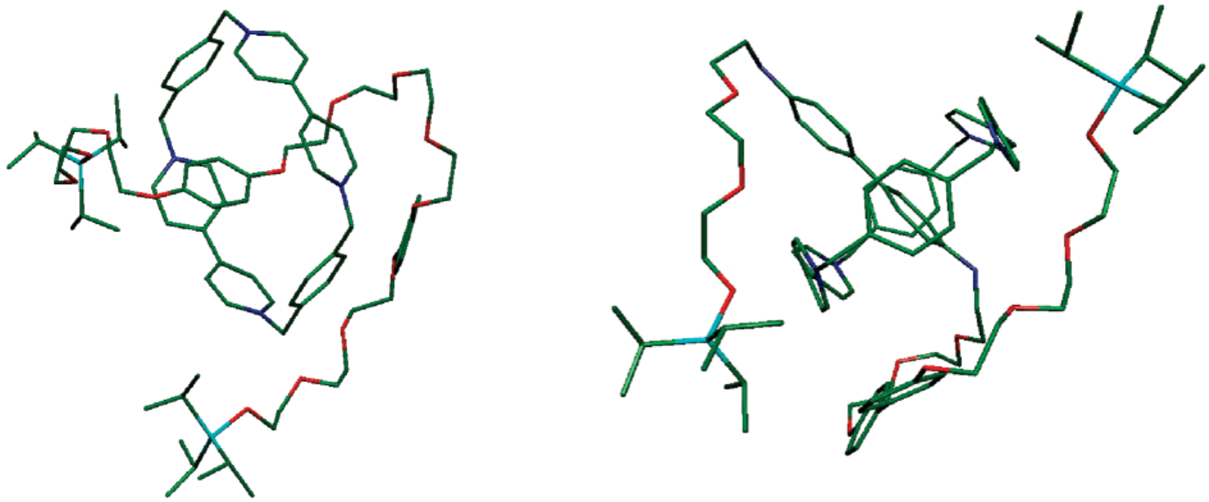


Figure 5. AM1 fully optimized structures of [2]rotaxane **1** (shaft **a**) on the left and [2]rotaxane **2** (shaft **b**) on the right.

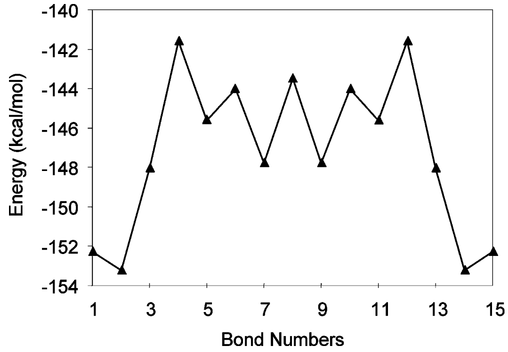


Figure 6. Electrostatic interaction energy between the ring and shaft in rotaxane **1** as calculated with eq 2.

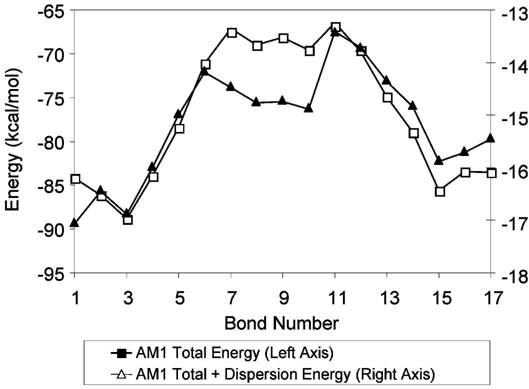


Figure 7. AM1 binding energy and AM1 binding energy plus dispersion energy for the shuttling process in rotaxane **2**. The bond numbers correspond to the number on the shaft as shown in Figure 2b.

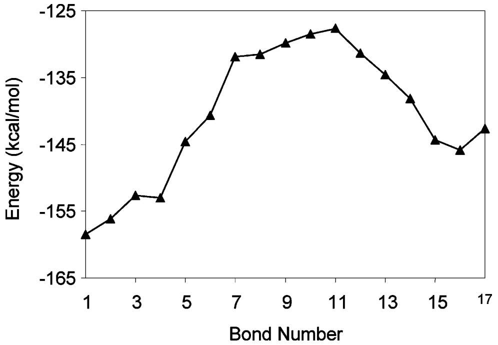


Figure 8. Electrostatic interaction energy between the ring and shaft in rotaxane **2** as calculated with eq 2.

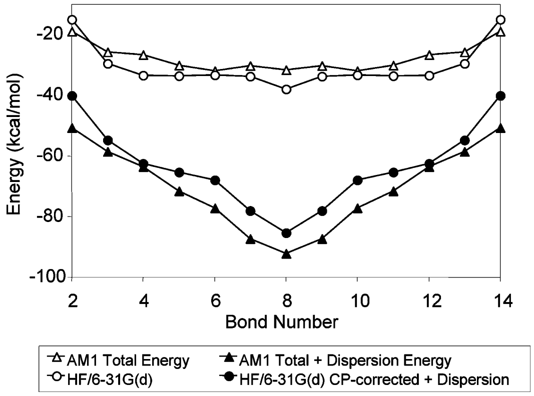


Figure 9. Upper curves - AM1 binding energy and AM1 binding energy plus dispersion energy for the shuttling process in pseudorotaxane **3**. The bond numbers correspond to the number on the shaft as shown in Figure 3a. Lower curves - HF/6-31G(d) binding energy and HF/6-31G(d) binding energy plus dispersion energy of pseudorotaxane **3** as a function of bond number (shown in Figure 3a).

preferentially at the benzidine station both with and without the dispersion contribution included. The predicted binding site preference is in agreement with experimental data.[30] The benzidine station was found to be favored by 3.2 kcal/mol at the AM1 level and by 11.9 kcal/mol when the dispersion interaction was included. A Boltzmann distribution based on either result reveals that the ring resides at the benzidine site virtually 100% of the time at 300 K. The lowest energy structure obtained from the AM1 calculations is shown in Figure 5, where it can be seen how the ring orients relative to the benzidine station. The electrostatic interaction energy, as calculated using the empirical eq 2 for various positions of the ring relative to the shaft, is shown in Figure 8. As for rotaxane **1**, The electrostatic interaction predicts the same preferred binding site (benzidine).

In this [2]rotaxane system, the two binding sites are nonidentical. This provides the opportunity to control the

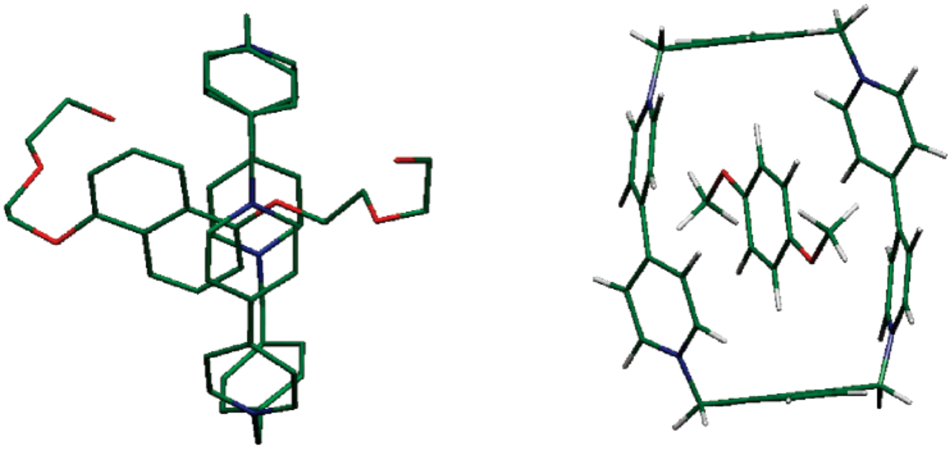


Figure 10. AM1 fully optimized structures of [2]pseudorotaxane **3** (shaft **c**) on the left and [2]pseudorotaxane **4** (shaft **d**) on the right.

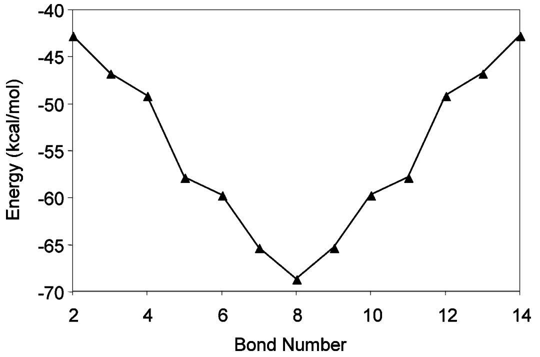


Figure 11. Electrostatic interaction energy between the ring and shaft in pseudorotaxane **3** as calculated with eq 2.

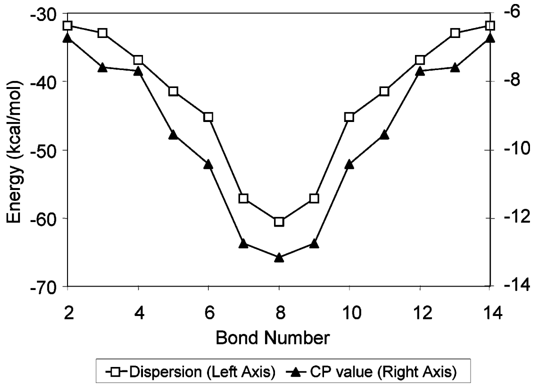


Figure 12. CP-correction energy and dispersion interaction energy as a function of the bond number (shown in Figure 3a) for pseudorotaxane **3**. Note that while the magnitudes differ significantly (the scales on the left and right vertical axes differ), qualitatively the CP correction tracks the dispersion energy very closely.

binding site preference by some external stimulus. Upon oxidation of this system, the benzidine station becomes positively charged, repelling the tetracationic ring to the 4,4′-biphenol station. This has been shown to occur experimentally by Bissell et al.[31] Semiempirical AM1 electronic structure calculations indeed show a very strong repulsion of the ring from the positively charged benzidine site when the system is in its oxidized state, resulting in restabilization of the ring around the 4,4′-biphenol station.

4.3. Pseudorotaxane 3. A schematic of the shaft of pseudorotaxane **3** is shown in Figure 3a. This system contains a 1,5-dioxynaphthalene binding station. Potential energy curves were generated for movement of the tetracationic ring relative to this shaft using both the AM1 and HF-SCF methods. The AM1 PEC with/without the inclusion of the empirical dispersion term is shown in Figure 9. This figure shows that the lowest energy structure corresponds to the ring positioned at the binding station (AM1 optimized structure, Figure 10). The electrostatic interaction as estimated with eq 2 also predicts the ring to bind at the 1,5-dioxynaphthalene binding station. The variation in the electrostatic interaction as the ring is shuttled along the shaft is shown in Figure 11. The geometries used for these electrostatic interaction calculations were obtained from the AM1 calculations.

To further investigate the root of the intercomponent binding, HF/6-31G(d) optimizations were carried out starting from the lowest energy AM1 structures at each position of the ring along the shaft. The PEC at the HF level of theory is shown in Figure 9. This figure also shows the CP-corrected HF energy plus the calculated empirical dispersion contribution. Both cases predict the ring to lie preferentially at the binding station. Note also the striking similarity between the AM1 and HF results.

Table 1. Calculated Binding Energies of Pseudorotaxane **4** Resulting from Full Optimizations at the AM1, HF/6-31G(d), and DFT/6-31G(d) Levels of Theory[a]

	AM1	HF/6-31G(d)	B3LYP/ 6-31G(d)
ring	−214.42887	−1598.46148	−1607.77949
sp. ring w/ring basis	N/A	−1598.40103	−1607.77831
sp. ring w/system basis	N/A	−1598.40599	−1607.78329
SHAFT	−66.22266	−458.45968	−461.01456
sp. shaft w/ring basis	N/A	−458.45984	−461.01785
sp. shaft w/system basis	N/A	−458.46420	−461.02378
pseudorotaxane	−280.66075	−2056.93618	−2068.81641
Δ***E*** (kcal/mol)	−5.79	−9.43	−14.03
Δ***E*** CP-corrected (kcal/mol**)**	N/A	−3.58	−7.19

[a] Energies in hartrees except as otherwise noted.

Table 2. Raw SP Energies, Binding Energies, and CP-Correction Values for Pseudorotaxane **4** Based on Coordinates Obtained by Full Optimization at Four Different Levels of Theory

	HF/6-31G	HF/6-31G(d)	HF/6-31G(d,p)	HF/6-311G(d,p)	HF/6-311+G(d,p)
AM1 Coordinates					
ring	−1597.79922	−1598.40285	−1598.46088	−1598.72567	−1598.73577
ring[a]	−1597.79746	−1598.40103	−1598.45902	−1598.72402	−1598.73405
ring[b]	−1597.80215	−1598.40599	−1598.46408	−1598.72682	−1598.73728
shaft	−458.26112	−458.44792	−458.46483	−458.55770	−458.56447
shaft[a]	−458.26137	−458.44766	−458.46460	−458.55750	−458.56428
shaft[b]	−458.26547	−458.45188	−458.46876	−458.56139	−458.56714
pseudorotaxane	−2056.07364	−2056.86277	−2056.93790	−2057.29364	−2057.30954
ΔE (kcal/mol)	−8.35	−7.53	−7.65	−6.45	−5.84
CP-correction (kcal/mol)	−5.51	−5.76	−5.79	−4.21	−3.82
ΔE CP-corrected (kcal/mol)	−2.84	−1.77	−1.86	−2.24	−2.02
HF/6-31G(d) Coordinates					
ring	−1597.85461	−1598.46148	−1598.51854	−1598.78375	−1598.79347
ring[a]	−1597.85333	−1598.45998	−1598.51700	−1598.78254	−1598.79218
ring[b]	−1597.85788	−1598.46483	−1598.52196	−1598.78514	−1598.79523
shaft	−458.26971	−458.45968	−458.47628	−458.56935	−458.57617
shaft[a]	−458.27238	−458.45984	−458.47646	−458.56950	−458.57624
shaft[b]	−458.27653	−458.46420	−458.48082	−458.57337	−458.57901
pseudorotaxane	−2056.14286	−2056.93618	−2057.01003	−2057.36640	−2057.38170
ΔE (kcal/mol)	−11.63	−9.43	−9.54	−8.34	−7.57
CP-correction (kcal/mol)	−5.46	−5.78	−5.84	−4.06	−3.66
ΔE CP-corrected (kcal/mol)	−6.17	−3.65	−3.70	−4.28	−3.91
B3LYP/6-31G(d) Coordinates					
ring	−1597.84630	−1598.44991	−1598.50706	−1598.77161	−1598.78146
ring[a]	−1597.84540	−1598.44895	−1598.50608	−1598.77075	−1598.78065
ring[b]	−1597.85002	−1598.45386	−1598.51109	−1598.77350	−1598.78376
shaft	−458.26870	−458.45630	−458.47303	−458.56592	−458.57278
shaft[a]	−458.27068	−458.45592	−458.47266	−458.56553	−458.57232
shaft[b]	−458.27544	−458.46092	−458.47764	−458.56997	−458.57524
pseudorotaxane	−2056.13369	−2056.92159	−2056.99566	−2057.35096	−2057.36600
ΔE (kcal/mol)	−11.73	−9.65	−9.78	−8.42	−7.38
CP-correction (kcal/mol)	−5.89	−6.22	−6.27	−4.51	−3.78
ΔE CP-corrected (kcal/mol)	−5.84	−3.43	−3.51	−3.91	−3.60
Ercolani Coordinates B3LYP/6-31G(d,p)					
ring	−1597.84745	−1598.45117	−1598.50832	−1598.77284	−1598.78273
ring[a]	−1597.84636	−1598.44996	−1598.50709	−1598.77178	−1598.78168
ring[b]	−1597.85099	−1598.45488	−1598.51210	−1598.77454	−1598.78482
shaft	−458.27057	−458.45732	−458.47404	−458.56689	−458.57363
shaft[a]	−458.27089	−458.45611	−458.47284	−456.17202	−458.57249
shaft[b]	−458.27573	−458.46117	−458.47789	−456.17545	−458.57542
pseudorotaxane	−2056.13489	−2056.92276	−2056.99681	−2057.35213	−2057.36710
ΔE (kcal/mol)	−10.58	−8.95	−9.07	−7.78	−6.73
CP-correction (kcal/mol)	−5.94	−6.27	−6.31	−3.88	−3.82
ΔE CP-corrected (kcal/mol)	−4.64	−2.69	−2.76	−3.90	−2.92

[a] Ring/shaft energy in system geometry with ring/shaft basis. [b] Ring/shaft energy in system geometry with pseudorotaxane basis. Energies in hartrees except as otherwise noted. Ercolani's coordinates can be found in ref 32.

To provoke a discussion of why calculations that neglect dispersion predict the correct binding site preference, a graph of the dispersion interaction (as obtained from the empirical eq 3 using the AM1 optimized geometries) and the CP-correction values (HF-SCF calculations) is shown in Figure 12. While the magnitudes differ significantly (note the different energy scales on the left and right vertical axes of Figure 12), this figure shows that the CP-correction qualitatively tracks the dispersion interaction very closely.

4.4. Pseudorotaxane 4. As mentioned earlier, pseudorotaxane **4** was only studied in one relative position (associated). The smaller size of this system allowed higher order calculations to be performed. The system and its components were optimized at the AM1, HF/6-31G(d), and DFT/6-31G(d) levels of theory, with the HF and DFT calculations being CP-corrected. The energy of the system and each component as well as the calculated binding energies are reported in Table 1. The effect of performing single-point calculations on structure geometries obtained at another level of theory was then investigated. Such comparisons are shown in Tables 2 and 3. In Table 2, four optimized geometries (AM1, HF/6-31G(d), B3LYP/6-31G(d), and coordinates

Table 3. Effect of Performing Single-Point Calculations on Pseudorotaxane **4** at Different Optimized Geometries[a]

	system	ring	shaft
AM1	−280.66075	−214.42887	−66.22266
AM1//HF/6-31G*	−280.57939	−214.36381	−66.20724
energy difference (kcal/mol)	−51.05	−40.83	−9.68
AM1	−280.66075	−214.42887	−66.22266
AM1//B3LYP/6-31G*	−280.62485	−214.40041	−66.21672
energy difference (kcal/mol)	−22.53	−17.86	−3.73
HF/6-31G*	−2056.93618	−1598.46148	−458.45968
HF/6-31G*//B3LYP/ 6-31G*	−2056.92159	−1598.44991	−458.45630
energy difference (kcal/mol)	−9.16	−7.26	−2.12
B3LYP/6-31G(d)	−2068.81641	−1607.77949	−461.01456
B3LYP/6-31G(d)//HF/ 6-31G(d)	−2068.80261	−1607.76882	−461.01141
energy difference (kcal/mol)	−8.66	−6.70	−1.97

[a] Energies in hartrees except as otherwise noted.

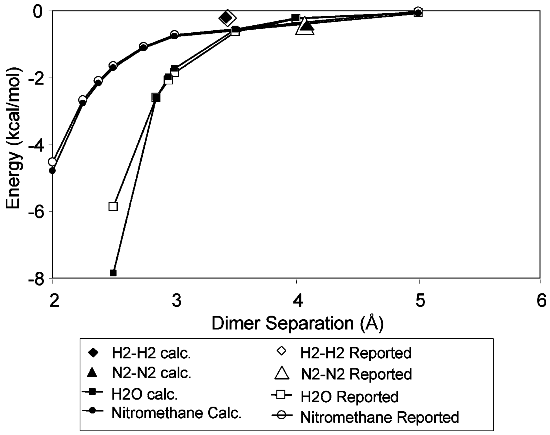


Figure 13. Dispersion interaction energy in the hydrogen, nitrogen, water, and nitromethane dimers as calculated using the empirical approximation of Grimme[19] with comparison to accurate values from the literature: $(H_2)_2$,[23] $(N_2)_2$,[24] $(H_2O)_2$,[25] $(CH_3NO_2)_2$.[26,27]

reported by Ercolani and Mencarelli,[32] who used the B3LYP/ 6-31g(d,p) method) were subjected to single-point calculations using various basis sets at the HF-SCF level: HF/6-31G, HF/6-31G(d), HF/6-31G(d,p), HF/6-311G(d,p), and HF/ 6-311+G(d,p). From these data, the difference in energy between a calculation on a fully optimized structure and single-point calculations based on structures obtained from optimization at a different level of theory may be determined. As shown in Table 3, this energy difference typically exceeds the binding energy of the complex. In the most extreme case the energy difference is 51 kcal/mol. This large shift is potentially critical as will be discussed later.

4.5. Dimer Systems. To validate the empirical dispersion expression due to Grimme,[19] Figure 13 compares the *inter*component dispersion interaction for four dimer systems (hydrogen, nitrogen, water, and nitromethane) as calculated by eq 3 with accurate values from the literature. For the $(H_2)_2$[23] and $(N_2)_2$[24] systems, the experimental binding energy is taken to be the "correct" value of the dispersion interaction since the intermonomer interaction is essentially pure dispersion in these cases. For the water[25] and nitromethane[26,27] dimers, accurate dispersion interaction energies were taken from SAPT calculations in the cited references. These results suggest that this empirical dispersion expression will predict with good accuracy the dispersion interaction between the ring and the shaft, as long as there are no close contacts, a condition that is expected to be fulfilled in the present case, i.e., *inter*component interactions in nonbonded rotaxane and pseudorotaxane complexes.

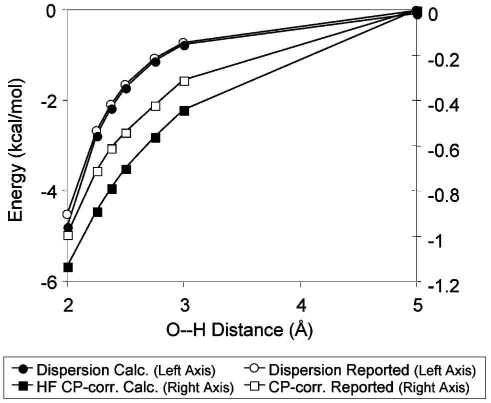


Figure 14. Dispersion interactions and CP correction as a function of the nitromethane dimer separation as calculated using the empirical approximation of Grimme[19] with comparison to accurate values from the literature.[26,27]

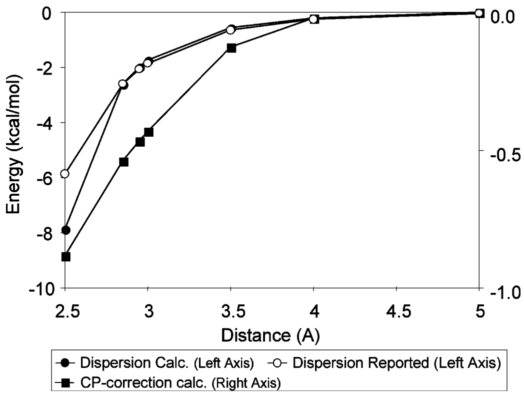


Figure 15. Dispersion interactions and CP correction as a function of the water dimer separation as calculated using the empirical approximation of Grimme[19] with comparison to accurate values from the literature.[25]

To help better understand the different contributions to intercomponent binding in the rotaxane and pseudorotaxane systems, several different types of calculations were performed on the water and nitromethane dimers. Single-point calculations (CP-corrected) were carried out at the HF/6-31G(d) level based on previously reported[25−27] dimer and monomer geometries. Figure 14 shows a graph of the calculated CP-correction, and dispersion interaction as estimated with eq 3, as a function of nitromethane dimer separation and compares these to accurate values from the literature (from refs 26 and 27). Figure 15 shows the same information for the water dimer; however, no CP-correction values were found in the literature for comparison. The

Table 4. Binding Energies of the Nitromethane Dimer at Different Level Theory (Optimized Calculations)[a]

	AM1	HF/6-31G(d)	B3LYP/6-31G(d)
monomer	−37.27434	−243.66198	−244.89022
dimer	−74.55654	−487.33076	−489.78855
Δ*E* (kcal/mol)	−4.93	−4.26	−5.09
Δ*E* CP-corrected (kcal/mol)	N/A	−3.46	−3.21

[a] Energies in hartrees except as otherwise noted.

reported dispersion interaction energies for the various geometries were obtained from refs 33 and 25. In both the water and nitromethane cases, the CP-correction qualitatively tracks the dispersion interaction, although it is quantitatively smaller. This tracking is similar to the result obtained for pseudorotaxane **3**.

Further calculations were undertaken to determine the consequences of performing SP calculations using structures optimized with a different level of theory and to explore the degree to which the calculations are converged with respect to basis set completeness. The water and nitromethane dimer systems were optimized at the AM1, HF/6-31G(d), and B3LYP/6-31G(d) levels, and the intercomponent binding energies were determined for comparison. The water dimer was also optimized at the MP2/6-31G(d,p) level. The calculated binding energies for nitromethane and water, both with and without CP-correction, are reported in Tables 4 and 6, respectively. The several optimized geometries were then subjected to single-point calculations using various basis sets at the HF-SCF level: HF/6-31G, HF/6-31G(d), HF/6-31G(d,p), HF/6-311G(d,p), and HF/6-311+G(d,p). The intercomponent interaction energies at the different levels of theory, for the nitromethane and water dimers, are tabulated in Tables 5 and 7, respectively.

To assess the reliability of structures obtained by performing optimization with a different level of theory, single-point calculations were carried out at the B3LYP/6-311G(dp) level on optimized geometries of the water and nitromethane dimers obtained at different levels of theory: AMBER, AM1, and HF/6-31G(d). Highly accurate published geometries of the water[25] and nitromethane[26,27] dimers were also considered. In Table 8, the total energies (B3LYP/6-311G(dp) single-point calculations) of the different dimer structures are reported. Since the literature geometries are assumed to be the most accurate, the absolute energy differences between the energy of the best published structure and that obtained at each other level of theory was calculated. The largest magnitude difference therefore corresponds to the least

Table 5. Binding Energies of the Nitromethane Dimer Based on Coordinates Obtained by Full Optimization at Four Different Levels of Theory

	HF/6-31G	HF/6-31G(d)	HF/6-31G(d,p)	HF/6-311G(d,p)	HF/6-311+G(d,p)
			AM1 Coordinates		
monomer	−243.52022	−243.65669	−243.66170	−243.72443	−243.73118
monomer A[a]	−243.51984	−243.65606	−243.66109	−243.72377	−243.73055
monomer A[b]	−243.52064	−243.65675	−243.66179	−243.72456	−243.73090
monomer B[a]	−243.51983	−243.65607	−243.66109	−243.72377	−243.73055
monomer B[b]	−243.52065	−243.65675	−243.66179	−243.72456	−243.73090
dimer (hartree)	−487.04999	−487.31917	−487.32954	−487.45532	−487.46795
Δ*E* (kcal/mol)	−6.00	−3.63	−3.85	−4.06	−3.52
CP-correction (kcal/mol)	−1.01	−0.86	−0.88	−0.99	−0.43
Δ*E* CP-corrected (kcal/mol)	−4.99	−2.77	−2.97	−3.07	−3.08
			HF/6-31G(d) Coordinates		
monomer	−243.52336	−243.66198	−243.66688	−243.72983	−243.73633
monomer A[a]	−243.52349	−243.66190	−243.66680	−243.72972	−243.73626
monomer A[b]	−243.52423	−243.66254	−243.66744	−243.73038	−243.73654
monomer B[a]	−243.52349	−243.66190	−243.66680	−243.72972	−243.73626
monomer B[b]	−243.52422	−243.66254	−243.66744	−243.73038	−243.73654
dimer (hartree)	−487.05615	−487.33076	−487.34075	−487.46683	−487.47914
Δ*E* (kcal/mol)	−5.92	−4.26	−4.39	−4.51	−4.06
CP-correction (kcal/mol)	−0.93	−0.80	−0.80	−0.83	−0.35
Δ*E* CP-corrected (kcal/mol)	−4.99	−3.46	−3.59	−3.68	−3.71
			B3LYP/6-31G(d) Coordinates		
monomer	−243.52562	−243.65690	−243.66182	−243.72348	−243.73021
monomer A[a]	−243.52558	−243.65655	−243.66148	−243.72313	−243.72989
monomer A[b]	−243.52644	−243.65731	−243.66224	−243.72395	−243.73028
monomer B[a]	−243.52558	−243.65656	−243.66148	−243.72313	−243.72990
monomer B[b]	−243.52643	−243.65730	−243.66224	−243.72394	−243.73028
dimer (hartree)	−487.06087	−487.32047	−487.33063	−487.45418	−487.46678
Δ*E* (kcal/mol)	−6.05	−4.19	−4.39	−4.53	−3.99
CP-correction (kcal/mol)	−1.08	−0.94	−0.95	−1.03	−0.48
Δ*E* CP-corrected (kcal/mol)	−4.97	−3.25	−3.43	−3.50	−3.51

[a] Monomer energy with dimer geometry and monomer basis. [b] Monomer energy with dimer geometry and dimer basis. Energies in hartrees except as otherwise noted.

Table 6. Binding Energies of the Water Dimer at Different Level Theory (Optimized Calculations)[a]

	AM1	HF/6-31G(d)	B3LYP/6-31G(d)	MP2/6-31G(dp)
monomer	−12.80931	−76.01075	−76.37192	−76.21979
dimer	−25.62733	−152.03046	−152.75596	−152.45080
Δ***E*** (kcal/mol)	−5.46	−5.62	−7.61	−7.05
Δ***E*** CP-corrected (kcal/mol)	N/A	−4.70	−6.84	−5.74

[a] Energies in hartrees except as otherwise noted.

Table 7. Binding Energies of the Water Dimer at Based on Coordinates Obtained by Full Optimization at Four Different Levels of Theory

	HF/6-31G	HF/6-31G(d)	HF/6-31G(d,p)	HF/6-311G(d,p)	HF/6-311+G(d,p)
		AM1 Coordinates			
monomer	−75.98352	−76.01028	−76.02283	−76.04620	−76.05239
monomer A[a]	−75.98341	−76.01019	−76.02272	−76.04608	−76.05227
monomer A[b]	−75.98571	−76.01232	−76.02509	−76.04902	−76.05305
monomer B[a]	−75.98374	−76.01027	−76.02281	−76.04615	−76.05237
monomer B[b]	−75.98496	−76.01155	−76.02419	−76.04784	−76.05268
dimer (hartree)	−151.97218	−152.02163	−152.04623	−152.09342	−152.10253
Δ*E* (kcal/mol)	−3.23	−0.68	−0.36	−0.65	1.41
CP-correction (kcal/mol)	−2.21	−2.14	−2.36	−2.90	−0.69
Δ*E* CP-corrected (kcal/mol)	−1.02	1.46	2.00	2.25	2.10
		HF/6-31G(d) Coordinates			
monomer	−75.98429	−76.01075	−76.02357	−76.04700	−76.05325
monomer A[a]	−75.98430	−76.01072	−76.02352	−76.04693	−76.05318
monomer A[b]	−75.98470	−76.01107	−76.02369	−76.04725	−76.05338
monomer B[a]	−75.98442	−76.01074	−76.02355	−76.04697	−76.05322
monomer B[b]	−75.98565	−76.01187	−76.02496	−76.04900	−76.05397
dimer (hartree)	−151.98031	−152.03046	−152.05597	−152.10284	−152.11387
Δ*E* (kcal/mol)	−7.36	−5.62	−5.54	−5.54	−4.63
CP-correction (kcal/mol)	−1.02	−0.92	−0.99	−1.47	−0.60
Δ*E* CP-corrected (kcal/mol)	−6.34	−4.70	−4.55	−4.06	−4.03
		B3LYP/6-31G(d) Coordinates			
monomer	−75.98324	−76.00975	−76.02216	−76.04546	−76.05167
monomer A[a]	−75.98306	−76.00939	−76.02173	−76.04500	−76.05122
monomer A[b]	−75.98357	−76.00987	−76.02205	−76.04548	−76.05147
monomer B[a]	−75.98326	−76.00963	−76.02201	−76.04529	−76.05151
monomer B[b]	−75.98510	−76.01122	−76.02404	−76.04794	−76.05217
dimer (hartree)	−151.97704	−152.02776	−152.05236	−152.09864	−152.10900
Δ*E* (kcal/mol)	−6.63	−5.18	−5.04	−4.84	−3.55
CP-correction (kcal/mol)	−1.48	−1.30	−1.47	−1.97	−0.57
Δ*E* CP-corrected (kcal/mol)	−5.15	−3.88	−3.57	−2.87	−2.98
		MP2/6-31G(d,p) Coordinates			
monomer	−75.98364	−76.01029	−76.02284	−76.04620	−76.05240
monomer A[a]	−75.98363	−76.01013	−76.02263	−76.04596	−76.05218
monomer A[b]	−75.98407	−76.01053	−76.02286	−76.04635	−76.05242
monomer B[a]	−75.98368	−76.01019	−76.02271	−76.04604	−76.05225
monomer B[b]	−75.98535	−76.01166	−76.02455	−76.04852	−76.05292
dimer (hartree)	−151.97837	−152.02926	−152.05417	−152.10065	−152.11120
Δ*E* (kcal/mol)	−6.96	−5.45	−5.33	−5.18	−4.02
CP-correction (kcal/mol)	−1.32	−1.17	−1.31	−1.81	−0.57
Δ*E* CP-corrected (kcal/mol)	−5.63	−4.28	−4.02	−3.38	−3.45

[a] Monomer energy with dimer geometry and monomer basis. [b] Monomer energy with dimer geometry and dimer basis. Energies in hartrees except as otherwise noted.

reliable structure. The results are included in Table 8. The results reveal that the AMBER geometries are the most inaccurate in both cases by more than an order of magnitude. The AM1 results were much better but (unsurprisingly) not as accurate as those obtained with HF or B3LYP. (AM1+E_{disp})//AM1 is therefore preferred over (AM1+E_{disp})//AMBER.

5. Discussion

The preferred binding sites of the cyclobis(paraquat-*p*-phenylene) ring in the rotaxane and pseudorotaxane systems considered here are known experimentally.[28,30,34,35] We have recovered the same results by carrying out a systematic series of AM1 calculations. In all cases, the lowest energy structures

Table 8. Single-Point Energy Values at the B3LYP/6-311G(dp) Level for the Water and Nitromethane Dimers at Different Optimized Geometries[a]

	energy (hartree)	abs. energy diff. from reported coordinates (hartree)
Nitromethane Dimer		
B3LYP/6-311G(dp)	−489.9393375	0.0033
B3LYP/6-311G(dp)//HF/6-31G(d)	−489.9320871	0.0039
B3LYP/6-311G(dp)//AM1	−489.9298954	0.0061
B3LYP/6-311G(dp)//AMBER	−489.7303006	0.2057
B3LYP/6-311G(dp)//reported_coords	−489.9359884	0
Water Dimer		
B3LYP/6-311G(dp)	−152.8091787	0.0017
B3LYP/6-311G(dp)//HF/6-31G(d)	−152.8073508	0.0001
B3LYP/6-311G(dp)//AM1	−152.8033472	0.0041
B3LYP/6-311G(dp)//AMBER	−152.7967612	0.0107
B3LYP/6-311G(dp)//reported_coords	−152.8074638	0

[a] The absolute difference in energy between each structure and highly accurate published geometries (reported coordinates) are also shown. The reported geometries can be found in the following refs: nitromethane[26,27] and water.[33]

found exhibit ring binding at the experimentally observed binding site positions as shown in Figures 4, 7, and 9. In each system considered, the total energy of the associated complex was found to be lower than that of the dissociated form. This predicts that the components would spontaneously associate in vacuo (and also in a solvent given sufficiently weak solvent effects). This approach to synthesis is referred to as self-assembly and is the preferred synthesis procedure for these systems.[28,30,34,35]

The success of AM1 in making qualitative predictions of binding site preference is notable because $\pi-\pi$ stacking has been assumed to be the dominant intercomponent interaction in these cases and AM1 neglects the dispersion interaction.

The empirical dispersion approximation developed by Grimme[19] was found to be very accurate for estimating the dispersion interaction in several dimer systems. Given this success, it is assumed to provide a reliable description of the intercomponent dispersion interaction in the rotaxane and pseudorotaxane systems. Figures 4 and 9 show that when the contribution of dispersion is added to the AM1 binding energies using the empirical expression of Grimme, the same preferred binding site(s) was predicted as provided by AM1 only. This suggests that inclusion of the dispersion interaction is required for correct quantitative, but not qualitative, description of the intercomponent binding.

To gain insight into *why* the AM1 calculations correctly predict the correct co-conformational preference even though dispersion is neglected, higher order calculations were performed on the pseudorotaxane systems and on the nitromethane and water dimers.

One possible source of the unexpected binding is Basis Set Superposition Error (BSSE). The BSSE is known to artificially increase the predicted strength of binding in nonbonded (van der Waals) complexes because the basis set for the complex is effectively more complete than that for the monomers.[36] While there is no canonical method of correcting for BSSE in an AM1 calculation because there are no standard semiempirical parameters for integrals involving basis functions that are not on atomic centers, the higher order calculations (HF and DFT) allow for the usual CP-correction to be determined. Figure 12 shows the HF/6-31G(d) CP-correction in pseudorotaxane **3** for various positions of the ring relative to the shaft and compares the CP correction to the magnitude of the dispersion interaction as estimated with eq 3. It can be seen that while the CP correction and the dispersion interaction differ significantly in magnitude, they qualitatively track each other very closely. The same result was also found in the nitromethane and water dimers as shown in Figures 14 and 15. Since the BSSE is not corrected for in the AM1 calculation, it makes an artificial contribution to the intercomponent binding, and this contribution has the same qualitative behavior as dispersion. This result, in part, explains why the correct co-conformational preference is recovered with the AM1 method: If dispersion is neglected, then in the absence of other stronger bonding interactions BSSE will lead to a similar structural preference as would the neglected dispersion term.

A second possible reason for the unexpected success of AM1 in predicting co-conformational preference is that van der Waals interactions may not be solely responsible for the binding preference. First, note that our empirical approximation of the electrostatic interaction predicts the correct binding site preference also. (See Figures 6 and 8.) While this empirical approximation to the electrostatic interaction is not quantitatively correct, it does possess the qualitatively correct long-range functional form. This suggests that there is possibly a nondispersive component to the intercomponent binding. Ercolani and Mencarelli[32] computationally studied several pseudorotaxanes, including pseudorotaxane **4**, by performing B3LYP/6-31G(d,p) optimizations and MP2/6-31G(d,p) single-point calculations. They concluded that “face-to-face interactions depend about one-half on electrostatic and frontier orbital contributions (the latter being more important) and the other half on London dispersion forces”.[32] On the other hand, they found the edge-to-face interaction to be solely due to dispersion. These results support our finding that dispersion is the dominant intercomponent interaction but is augmented by a nondispersive component. The binding interaction obtained from the empirical dispersion equation greatly exceeds the binding energy determined by AM1 or HF.

A second indication that the intercomponent interaction is not purely due to dispersion comes from the series of first-principles calculations reported in Tables 1−3 (and supported by the results from the calculations on the water and nitromethane dimers reported in Tables 4−6). Table 1 shows the interaction energy ΔE as computed with several levels of theory. (ΔE is defined as the energy of the association reaction: the total energy of the associated complex minus the sum of the total energies of the dissociated components. Therefore, $\Delta E < 0$ implies a bound complex.) The results show that HF/6-31G(d) predicts a negative interaction energy and that there is some residual binding even after application of the CP correction. Since HF neglects dispersion, it is reasonable to conclude that there is a nondispersive component to the interaction energy. A similar result is obtained at the B3LYP/6-31G(d) level of theory. Since most popular density functionals are local in nature and therefore neglect long-range dispersion,[37] again we can conclude that there is a nondispersive contribution to the intercomponent interaction.

The above conclusion is in some disagreement with the conclusions of a study by Romero et al.,[13] in which they studied pseudorotaxane **4**, among other systems. Romero et al. claimed that "...the origin of stability of cyclobis(paraquat-*p*-phenylene) inclusion complexes with donor molecules is a dispersion interaction which can contribute up to 100% to the binding".[13] To support this claim they conducted DFT optimization calculations using the BHandHLYP/6-31G(d) functional/basis and performed single-point calculations on the resulting structures using both the HF and LMP2 methods with the 6-311+G(d,p) basis (all reported values were CP-corrected). They reported that the HF/6-311+G(d,p)//BH and HLYP/6-31G(d) calculation on pseudorotaxane **4** reveals a repulsive interaction. Their interpretation is that, since the interaction is attractive ($\Delta E < 0$) at the DFT and MP2 levels but becomes repulsive ($\Delta E > 0$) when the strictly nondispersive HF method is applied to the same structure, the entire binding must result from dispersion interactions. By contrast, we show in Table 2 that the binding is still attractive when HF single-point calculations are carried out on a variety of geometries. Exact coordinates for the structures of Romero et al.[13] were not published, but we found $\Delta E < 0$ for HF−SP calculations carried out with five different basis sets on structures obtained by full optimization at four different levels of theory.

There are at least two indications that pseudorotaxane **4** possesses a nondispersive component to the intercomponent interaction, despite the interpretation of Romero et al.[13] First, the fact $\Delta E < 0$ is predicted at the BH and HLYP/6-31G(d) level of theory indicates that there is a nondispersive component to the intercomponent interaction. Since this interaction is a nonbonding (van der Waals) interaction, it occurs on length scales that are generally longer than the local interaction included in most common density functionals.[37,38] Second, we show in Table 3 that the difference in energy between a fully optimized structure and a single-point calculation based on a structure that was optimized at a different level of theory is significant in all cases considered, in some cases larger than the obtained binding energy. Error occurs in both the components and the system, and often these errors cancel each other out, but reliable prediction of the binding energy relies on such fortuitous cancellation of errors.

6. Conclusions

The AM1 Hamiltonian is computationally efficient yet incorporates an explicit description of the electronic structure of a system. These features render it attractive for application to large systems where multiple charge and electronic states must be considered, as is the case for switchable rotaxanes and pseudorotaxanes. Many rotaxanes exhibit $\pi-\pi$ stacking interactions between the ring and the binding stations on the shaft. $\pi-\pi$ stacking results from dispersion forces. Unfortunately, since AM1 is a HF-based technique, AM1 calculations neglect dispersion. Remarkably, AM1 calculations often still recover the correct binding site preference. We have discovered that this is in part due to basis set superposition error (BSSE). Typically one corrects for BSSE in an intermolecular interaction using the counterpoise (CP) correction, but there is no canonical way to perform the CP correction in an AM1 calculation. AM1 calculations of intercomponent interactions therefore include BSSE. We have shown that the BSSE qualitatively mimics the dispersion that is neglected in the AM1 calculation. While the magnitude of the intercomponent binding is not properly "predicted" by the BSSE, qualitatively correct binding site preference can result. Clearly, this is a theoretically unreliable way to predict the structures of interlocked macromolecular complexes where dispersion dominates the intercomponent interaction. AM1 is more defensible as a technique to model interlocked macromolecular complexes where hydrogen-bonding governs the intercomponent interaction, because AM1 is known to predict qualitatively correct trends among hydrogen-bonding interactions.[14]

This research was supported by National Science Foundation grant CHE0449595 and DuPont Inc.

References

(1) Wenz, G.; Han, B.-H.; Muller, A. Cyclodextrin Rotaxanes and Ployrotaxanes. *Chem. Rev.* **2006**, *106*, 782−817.

(2) Credi, A. Artificial nanomachines based on interlocked molecules. *J. Phys.: Condens. Matter* **2006**, *18*, 1779−1795.

(3) Tian, H.; Wang, Q.-C. Recent progress on switchable rotaxanes. *Chem. Soc. Rev.* **2006**, *35*, 361−374.

(4) Frankfort, L.; Zheng, X.; Sohlberg, K. Mechanical Molecular Nanodevices. *Encycl. Nanosci. Nanotech.* **2004**, *5*, 73−89.

(5) Sohlberg, K.; Lee, K.-H. An introduction to modeling interlocked molecule systems and application to a "molecular elevator". *J. Comput. Theor. Nanosci.* **2006**, *3* (6), 865−873.

(6) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(7) de Federico, M.; Jaime, C.; Free Energy Calculations (FEP and TI): Conformational Preference of a Cyclodextrinic [2]Catenane: A Case Study. *J. Comput. Theor. Nanosci.* **2006**, *3* (6), 874−879.

(8) Shirts, R. B.; Stolworthy, L. D. Conformational Sensitivity of Polyether Macrocycles to Electrostatic Potential: Partial Atomic Charges, Molecular Mechanics, and Conformational Predictions. *J. Inclusion Phenom.* **1995**, *20*, 297−321.

(9) Rappe, A. K.; Goddard, W. A, III. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem* **1991**, *95*, 3358−3363.

(10) Zheng, X.; Sohlberg, K. Origin of Co-Conformational Selectivity in a [3]rotaxane. *J. Phys. Chem. A* **2006**.

(11) Grabuleda, X.; Ivanov, P.; Jaime, C. Shuttling Process in [2]Rotaxanes. Modeling by Molecular Dynamics and Free Energy Perturbation Simulations. *J. Phys. Chem. B* **2003**, *31*, 7582−7588.

(12) Grabuleda, X.; Jaime, C. Molecular Shuttles. A Computational Study (MM and MD) on the Translational Isomerism in Some [2] Rotaxanes. *J. Org. Chem.* **1998**, *63*, 9635−9643.

(13) Romero, C.; Fomina, L.; Fomine, S. How Important Is the Dispersion Interaction for Cyclobis(paraquat-pphenylene)-Based Molecular "Shuttles"? A Theoretical Study. *Int. J. Quantum Chem.* **2005**, 102.

(14) Buemi, G.; Zuccarello, F.; Raudino, A. Hydrogen bonding and rotation barriers: A comparison between MNDO and AM1 results. *J. Mol. Struct. THEOCHEM* **1988**, *164* (3−4), 379−389.

(15) Frankfort, L.; Sohlberg, K. Semiempirical study of a pH-switchable [2]-rotaxane. *J. Mol. Struct. THEOCHEM* **2003**, *621*, 253−260.

(16) Zheng, X.; Sohlberg, K. Modeling of a Rotaxane-based Molecular Device. *J. Phys. Chem.* **2003**, *107*, 1207.

(17) Zheng, X.; Sohlberg, K. Modeling bistability and switching in a [2]catenane. *Phys. Chem. Chem. Phys.* **2004**, *6*, 809 - 815.

(18) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(19) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787−1799.

(20) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, V. C.; Ghio, G.; Alagona, G.; Profeta, S., Jr.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

(21) Hyperchem. *Hyperchem*, *Release 5.01 for Windows*; Hypercube Inc.: 419 Phillip Street, W., Ontario N2L3X2, Canada, 1996.

(22) Flükiger, P. H. P. L.; Portmann, S.; Weber, J. *MOLEKEL 4.0*; 2000.

(23) Diep, P.; Johnsona, J. K. An accurate H2−H2 interaction potential from first principles. *J. Chem. Phys.* **2000**, *112* (10), 4465−4473.

(24) Aquilanti, V.; Bartolomei, M.; Cappelletti, D.; Carmona-Novillo, E.; Pirani, F. The N2−N2 system: An experimental potential energy surface and calculated rotovibrational levels of the molecular nitrogen dimer. *J. Chem. Phys.* **2002**, *117* (2), 615−627.

(25) Rybak, S.; Jeziorski, B.; Szalewicz, K. Many-body symmetry-adapted perturbation theory of intermolecular interactions. H20 and HF dimers. *J. Chem. Phys.* **1991**, *95* (9), 6576−6601.

(26) Cole, S. J.; Szalewicz, K. Correlated calculation of the interaction in the nitromethane dimer. *J. Chem. Phys.* **1986**, *84* (12), 6833−6836.

(27) Cole, S. J.; Szalewicz, K.; Bartlett, R. J. Nitromethane Dimer Potential Energy Surface Studies. *Int. J. Quantum Chem.* **1986**, *30*, 695−711.

(28) Anelli, P. L.; Spencer, N.; Stoddart, J. F. A molecular shuttle. *J. Am. Chem. Soc.* **1991**, *113*, 5131.

(29) Levine, I. N. Comparisons of Methods. In *Quantum Chemistry*, 4th ed.; Prentice Hall: Englewood Cliffs, NJ, 1991; pp 594−607.

(30) Córdova, E.; Bissel, R. A.; Spencer, N.; Ashton, P. R.; Stoddart, J. F.; Kaifer, A. E. Novel Rotaxanes Based on the Inclusion Complexation of Biphenyl Guests by Cyclobis(paraquat-*p*-phenylene). *J. Org. Chem.* **1993**, *58*, 6550−6552.

(31) Bissell, R. A.; Cordova, E.; Kaifer, A. E.; Stoddart, J. F. A Chemically and Electrochemically Switchable Molecular Shuttle. *Nature* **1994**, *369*, (6476), 133−137.

(32) Ercolani, G.; Mencarelli, P. Role of Face-to-Face and Edge-to-FaceAromatic Interactions in the Inclusion Complexation of Cyclobis(paraquat-p-phenylene): A Theoretical Study. *J. Org. Chem.* **2003**, *68*, 6470−6473.

(33) Szalewicz, K.; Cole, S. J.; Kolos, W.; Bartlett, R. J. A Theoretical Study of the Water Dimer Interaction. *J. Chem. Phys.* **1988**, *89*, 3662−3673.

(34) Anelli, P. L.; Ashton, P. R.; Ballardini, R.; Balzani, V.; Delgado, M.; Gandolfi, M. T.; Goodnow, T. T.; Kaifer, A. E.; Philp, D.; Pietraszkiewicz, M.; Prodi, L.; Reddington, M. V.; Slawin, A. M. Z.; Spencer, N.; Stoddart, J. F.; Vicent, C.; Williams, D. J. Molecular Meccano. 1. [2]Rotaxanes and a [2]Catenane Made to Order. *J. Am. Chem. Soc.* **1992**, *114* (1), 193−218.

(35) Ballardini, R.; Balzani, V.; Gandolfi, M. T.; Prodi, L.; Venturi, M.; Philp, D.; Ricketts, H. G.; Stoddaart, J. F. A photochemically driven molecular machine. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1301−1302.

(36) Clark, T. Ab Initio Methods. In *A Handbook of Computational Chemistry*; Wiley: New York, 1985; pp 233−317.

(37) Gerber, I. C.; Ángyán, J. G. London dispersion forces by range-separated hybrid density functional with second order perturbational corrections: The case of rare gas complexes. *J. Chem. Phys.* **2007**, *126* (044103), 1−14.

(38) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of π-π Interactions. *J. Phys. Chem. A* **2006**, *110*, 10345−10354.

CT7001623

JCTC Journal of Chemical Theory and Computation

Geometries of Second-Row Transition-Metal Complexes from Density-Functional Theory

Mark P. Waller,[†] Heiko Braun,[‡] Nils Hojdis,[‡] and Michael Bühl*[,†]

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1,
D-45470 Mülheim an der Ruhr, Germany, and Bergische Universität Wuppertal,
Fachbereich Mathematik und Naturwissenschaften, D-42097 Wuppertal, Germany

Received July 19, 2007

Abstract: A data set of 19 second-row transition-metal complexes has been collated from sufficiently precise gas-phase electron-diffraction experiments and used for evaluating errors in DFT optimized geometries. Equilibrium geometries have been computed using 15 different combinations of exchange-correlation functionals in conjunction with up to three different effective core potentials. Most DFT levels beyond the local density approximation can reproduce the 29 metal−ligand bond distances selected in this set with reasonable accuracy and precision, as assessed by the mean and standard deviations of optimized vs experimentally observed bond lengths. The pure GGAs tested in this study all have larger standard deviations than their corresponding hybrid variants. In contrast to previous findings for first-row transition-metal complexes, the TPSSh hybrid meta-GGA is slightly inferior to the best hybrid GGAs. The ranking of some popular density functionals, for second-row transition-metal complexes, ordered according to decreasing standard deviation, is VSXC ≈ LSDA > BLYP > BP86 > B3LYP ≈ TPSSh > PBE hybrid ≈ B3PW91 ≈ B3P86. When zero-point vibrational corrections, computed at the BP86/SDD level, are added to equilibrium bond distances obtained from a number of density-functional/basis-set combinations, the overall performance in terms of mean and standard deviations from experiment is not improved. For a combined data set comprised of the first- and second-row transition-metal complexes the hybrid functionals B3P86, B3PW91, and the meta-GGA hybrid TPSSh afford the lowest standard deviations.

Introduction

More than 40 years after the birth of modern density functional theory (DFT) the exact exchange-correlation functional remains ever elusive. Therefore DFT currently employs a heuristic approach, spawning an extraordinarily large number of approximate functionals being proposed in the literature. As DFT remains reliant upon judicious validation against experiment (i.e., parametrization), accurate experimental data is vital, and the quality of a particular functional is ultimately connected to the quality of experimental data available. A growing body of literature for a posteriori estimates of errors for particular exchange-correlation functionals forms the basis for critical assessment of conclusions drawn from computational results within a DFT framework.

It is particularly important to evaluate the quality of DFT-derived geometries, as their accuracy may be crucial for further computations of energies or properties. For the important class of transition-metal complexes (a stronghold of modern DFT) this validation is hampered by a scarcity of accurate structure determinations in the gas phase, to which the overwhelming majority of DFT applications would refer. Quite frequently, parameters optimized for pristine molecules are compared to those obtained from X-ray crystallography or neutron diffraction, that is, for structures

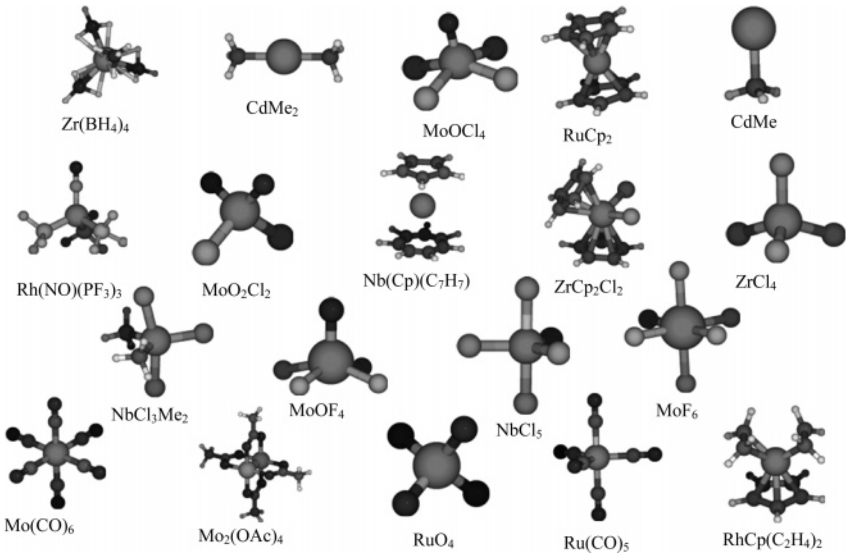
* Corresponding author fax: + (0)208-306 2996; e-mail: buehl@mpi-muelheim.mpg.de.

[†] Max-Planck-Institut für Kohlenforschung.

[‡] Universität Wuppertal.

10.1021/ct700178y CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/19/2007

Chart 1. Data Set 2: Second-Row Transition-Metal Complexes



in the solid with unknown effects from packing forces and intermolecular interactions.

Occasionally, newly developed functionals are also tested against gas-phase geometries but usually only for a small number of complexes (see refs 1−6 for a few illustrative examples). The overall experience with DFT-optimized geometries is that most gradient-corrected (GGA) or hybrid functionals perform reasonably well, albeit with a tendency to overestimate metal−ligand bond distances by several pm, and with deviations typically increasing from metal−C to metal−P bonds.[7] We recently published a study evaluating the ability of DFT to reproduce experimental gas-phase geometries for a test set containing complexes from the first transition row, hereafter paper 1.[8] This study assessed popular density functionals and basis sets in terms of mean and standard deviations between optimized and experimental metal−ligand bond lengths in the gas phase. Drawing from a large compilation of gas-phase structures, from gas-phase electron diffraction (GED) and/or microwave spectroscopy (MW), we proposed a data set comprised of first-row transition-metal complexes. This data set is now referred to as data set 1, encompassing complexes of all 3d metals from Sc to Cu. Only metal−ligand bond lengths that have been determined with a precision better than 1 pm were incorporated, affording a test set of 32 molecules with 50 individual bond distances. Statistical analysis allowed the ranking of a number of popular functionals according to mean and standard deviations from experiment over all these distances.

The comparison between experimental and optimized bond lengths is hampered by the fact that the former refer to thermally averaged quantities, whereas the latter are equilibrium values, i.e., pertaining to vibrationless entities at 0 K.[9] Even if the experiments could be conducted at (or extrapolated to) that temperature, they would still yield structures averaged over the zero-point motion (r_g^0) and could not be directly compared to equilibrium geometries (r_e) from simple energy minimization. There is evidence for small first-row molecules that the zero-point motion actually affords the largest correction to equilibrium distances and that thermal effects on top of them (i.e., the difference between zero and finite T) tend to be much smaller.[10] If this holds also for the transition-metal complexes, computed zero-point corrected geometries would be much better suited for direct comparison with experiment than the raw equilibrium structures.

In a follow-up study for data set 1, we applied vibrational corrections to the equilibrium geometries,[11] using two perturbation methods that have been devised to compute such corrections.[12,13] The standard deviations were not reduced, however, and the relative ordering of functionals did not change with the addition of such vibrational corrections.

We now extend these studies to compounds from the second transition row. In an analogous fashion to data set 1, we selected data set 2 comprising sufficiently precise distances (again better than 1 pm) that have been determined experimentally at room temperature or slightly above (see Chart 1). In many cases, not all degrees of freedom have been refined experimentally, and only mean values for formally nonequivalent distances are known. The final selected experimental parameters are collected in Table 1. Data set 2 does have some notable absences, namely, no complexes containing yttrium, technetium, or silver and, perhaps most unfortunate, no palladium containing complexes. Although with just 19 molecules and 29 bonds, data set 2 is smaller than data set 1, the molecules collated in data set 2 should provide some insight into the relative performance of different density functionals for second-row transition-metal complexes. A number of popular local, gradient-corrected, hybrid, and meta-GGA functionals, together with a variety of effective core potentials (ECPs) and basis sets, are assessed in terms of mean and standard deviation from the corresponding experimental reference values for data set 2, in an analogous fashion to paper 1. We also report computed zero-point corrections to the bond distances for data set 2 in order to furnish increments to estimate r_g^0 from r_e values, thus facilitating the comparison between theory and experiment.

The broad aims of this paper are twofold: first, it is directed toward the identification of functionals, which perform well in terms of optimized geometry, specifically

Table 1. Bond Lengths *r* (in pm) of Second-Row Transition-Metal Complexes in the Gas Phase, as Derived by GED,[a] and Vibrational Corrections Δr_{vib} to Equilibrium Values, Computed at the BP86/SDD Level

ref	compd *symmetry*	parameter	[bond no.][b]	$r_{a/g}$	Δr_{vib}
39	$ZrCl_4$ T_d	r(Zr−Cl)	[1]	232.8(5)	0.18
40	$Zr(BH_4)_4$ T	r(Zr-B)	[2]	232.4(5)	2.74
		$r(\mathrm{Zr{-}H^{br}})$	[3]	214.4(6)	3.22
41	$ZrCp_2Cl_2$ C_2	$r(\mathrm{Zr\text{-}C})^{mean}$	[4]	249.2(9)	1.08
42	$NbCl_5$ D_{3h}	$r(\mathrm{Nb{-}Cl^{ax}})$	[5]	230.6(5)	0.26
		$r(\mathrm{Nb{-}Cl^{eq}})$	[6]	227.5(4)	0.21
43	$NbCl_3Me_2$ C_{2v}	$r(\mathrm{Nb{-}Cl^{ax}})$	[7]	230.4(5)	0.37
		$r(\mathrm{Nb{-}Cl^{eq}})$	[8]	228.8(4)	0.59
		r(Nb-C)	[9]	213.5(9)	0.11
44	$Nb(Cp)(C_7H_7)$ C_s	$r(\mathrm{Nb\text{-}C})^{mean\ c}$	[10]	235.8(2)	0.45
45	MoF_6 O_h	r(Mo−F)	[11]	182.0(3)	0.17
46	$MoOF_4$ C_{4v}	r(Mo═O)	[12]	165.0(7)	0.23
		r(Mo−F)	[13]	183.6(3)	0.17
46	$MoOCl_4$ C_{4v}	r(Mo═O)	[14]	165.8(5)	0.01
		r(Mo-Cl)	[15]	227.9(3)	0.29
47	MoO_2Cl_2 C_{2v}	r(Mo═O)	[16]	168.6(4)	0.12
		r(Mo-Cl)	[17]	225.8(3)	0.32
48	$Mo_2(OAc)_4$ C_4	$r(\mathrm{Mo}\overset{4}{-}\mathrm{Mo})$	[18]	207.9(3)	0.26
		r(Mo−O)	[19]	210.8(3)	0.35
49	$Mo(CO)_6$ O_h	r(Mo−C)	[20]	206.3(3)	0.46
50	RuO_4 T_d	r(Ru═O)	[21]	170.6(3)	0.30
51	$Ru(CO)_5$ D_{3h}	$r(\mathrm{Ru{-}C^{ax}})$	[22]	195.0(9)	0.42
		$r(\mathrm{Ru{-}C^{eq}})$	[23]	196.9(3)	0.45
52	$RuCp_2$ D_{5h}	r(Ru-C)	[24]	219.6(3)	0.59
53	$Rh(NO)(PF_3)_3$ C_3	r(Rh-P)	[25]	224.5(5)	1.00
54	$RhCp(C_2H_4)_2$ C_s	$r(\mathrm{Rh{-}C^{Cp}})$	[26]	226.3(2)	0.87
		$r(\mathrm{Rh{-}C^{C2H4}})$	[27]	210.9(2)	0.77
55	CdMe C_{3v}	$r(\mathrm{Cd\text{-}C})^{d}$	[28]	222.1(7)	2.01
56	$CdMe_2$ D_3	r(Cd-C)	[29]	211.2(4)	0.45

[a] br = bridging, Cp = cyclopentadienyl, ax = axial, eq = equatorial, Me = methyl, OAc = acetate. [b] In brackets: running number of bonds. [c] Doublet ground state. [d] $^2A_1$ state.

for second row-transition-metal complexes; second, we present a combined set of 3d- and 4d-transition-metal complexes in order to assess the performance of functionals over a more diverse range. The consequences of such analysis should serve to assist future DFT studies in preselecting candidate functionals for further more rigorous and system-specific validation. Furthermore, the data set may be useful for the future refinements of density functionals in order to attain better performance for transition-metal complexes.

Computational Details

Geometries were fully optimized in the given symmetry (as given in Table 1) using Gaussian 03[14]and several local (LSDA)[15] and gradient-corrected density functional combinations as implemented therein. Most functionals are composed of one of several exchange parts, namely Becke (B),[16] Becke hybrid (B3),[17] OPTX(O),[18] or OPTX hybrid (O3),[19] together with one of several correlation parts, namely P86,[20] PW91,[21] or LYP[22] (in parentheses: symbols used in combined forms). Other functionals comprise HCTH/407 (denoted HCTH)[3,23] and the PBE hybrid functional[24] (denoted PBE1, Gaussian keyword PBE1PBE) as well as the meta-GGAs BMK,[25] VSXC,[26] TPSS,[27] and TPSS hybrid (denoted TPSSh).[28] A fine integration grid (75 radial shells with 302 angular points per shell) has been used, except for VSXC, which has been shown to require finer grids[29] (here we used 99 radial shells with 590 angular points). The following relativistic small-core ECPs with the corresponding valence basis sets were employed on the metals: LANL2DZ[30] (with [3s3p2d] valence basis), SDD,[31] i.e., the Stuttgart-Dresden ECP (together with the [6s5p3d] valence basis), and CEP-121G (with a [4s4p3d] valence basis).[32] On the ligands, the 6-31G* basis[33] was used throughout, except for the selected cases, where Dunning's double-ζ basis[34] was used on the ligands. In addition, we tested Ahlrichs-type valence basis sets that had been designed for the use with the SDD ECPs,[35] denoted svp, tzvp, and qzvp (with [5s3p2d1f], [6s4p3d1f], and [7s5p4d3f1g] contractions for the metals, respectively), together with the corresponding all-electron bases on the ligands.[36−38] For essentially all combinations of functionals, ECPs, and ligand basis sets, the minimum character of all optimized structures was verified by evaluation of the harmonic vibrational frequencies. Closed- and open-shell species were treated with restricted and unrestricted formalisms, respectively. For the computation of effective geometries via the cubic force field, the Barone method[13] was invoked at the BP86/SDD level within Gaussian 03 revision D.01.[14] The default values were used for step size in the numerical differentiation (0.025 Å) and integration grid (SG1).

Results and Discussion

Equilibrium Bond Lengths. The 29 selected experimental metal−ligand bond distances (mostly r_a and r_g values from GED) are collated in Table 1. In a first step, the corresponding equilibrium bond lengths (r_e), optimized at various levels of DFT, are directly compared to these experimental data. In this first assessment, the accuracy of DFT is investigated via statistical analysis of the error, where this error is defined as r_e-r_{exp}, i.e., a positive value indicates overestimation of the optimized distance compared to experiment. The mean signed and unsigned deviations are given in Table 2, for a number of popular density functionals, together with standard and maximum absolute deviations.

Even though the nature of theoretical and experimental distances is different, a number of conclusions can be drawn from Table 2:

1. Optimized bond distances are, on average, always overestimated (cf. $\bar{\Delta}^{equil}$ values), except with LSDA, which, due to its significant overbinding, produces bond lengths shorter than refined from experiment. This was also clearly observed for data set 1 and has been noted in numerous other studies. Therefore LSDA is not recommended for complexes containing either first- or second-row transition metals. The performance of the VSXC functional is also particularly poor.

2. In agreement with our previous finding for data set 1, the BP86 variant is a promising pure GGA functional in terms of mean and standard deviation, and BPW91 is found to be comparable. Therefore both of these functionals should prove useful in conjunction with the popular resolution of identity (RI) approximation that has been implemented in a

Table 2. Statistical Assessment of Equilibrium (r_e) Metal−Ligand Bond Distances Computed at a Number of Levels of Theory Relative to Experimentally Reported Values (r_{exp})[a]

entry	functional	basis set[b]	$\bar{\Delta}^{equil}$	$\lvert\bar{\Delta}\rvert^{equil}$	$\bar{\Delta}_{std}^{equil}$	Δ_{max}^{equil}
1	PBE1	SDD	0.23	1.26	1.65	4.46 [8]
2	B3P86	SDD	0.54	1.28	1.57	4.10 [4]
3	BP86	SDD	2.41	2.57	2.03	6.51 [4]
4	B3PW91	SDD	0.88	1.42	1.59	4.64 [4]
5	BPW91	SDD	2.33	2.52	2.00	6.35 [4]
6	TPSSh	SDD	1.39	1.79	1.84	4.37 [4]
7	TPSS	SDD	2.10	2.31	1.96	5.19 [4]
8	O3LYP	SDD	1.34	1.88	1.93	5.79 [4]
9	OLYP	SDD	2.13	2.48	2.19	6.86 [4]
10	B3LYP	SDD	2.52	2.63	1.97	7.12 [28]
11	BLYP	SDD	4.53	4.53	2.48	11.47 [28]
12	HCTH	SDD	1.83	2.35	2.49	9.25 [28]
13	BMK	SDD	1.64	2.10	2.11	6.07 [7]
14	VSXC	SDD	3.18	3.18	2.93	16.90 [28]
15	LSDA	SDD	−1.49	2.55	2.70	5.00 [18]
16	B3P86	CEP-121G	1.36	1.58	1.48	4.80 [8]
17	BP86	CEP-121G	3.29	3.29	1.89	6.65 [4]
18	B3P86	LANL2DZ	1.39	1.81	2.55	10.34 [28]
19	B3P86	LANL2DZ[c]	3.61	4.12	3.35	9.06 [28]
20	BP86	LANL2DZ	3.59	3.65	3.39	16.12 [28]
21	BP86	LANL2DZ[c]	5.59	5.76	3.96	14.53 [28]
22	B3P86	svp	−0.40	1.35	1.70	−3.68 [21]
23	BP86	svp	1.52	1.85	1.93	6.00 [28]
24	B3P86	tzvp	−0.38	1.53	1.83	3.55 [4]
25	BP86	tzvp	1.64	2.03	2.10	6.84 [28]
26	BP86	qzvp	1.09	1.65	1.91	5.93 [4]

[a] All units are in picometers. $\bar{\Delta}^{equil}$, $\lvert\bar{\Delta}\rvert^{equil}$, $\bar{\Delta}_{std}^{equil}$, and Δ_{max}^{equil} denote mean, mean absolute, standard, and maximum absolute deviations, respectively. In square brackets: bond numbers from Table 1 for which the maximum error occurs. [b] 6-31G* basis for the ligands, except where otherwise noted. [c] D95 for the ligands.

number of computational chemistry packages.[57] The BLYP and OLYP functionals however have relatively larger errors for data set 2 and are therefore not particularly recommended for geometry optimization of transition-metal complexes.

3. The hybrid functionals perform always better than the corresponding pure GGAs. This is an important point, considering the active research into adapting the RI-approximation to hybrid-GGAs.[58] B3LYP is somewhat inferior to PBE1, B3PW91, and B3P86.

4. The meta-GGAs are not necessarily outperforming the hybrid functionals (such as B3P86 or B3PW91), for example compare entries 2 and 6. This is in direct contrast to previous observations obtained from analysis of data set 1, where the meta-GGAs showed significant improvement over the hybrid-GGAs. Although again here, for data set 2, these functionals do have a low mean deviation and a low standard deviation.

5. The three tested ECPs show increasing mean and standard deviations in the order SDD < CEP-121G < LANL2DZ, both at BP86 and B3P86 levels of theory. Even though the discrimination between SDD and CEP-121G is not very pronounced, the former appears to be the ECP of choice for complexes from the second transition row. In keeping with the findings for data set 1, the LANL2DZ ECP, is found to be inadequate for data set 2, in particular when used with small basis sets on the ligands, and is therefore not recommended for use in geometry optimization of transition-metal complexes.

6. The basis sets employed on the ligands can affect the bond lengths in gas-phase optimization using DFT. In conjunction with the LANL2DZ ECP, the smaller D95 Dunning basis set instead of the more common 6-31G* basis set does result in significant deterioration of agreement between theory and experiment.

7. The difference between the Ahlrichs-type svp, tzvp, and qzvp bases, as assessed by MAD and SD, is rather small when employing the BP86 functional, (compare entries 23, 25, and 26). The difference between the BP86 and B3P86 functionals using a particular Alrichs-type basis set is significantly larger in comparison (e.g. entries 22 vs 23 or 24 vs 25). Despite the rather large size of the qzvp basis set, the results of Table 2 show that it does not necessarily outperform the SDD/6-31G* combination for this data set of second-row transition-metal complexes.

To provide a easy interpretation of the data in Table 2, normal distributions with the same mean and standard deviations are plotted for selected functionals (BP86, B3LYP, TPSS, and TPSSh together with SDD and 6-31G* basis) in Figure 1a. The different ECPs and basis sets (SDD, LANL2DZ, and LANL2DZ:D95) combined with the BP86 functional are plotted in Figure 2b. This nicely summarizes the ability of optimized geometries computed using DFT to reproduce experimentally reported bond lengths for second-row transition-metal complexes.

The mean deviation for the 15 functionals and 3 basis sets (as given in Table 2) averaged over all bonds containing a particular metal are provided in Figure 2, in order to highlight particular transition metals that are particularly challenging for DFT. The gross overestimation of the distances involving Cd can be traced back to the Cd−C bond in CdMe, which, apparently, poses a special problem for DFT (note that the largest Δ_{max} values in Table 2 refer to this bond). Bond lengths involving Ru are particularly well reproduced. The remaining deviations are all around 2−3 pm, DFT is typically overestimating the experimental bond lengths.

Alternatively the ability of DFT to reproduce experimental bond lengths can be subdivided based on the identity of the ligand atom directly coordinated to the metal center. Figure 3 shows the mean deviation over the same levels of theory in Table 2 for the different ligand atom types. For the bonds to hydrogen, boron, and phosphorus, with only one representative, no conclusions can be drawn. The metal−O and metal−C bond distances are around 1 and 2 pm, respectively, longer than experiment on average. The metal−F and metal−Cl distances are slightly worse being, on average, around 3 pm longer than experiment; therefore, complexes which contain a metal halogen bond are more challenging for DFT.

Vibrationally Averaged Bond Lengths. We commence this section with a brief summary of findings from our recently published study of vibrational corrections applied to first-row transition-metal complexes in paper 2:[11] (a) the vibrational corrections are essentially transferable among different density functional and basis set combinations; (b) the corrections are almost exclusively positive (i.e., elonga-

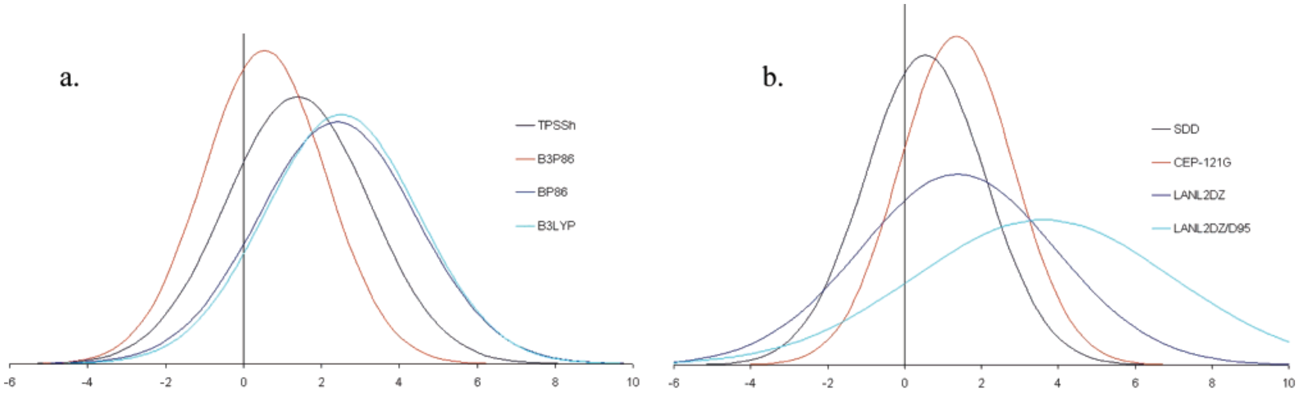


Figure 1. Normal distributions for the errors in the estimated metal−ligand bond lengths for data set 2: (a) displays the effects of different density functional using the SDD pseudopotential and (b) the ECP and basis-set dependence using the B3P86 functional (6-31G* on the ligand except where otherwise noted).

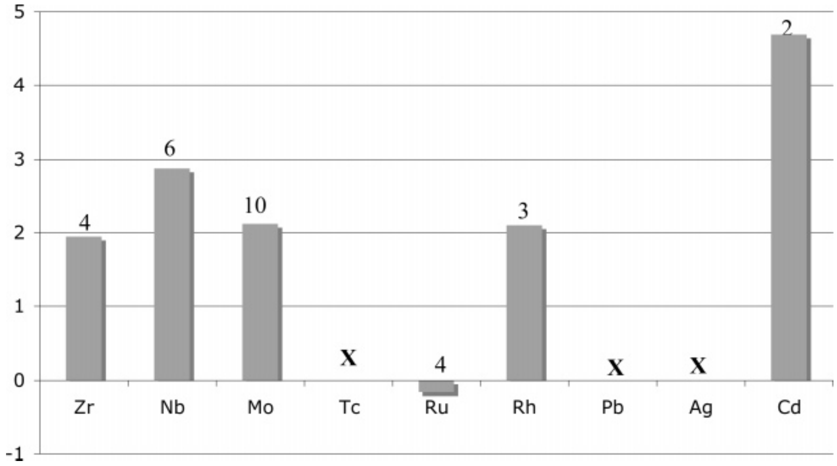


Figure 2. Mean deviation for different density functionals and basis sets combinations for complexes subdivided based on metal center, the number above the columns indicating the number of bonds containing the metal.

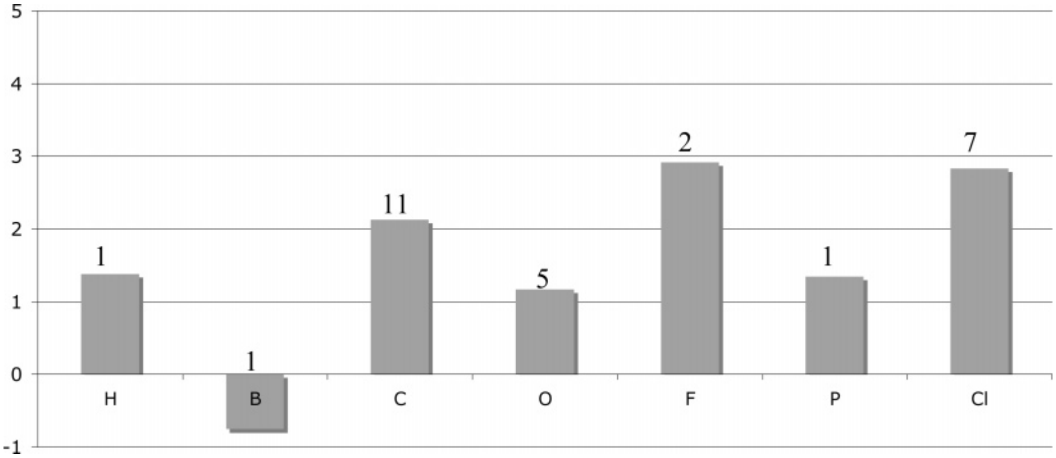


Figure 3. Mean deviation for different density functionals and basis sets combinations for bond lengths grouped according to coordinated ligand atom type, the number above columns indicating the number of bonds containing the element.

tion of equilibrium bond lengths); (c) the corrections do not reduce the standard deviation for data set 1 but simply increase the mean deviation (by ca. 0.5 pm).

For data set 2, the vibrational corrections obtained at the BP86/SDD level are collated in Table 1 (Δr_{vib} values), ranging essentially from zero ($MoOCl_4$: Mo═O) to ca. 2 pm (CdMe: Cd−C), i.e., similar in magnitude to those for data set 1. All effective metal−ligand bonds studied show an increase in bond length relative to the corresponding equilibrium structure, due to the anharmonicity of the potential energy surface. The mean deviation between effective and equilibrium bond lengths (i.e., the vibrational correction) amounts to 0.64 pm, which is 0.15 pm larger in magnitude than for the 3*d* transition-metal data set 1. In all but the LSDA case (using SDD/6-31G*) and the B3P86 functional (in conjunction with the Ahlrichs basis sets), all

Table 3. Statistical Assessment[a] of the Deviation between Estimated Effective Geometries (BP86/SDD Vibrational Correction Added to the Equilibrium Bond Lengths, Which Were Optimized at the Level of Theory Stated in Each Row) and Experimental Bond Lengths for Data Set 2 (Bond Lengths in pm)

entry	functional	basis set[b]	$\bar{\Delta}^{eff}$	$\lvert\bar{\Delta}\rvert^{eff}$	$\bar{\Delta}_{std}^{eff}$
1	PBE1	SDD	0.83	1.44	1.72
2	B3P86	SDD	1.19	1.55	1.57
3	BP86	SDD	3.02	3.13	1.99
4	B3PW91	SDD	1.45	1.71	1.62
5	BPW91	SDD	2.92	3.03	1.97
6	TPSSh	SDD	2.00	2.16	1.77
7	TPSS	SDD	2.57	2.74	1.99
8	O3LYP	SDD	1.94	2.23	2.01
9	OLYP	SDD	2.70	2.88	2.25
10	B3LYP	SDD	3.15	3.20	2.12
11	BLYP	SDD	5.02	5.02	2.69
12	HCTH	SDD	2.44	2.72	2.61
13	BMK	SDD	2.27	2.64	2.33
14	VSXC	SDD	3.49	3.74	3.55
15	LSDA	SDD	−0.78	2.26	2.58
16	B3P86	CEP-121G	2.00	2.02	1.45
17	BP86	CEP-121G	3.95	3.95	1.84
18	B3P86	LANL2DZ	2.12	2.36	2.68
19	B3P86	LANL2DZ[c]	4.08	4.47	3.55
20	BP86	LANL2DZ	4.18	4.21	3.53
21	BP86	LANL2DZ[c]	6.12	6.16	4.12
22	B3P86	svp	0.19	1.38	1.86
23	BP86	svp	2.10	2.30	2.08
24	B3P86	tzvp	0.21	1.54	1.97
25	BP86	tzvp	2.23	2.42	2.21
26	BP86	qzvp	1.66	1.98	2.00

[a] $\bar{\Delta}^{eff}$, $\lvert\bar{\Delta}\rvert^{eff}$, and $\bar{\Delta}_{std}^{eff}$ denote mean, mean absolute, and standard deviations, respectively. [b] 6-31G* for the ligands except otherwise stated. [c] D95 for the ligands.

errors in the equilibrium bond lengths were shown in the previous section to overestimate (to varying extent) experimental bond lengths on average. Therefore the agreement between effective geometries and experimentally refined gas-phase geometries in terms of mean and standard deviation is not improved.

In accord with the findings from data set 1, we assume transferability among functionals and use the vibrational corrections (Δr_{vib}) computed at the BP86/SDD level (Table 1) as increments, which are added to equilibrium bond lengths computed at a number of different levels of theory, affording a set of estimated effective geometries ($r^{est}_{eff} = r_e + \Delta r_{vib}$). In an analogous fashion to paper 2, effective bond lengths are assessed in terms of mean and standard deviations (Table 3). The key result from Table 3 is that applying vibrational corrections does not necessarily improve agreement between experiment and theory and, in fact, decreases agreement for all density functionals except with the LSDA/SDD/6-31G* and BP86/SDD/tzvp combinations. It appears that the vibrational corrections are simply shifting the error distribution in the positive direction without greatly affecting its width.

Combined Data Sets. The selection of a density functional for geometry optimization of molecules in the gas phase is clearly an important issue for computational chemistry. A functional with a narrow error distribution across a wider range of the periodic table is clearly advantageous. To aid in the selection of currently available functionals and to provide benchmarks for future development of functionals, analysis of data set 2 (second-row transition-metal complexes) is combined with the previous data set 1 (first-row transition-metal complexes) and is hereafter referred to as data set 3. As there appears to be no significant improvement between computed and experimentally observed bond lengths upon inclusion of vibrational effects into the theoretical calculations, we now focus on the raw equilibrium values. The corresponding statistical analysis of data set 3 is given in Table 4. The combined data set affords more general conclusions to be drawn regarding the accuracy and precision of modern DFT in transition-metal chemistry.

An overall ranking of functionals for the first- and second-row transition-metal complexes is however dominated by first-row complexes, as the first-row transition-metal complexes are over-represented in the combined data set (data set 3). In order to interpret Table 4 for the ranking of density functionals and basis sets based on their ability to reproduce experimental gas-phase geometries, we primarily consider the standard deviation. The convergence of the Alrichs-type basis sets is remarkably good for the combined data sets, compare entries 17−19, and these basis sets do significantly outperform both the 6-31G* and D95 basis sets, compare entries 2, 3, and 17−19, in conjunction with the BP86 functional. We crudely group the performance of functionals into three categories: not recommended, recommended, and highly recommended.

Not recommended: The LSDA and VSXC functionals again prove inadequate in providing reasonable geometries for first- and second-row transition-metal containing complexes. BLYP and OLYP are found to be inferior to other pure functionals tested within this study and therefore are also not recommended.

Recommended: The BP86 and BPW91 functionals perform comparatively well, and, therefore in cases where it is desirable to invoke the RI approximation, these functionals can be employed with only a marginally greater deviation from experiment, compared to their hybrid counterparts. These GGAs even perform slightly better than the popular B3LYP hybrid functional.

Highly recommended: B3P86, B3PW91, and TPSSh perform almost equally well in terms of standard deviation. The former two functionals are somewhat underestimating the experimental bond lengths (which would be improved upon inclusion of the zero-point corrections discussed above), and the hybrid meta-GGA is slightly overestimating. Therefore we conclude that these three functionals are highly recommended, as they appear most appropriate for geometry optimization of first- or second-row transition-metal complexes.

The normal distributions for three representative functionals are plotted in Figure 4. The normal distributions of a highly recommended (TPSSh), recommended (BP86), and nonrecommended (BLYP) functional provides a good visual indication of the performance.

Table 4. Statistical Assessment[a] of the Deviation between the Equilibrium Bond Lengths, Optimized at the Level of Theory Stated in Each Row, and Experimental Ones for Data Set 3 (Bond Lengths in pm)[b]

entry	functional	3d basis[c]	4d basis[c]	$\bar{\Delta}^{equil}$	$\lvert\bar{\Delta}\rvert^{equil}$	$\lvert\bar{\Delta}\rvert_{std}^{equil}$	$\lvert\Delta\rvert_{max}$
1	B3P86	AE1	SDD	−0.63	1.48	1.78	5.94 [I:35]
2	BP86	AE1	SDD	1.13	1.84	2.08	6.51 [II:4]
3	BP86	SDD	SDD	0.52	1.89	2.39	6.51 [II:4]
4	B3PW91	AE1	SDD	−0.26	1.47	1.81	5.61 [I:35]
5	BPW91	AE1	SDD	1.13	1.84	2.06	6.35 [II:4]
6	TPSSh	AE1	SDD	0.22	1.35	1.79	5.18 [II:8]
7	TPSS	AE1	SDD	0.88	1.54	1.89	5.96 [II:18]
8	O3LYP	AE1	SDD	0.28	1.70	2.07	5.79 [II:4]
9	OLYP	AE1	SDD	1.08	1.96	2.26	6.86 [II:4]
10	B3LYP	AE1	SDD	1.30	1.99	2.11	7.12 [II:28]
11	B3LYP	SDD	SDD	0.66	2.05	2.47	7.12 [II:28]
12	BLYP	AE1	SDD	3.16	3.23	2.42	11.47 [II:28]
13	HCTH	AE1	SDD	0.57	1.92	2.45	9.25 [II:28]
14	BMK	AE1	SDD	0.65	1.99	2.36	6.07 [II:7]
15	VSXC	AE1	SDD	2.19	2.34	2.37	16.90 [II:28]
16	LSDA	AE1	SDD	−2.92	3.32	2.54	9.46 [I:35]
17	BP86	svp	SDD/svp	0.30	1.68	2.11	6.00 [II:28]
18	BP86	tzvp	SDD/tzvp	0.87	1.74	2.05	6.84 [II:28]
19	BP86	qzvp	SDD/qzvp	0.41	1.49	1.87	5.14 [I:1]

[a] $\bar{\Delta}^{equil}$, $|\bar{\Delta}|^{equil}$, $\bar{\Delta}_{std}^{equil}$, and Δ_{max}^{equil} denote mean, mean absolute, standard, and maximum absolute deviations, respectively. [b] The corresponding mean and standard deviations for equilibrium bond lengths (uncorrected) are also shown for comparison. [c] 6-31G* for the ligands, unless otherwise stated. In square brackets: bond numbers from Table 1 in this paper or in paper 1 (labeled II and I, respectively), for which the maximum error occurs.

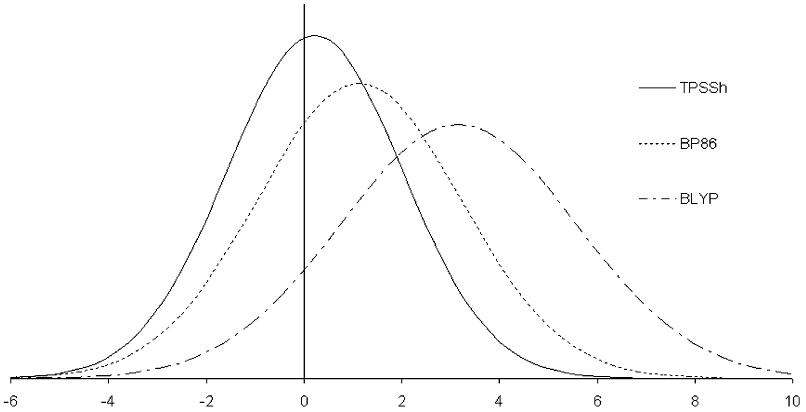


Figure 4. Normal distributions for the errors in the estimated metal−ligand bond lengths for data set 3, illustrating the effects of different density functionals in conjunction with the SDD ECP and 6-31G* basis set.

Conclusions

A new data set of 29 bond lengths comprised of second-row transition-metal complexes is proposed to be a good testing ground for existing density functionals and should be useful for validation of future density functionals or indeed for parametrization or reparametrization of hybrid functionals. Zero-point vibrational corrections (obtained at the BP86/SDD level) serve to increase equilibrium distances by ca. 0−2 pm, depending on the nature of the particular bond. Based on results obtained previously for 3d metals complexes, the vibrational corrections were transferred to the various density-functional/basis-set combinations to create estimated effective (vibrationally averaged) bond parameters. These corrections do not affect the widths of the error distributions (assessed as standard deviations between theoretical and experimental bond lengths) but simply shift these distributions to more positive values. As nearly all the functional trialed in this study on average overestimate the experimental bond lengths, this shift in bond length decreases agreement with experiment.

The combination of data set 2 which contains second-row transition-metal complexes present herein with the previous data set 1 (first-row transition-metal complexes) has created a larger more diverse set which should prove to be a useful testing suite for newly developed density functionals or ab initio methods, in order to assess the relative performance in reproducing the geometries of transition-metal complexes in the gas phase.

Acknowledgment. M.B. wishes to thank Prof. W. Thiel, the MPI Mülheim, and the Deutsche Forschungsgemeinschaft for continuing support. Computations were performed on a local computer cluster of Intel Xeon and Opteron PCs at the MPI Mülheim, maintained by Horst Lenk.

Supporting Information Available: Tables with individual optimized bond distances and BP86/SDD optimized Cartesian coordinates of all complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Rosa, A.; Ehlers, A. W.; Baerends, E. J.; Snijders, J. C.; te Velde, G. J. *J. Phys. Chem.* **1996**, *100*, 5690−5696.

(2) Filatov, M.; Thiel, W. *Phys. Rev. A* **1998**, *57,* 189−199.

(3) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264−6271.

(4) (a) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *Phys. Chem. A* **2005**, *109*, 4388−4403. (b) Schultz, N. E.; Zhao Y.; Truhlar D. G. *Phys. Chem. A* **2005**, *109*, 11127−11143.

(5) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.

(6) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys*. **2007**, *126*, 124115.

(7) For example: Bray, M. R.; Deeth, R. J.; Paget, V. J.; Sheen, P. D. *Int. J. Quantum Chem.* **1997**, *61*, 85−87.

(8) Bühl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282−1290.

(9) The equilibrium distance, r_e, is the distance between the positions of the nuclei on the potential energy surface, as obtained from standard geometry optimizations; r_g is the average internuclear distance at temperature *T*, r_g^0 that at zero K. It is the latter value that our computed effective geometries refer to. Typical quantities derived experimentally are r_a (the effective internuclear distance as derived from electron scattering intensity), r_α (the distance between average nuclear positions in the thermal equilibrium at temperature *T*), r_z (the distance between average nuclear positions in the ground vibrational state), or r_0 (the effective internuclear distance obtained from the rotational constants), see e.g.: Hargittai, I. In *Stereochemical Applications of Gas-Phase Electron Diffraction, Part A: The Electron Diffraction Technique*; Hargittai, I., Hargittai, M., Eds.; VCH Publisher: Weinheim, 1988; pp 1−54.

(10) Toyama, M.; Oka, T.; Morino, Y. *J. Mol. Spectrosc.* **1994**, *13*, 193−213.

(11) Waller, M. P.; Bühl, M. *J. Comput. Chem.* **2007**, *28*, 1531−1537.

(12) (a) Ruud, K.; Åstrand, P.-O.; Taylor, P. R. *J. Chem. Phys.* **2000**, *112*, 2668−2683. (b) Ruud, K.; Åstrand, P.-O.; Taylor, P. R. *J. Am. Chem. Soc.* **2000**, *123*, 4826−4833. (c) Ruden, T.; Lutnæss, O. B.; Helgaker, T. *J. Chem. Phys.* **2003**, *118*, 9572−9581.

(13) (a) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059−3065. (b) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.

(14) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(15) Vosko, S. H.; Wilk L.; Nusair, M. *Can. J. Phys*. **1980**, *58*, 1200−1211. Functional III of that paper used.

(16) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(17) Becke, A. D. *J. Chem. Phys.* **1996**, *98*, 5648−5642.

(18) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403−412.

(19) Cohen, A. J.; Handy, N. C. *Mol. Phys.* **2001**, *99*, 607−615.

(20) (a) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822−8824. (b) Perdew, J. P. *Phys. Rev. B* **1986**, *34*, 7406.

(21) (a) Perdew, J. P. In *Electronic Strucure of Solids;* Ziesche, P., Eischrig, H., Eds.; Akademie Verlag: Berlin, 1991. (b) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244−13249.

(22) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(23) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497−5503.

(24) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396. (c) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys*. **1996**, *105*, 9982−9985. (d) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys*. **1999**, *110*, 5029−5036.

(25) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405−3416.

(26) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400−410.

(27) (a) Tao, J.; Perdew, J. P.; Staroverov V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401. (b) Tao, J.; Perdew, J. P.; Staroverov V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2004**, *120*, 6898−6911.

(28) (a) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 146401. (b) Staroverov, V. N.; Scuseria, G. E.; Tao J.; Perdew, J. P. *J. Chem. Phys.* **2004**, *121*, 11507.

(29) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett*. **2004**, *394*, 334−338.

(30) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299−310.

(31) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.

(32) (a) Stevens, W.; Basch, H.; Krauss, J., *J. Chem. Phys.* **1984**, *81*, 6026. (b) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Che*m. **1992**, *70*, 612. (c) Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993**, *98*, 5555.

(33) (a) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257−2261. (b) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213−222.

(34) Dunning, T. H. In *Modern Theoretical Chemistry*; Schaefer, H. F., Ed.; Plenum Press: New York, 1977; Vol 4, pp 1−27.

(35) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys*. **2005**, *7*, 3297−3305.

(36) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys*. **1992**, *97*, 2571−2577.

(37) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys*. **1994**, *100*, 5829−5835.

(38) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys*. **2003**, *119*, 12753−12762.

(39) Utkin, A. N.; Petrova, V. N.; Girichev, G. V.; Petrov, V. M. *Russ. J. Struct. Chem. (Engl. Transl.*) **1982**, *27*, 660−661.

(40) Haaland, A.; Shorokhov, D. J.; Tutukin, A. V.; Volden, H. V. *Inorg. Chem.* **2002**, *41*, 6646−6655.

(41) Ronova, I. A.; Alekseev, N. V. *Russ. J. Struct. Chem. (Engl. Transl.*) **1977**, *18*, 180−182.

(42) Gove, S. K.; Gropen, O.; Fægri, K.; Haaland, A.; Martinsen, K.-G.; Strand, T. G.; Volden, H. V.; Swang, O. *J. Mol. Struct*. **1999**, *485*−*486*, 115−119.

(43) (a) McGrady, S. G.; Haaland, A.; Verne, H. P.; Volden, H. V.; Downs, A. J.; Shorokhov, D.; Eickerling, G.; Scherer, W. *Chem. Eur. J.* **2005**, *11*, 4921−4934. Reinvestigation from the following: (b) Ischenko, A. A.; Strand, T. G.; Demidov, A. V.; Spiridonov, V. P. *J. Mol. Struct*. **1978**, *43*, 227−243.

(44) Mawhorter, R. J.; Rankin, D. W. H.; Robertson, H. E.; Green, M. L. H.; Scott, P. *Organometallics* **1994**, *13*, 2401−2404. Even though the mean bond length between Nb and the C atoms of the seven-membered ring is reported with an uncertainty below 1 pm in that paper, the difference between the mean Nb−C^{C7H7} and Nb−C^{C5H5} distances is associated with a much larger error. We therefore included the mean value over all Nb−C distances, which could be refined with the quoted precision.

(45) Seip, H. M.; Seip, R. *Acta Chem. Scand*. **1966**, *20*, 2698.

(46) Ijima, K. *Bull. Chem. Soc. Jpn*. **1977**, *50*, 373−375.

(47) Thomassen, H.; Hedberg, K. *J. Mol. Struct*. **1992**, *273*, 197−206.

(48) Kelley, M. H.; Fink, M. *J. Chem. Phys*. **1982**, *76*, 1407−1416.

(49) Seip, S. P.; Seip, H. M. *Acta Chem. Scand*. **1966**, *20*, 2711.

(50) Schäfer, L.; Seip, H. M. *Acta Chem. Scand*. **1967**, *21*, 737.

(51) Huang, J.; Hedberg, K.; Davis, H. B.; Pomeroy, R. K. *Inorg. Chem*. **1990**, *29*, 3923−3925.

(52) Haaland, A.; Nilsson, J. E. *Acta Chem. Scand*. **1968**, *22*, 2653−2670.

(53) Bridges, D. M.; Rankin, D. W. H.; Clement, D. A.; Nixon, J. F. *Acta Crystallogr., Sect. B: Struct. Sci*. **1972**, *B28*, 1130.

(54) Blom, R.; Rankin, D. W. H.; Robertson, H. E.; Perutz, R. N. *J. Chem. Soc*., *Dalton Trans.* **1993**, 1983−1986.

(55) Rotational spectroscopy from high-resolution laser-induced fluorescence spectra, r_0 value: Cerny, T. M.; Tan, X. Q.; Williamson, J. M.; Robles, E. S. J.; Ellis, A. M.; Miller, T. A. *J. Chem. Phys.* **1993**, *99*, 9376−9388.

(56) High-resolution Raman spectroscopy, r_0 value: Suryanarayana Rao, K.; Stoicheff, B. P.; Turner, R. *Can. J. Phys*. **1960**, *38*, 1516.

(57) (a) Dunlap, B. I. *J. Chem. Phys*. **1983**, *78*, 3140. (b) Dunlap, B. I. *J. Mol. Struct.* **2000**, *529*, 37. (c) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652−660.

(58) Weigend, F. *Phys. Chem. Cem. Phys*. **2002**, *4*, 4285−4291.

CT700178Y

JCTC Journal of Chemical Theory and Computation

Charge and Spin Currents in Open-Shell Molecules: A Unified Description of NMR and EPR Observables

Alessandro Soncini*

Department of Chemistry, Laboratory of Quantum Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium

Received July 10, 2007

Abstract: The theory of EPR hyperfine coupling tensors and NMR nuclear magnetic shielding tensors of open-shell molecules in the limit of vanishing spin−orbit coupling (e.g., for organic radicals) is analyzed in terms of spin and charge current density vector fields. The ab initio calculation of the spin and charge current density response has been implemented at the Restricted Open-Shell Hartree−Fock, Unrestricted Hartree−Fock, and unrestricted GGA-DFT level of theory. On the basis of this formalism, we introduce the definition of nuclear hyperfine coupling density, a scalar function of position providing a partition of the EPR observable over the molecular domain. Ab initio maps of spin and charge current density and hyperfine coupling density for small radicals are presented and discussed in order to illustrate the interpretative advantages of the newly introduced approach. Recent NMR experiments providing evidence for the existence of diatropic ring currents in the open-shell singlet pancake-bonded dimer of the neutral phenalenyl radical are directly assessed via the visualization of the induced current density.

1. Introduction

Open-shell molecules are basic components of chemical and biological systems.[1] They are at the heart of many processes, ranging from simple organic reactions[1] up to more complex chemical processes involving enzymes or nucleic acids.[2] Owing to the extremely short-lived nature of these species, magnetic spectroscopies, such as Electron Paramagnetic Resonance (EPR) and Nuclear Magnetic Resonance (NMR), are primary tools for elucidating their structure.[3,4] Whereas the theoretical and computational investigation of magnetic response properties of open-shell molecules has always played a fundamental role in the interpretation of EPR spectra,[5−10] the theory and computation of molecular properties related to the NMR spectroscopy of paramagnetic species has received comparatively much less attention, although progress in computational capabilities and successful investigations of paramagnetic NMR of coordination compounds and metalloproteins have given new impetus to research in this area.[4,11−14]

Among the theoretical strategies that have been devised and applied to unravel the rich phenomenology underlying magnetic spectroscopies, the representation and analysis of magnetic properties of closed-shell molecules via the induced quantum mechanical current density has been pursued since the early days of NMR spectroscopy.[15−17] In fact, direct visualization of the induced current density has been shown to embody several interpretative advantages. For instance, this approach has introduced very successful concepts in chemistry, such as the ring current model[15−18] explaining anomalous proton chemical shifts of aromatic and antiaromatic molecules, and a definition of chemical aromaticity and antiaromaticity based on the magnetic criterion.[19−25]

An analogous approach to the study of NMR or EPR spectra of open-shell molecules has always been discussed mostly on a theoretical basis. Theoretical predictions concerning the existence and the (diatropic) sense of circulation of the induced current density $\mathbf{J}^{(1)}$ in putative [4*n*] aromatic triplets have been proposed by Fowler et al.,[23] on the basis of the ipsocentric model[24,25] for the analysis of ring currents in π-conjugated networks. The zeroth-order spin-current $\mathbf{J}^{(0)}$

* Corresponding author e-mail: Alessandro.Soncini@chem.kuleuven.be.

10.1021/ct700169h CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/17/2007

associated with the ground state of an open-shell system has been considered as a formal starting point to deduce the expression for effective hyperfine spin Hamiltonians in the nonrelativistic limit (see ref 26, Chapter 17, pp 690−692) although no explicit formulation of the property itself (i.e., the tensor components of the hyperfine coupling defined as energy derivatives) in terms of $\mathbf{J}^{(0)}$ has been discussed. A formulation of the EPR *g* tensor up to second order (i.e., the leading contribution for systems whose ground state is orbitally nondegenerate) in terms of the first order induced current density $\mathbf{J}^{(1)}$ has been introduced in ref 27 (see Chapter 11, pp 398−402). This definition, based on second-order perturbation theory, has been extended to the spin-other-orbit contribution to the *g* tensor,[28−30] and a computational recipe for its evaluation using London orbitals to accelerate the convergence toward origin-independent results within the spin-polarized Kohn−Sham (KS) DFT formalism has been provided and implemented.[30] Despite these preliminary efforts, to date, to the best of our knowledge, current densities in open-shell systems have always been used at best as computational byproducts of the calculation of second-order response properties, and no attempt has been made to *visualize* and *rationalize* the magnetic response of open-shell systems in terms of current density vector fields.

It is the purpose of this work to introduce the theoretical and computational framework aimed at the representation of the magnetic properties of open-shell molecules in terms of current density vector fields and show how this theoretical approach can be of great use in providing a unified interpretative model for the rationalization of EPR and NMR parameters. One of the basic features that diversifies the response of a closed-shell from that of an open-shell system consists of the presence of the zeroth-order spin current density $\mathbf{J}^{(0)}$ in the latter.[26] In the limit of very low temperatures, this zeroth-order contribution is independent of the strength of the applied magnetic field, which merely provides the direction of a preferred quantization axis for the total spin of the molecule. Accordingly, neglecting spin−orbit coupling in this first introductory study (a very plausible approximation for a vast class of open-shell molecules, e.g., organic radicals, triplets, and open-shell singlets), any attempt at the visualization of the magnetic response of an open-shell molecule must be inclusive of two main contributions: a zeroth-order spin current density $\mathbf{J}^{(0)}$ related to the unperturbed wave function[26] and the usual first-order charge current density $\mathbf{J}^{(1)}$, which contains the full information related to the linear magnetic response of the system.[27,18] In order to compute these quantities, we propose here an efficient ab initio computational procedure based on the Restricted Open-Shell Hartree−Fock (ROHF) wave function. The calculation $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ has also been implemented at the Unrestricted Hartree−Fock (UHF) level of theory. Finally, once the unperturbed KS α and β spin orbitals are obtained from standard DFT calculation packages within unrestricted GGA-DFT approaches, it is possible to obtain $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ by means of straightforward sum-over-states expressions, since in GGA-DFT theories the response to a magnetic field is rigorously computed within a so-called uncoupled formalism.[31,32]

The unified description in terms of $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ of the magnetization density arising (i) in zero order, from the ground-state electronic spin, and (ii) in first order from the molecular response to an external field allows one to make straightforward use of the Biot-Savart law from classical electrodynamics to reformulate the magnetic response properties of open-shell systems.[15−18,26−30,33−36] In particular, whereas the detailed tensor expression for the temperature-independent contribution to the NMR nuclear magnetic shielding tensor in terms of $\mathbf{J}^{(1)}$ is very well-known (see, e.g., ref 18), we show here that, in the limit of vanishing spin−orbit coupling, (i) the temperature-dependent contribution to the NMR nuclear magnetic shielding of open-shell molecules and (ii) the EPR hyperfine coupling tensor $A^{\mathrm{I}}_{\alpha\beta}$ can both be reformulated in terms of three-dimensional space integrals of a second-rank spin-current density tensor, defined as the formal derivative of $\mathbf{J}^{(0)}$ with respect to the electronic spin component along the direction of the chosen quantization axis, thus providing a rigorous link between the maps of $\mathbf{J}^{(0)}$ and the calculated components of $A^{\mathrm{I}}_{\alpha\beta}$. The current density formulation of the *g* tensor based on second-order perturbation theory introduced in ref 27 and extended to include the spin-other-orbit contributions in refs 28−30 clearly provides the natural theoretical basis for the extension of the present analysis to include spin−orbit coupling, a formulation which would need to be further generalized beyond second-order perturbation theory in order to properly describe, e.g., systems whose ground state is orbitally degenerate. But this goes beyond the scope of the present work.

Finally, following on from the successful application of shielding density functions and spin−spin coupling density functions to the interpretation of NMR observables in closed-shell molecules,[33−36] we introduce here the theoretical definition of hyperfine coupling density $A^{\mathrm{I}}_{\alpha\beta}(\mathbf{r})$, a set of scalar functions of position that can be easily computed from $\mathbf{J}^{(0)}$ and plotted all over the molecular domain to assess the contribution of each point in space to the integrated $A^{\mathrm{I}}_{\alpha\beta}$. As an application of the newly developed methodology, we present (i) ab initio calculations of $\mathbf{J}^{(0)}$, $\mathbf{J}^{(1)}$, and $A^{\mathrm{I}}_{\alpha\beta}(\mathbf{r})$ for three small radicals, BH_2, CH_2^-, and NH_2, to discuss and interpret the physical origin their NMR and EPR observables, and (ii) ab initio and broken symmetry GGA-DFT calculations of $\mathbf{J}^{(1)}$ for the neutral phenalenyl radical and its pancake-bonded dimer (an open-shell singlet), to interpret recent NMR experiments and NICS computational analysis concerning the magnetic aromaticity of these open-shell molecules.[37]

2. Current Density Representation of an Open-Shell Molecule in a Magnetic Field

The ground state of a molecule characterized by a spinless electron density matrix $P(\mathbf{r},\mathbf{r}')$ and a spin density $Q_\gamma(\mathbf{r})$, (γ =x, y, z) is associated with a current density distribution $J_\alpha(\mathbf{r})$ given by[27,38,39]

$$J_\alpha(\mathbf{r}) = -\frac{e}{m}\mathscr{R}\,[p_\alpha P(\mathbf{r},\mathbf{r}')]_{\mathbf{r}'=\mathbf{r}} - \frac{e}{m}\,\epsilon_{\alpha\beta\gamma}\,\nabla_\beta Q_\gamma(\mathbf{r}) \qquad (1)$$

where *m* and −*e* are the mass and charge of the electron, p_α is the electronic linear momentum, $\epsilon_{\alpha\beta\gamma}$ is the Levi-Civita third rank skew tensor, and the Einstein's convention for summation over repeated Greek indices is in force. Let us consider an open-shell molecule. In the absence of orbital degeneracy the first term on the rhs of (1) describes the response of the system to an external magnetic field. Accordingly, the spinless density matrix can be expanded to first order in the field as $P(\mathbf{r},\mathbf{r}')= P^{(0)}(\mathbf{r},\mathbf{r}')+ P^{(1)}(\mathbf{r},\mathbf{r}')$, and an explicit expression for the first term on the rhs of (1) can be provided according to the well-known equations for the first-order induced current density[27,18]

$$J_\alpha^{(1)}(\mathbf{r}) = \frac{e}{mc} A_\alpha(\mathbf{r})P^{(0)}(\mathbf{r}) - \frac{e}{m} \mathscr{R}\,[p_\alpha P^{(1)}(\mathbf{r},\mathbf{r}')]_{\mathbf{r}'=\mathbf{r}} \quad (2)$$

where $A_\alpha(\mathbf{r})$ is the vector potential associated with the external magnetic field, and c is the speed of light.

Let us now turn to the second term[26,38,39] on the rhs of (1). The ground state of an open-shell molecule with total spin quantum number $S\neq 0$ consists of a $2S+1$ degenerate multiplet. Each state of the degenerate set can be characterized by a spin projection quantum number M_S along an arbitrary quantization axis; accordingly, the density of spin angular momentum $Q_\gamma(\mathbf{r})$ will depend on which state we consider. However, in accordance with the Wigner-Eckart theorem, the spin densities are in fact all the same except for a proportionality constant.[27] It is therefore expedient to introduce a reduced spin density scalar function, $Q(\mathbf{r})$, common to all components of the multiplet, and an effective spin density operator $Q_{\mathrm{op},\gamma}(\mathbf{r})$ proportional to $Q(\mathbf{r})$ that differentiates the various components when averaged over the corresponding states. Thus

$$Q_{\mathrm{op},\gamma}(\mathbf{r}) = \frac{Q_S(\mathbf{r})}{S}\, \delta_{\gamma z} S_{\mathrm{op},z} = Q(\mathbf{r})\delta_{\gamma z} S_{\mathrm{op},z} \quad (3)$$

where $Q_{\mathrm{S}}(\mathbf{r})$ is the spin density component along the quantization axis z corresponding to the electronic state with highest spin projection $M_S = S$, and $S_{\mathrm{op},z}$ is the total spin projection operator along z. If $|SM_S\rangle$ denotes a spin eigenfunction, the following relationships hold

$$\int \langle SM_S|Q_{\mathrm{op},\gamma}(\mathbf{r})|SM_S\rangle d^3r = \int Q_\gamma(\mathbf{r}) d^3r = \langle S_{\mathrm{op},\gamma}\rangle = M_S\delta_{\gamma z} \quad (4)$$

where $\delta_{\gamma \mathrm{z}}$ is the Kronecker delta. In the absence of a magnetic field, the averaging of (3) over all $2S+1$ degenerate spin components clearly results in $Q_\gamma(\mathbf{r}) = 0$. In a magnetic field, a physical choice for the quantization axis consists of the direction of the field itself, along which the Zeeman interaction induces a nonzero spin density polarization by splitting the degenerate multiplet into its $2S+1$ components.

With the understanding that the reference frame is always chosen so that the quantization axis γ coincides with the direction of the magnetic field, from the second term on the rhs of (1) we can immediately define an effective spin-current density diagonal operator $\mathbf{J}^{(0)}$, solely acting on spin variables within a given multiplet, as

$$J_\alpha^{(0)}(\mathbf{r}) = \mathscr{F}_\alpha^{S_\gamma}(\mathbf{r}) S_{op,\gamma}, \quad \mathscr{F}_\alpha^{S_\gamma}(\mathbf{r}) = -\frac{e}{m}\, \epsilon_{\alpha\beta\gamma} \nabla_\beta Q(\mathbf{r}) \quad (5)$$

where we have introduced the second rank zeroth-order spin current density tensor $\mathscr{F}_\alpha^{S_\gamma}(\mathbf{r})$ formally defined as $\mathscr{F}_\alpha^{S_\gamma}(\mathbf{r}) = \partial J_\alpha^{(0)}/\partial S_\gamma$. Note that $\mathbf{J}^{(0)}$, since defined as the curl of the spin density, is identically divergenceless. It immediately follows that the continuity equation is fulfilled independently of the approximation and basis set used to compute the wave function.

3. Current Density Formulation of the Nuclear Hyperfine Coupling Tensor and Density of Hyperfine Coupling

Let us consider an open-shell molecule characterized by total spin S and a magnetic nucleus J with spin I. The corresponding nuclear hyperfine coupling constant (HCC) measured in EPR can be rationalized in terms of the splitting of the $(2S+1) \times (2I+1)$ degenerate ground state that is induced by the interaction between nuclear and electron spin magnetic moments.[6,26] The underlying electron-nucleus spin–spin coupling mechanism is described in the nonrelativistic limit by the isotropic Fermi contact (FC) and the traceless anisotropic spin dipolar (SD) Hamiltonians.[6,26,27] If the distance between electron i and nucleus J is called $\mathbf{r}_{iJ}$, the FC and SD Hamiltonians are given by[26,27]

$$H^J_{\mathrm{FC}} = \frac{8\pi}{3}\, \beta_e\beta_N g_e g_J \sum_{i=1}^{n} \delta(\mathbf{r}_{iJ})\boldsymbol{\sigma}_i\cdot\mathbf{I}_J$$

$$H^J_{\mathrm{SD}} = \beta_e\beta_N g_e g_J \sum_{i=1}^{n} \sigma_{i\lambda} r_{iJ}{}^{-5}(3r_{iJ\lambda}r_{iJ\mu} - r_{iJ}{}^{2}\delta_{\lambda\mu})I_{J\mu} \quad (6)$$

where $\beta_e = e\hbar/2mc$ is the Bohr magneton (in cgs emu units), $\beta_N = e\hbar/2m_pc$ is the nuclear magneton with m_p the proton mass, $\delta(\mathbf{r}_{iJ})$ is the Dirac delta function, $\boldsymbol{\sigma}_i$ is a vector whose components are given by the Pauli matrices associated with electron i, and g_e and g_J are the isotropic g-factors for the electron and the nucleus J, respectively. The first-order change in the energy of the system caused by (6) is completely equivalent to the splitting of the $(2S+1) \times (2I+1)$ degenerate multiplet produced by the following effective spin Hamiltonians (see, e.g., ref 27 p 389)

$$H^{\mathrm{FC}}_{\mathrm{eff}} = h\, A^J_{\mathrm{iso}}\mathbf{I}_J\cdot\mathbf{S} \equiv W^{\mathbf{I}_J\mathbf{S}}_{\mathrm{iso}}, \quad H^{\mathrm{SD}}_{\mathrm{eff}} = h\, I_{J\lambda}A^J_{\lambda\mu}S_\mu \equiv W^{\mathbf{I}_J\mathbf{S}}_{\mathrm{SD}} \quad (7)$$

where A^J_{iso} and $A^J_{\lambda\mu}$ are, respectively, the isotropic and dipolar components of the hyperfine coupling tensor. These can be defined in terms of the scalar spin density function (3) as[27] (in Hz):

$$A^J_{\mathrm{iso}} = \frac{8\pi\beta_e\beta_N g_e g}{3h} \int d\mathbf{r}_1 \delta(\mathbf{r}_{1J})Q(\mathbf{r}_1) = \frac{8\pi\beta_e\beta_N g_e g}{3h}\, Q(\mathbf{R}_J) \quad (8)$$

$$A^J_{\lambda\mu} = \frac{\beta_e\beta_N g_e g}{h} \int d\mathbf{r}_1 r_{1J}{}^{-5}(3r_{1J\lambda}r_{1J\mu} - r_{1J}{}^{2}\delta_{\lambda\mu})Q(\mathbf{r}_1) \quad (9)$$

Within the effective spin Hamiltonian formalism, the terms in (7) can now be interpreted as interaction energies between classical magnetic dipole moments associated with the electronic and nuclear spins, with a total interaction energy given by $W^{\mathbf{I}_J\mathbf{S}} = W^{\mathbf{I}_J\mathbf{S}}_{\mathrm{iso}} + W^{\mathbf{I}_J\mathbf{S}}_{\mathrm{SD}}$.

An alternative expression for $W^{\mathbf{I}_J\mathbf{S}}$ can be formulated within the current density formalism (see, e.g., ref 26). In fact, according to classical electrodynamics, the interaction energy between the spin current density $\mathbf{J}^{(0)}$ and a nuclear magnetic dipole $\mu^J = g_J\beta_N\mathbf{I}_J$ reads as[26,18]

$$W^{\mathbf{I}_J\mathbf{S}} = -\frac{1}{c}\int d\mathbf{r}\mathbf{A}^{\boldsymbol{\mu}J}\cdot\mathbf{J}^{(0)}(\mathbf{r}) \tag{10}$$

where the classical vector potential associated with nucleus J is $A_\alpha^{\mu_J}(\mathbf{r}) = g_J\beta_N\epsilon_{\alpha\beta\gamma}I_{J\beta}(r_\gamma - R_{J\gamma})/|\mathbf{r} - \mathbf{R}_J|^3$. Using the definition for $\mathbf{J}^{(0)}$, we can write (10) more explicitly as

$$W^{\mathbf{I}_J\mathbf{S}} = -\frac{g_J\beta_N}{c}\epsilon_{\alpha\beta\gamma}I_{J\beta}S_\lambda\int \mathscr{F}_\alpha^{S_\lambda}(\mathbf{r})\frac{(r_\gamma - R_{J\gamma})}{|\mathbf{r} - \mathbf{R}_J|^3}d\mathbf{r} \tag{11}$$

The consistency between definitions (11) and (7) can be easily checked via the procedure adopted, for instance, in ref 36. Definition (11) was implicit, e.g., in the derivation of the effective spin Hamiltonians (17) via $\mathbf{J}^{(0)}$ given in ref 26, although the present form, via the introduction of the components of a second-rank spin current density tensor (5), delivers an expression that is explicitly bilinear in $\boldsymbol{I}$ and $\boldsymbol{S}$, a fact that is well-known to be useful in defining molecular properties in terms of analytic energy derivatives, rather than via finite-field approaches.

Given the fact that (7) provides an exact factorization of the problem into separate space and spin manifolds, it is now possible to provide a formal definition for the HCCs in terms of energy derivatives with respect to the components of the electron spin $\boldsymbol{S}$ and the nuclear spin $\boldsymbol{I}_J$, as

$$A_{\lambda\mu}^J = A_{\mathrm{iso}}^J\delta_{\lambda\mu} + A_{\lambda\mu}^J = \frac{1}{h}\frac{\partial^2 W^{\mathbf{I}_J\mathbf{S}}}{\partial I_{J\lambda}\partial S_\mu} \tag{12}$$

where $W^{\mathbf{I}_J\mathbf{S}}$ is given either by the sum of eqs 7 or by eq 11.

Thus, the formal definition (12), which trivially recovers (8) and (9) when used with (7), leads to a new expression for the nuclear hyperfine coupling tensor, when applied to the energy definition based on the current density formalism (11):

$$A_{\lambda\mu}^J = -\frac{g_J\beta_N}{hc}\epsilon_{\lambda\gamma\alpha}\int d\mathbf{r}\frac{(r_\gamma - R_{J\gamma})}{|\mathbf{r} - \mathbf{R}_J|^3}\mathscr{F}_\alpha^{S_\mu}(\mathbf{r}) \tag{13}$$

The advantage of this expression consists of the fact that it provides a formulation of each tensor component describing hyperfine coupling that is formally based on classical electrodynamics and thus puts it on an equal footing with analogous definitions introduced for NMR and EPR observables, such as the nuclear magnetic shielding,[15−18] the indirect nuclear spin−spin coupling tensor,[39,34,36] and the g-tensor.[27−30]

As with nuclear magnetic shielding[33,35] and nuclear spin−spin coupling tensors,[34,36] given the current density definition of hyperfine coupling constants (13), it is now straightforward to define a *density of hyperfine coupling* as

$$A_{\lambda\mu}^J(\mathbf{r}) = -\frac{g_J\beta_N}{hc}\epsilon_{\lambda\gamma\alpha}\frac{(r_\gamma \text{ - } R_{J\gamma})}{|\mathbf{r} - \mathbf{R}_J|^3}\mathscr{F}_\alpha^{S_\mu}(\mathbf{r}) \tag{14}$$

These scalar functions can easily be plotted over the molecular domain and interpreted as a pointwise partition of the Biot-Savart law, in that they provide a straightforward visualization of the point-by-point contribution from the spin current distribution to the magnetic field induced at the site of nucleus J. It is important to stress that in principle the set of scalar functions (14) is not uniquely defined. It is in fact possible to add to (14) an arbitrary scalar function that integrates to zero, and the resulting hyperfine density would exactly lead on integration to the same set of observables. It is however not straightforward to define a physically sensible transformation that would accomplish such a change in (14), also given the fact that the hyperfine density is given by the product of the derivatives of two functions (the spin current and the nuclear vector potential) that are both origin-invariant and, in the case of the current, even gauge invariant. At any event, although endowed with a certain degree of arbitrariness, it will be shown later on that the hyperfine coupling density (14) is indeed very well-defined from a conceptual point of view and useful to interpret spin current density maps.

4. Current Density Formulation of the Nuclear Magnetic Shielding Tensor in Paramagnetic Molecules

NMR chemical shifts for open-shell molecules, in the limit of small spin−orbit coupling, can be rationalized in terms of two main contributions.[11,12] The first consists of the usual temperature independent or orbital contribution, also accounting for NMR chemical shifts in closed-shell molecules. The current density formulation of the orbital term for an open-shell molecule can be obtained straightforwardly for any magnetic nucleus J from the analogous definition for the closed-shell case, given by the space integral[18]

$$\sigma_{J,\alpha\beta} = \frac{\partial^2 W^{\mu_J\mathbf{B}}}{\partial\mu_{J\alpha}\partial B_\beta} = -\frac{1}{c}\epsilon_{\alpha\lambda\mu}\int d\mathbf{r}\frac{r_\lambda - R_{J\lambda}}{|\mathbf{r} - \mathbf{R}_J|}\mathscr{F}_\mu^{B_\beta}(\mathbf{r}) \tag{15}$$

where the first-order current density tensor $\mathscr{F}_\gamma^{B_\delta}$ $(\mathbf{r})$is defined as $\mathscr{F}_\gamma^{B_\delta}(\mathbf{r}) = \partial J_\gamma^{(1)}/\partial B_\delta$. The induced current tensor $\mathscr{F}_\gamma^{B_\delta}$ $(\mathbf{r})$ is routinely calculated and visualized for closed-shell molecules.[18,22,24,25,41] On the other hand, in studies on open-shell molecules $\mathscr{F}_\gamma^{B_\delta}(\mathbf{r})$ has so far appeared just as an intermediate computational byproduct used to recover, on numerical integration over a three-dimensional grid, the value of g-tensor components,[28−30] and no attempt at the visualization of $\mathbf{J}^{(1)}$ and interpretation of the integrated magnetic properties based on current density plots appears to exist in the literature.

The second (and dominant) contribution to the observed NMR chemical shift in paramagnetic species is absent in closed-shell molecules, as it arises from the interaction between the electron spin density and the nuclear magnetic moments.[11−14] Let us briefly discuss the origin of this term. The nuclear shielding tensor is defined as a second derivative of the energy with respect to the nuclear dipole and the external field (see eq 15). Accordingly, since the interaction between the spin density distribution and the nuclear magnetic moment (see (7) or (11)) does not formally depend on the external field, it follows that the contribution from the electron-nucleus spin−spin coupling to the NMR shielding is zero for a pure spin state. However, as highlighted, e.g., in ref 11, owing to the rapid times associated with electron spin relaxation when compared to those involved in the NMR experiment, for high enough thermal energy k_BT (where T is the absolute temperature and k_B the Boltzmann constant) the electron spin density magnetization experienced by the nuclei consists of a thermal average over the field-dependent Zeeman-split spin states. Carrying out the statistical sum within the Van Vleck approximation[12] ($g\beta B \ll k_BT$), one obtains the following expression for the average component of the spin along the quantization axis[11,12]

$$\langle S_\gamma \rangle = -\, g\beta B_\gamma \frac{S(S+1)}{3k_BT} \tag{16}$$

which introduces a magnetic field dependence in the expression for the thermal average of $W^{\mathbf{I}_I\mathbf{S}}$. Thus, substitution of (16) into (11) leads to a new expression for the temperature-dependent contribution to $\sigma_{I,\alpha\beta}$ in terms of $\mathbf{J}^{(0)}$ as

$$\sigma^{\mathbf{S}}_{I,\alpha\beta} = \frac{\partial^2 \langle W^{\mathbf{I}_I\mathbf{S}} \rangle}{\partial \mu_{I\alpha} \partial \mathrm{B}_\beta} = \frac{g\beta S(S+1)}{3k_BTc}\, \epsilon_{\alpha\lambda\mu} \int d\mathbf{r}\, \frac{(r_\lambda - R_{I\lambda})}{|\mathbf{r} - \mathbf{R}_I|^3}\, \mathscr{F}^{S_\beta}_{\mu}(\mathbf{r}) \tag{17}$$

Equations 13, 15, and 17 share the same three-dimensional integral structure, involving a spin or charge current density distribution, thereby providing a common framework where to describe NMR and EPR observables in terms of maps of spin and charge current density and, accordingly, in terms of concepts based on classical electrodynamics. It is important to stress that these expressions do not define an improved methodology to obtain the numerical value of the response tensor, as they would recover on integration exactly the same results as those obtained via ordinary response theory, provided the same choice of method and basis set is made. Vice versa, the current density tensors appearing in (13), (15), and (17) are not trivial byproducts of normal response calculations, as they involve the explicit buildup of the perturbed and unperturbed wave functions and their spatial gradients over a defined grid of points. The usefulness of expressions like (13), (15), and (17) lies in the fact that they provide an exact connection between the maps of spin and charge current density and the integrated response properties, thus setting a rigorous basis for analyzing and rationalizing the results within classical electromagnetism theory. Next, we describe an ab initio computational procedure to evaluate $\mathbf{J}^{(0)}$ and $\mathbf{J}^{(1)}$ and present a few applications.

5. (Coupled)-Hartree−Fock Calculation of Charge and Spin Currents in Open-Shell Molecules

Let the one-row matrix χ contain a set of atomic basis functions and the matrices $\mathbf{c}_i^{(0)}$ contain on the columns the coefficients for the ROHF doubly occupied ($i = 1$), singly occupied ($i = 2$), and virtual ($i = 3$) molecular orbitals (MOs) represented in the basis χ. The MO's $\mathbf{c}_i^{(0)}$are self-consistent solutions to[27]

$$\mathbf{F}_{\mathrm{eff}}{}^{(0)}\mathbf{R}_i{}^{(0)} - \mathbf{R}_i{}^{(0)}\mathbf{F}_{\mathrm{eff}}{}^{(0)} = \mathbf{0}, \quad i = 1,2$$

$$\mathbf{F}_{\mathrm{eff}}{}^{(0)} = a\mathbf{R'}_2{}^{(0)}\mathbf{F}_1{}^{(0)}\mathbf{R'}_2{}^{(0)} + b\mathbf{R'}_1{}^{(0)}\mathbf{F}_2{}^{(0)}\mathbf{R'}_1{}^{(0)} + c\mathbf{R'}_3{}^{(0)}(2\mathbf{F}_1{}^{(0)} - \mathbf{F}_2{}^{(0)})\mathbf{R'}_3{}^{(0)} \tag{18}$$

where $\mathbf{R'}_i^{(0)} = \mathbf{1} - \mathbf{R}_i^{(0)}, \mathbf{R}_i^{(0)} = \mathbf{c}_i^{(0)}\,\mathbf{c}_i^{(0)\dagger}$ (i =1,2,3) are the density matrices represented in the basis χ; $\mathbf{F}_i^{(0)} = \mathbf{h}^{(0)} + \mathbf{G}_i^{(0)}$ are the Fock Hamiltonians for the doubly and singly occupied subspaces ($i = 1,2$) represented on χ and defined, e.g., in refs 27, 40, 43; and a, b, and c are arbitrary nonzero convergence parameters. Since in ROHF the spin density is completely described by the singly occupied MO's space, the spin current density tensor defined by eq 5 can be straightforwardly written in terms of $\mathbf{R}_2^{(0)}$ alone as

$$\mathscr{F}^{S_\gamma}_{\alpha}(\mathbf{r}) = -\frac{e}{m}\, \epsilon_{\alpha\beta\gamma} \{ \nabla_\beta \chi \mathbf{R}_2{}^{(0)} \chi^\dagger + \chi\, \mathbf{R}_2{}^{(0)} \nabla_\beta \chi^\dagger \} \tag{19}$$

In order to compute $\mathbf{J}^{(1)}$, the ROHF equations (18) are expanded to first order in the magnetic field, resulting in[43]

$$\mathbf{F}_{\mathrm{eff}}{}^{(0)}\mathbf{R}_i{}^{(1)} - \mathbf{R}_i{}^{(1)}\mathbf{F}_{\mathrm{eff}}{}^{(0)} + \mathbf{F}_{\mathrm{eff}}{}^{(1)}\mathbf{R}_i{}^{(0)} - \mathbf{R}_i{}^{(0)}\mathbf{F}_{\mathrm{eff}}{}^{(1)} = \mathbf{0}, \quad i = 1,2 \tag{20}$$

which defines the Coupled-ROHF (CROHF) procedure. It can be shown, using the projector properties of the unperturbed density matrices,[43] that the first-order densities can be written as

$$\mathbf{R}_1{}^{(1)} = (\mathbf{x}_1 + \mathbf{x}_1{}^\dagger) + (\mathbf{x}_0 + \mathbf{x}_0{}^\dagger)$$

$$\mathbf{R}_2{}^{(1)} = (\mathbf{x}_2 + \mathbf{x}_2{}^\dagger) - (\mathbf{x}_0 + \mathbf{x}_0{}^\dagger) \tag{21}$$

with $\mathbf{x}_0 = \mathbf{R}_1^{(0)}\,\mathbf{R}_1^{(1)}\,\mathbf{R}_2^{(0)}$, $\mathbf{x}_1 = \mathbf{R}_1^{(0)}\,\mathbf{R}_1^{(1)}\,\mathbf{R}_3^{(0)}$, and $\mathbf{x}_2 = \mathbf{R}_2^{(0)}\,\mathbf{R}_2^{(1)}\,\mathbf{R}_3^{(0)}$. A method for the iterative solution of (20), based on three separate iterative procedures for the calculation of $\mathbf{x}_0$, $\mathbf{x}_1$, and $\mathbf{x}_2$, was described in ref 43.

Here we propose an alternative computational strategy, based on the direct evaluation of the perturbed MO's coefficients for the doubly and singly occupied subspaces. Let us define $\phi_j^{(1)}(\mathbf{r}) = \chi\mathbf{d}_j^{(1)}$ and $\phi_k^{(1)}(\mathbf{r}) = \chi\mathbf{s}_k^{(1)}$, with j (k) denoting a doubly (singly) occupied MO. It can easily be shown that equations (21) can be recast as

$$\mathbf{R}_1{}^{(1)} = \sum_{j}^{\text{doubly-occ}} (\mathbf{c}_j{}^{(0)}\mathbf{d}_j{}^{(1)\dagger} + \mathbf{d}_j{}^{(1)}\mathbf{c}_j{}^{(0)\dagger})$$

$$\mathbf{R}_2{}^{(1)} = \sum_{k}^{\text{singly-occ}} (\mathbf{c}_k{}^{(0)}\mathbf{s}_k{}^{(1)\dagger} + \mathbf{s}_k{}^{(1)}\mathbf{c}_k{}^{(0)\dagger}) \tag{22}$$

Accordingly, (20) can now be solved by setting up a single iterative procedure for the unknown perturbed coefficient matrices $\mathbf{d}_j^{(1)}$ and $\mathbf{s}_k^{(1)}$, by defining the Hartree−Fock propagators for the doubly and singly occupied subspaces as

$$\mathbf{M}_1^j = \sum_k^{\text{singly-occ}} \frac{\mathbf{c}_k^{(0)}\mathbf{c}_{\boldsymbol{k}}^{(0)\dagger}}{\epsilon_j - \epsilon_j} + \sum_m^{\text{vir}} \frac{\mathbf{c}_m^{(0)}\mathbf{c}_m^{(0)\dagger}}{\epsilon_j - \epsilon_m}, \quad j \in \text{doubly-occ}$$

$$\mathbf{M}_2{}^k = \sum_j^{\text{doubly-occ}} \frac{\mathbf{c}_j^{(0)}\mathbf{c}_j^{(0)\dagger}}{\epsilon_k - \epsilon_i} + \sum_m^{\text{vir}} \frac{\mathbf{c}_m^{(0)}\mathbf{c}_m^{(0)\dagger}}{\epsilon_k - \epsilon_m}, \quad k \in \text{singly-occ} \tag{23}$$

and iteratively propagating the unperturbed MOs according to

$$\mathbf{d}_j^{(1)} = \mathbf{M}_1{}^j\mathbf{F}_{\text{eff}}{}^{(1)}\mathbf{c}_j^{(0)}, \quad \mathbf{s}_k^{(1)} = \mathbf{M}_2{}^k\mathbf{F}_{\text{eff}}{}^{(1)}\mathbf{c}_k^{(0)} \tag{24}$$

From the converged value of $\mathbf{d}_j^{B_\beta}$ and $\mathbf{s}_k^{B_\mathrm{b}}$ (the superscript B_β labels the components of the perturbing magnetic dipole operator) the CROHF induced current density (2) can finally be computed as

$$\mathscr{F}_\alpha^{B_\beta}(\mathbf{r}) = -\frac{e}{2mc}\epsilon_{\alpha\beta\gamma}r_\gamma\chi\mathbf{D}\chi^\dagger + \frac{4ie\hbar}{m}\sum_j^{\text{doubly-occ}}\{\mathbf{d}_j^{B_\beta\dagger}\chi^\dagger\nabla_\alpha\chi\mathbf{c}_j^{(0)} - \mathbf{c}_j^{(0)\dagger}\chi^\dagger\nabla_\alpha\chi\mathbf{d}_j^{B_\beta}\} + \frac{2ie\hbar}{m}\sum_k^{\text{singly-occ}}\{\mathbf{s}_k^{B_\beta\dagger}\chi^\dagger\nabla_\alpha\chi\mathbf{c}_k^{(0)} - \mathbf{c}_k^{(0)\dagger}\chi^\dagger\nabla_\alpha\chi\,\mathbf{s}_k^{B_\beta}\} \tag{25}$$

where $\mathbf{D} = 2\mathbf{R}_1 + \mathbf{R}_2$. The routines for the CROHF evaluation of magnetic response have been implemented in the SYSMO package.[42]

A procedure for (i) the Coupled-UHF (CUHF) magnetic response calculation and (ii) the Unrestricted-GGA-DFT (UDFT) magnetic response calculation based on zeroth-order KS spin-orbitals obtained from the program Gaussian 03[44] has also been implemented. The CUHF and UDFT computational schemes will not be described here, as they have been implemented following well-known procedures.[30] The CROHF, CUHF, and UDFT schemes for the calculation of the first-order current (25) have been implemented within the four distributed-origin approaches generally known as Continuous distribution of The Origin of the Current Density (CTOCD)-methods.[45−47] These methods have been shown to converge to origin-independent results for the magnetic response with relatively small basis sets.[47] Also, for closed-shell molecules, it has been shown that the DZ or *ipsocentric* variant of the CTOCD methods allows an optimal orbital partition of the induced current density, thus providing a frontier orbital model for the rationalization of the magnetic response of π-conjugated systems.[24,25] The question of open-shell aromaticity[48,49] and its relation with the induced current density[23] can now be quantitatively assessed via plots of (25) or similar expressions corresponding to the CTOCD methods.

6. Results and Discussion

As preliminary applications of the developed methodology, first we report here the calculation of the zero- and first-order current density and of the density of hyperfine coupling in the open-shell molecules BH_2, CH_2^-, and NH_2. Next, as further applications of the newly developed methodology, we present the calculation of maps of induced current density $\mathbf{J}^{(1)}$ for (i) the neutral phenalenyl radical and (ii) the pancake-bonded dimer of the neutral phenalenyl radical. This study stems from the recent discussion of ^{1}H NMR experimental data and NICS calculations performed on the dimer compound, in which the measured proton chemical shifts have been interpreted in terms of the existence of a deshielding *ring current* in the dimer singlet diradical, a molecule that NICS calculations confirm as an open-shell aromatic π-complex.[37]

Table 1. Hyperfine Coupling Constant Tensor Components Relative to the Heavy Atoms B, C, and N, Resulting from Numerical Integration of $\mathbf{J}^{(0)}$ for the Three Molecules BH_2, CH_2^-, and NH_2[a]

BH_2	A_{iso}	A_{C2}	$A_{\parallel}$	$A_{\perp}$
ROHF	324.1	78.4	−39.5	−38.9
UHF	365.2	80.4	−38.9	−41.4
RAS-II[b]	323.3	79.7	−38.8	−40.9
CH_2	A_{iso}	A_{C2}	$A_{\parallel}$	$A_{\perp}$
ROHF		−58.3	−58.1	116.4
UHF	135.5	−56.4	−62.7	119.1
RAS-II[b]	59.5	−56.0	−59.3	115.3
NH_2	A_{iso}	A_{C2}	$A_{\parallel}$	$A_{\perp}$
ROHF		−42.3	−42.2	84.5
UHF	42.5	−43.0	−41.2	84.2
RAS-II[b]	26.6	−41.6	−40.9	82.5

[a] All results are in MHz. [b] From calculations reported in ref 11.

6.1. Current Density Maps for Small Radicals. The experimental geometries reported in ref 11 have been employed in the present calculations. The three C_{2v} molecules have $S = 1/2$ ground states with symmetries $^2A_1$ (BH_2),$^2B_1$ (CH_2^-), and $^2B_1$ (NH_2). All the calculations were performed at the (Coupled) ROHF and UHF level of theory using the aug-cc-pVTZ basis set. The spin and charge current densities have been computed and integrated over the whole molecular domain using (i) eq 13 to obtain the isotropic and dipolar components of the HCC for the B, C, and N nuclei (see Table 1) and (ii) eq 15 to obtain the orbital (temperature independent) contribution to the paramagnetic shielding tensor components for B, C, and N (see Table 2). In the same tables we also report more accurate calculations performed at the RAS-II level of theory, taken from ref 11, for the sake of comparison. The results in Table 1 show that the uncorrelated ROHF and UHF approaches perform quite well for these simple systems. It can be seen that the uncorrelated calculations reproduce at a quantitative level the RAS-II dipolar components of $A^{\mathrm{I}}{}_{\alpha}\beta$ for all three molecules and A_{iso} for the BH_2. However, the isotropic part of the hyperfine coupling for CH_2^- and NH_2 shows only qualitative agreement with the more accurate RAS-II results. The HCCs for CH_2^- and NH_2 were calculated at the UHF level of theory in order to account for the spin density polarization at the site of the heavy nucleus. This effect cannot be recovered within the ROHF approach, because of the nodal character at the nuclear site of the *p* atomic orbital hosting the unpaired electron density and the absence of spin polarization effects within this approach. Accordingly, the overestimation of spin density at the C and N nuclei can be reasonably ascribed to higher-spin contamination of the UHF wave function. Since

Table 2. Paramagnetic Nuclear Magnetic Shielding Tensor Components (Orbital Contribution Only) Relative to the Heavy Atoms B, C, and N, Resulting from Numerical Integration of $\mathbf{J}^{(1)}$ Computed Both Using a Common Origin (CO) and Using the Distributed Origin Method CTOCD-PZ2 (PZ2) for the Three Molecules BH_2, CH_2^-, and NH_2, Using an aug-cc-pVTZ Basis Set[a]

BH_2	σ_{iso}	σ_{C2}	$\sigma_{\parallel}$	$\sigma_{\perp}$
CROHF−CO	−107.5	−17.9	−409.7	105.3
CROHF-PZ2	−115.5	−22.8	−426.5	102.7
CUHF−CO	−85.1	−18.5	−343.0	106.2
CUHF-PZ2	−92.8	−23.4	−358.6	103.6
RAS-II[b]	−167.6	−29.3	−569.5	95.9
CH_2^-	σ_{iso}	σ_{C2}	$\sigma_{\parallel}$	$\sigma_{\perp}$
CROHF−CO	7.8	74.7	−273.5	222.1
CROHF-PZ2	5.1	72.8	−279.3	221.6
UHF−CO	30.3	83.4	−216.4	223.8
UHF-PZ2	27.8	81.6	−221.5	223.3
RAS-II[b]	−19.4	58.3	−350.7	233.8
NH_2	σ_{iso}	σ_{C2}	$\sigma_{\parallel}$	$\sigma_{\perp}$
ROHF−CO	−237.6	−78.9	−891.9	257.9
ROHF-PZ2	−241.3	−81.4	−899.7	257.3
UHF−CO	−184.2	−54.6	−756.0	258.1
UHF-PZ2	−187.5	−57.0	−763.0	257.5
RAS-II[b]	−291.1	−109.6	−1029.6	265.9

[a] All results are in ppm. [b] From calculations reported in ref 11.

in this introductory study the focus is on a novel method providing a pictorial representation of physical mechanisms, rather than on numerical accuracy, we did not implement procedures to project out higher spin components from the UHF wave function. However, it is clear that the qualitative agreement of the results for A_{iso} and the quantitative agreement of the calculated values for the anisotropic components of $A^{I}_{\alpha\beta}$ represent a sufficiently reliable framework whereon to base a sound analysis of the corresponding spin current density functions.

The results for the orbital part of the paramagnetic shielding tensors computed at the CROHF and CUHF level of theory within the current density formalism are reported in Table 2. They display overall good agreement with the RAS-II results. As for the closed-shell case, the computation of these quantities within the finite basis set approach can suffer from gauge-origin dependence. We explored this possibility by computing $\mathbf{J}^{(1)}$ both within the common origin (CO) approximation and within the distributed origin approach CTOCD in its PZ2 variant, which performs the best among the CTOCD schemes.[47] As it is evident from the results reported in Table 2, the agreement between CO and PZ2 is reasonably good, showing that for these simple molecules the aug-cc-pVTZ basis set already leads to results that do not display significant gauge-origin dependence. The best noncorrelated method appears to be CROHF-PZ2. The worst performance when CROHF and CUHF results are compared with RAS-II is observed for CH_2^-. Although the sign and relative magnitudes of the individual tensor components computed at the CROHF level are in qualitative agreement with the RAS-II results, the average shielding has a different sign from that computed at the RAS-II level.

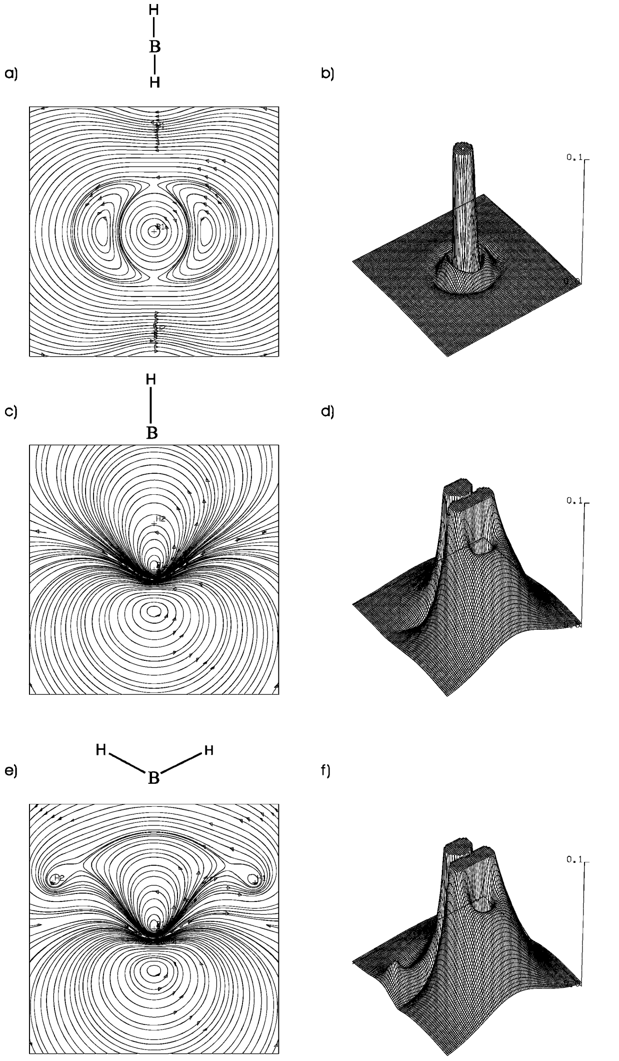


Figure 1. Streamlines (left column) and modulus (right column) of the ROHF spin-current density $\mathbf{J}^{(0)}$ characterizing the magnetic field-split ground-state doublet of BH2, plotted for three different orientations of the field (quantization axis): (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the boron atom.

However, even within the RAS-II approximation, the isotropic shielding is relatively small, and the main reason for this is the cancellation between the strong paramagnetic component along the direction perpendicular to the C_2 axis and contained in the molecular plane ($\sigma_{\parallel}$) and the strong diamagnetic component perpendicular to the molecular plane ($\sigma_{\perp}$). This behavior is well reproduced within the CROHF methods. Let us now turn to the visualization and discussion of the physical mechanisms underlying the results obtained for the EPR and NMR observables in terms of the maps of $\mathbf{J}^{(0)}$ (Figures 1−3), maps of $A^{J}{}_{\alpha}\beta(\mathbf{r})$ (Figure 4), and maps of $\mathbf{J}^{(1)}$ (Figures 5−7).

6.1.1. Spin Currents and Density of Hyperfine Coupling Constant. BH_2. The ROHF spin current density maps for the BH2 radical are shown in Figure 1. For an external magnetic

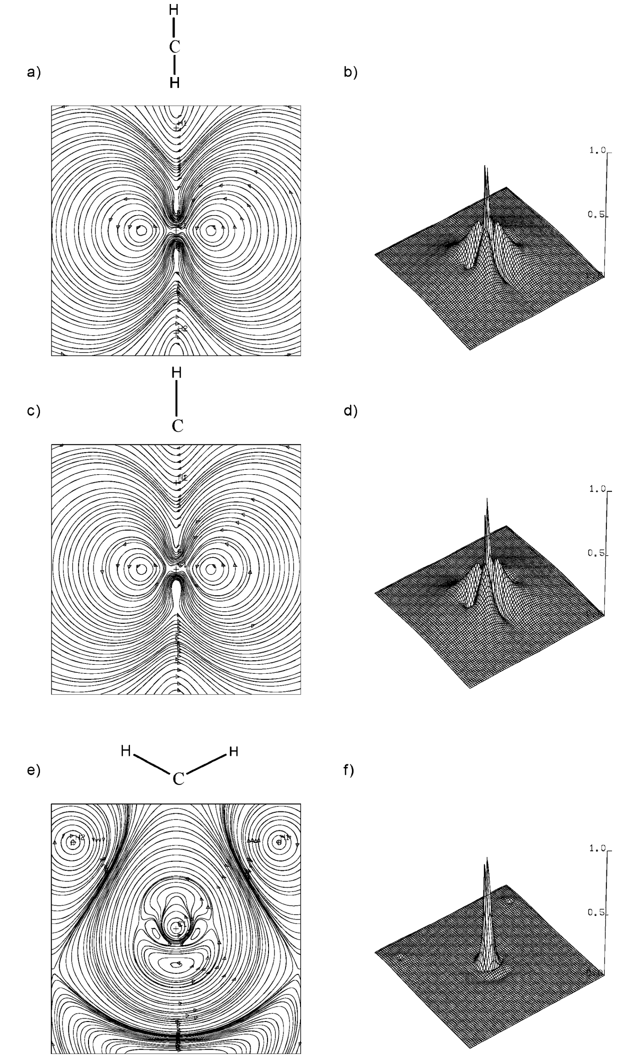


Figure 2. Streamlines (left column) and modulus (right column) of the ROHF spin-current density $\mathbf{J}^{(0)}$ characterizing the magnetic field-split ground-state doublet of CH_2^-, plotted for three different orientations of the field (quantization axis): (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the carbon atom.

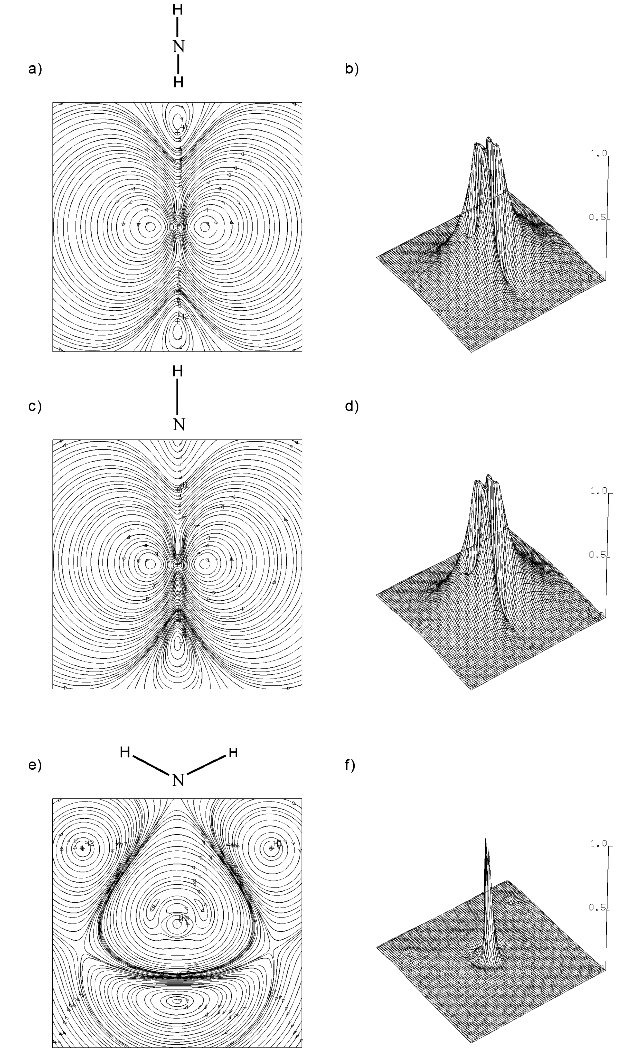


Figure 3. Streamlines (left column) and modulus (right column) of the ROHF spin-current density $\mathbf{J}^{(0)}$ characterizing the magnetic field-split ground-state doublet of NH_2, plotted for three different orientations of the field (quantization axis): (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the nitrogen atom.

field oriented along the C_2 symmetry axis, the current density plotted in the perpendicular plane at the height of the boron atom (Figure 1a) consists of three paramagnetic (anticlockwise) vortices: one centered on the boron atom and characterized by a peak of very large magnitude (see high positive peak in Figure 1b) and two other vortices, which are very weak in magnitude, and are displaced symmetrically at the sides of the central atom. When the current is plotted at different heights further away from B (not shown), the current density map appears very similar to that plotted in Figure 1a, although only the central vortex survives with significant intensity. For the two orientations of the field perpendicular to the C_2 axis, the corresponding maps of spin current density display a pattern that resembles a distorted p atomic orbital, carrying two paramagnetic vortices (Figure 1c,e) of similar intensity (Figure 1d,f) centered on its lobes. The circulation on one of the two lobes is centered on the boron atom. For both orientations perpendicular to the C_2 axis, at different heights further away from B (not shown here), the current density maps are similar in appearance to those reported in Figure 1c,e, although the modulus of the paramagnetic vortex centered on B quickly becomes very small in magnitude, so that the dominant feature is represented by a localized paramagnetic circulation centered on the lobe that does not enclose the boron atom.

The maps can be readily understood in terms of the valence electronic structure of BH2. This can be described on the basis of three sp^2 hybrid atomic orbitals (AOs) contained in the molecular plane. With respect to a plane containing the C_2 axis and perpendicular to the molecular plane, two of the three sp^2 hybrids form two doubly occupied MOs in combination with the hydrogen s AO, one symmetric (a_1) and one antisymmetric (b_2). The third sp^2 hybrid (a_1) lies along the C_2 axis and hosts the unpaired electron density,

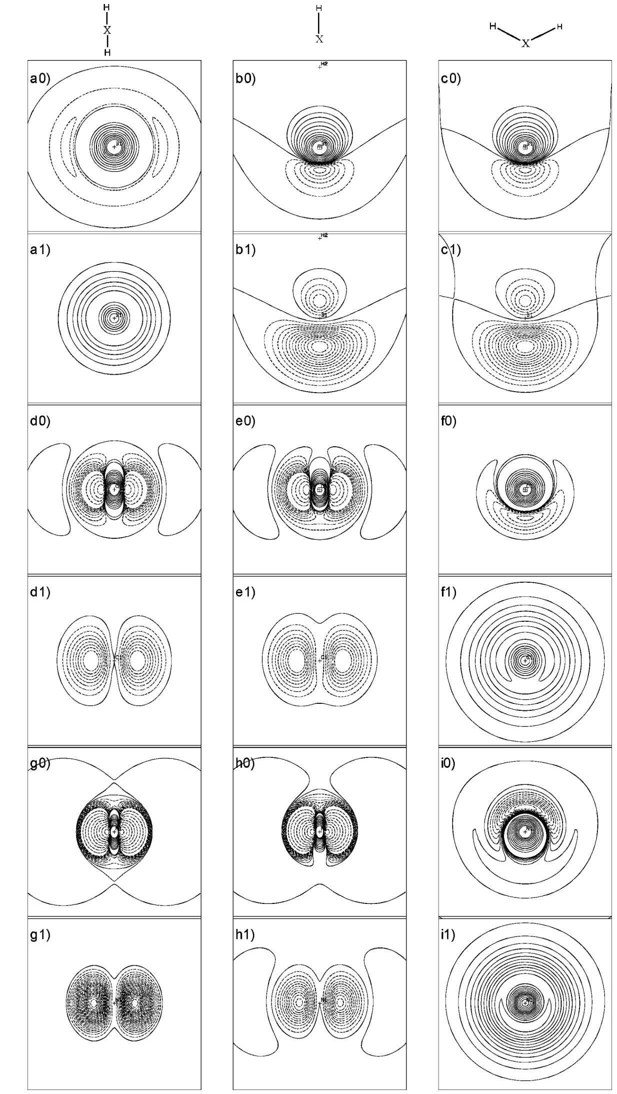


Figure 4. Contour maps of nuclear hyperfine coupling density $A^{X}_{\alpha\beta}(\mathbf{r})$ (X = B, C, N) for the three radicals BH_2 (first two rows, a0, a1, b0, b1, c0, c1) CH_2^- (third and fourth rows, d0, d1, e0, e1, f0, f1), and NH_2 (last two rows, g0, g1, h0, h1, i0, i1), corresponding to the total (isotropic plus dipolar terms) tensor components A_{C2} (left column), $A_{\|}$ (central column), and $A_{\perp}$ (right column). Solid (dashed) lines correspond to positive (negative) density. Figures labeled x0 (x = a, b, c, d, e, f, g, h, i) correspond to planes containing the heavy nucleus, whereas the ones labeled x1 correspond to planes at 0.4 a0 from the heavy nucleus. The quantization axis corresponds to the C_2 axis (a−c); the axis perpendicular to C_2 and contained in the molecular plane (d−f); and the axis perpendicular to the molecular plane (g−i). The contour lines are plotted from −1 to 1 au, in steps of 0.1 au. For planes containing the heavy nuclei, 20 contour lines ± e^r au were added, with r ranging from 0.5 to 5.5 in steps of 0.5.

i.e., it represents the singly occupied molecular orbital (SOMO). Seen along the C_2 axis and projected onto a perpendicular plane, the a_1 SOMO can be seen as an s atomic orbital hosting one unpaired electron. Accordingly, the corresponding spin current map shown in Figure 1a consists mainly of a single paramagnetic vortex centered on the B nucleus.

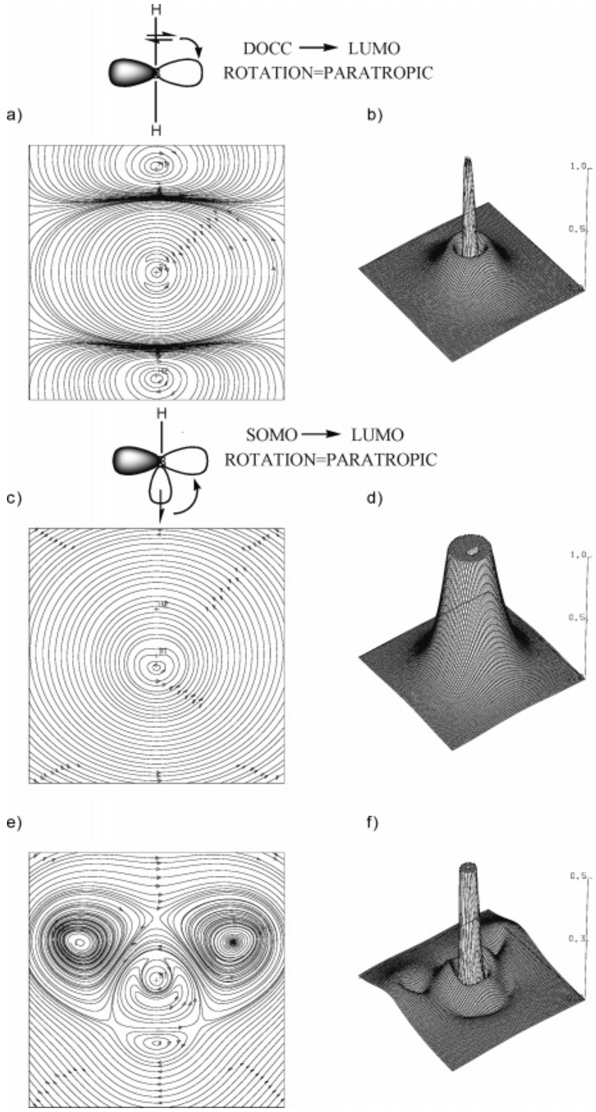


Figure 5. Streamlines (left column) and modulus (right column) of the CROHF-DZ first-order current density $\mathbf{J}^{(1)}$ induced in BH_2, by an external magnetic field plotted for three different orientations of the field: (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the boron atom.

The corresponding map of HCC density $A_{\alpha\beta}(\mathbf{r})$ reported in Figure 4a0 provides a clear picture of the contribution of $\mathbf{J}^{(0)}$ to the integrated HCC tensor components. As detailed in ref 35, magnetic property density functions provide a map of the contribution to the magnetic field calculated at a given point in space (e.g., at the B site in this case) arising from current density patterns distributed all over the molecular domain. The contribution to the integrated HCC arising from a given current density pattern is ruled by the Biot-Savart law of classical electrodynamics. In particular, such a contribution is positive (negative) if the field induced at the probe site is reinforcing (opposing) the external magnetic field. At the height of the B atom, the current is dominated by the isotropic Fermi contact contribution to the HCC (see eq 8). This contribution is represented by the large peak of positive HCC density in Figure 4a0, which indicates that

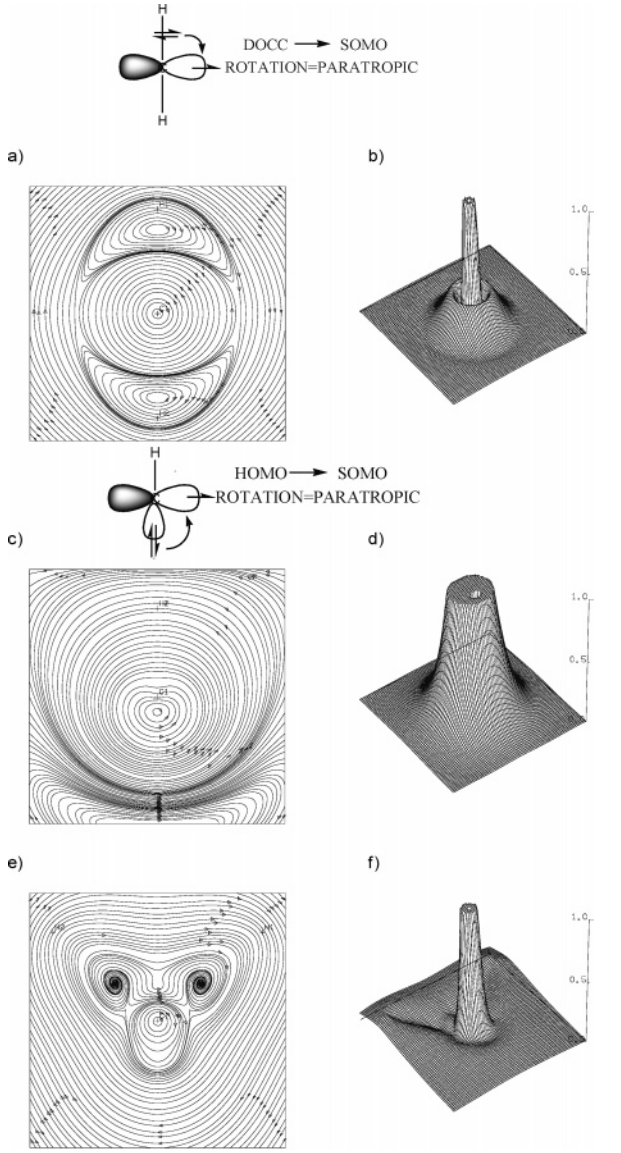


Figure 6. Streamlines (left column) and modulus (right column) of the CROHF-DZ first-order current density $\mathbf{J}^{(1)}$ induced in CH_2^- by an external magnetic field, plotted for three different orientations of the field: (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the carbon atom.

the paramagnetic vortex centered on B produces a local magnetic field parallel to the external field. The fact that plots at different heights are characterized by similar current density maps can be clearly seen on the map of HCC density plotted at 0.4 a_0 distance from the plane containing B (Figure 4a1), from which it is evident that the boron atom is still fully enclosed within a shielding cone produced by the dominant paramagnetic vortex. This explains the positive sign of A_{C2} for boron.

The situation is rather different for the other two orientations of the quantization axis (Figure 1c,e). In these two cases the unpaired electron density is polarized along directions perpendicular to the SOMO axis. The associated spin current consists of two paratropic circulations centered on the two lobes of the sp^2 hybrid SOMO. One of the two paramagnetic vortices is centered on B, producing a local field parallel to the external one, thus providing a positive contribution to the integrated HCC (see the high positive peak centered on B in the maps of hyperfine coupling density, Figure 4b0,4c0). On the other hand, the paratropic circulation centered on the lobe of the sp^2 hybrid that does not enclose the boron atom provides a net negative contribution to the integrated HCC (see the negative peak in Figure 4b0,4c0). This fact can easily be understood in terms of the Biot-Savart law. The boron atom lies outside the shielding cone characterizing the Biot-Savart magnetic field distribution induced by the neighboring paramagnetic circulation, and, accordingly, it experiences an induced field that opposes the external one. In the plane containing the boron atom the positive contact contribution

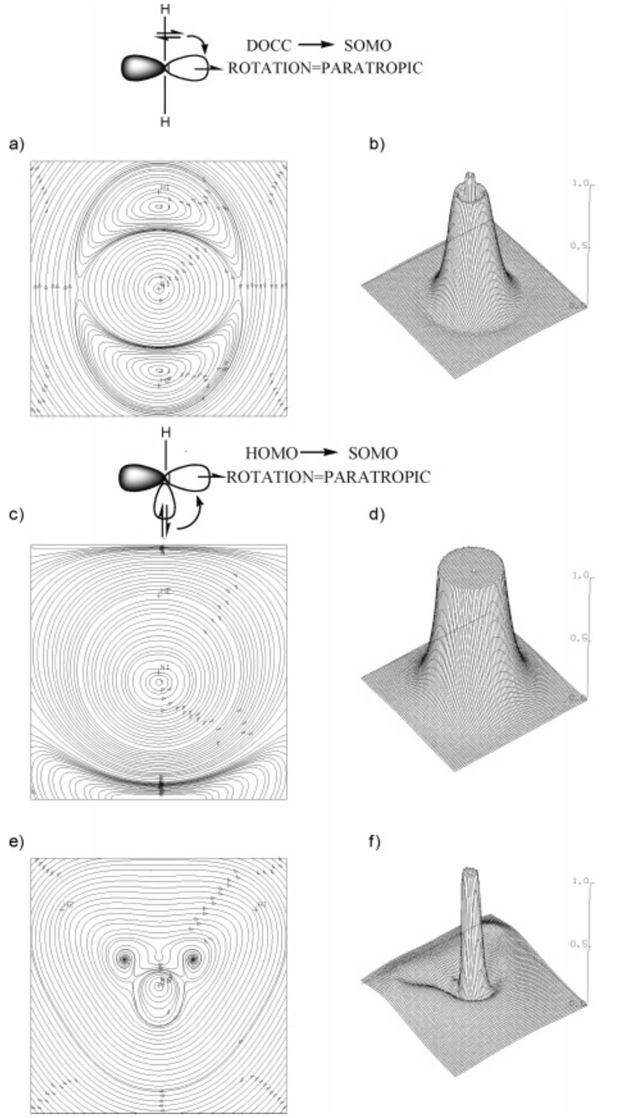


Figure 7. Streamlines (left column) and modulus (right column) of the CROHF-DZ first-order current density $\mathbf{J}^{(1)}$ induced in NH_2, by an external magnetic field, plotted for three different orientations of the field: (a) and (b) field along the C_2 axis; (c) and (d) field in the molecular plane and perpendicular to the C_2 axis; and (e) and (f) field perpendicular to the molecular plane. All maps are plotted over a plane containing the nitrogen atom.

prevails, leading to a positive A_{iso}. However, at heights further away from the plane containing B, the paramagnetic circulation that does not enclose the nuclear probe becomes the dominant feature (see Figure 4b1,4c1), a fact that rationalizes the negative values of $A_{\|}$ and $A_{\perp}$ reported in Table 1.

CH_2^- and NH_2. The radicals CH_2^- and NH_2 share almost identical spin current density maps (see Figures 2 and 3). This fact is a consequence of their virtually identical valence electronic structure. In particular, in both cases the valence space can be described on the basis of three sp^2 hybrid atomic orbitals (AOs) centered on the heavy nucleus and confined to the molecular plane and one p AO perpendicular to the molecular plane. The three MOs formed by the symmetric (a_1) and antisymmetric (b_2) combination of two sp^2 hybrid atomic orbitals, and the third sp^2 hybrid (a_1) pointing in the C_2 direction, are now all doubly occupied, the latter hosting a lone pair. In both radicals, the unpaired electron density occupies the perpendicular p AO (b_1), which thereby now represents the SOMO. When the external field is chosen along the C_2 axis, or perpendicular to it but contained in the molecular plane, the electron spin density distribution is polarized along directions perpendicular to the axis of the p SOMO. Accordingly, the associated spin current density plotted on planes perpendicular to the field displays the typical shape of a p orbital (see Figures 2a,c and 3a,c). In particular, the spin current density map consists of two paramagnetic circulations localized on the lobes of the p AO. Because of the spin polarization effects approximately described by the UHF wave function, also the s orbital on the heavy nucleus carries a nonzero spin density. This translates into a spin current connecting the two lobes, which is equivalent to an effective paramagnetic vortex centered on the heavy nucleus, associated with a Fermi contact contribution to $\mathbf{J}^{(0)}$.

At the height of the heavy nuclei, the Fermi contact circulation dominates the current density maps and results in a positive isotropic HCC. The signature of the contact current can be observed on the maps of $A_{\alpha\beta}(\mathbf{r})$ as a positive peak centered on the heavy nucleus (see Figure 4d0,e0,g0,h0). Further away from the plane containing the heavy nuclei, the two paramagnetic lobe-centered circulations become the dominant feature in the maps, and, since the central nucleus lies outside their anisotropy shielding cones, it experiences an induced field that opposes the external one, resulting in a negative contribution to the integrated HCC. This can be clearly seen in the maps of HCC density in terms of two steep negative peaks centered at the sides of the heavy atom, both at the height of the C and N nuclei (see Figure 4d0,e0,g0,h0), and, even more clearly, at 0.4 a_0 above (Figure 4d1,e1,g1,h1). The deshielding effect arising from the two paramagnetic circulations that do not enclose the nuclear probe rationalizes the negative signs of A_{C2} and $A_{\|}$ reported in Table 1 for CH_2^- and NH_2.

A magnetic field perpendicular to the molecular plane polarizes the spin density along the p SOMO axis. Accordingly, any plot on a plane perpendicular to this direction can be interpreted in terms of a projection of the unpaired electron density on an *s*-like orbital. The maps of spin current density reported in Figures 2e and 3e are indeed dominated by a single paramagnetic vortex centered on the heavy nucleus, as shown by the corresponding maps of HCC density in Figure 4f0,f1,i0,i1. This results in the positive $A_{\perp}$ are reported in Table 1.

6.1.2. First-Order Current Density and Temperature-Independent Contribution to Nuclear Magnetic Shielding. BH_2. Figure 5 shows the maps of current density induced in the BH_2 radical by an external magnetic field, calculated at the CROHF-DZ level of theory. For a field oriented along the C_2 axis (Figure 5a), the induced current density distribution is dominated by two concentric counter-rotating circulations: diatropic (clockwise) and strong the inner one, weak and paratropic the outer one. The inner circulation represents the diamagnetic response of the core electrons. Within the minimal valence space, the outer paratropic circulation can be rationalized as follows. It has been shown that diatropic and paratropic contributions to the first-order current density in closed-shell molecules are associated with virtual transitions from occupied to unoccupied MOs.[24,25]

A given transition provides a diatropic (paratropic) contribution to the total current, if by symmetry it is electric (magnetic) dipole allowed, with respect to those components of the electric (magnetic) dipole operator perpendicular (parallel) to the magnetic field. The intensity of the contribution is proportional to the inverse of the energy gap between the two intervening MOs (which implies that frontier orbital contributions will dominate the response) and depends on the degree of overlap between the occupied MO and the virtual MO transformed by the relevant electric or magnetic dipole operator. The phenomenology of magnetically active orbital transitions in open-shell molecules is enriched by the possibility of nonzero matrix elements between doubly occupied and singly occupied MOs and between singly occupied and virtual MOs. A detailed analysis of all possibilities goes beyond the scope of the present work. Nevertheless, some elementary pictorial concepts based on these symmetry selection rules can help the rationalization of the present results.

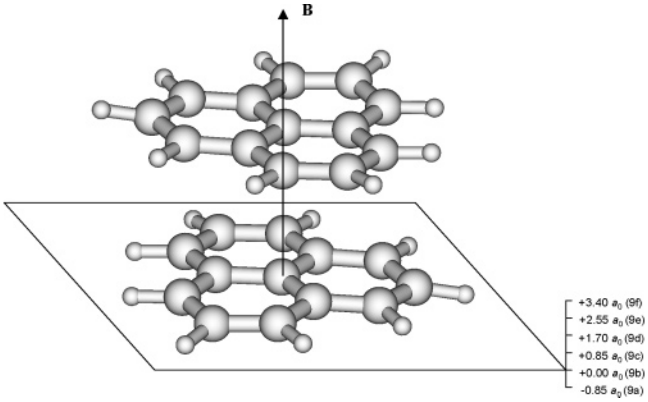
When the field lies along the C_2 axis, the transition from the doubly occupied antisymmetric combination b_2 of in-plane sp^2 hybrid AOs to the lowest unoccupied molecular orbital (LUMO) b_1 is magnetic dipole allowed, i.e., it is rotationally allowed with respect to the a_2 rotation about an axis parallel to the field (see the scheme above Figure 5a), since $b_2 \times a_2 \times b_1 \supset a_1$. Hence, the observed paratropic circulation can be rationalized in terms of the b_2 to b_1 rotationally allowed transition. However, the energy gap between the doubly occupied b_2 and the LUMO b_1 can be quite large. Also, because of the nonlinear geometry of BH_2, only a fraction of the b_2 doubly occupied MO overlaps with the b_1 LUMO after the action of the perpendicular magnetic dipole operator, which explains the weak paratropic response shown in Figure 5a,b, and the small negative value for σ_{C2} reported in Table 2.

A different situation is encountered for a field oriented perpendicular to the C_2 axis and contained in the molecular plane. As shown in Figure 5c, the induced current density map in this case is dominated by a strong paramagnetic vortex centered just below the B atom. This pattern can be understood in terms of the rotationally allowed transition from the a_1 SOMO to the b_1 LUMO, as illustrated in the scheme above Figure 5c. The large magnitude of this transition (and of the corresponding induced current) is due both to the small energy gap between SOMO and LUMO and to the strong overlap between the a_1 sp^2 hybrid after the action of a rotation and the coplanar p LUMO. This transition and the resulting paratropic current explains the large and negative value for $\sigma_{\|}$ reported in Table 2, which also dominates the isotropic response. Finally, it is easy to see that with respect to a rotation about an axis perpendicular to the molecular plane, within the frontier orbital space, there is no transition that is rotationally allowed. Accordingly, the response to a magnetic field oriented along this direction is described by a map of induced current density (Figure 5e,f) dominated by a diatropic circulation centered on the B nucleus, leading to a positive $\sigma_{\perp}$.

CH_2^- and NH_2. The maps of $\mathbf{J}^{(1)}$ for CH_2^- and NH_2 (Figures 6 and 7) present many features in common with those obtained for BH_2. In particular, also in this case the current density map for fields oriented along the C_2 axis (Figures 6a and 7a) or perpendicular to C_2 but contained in the molecular plane (Figures 6c and 7c), is characterized by a paratropic response, weak in the former case, and very strong in the latter. When the field lies along the C_2 axis, the current pattern can be described in terms of a strong, inner diatropic circulation centered on the heavy atom and an outer concentric counter-rotating partropic circulation. The paratropicity originates from a transition from the doubly occupied b_2 MO, to the b_1 p SOMO. The fact that this transition appears to dominate the response in BH_2 and NH_2 leading to a negative σ_{C2} but is overwhelmed by the core diatropic response in CH_2^-, leading to a positive σ_{C2}, can be qualitatively rationalized on the basis of pure geometrical arguments. The experimental geometries employed in the present calculations are characterized by H−X−H angles of 131° (X = B), 103° (X = N), and 99.7° (X = C). Clearly, the smaller the HXH angle (CH_2^- is characterized by the smallest value), the smaller the component of the antisymmetric combination of sp^2 hybrids (b_2 HOMO) that overlaps with the b_1 p-SOMO (or LUMO in the case of BH_2) when rotated by the relevant magnetic dipole operator, and, accordingly, the smaller the contribution of the paratropic transition to the total induced current.

A magnetic field perpendicular to the C_2 axis and contained in the molecular plane induces a strong paratropic current density vortex, as in the BH_2 case (see Figures 6c,d and 7c,d). The rotationally allowed transition at the heart of the observed paratropicity occurs between the a_1 HOMO (the lone pair) and the p-like b_1 SOMO, as for the BH_2 radical, although in that case the transition occurred between the SOMO and the LUMO. The resulting $\sigma_{\|}$, large and negative, is a clear consequence of the overwhelming paratropic circulation. Once again, since no transition within the frontier orbital space is rotationally allowed, the response to a magnetic field along the direction perpendicular to the molecular plane is dominated by a diatropic circulation centered on the heavy nucleus, which leads to positive and large $\sigma_{\perp}$ (see Table 2).

Scheme 1



6.2. Current Density Maps for the Pancake-Bonded Dimer of the Neutral Phenalenyl Radical. The NMR spectrum of the (open-shell singlet) pancake-bonded dimer of the neutral phenalenyl radical has been recently reported[37] and interpreted in terms of the existence of global diatropic ring currents, whose signature in the NMR spectrum is represented by the downfield chemical shift (6.47 ppm) assigned to the proton directly bonded to the aromatic phenalenyl rings (for a model system see Scheme 1). The model on which the assignment relies, i.e., the ring current model, would thus classify the π-dimeric complex as an aromatic molecule, according to the magnetic criterion,[19−22] a fact that was used to corroborate the evidence for the chemical stability of the experimentally characterized complex.[37] The ring current model was further tested and confirmed in ref 37 from a computational standpoint by means of NICS calculations.[21] However, there exists some dispute in the literature as to whether isotropic averages such as downfield proton chemical shifts and NICS calculations can in fact be always considered as reliable magnetic aromaticity indicators,[50,51] especially when the focus is on polycyclic systems as in the present case. Hence, direct visualization of $\mathbf{J}^{(1)}$ in order to test the existence of a ring current in this open-shell singlet represents an important piece of information to assess the magnetic aromaticity of this molecule and represents a natural application of the newly developed methodology. To ensure the open-shell singlet character of the phenalenyl dimer we employed a UDFT method with the GGA functional HCTH,[31] available in Gaussian 03,[44] which has been shown to improve on the calculation of magnetic properties over conventional GGA functionals. From the unperturbed KS orbitals obtained from a Gaussian 03 calculation with a cc-pVDZ basis set, we computed $\mathbf{J}^{(1)}$ using the newly implemented routines in the SYSMO package.[42]

Note that, due to the broken symmetry nature of the solutions obtained from the unrestricted DFT method, an unphysical spin current density $\mathbf{J}^{(0)}$ is obtained for the π-dimeric complex of the phenalenyl radical, despite the fact that its ground state should be a singlet. This is clearly a drawback of the chosen approach. However, two reasons can justify its use in the present investigation. First, the shielding and NICS calculations reported in ref 37 have been performed within the same general approach, i.e., they are broken symmetry DFT calculations, so that the visual

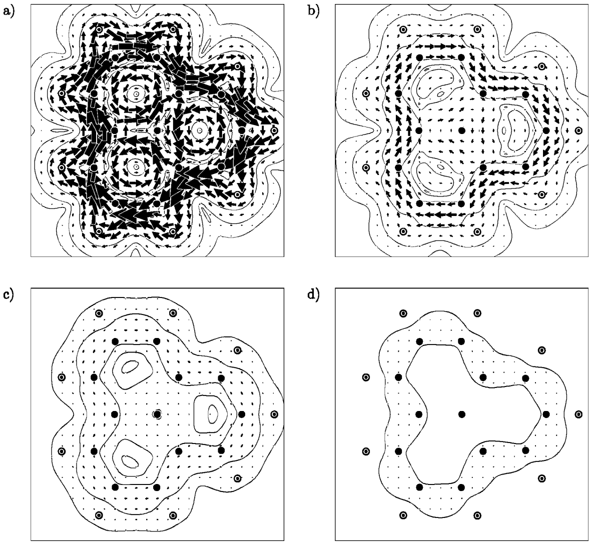


Figure 8. Maps of unrestricted HCTH-(GGA)DFT current density $\mathbf{J}^{(1)}$ induced in D_{3h} neutral phenalenyl radical by a magnetic field perpendicular to the molecular plane plotted as arrows whose area is proportional to the current modulus. The maps show plots on a plane parallel to the molecular plane and distant from it: (a) 0.85 a_0 ($j_{\max} = 0.0929$ c au), (b) 1.7a_0 ($j_{\max} = 0.0332$ c au), (c) 2.55 a_0 ($j_{\max} = 0.0073$ c au), and (d) 3.4 a_0 ($j_{\max} = 0.0018$ c au). Black filled circles (dotted circles) correspond to carbon (hydrogen) nuclei. (Anti)clockwise circulations correspond to diatropic (paratropic) currents.

analysis proposed here is fully consistent with the results that are the object of this analysis. Second, given that spin−orbit coupling can be neglected for this system, we can reasonably expect that the maps of $\mathbf{J}^{(1)}$ obtained from the present broken symmetry calculations will survive at least qualitatively unchanged to more accurate treatments.

First, we computed $\mathbf{J}^{(1)}$ for the neutral phenalenyl monomer. This organic radical is an odd-alternant hydrocarbon with high symmetry (D_{3h}) and is stable in solution under an inert gas atmosphere.[52] The molecule has the ability to form three redox species: cation, radical, and anion. Whereas the ring current aromaticity of the closed shell anion and cation has been assessed via ab initio calculations,[53] no such investigation has been undertaken to date on the ring-current response of the neutral radical system. Accordingly, we first proceeded with the optimization of its D_{3h} structure at the UB3LYP/6-31G* level of theory using the program Gaussian 03.[44] Then we computed the UHCTH/cc-pVDZ maps of current density induced by a magnetic field perpendicular to the molecular plane, over two-dimensional regular grids defined on planes parallel to the molecular plane, up to 3.4 a_0 distance from it (about half the separation between the two phenalenyl units in the pancake-bonded dimer). In Figure 8 the resulting maps of $\mathbf{J}^{(1)}$ are reported: it is evident that the dominant motif can be described in terms of a large diatropic ring current circulating over the 12-carbon perimeter, a clear-cut signature of the magnetic aromaticity of this neutral radical system. At 0.85 a_0 (Figure 8a), close to the maximum of π-electron density, we can observe the ring current at its maximal strength ($j_{\max} = 0.0929$ c au). Further away from the molecular plane the π-electron ring current is still visible (Figure 8b,c), although progressively dying off, so that at 3.4 a_0 (Figure 8d) hardly anything is still visible, the maximal modulus of the current density being only about $j_{\max} = 0.0018$ c au.

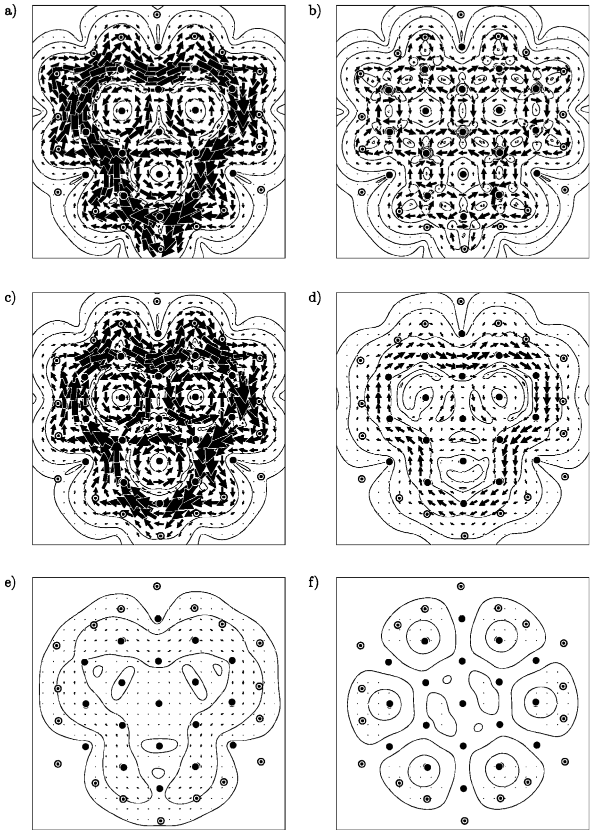


Figure 9. Maps of unrestricted HCTH-GGA-DFT current density $\mathbf{J}^{(1)}$ induced in the pancake-bonded dimer of the neutral phenalenyl radical by a magnetic field (**B**) oriented along the direction connecting the two carbon nuclei at the center of the two monomers (see Scheme 1). The maps show projections of $\mathbf{J}^{(1)}$ on planes at (a) −0.85 a_0 ($j_{\max} = 0.1065$ c au), (b) 0.0 a_0 ($j_{\max} = 1.957$ c au), (c) +0.85 a_0 ($j_{\max} = 0.1054$ c au), (d) +1.7 a_0 ($j_{\max} = 0.0433$ c au), (e) +2.55 a_0, ($j_{\max} = 0.0093$ c au), and (f) +3.4 a_0 ($j_{\max} = 0.0063$ c au) distance from the average height along **B** of the carbon nuclei belonging to one of the nearly planar monomers (see Scheme 1). A scaling (reduction) factor of about 0.3 has been applied to the plot on the molecular plane (b) to ease the visualization.

Next we performed the current density response calculations within the broken symmetry approximation for the pancake-bonded dimer cast in the C_i geometry optimized at the UB3LYP/6-31G* reported in ref 37, but here the *tert*-butyl groups were replaced by hydrogen nuclei and the six new H−C distances reoptimized keeping all other degrees of freedom frozen at the same level of theory. In Figure 9 we report the maps of $\mathbf{J}^{(1)}$ induced in the dimeric π-complex by a magnetic field **B** oriented along the direction connecting the two carbon atoms at the center of the two phenalenyl monomers (chosen as the z-axis) and plotted on planes perpendicular to **B**. The current density plotted on planes at ± 0.85 a_0 distance from the average height along **B** of the carbon nuclei belonging the nearly planar bottom-monomer (see Scheme 1) consists of a strong global diatropic ring current, almost undistinguishable from that plotted at the corresponding height for the D_{3h} monomer, although slightly stronger in magnitude in the dimer case ($j_{\max}$(dimer) ≈ 0.106 c au, $j_{\max}$(monomer) ≈ 0.093 *c* au). Moving toward the middle region of the pancake-bonded dimer, we can observe that the current dies off basically as quickly as in the monomer case (compare Figure 8b−d with Figure 9d−f) although the maximal current magnitude remains consistently larger in the dimer case (see $j_{\max}$ values in the captions to Figures 8 and 9). The slightly larger magnitude of the $\mathbf{J}^{(1)}$ modulus for the dimer in the region of space bracketed by the two monomeric units when compared with that at corresponding heights for the phenalenyl radical can be rationalized in terms of the additional electron density shifted in such region as a consequence of the formation of a weak pancake bond (see ref 37 for a detailed discussion). The increased electron density with respect to the monomer case gives rise to a larger diamagnetic contribution to the ring current.

However, it is interesting to note that the large aromatic NICS values computed in the region between the two monomers are clearly due both to the concerted action of the two bracketing ring currents, which are evidently "localized" above and below such region, and to the slight increase in electron density (and consequently in diatropic current density) in the region between the two monomers, but *no significant ring current exists in the pancake-bond region*. The stronger monomeric ring currents and the electron density shift lead on integration to NICS values that have been shown in ref 37 to be larger in magnitude than the value one would obtain by adding up the distinct contributions from each single monomer. The maps in Figure 9 clearly show that the NICS enhancement is indeed a signature of the aromatic character of the π-complex but also show as clearly that this is mostly due to an overall enhancement of the local aromaticity of the two monomers (which, in fact, are magnetically aromatic in their own right, as seen from Figure 8) rather than to the existence of a significant ring current in the pancake-bonding region. In this respect it appears evident how the actual plot of $\mathbf{J}^{(1)}$ in addition to recovering the information provided by the NICS scan reported in ref 37 leads to a richer picture of the actual space-distribution of the aromatic regions of the molecule, a kind of information that gets completely lost upon integration.

7. Conclusions

In this work we developed a consistent theoretical and computational approach to the representation of the magnetic response of open-shell molecules with small spin−orbit coupling in terms of ab initio spin and charge current density vector fields. Preliminary investigations show that the newly introduced methodology provides a powerful tool for the interpretation of the mechanisms underlying the observables measured in the NMR and ESR experiments, in terms of simple concepts from classical electrodynamics and basic molecular orbital theory.

Acknowledgment. The Francqui Foundation, the Flemish Research Council (FWO), and the University of Leuven are gratefully acknowledged for financial support.

Supporting Information Available: Cartesian coordinates of the 5 molecules for which the calculations presented in the present work have been performed. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mile, B. *Curr. Org. Chem.* **2000**, *4*, 55−83.

(2) Stubbe, J.; van der Donk, W. *Chem. Rev.* **1998**, *98*, 705−762.

(3) Turro, N. J.; Kleinman, M. H.; Karatekin, E. *Angew. Chem., Int. Ed.* **2000**, *39*, 4436−4461.

(4) Bertini, I.; Luchinat, C.; Parigi, G.; Pierattelli, R. *ChemBioChem* **2005**, *6*, 1536−1549.

(5) McConnell, H. M.; Chesnut, D. B. *J. Chem. Phys.* **1958**, *28*, 107−117.;

(6) Atherton, N. M. *Principles of Electron Spin Resonance*; Ellis Horwood and PTR Prentice Hall Press: New York, 1993.

(7) Lushington, G. H.; Grein, F. *Theor. Chim. Acta* **1996**, *93*, 259−267.

(8) Jayatilaka, D. *J. Chem. Phys.* **1998**, *108*, 7587−7594.

(9) Neese, F. *J. Chem. Phys.* **2001**, *115*, 11080−11096.

(10) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231−1253.

(11) Rinkevicius, Z.; Vaara, J.; Telyatnyk, L.; Vahtras, O. *J. Chem. Phys.* **2003**, *118*, 2550−2561.

(12) Moon, S.; Patchkovskii, S. In *Calculation of NMR and EPR parameters*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley: Weinheim, 2004; Part B, Chapter 20, pp 325−338.

(13) Pennanen, T. O.; Vaara, J. *J. Chem. Phys.* **2005**, *123*, 174102.

(14) Hrobàrik, P.; Reviakine, R.; Arbuznikov, A.; Malkina, O. L.; Malkin, V. G.; Köhler, F. H.; Kaupp, M. *J. Chem. Phys.* **2007**, *126*, 024107.

(15) London, F. *J. Phys. Radium* **1937**, *8*, 397−409.

(16) Pauling, L. *J. Chem. Phys.* **1936**, *4*, 673−677.

(17) Pople, J. A. *J. Chem. Phys.* **1956**, *24*, 1111.

(18) Lazzeretti, P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2000**, *36*, 1−88.

(19) Elvidge, J. A.; Jackman, L. M. *J. Chem. Soc.* **1961**, 859−866.

(20) Dauben, H. J.; Wilson, J. D.; Laity, J. L. *Nonbenzenoid aromatics;* Snijder, J. P., Ed.; Academic Press: New York, 1971; Vol. II.

(21) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema, Hommes, N. J. R. *J. Am. Chem. Soc.* **1996**, *118*, 6317−6318.

(22) Fowler, P. W.; Steiner, E.; Havenith, R. W. A.; Jenneskens, L. W. *Magn. Reson. Chem.* **2004**, *42*, S68−S78.

(23) Fowler, P. W.; Steiner, E.; Jenneskens, L. W. *Chem. Phys. Lett.* **2003**, *371*, 719−723.

(24) Steiner, E.; Fowler, P. W. *Chem. Commun. (Cambridge)* **2001**, 2220−2221.

(25) Steiner, E.; Fowler, P. W. *J. Phys. Chem. A.* **2001**, *105*, 9553−9562.

(26) Abragam, A.; Bleaney, B. *Electronic Paramagnetic Resonance of Transition Ions*; Oxford University Press: London, 1970.

(27) McWeeny, R. *Methods of Molecular Quantum Mechanics*; Academic Press: London, 1989.

(28) Pickard, C. J.; Mauri, F. *Phys. Rev. Lett*. **2002**, *88*, 086403.

(29) Patchkovskii, S.; Schreckenbach, G. In *Calculation of NMR and EPR parameters*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley: Weinheim, 2004; Part D, Chapter 32, pp 505−530.

(30) Patchkovskii, S.; Strong, R. T.; Pickard, C. J.; Un, S. *J. Chem. Phys.* **2005**, *122*, 214101.

(31) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264−6278.

(32) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 3015−3023.

(33) Jameson, C. J.; Buckingham, A. D. *J. Phys. Chem.* **1979**, *83*, 3366.

(34) Soncini, A.; Lazzeretti, P. *J. Chem. Phys.* **2003**, *118*, 7165−7173.

(35) Soncini, A.; Fowler, P. W.; Lazzeretti, P.; Zanasi, R. *Chem. Phys. Lett.* **2005**, *401*, 164−169.

(36) Soncini, A.; Lazzeretti, P. *ChemPhysChem.* **2006**, *7*, 679−684.

(37) Suzuki, S.; Morita, Y.; Fukui, K.; Sato, K.; Shiomi, D.; Takui, T.; Nakasuji, K. *J. Am. Chem. Soc.* **2006**, *128*, 2530−2531.

(38) Landau, L. D.; Lifshitz, E. M. *Quantum Mechanics*; Pergamon: Oxford, 1981.

(39) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *J. Mol. Struct. THEOCHEM* **1994**, *313*, 299−312.

(40) McWeeny, R.; Steiner, E. *Adv. Quantum Chem.* **1965**, *2*, 93−117.

(41) Juselius, J.; Sundholm, D.; Gauss, J. *J. Chem. Phys*. **2004**, *121*, 3952−3963.

(42) Lazzeretti, P.; Zanasi, R. *SYSMO package*; University of Modena: 1980. Additional routines for the evaluation and plotting of current density: E. Steiner, P. W. Fowler, R. W. A. Havenith, A. Soncini. For (C)ROHF, (C)UHF, and (U)GGA-DFT calculations: A. Soncini.

(43) McWeeny, R.; Diercksen, G. *J. Chem. Phys.* **1968**, *49*, 4852−4856.

(44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(45) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, *210*, 223−231.

(46) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Chem. Phys. Lett.* **1994**, *200*, 299−304.

(47) Zanasi, R. *J. Chem. Phys.* **1996**, *105*, 1460−1469.

(48) Baird, N. C. *J. Am. Chem. Soc.* **1972**, *94*, 4941−4948.

(49) Gogonea, V.; Schleyer, P. V. R.; Schreiner, P. R. *Angew. Chem., Int. Ed.* **1998**, *37*, 1945−1948.

(50) Wannere, C. S.; Corminboeuf, C.; Allen, W. D.; Schaefer, H. F., III; Schleyer, P. v. R. *Org. Lett.* **2005**, *7*, 1457 -1460.

(51) Faglioni, F.; Ligabue, A.; Pelloni, S.; Soncini, A.; Viglione, R. G.; Ferraro, M. B.; Zanasi, R.; Lazzeretti, P. *Org. Lett.* **2005**, *7*, 3457 -3460.

(52) Reid, D. H. *Q. Rev.* **1965**, *19*, 274.

(53) Cyranski, M. K.; Havenith, R. W. A.; Dobrowolski, M. A.; Gray, B. R.; Krygowski, T. M.; Fowler, P. W.; Jenneskens, L. W. *Chem. Eur. J.* **2007**, *13*, 2201−2207.

CT700169H

JCTC Journal of Chemical Theory and Computation

7-Norbornyl Cation − Fact or Fiction? A QTAIM-DI-VISAB Computational Study

Nick H. Werstiuk*

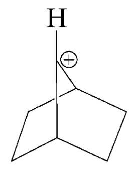
Department of Chemistry, McMaster University, Hamilton ON L8S 4M1, Canada

Received July 15, 2007

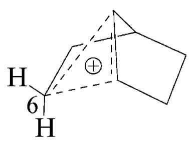
Abstract: QTAIM-DI-VISAB analyses were used to characterize the bonding of the ‘nonclassical’ 7-norbornyl cation and its rearrangement transitions states. These analyses involved obtaining QTAIM molecular graphs and delocalization indexes (DIs) that were correlated with the proximities of atomic basins (VISAB). This study showed that the so-called 7-norbornyl cation actually exhibits the molecular graph of the bicyclo[3.2.0]heptyl cation at its equilibrium geometry. Dynamical aspects of its molecular graph/density were explored with QTAIM by analyzing the nuclear motions of the 206 cm^{-1} normal mode. This study cements the QTAIM-DI-VISAB analysis as a method of choice for establishing the nature of the bonding in so-called nonclassical carbocations while obviating the need for dotted-line representations of bonding.

Introduction

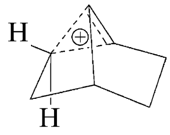
The structure of the 7-norbornyl cation (**1**) was the focus of many experimental and theoretical studies for several decades, one of the latest being the work of Mesić et al.[1] This activity followed the suggestion by Winstein[2] in 1958 that it should be considered as a tricycloheptonium nonclassical cation shown in its usual dashed-line representation as **7-like**; this cation was considered as the common intermediate generated in solution by solvolysis of 7-bicyclo[2.2.1]heptyl *p*-bromobenzenesulfonate and exo-2-bicyclo[3.2.0]heptyl *p*-bromobenzenesulfonate.[2−4] The hypothetical 2-bicyclo[3.2.0]heptyl cation is shown as **2**. The first standard high-level computational study on the so-called 7-norbornyl cation was carried out by Sieber et al.[5] in 1993, and it recently has been implicated in reaction mechanisms of the fragmentation of 7-norbornyloxy(chloro)carbene.[6] In all publications to this point the species in question has been named the 7-norbornyl cation, and a variety of dashed-line structural formulas was used, including the usual ones that are 7-norbornyl-like—shown as **7-like**—and 2-bicyclo[3.2.0]-like shown as **[3.2.0]-like**.



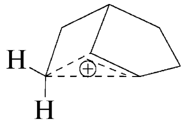
1



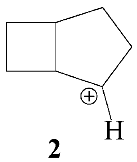
7-like



7-like-e



[3.2.0]-like



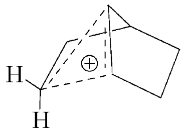
2

In fact, dotted/dashed lines, hollow tubes, and solid tubes of ORTEP drawings, and combinations thereof have been used in publications in attempts to represent the bonding of this so-called nonclassical carbocation with the implication being that C6 of **7-like** is a pentacoordinate atom. In exploring the 7-norbornyl potential energy surface computationally, Sieber et al.[5] located three species

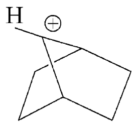
* Corresponding author phone: (905)525-9140; fax: (905)522-2509; e-mail: werstiuk@mcmaster.ca.

10.1021/ct700176d CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/24/2007

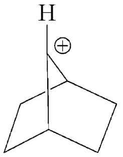
as stationary points—**1-*C₁***, **1-*C_S***



1-*C₁*



1-*C_S*



1-*C_{2V}*

in which C7 leaned toward one ethano bridge, and **1-C_{2V}** at the MP2(full)/6-31G(d) level. However, **1-C_S** and **1-C_{2V}** possessed an imaginary frequency; based on the nature of the vibrational mode corresponding to the imaginary frequency, **1-C_S** was viewed as the transition state for the same-face degenerate rearrangement of **1-C_1** to its enantiomer **7-like-e**, and **1-C_{2V}** the transition state for bridge flapping of **1-C_S**; the details of the potential energy surface for the **1-C_1**, **1-C_S**, and **1-C_{2V}** interconversion were not defined.

Our recent computational studies on a number of cations, including 2-norbornyl, established that coordination based on the number of bond paths—as defined in a QTAIM (Quantum Theory of Atoms in Molecules)[7] molecular graph—terminating at a nucleus in any species—cation, carbanion, radical, or carbene—should be used as the criterion of hypercoordination and hypervalency.[8−11] We argued that this approach should be used regardless of the nature of the intermediate to obviate the confusion and inaccuracies associated with using indicators such as dashed lines, dotted lines, cross-hatched lines, hollow tubes, and solid tubes in structural formulas. In addition to using QTAIM molecular graphs to show *molecular* structure we recently showed that QTAIM-DI-VISAB analyses—a combination of QTAIM molecular graphs, an evaluation of delocalization indexes (DIs), and a visualization of the closeness of atomic basins (VISAB)—are useful for characterizing the bonding in molecules at their equilibrium geometries.[12,13] This paper reports the results of a QTAIM-DI-VISAB study on the bonding of the so-called 7-norbornyl cation.

Computational Methods

Our previous experiences with DFT calculations on carbocations clearly showed that the B3PW91 hybrid functional is superior to B3LYP in computing the geometries of delocalized, so-called nonclassical species.[8−11] To provide additional support for this finding, calculations were carried out on O-protonated 2,2-dimethyloxirane—studied recently by Carlier et al.[14] and described as a particularly challenging computational problem—to compare results from B3LYP, B3PW91, PBE1PBE, and CCSD calculations at the 6-311G(d,p) level as implemented in G03.[15] The results—including Carlier's data obtained at the 6-311++G(d,p) level shown in italics—are summarized in Figure 1. It is clear that B3PW91 and PBE1PBE are expected to be superior to B3LYP in cases where relatively weak polar bonds are involved. Cation geometries were optimized at B3PW91/6-311G(d,p), PEB1PBE/6-311G(d,p), and CCSD(full)/6-311G(d,p) levels with GaussView[16] being used to fix C_S and C_{2V} symmetries where necessary. Selected internuclear distances are collected in Figure 2, and the Cartesian coordinates of the optimized geometries are given in Tables 1S−9S (Supporting Information). The average geometry—Cartesian coordinates are given in Table 10S (Supporting Information)—of **1-C_1** at 0 K was obtained with a G03 FREQ=ANHARM calculation at B3PW91/6-311G(d,p). Cation **1-C_1** was also studied at the B3PW91 and PBE1PBE levels with the Carlier basis set (6-311++G(d,p)) to probe the effect of increasing basis-set size; the results—values in brackets—given in Figure 2 show that the geometrical effects are negligible. Frequency calculations were carried out on the stationary points at the 6-311G(d,p) level to confirm them as energy minima or transitions states. CCSD minima were confirmed with MP2 frequency calculations. Thermochemical data are collected in Table 1 (B3PW91), Table 2 (PBE1PBE), and Table 3 (CCSD). While $\Delta E^{\ddagger}_{\mathrm{lec}}$, $\Delta E^{\ddagger}_{0}$, $\Delta E^{\ddagger}_{298}$, and $\Delta H^{\ddagger}_{298}$ were computed, only $\Delta E^{\ddagger}_{\mathrm{elec}}$, $\Delta H^{\ddagger}_{298}$, ΔE_{elec}, and ΔH_{298} are included in the discussion.

Method	**C-O(Å)**
CCSD	*1.599*
MP2	*1.598*
PBE1PBE	*1.634*
B3PW91	*1.671*
B3LYP	*1.790*
CCSD/6-311G(d,p)	1.593
PBE1PBE/6-311G(d,p)	1.628
B3PW91/6-311G(d,p)	1.660
B3LYP/6-311G(d,p)	1.731

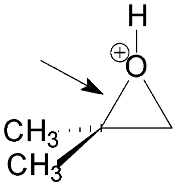
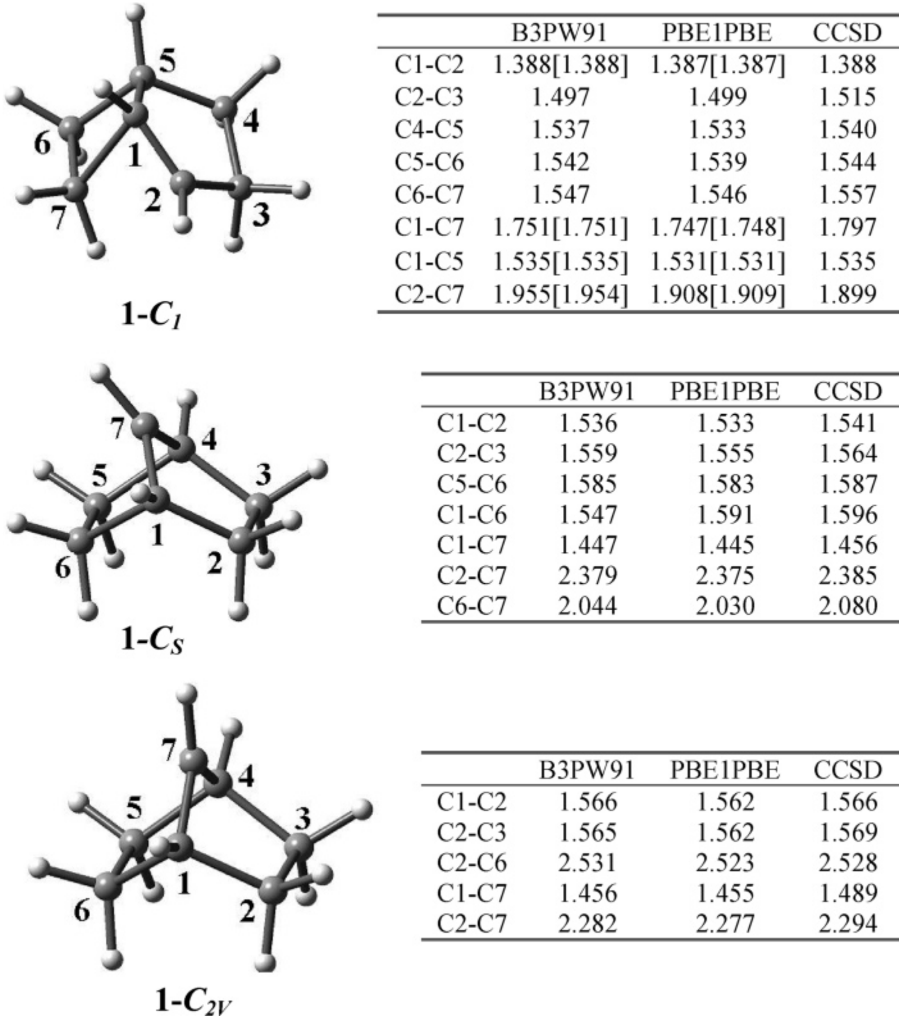


Figure 1. C−O distances of O-protonated 2,2-dimethyloxirane at various levels of theory. Data in italics obtained by Carlier et al. at 6-311++G(d,p).

QTAIM analyses of the wave functions to investigate the topologies of the electron densities were carried out with AIM2000,[17] and values of $\rho(\mathbf{r}_c)$ at selected bond critical points are collected in Figure 3. AIMALL97[18] was used to integrate atomic basins, to obtain atomic populations, total charges, and atomic overlap matrices required for DI calculations. That the total charges of **1-C_1**, **1-C_S**, and **1-C_{2V}** obtained at the various levels of theory (Figure 3) were less than 1% higher than the expected value of 1.0 onfirmed the quality and validity of the QTAIM data. The program LIDICALC[19,20] was used to obtain DIs; selected values for pairs of atoms are listed in Figure 4. Isosurface plots of the density (Figure 5(b),(c)) and the Laplacian of the density (Figure 5(d)) were obtained with the B3PW91/6-311G(d,p) wave function using NABLA[21] to obtain the grid points and OpenDX[22] to generate the plots. Atomic basins were obtained with AIM2000 at a contour value of 0.005 au—this includes >95% of the electrons—using a mesh grid size of 0.125 and plotted with a sphere size of 0.15. GaussView[16] was used to simulate an IR spectrum of **1-C_1** and obtain nuclear displacement vectors. GaussView and ChemCraft[23] were used to animate the normal modes; the 206 cm^{-1} mode, the one that appeared to bring C2 and C7 within a distance where a BP could be formed, was selected for a detailed analysis. Using ChemCraft, the nuclear motions of this mode—the G03 displacement coordinates were scaled by 0.35—were frozen



	B3PW91	PBE1PBE	CCSD
C1-C2	1.388[1.388]	1.387[1.387]	1.388
C2-C3	1.497	1.499	1.515
C4-C5	1.537	1.533	1.540
C5-C6	1.542	1.539	1.544
C6-C7	1.547	1.546	1.557
C1-C7	1.751[1.751]	1.747[1.748]	1.797
C1-C5	1.535[1.535]	1.531[1.531]	1.535
C2-C7	1.955[1.954]	1.908[1.909]	1.899

	B3PW91	PBE1PBE	CCSD
C1-C2	1.536	1.533	1.541
C2-C3	1.559	1.555	1.564
C5-C6	1.585	1.583	1.587
C1-C6	1.547	1.591	1.596
C1-C7	1.447	1.445	1.456
C2-C7	2.379	2.375	2.385
C6-C7	2.044	2.030	2.080

	B3PW91	PBE1PBE	CCSD
C1-C2	1.566	1.562	1.566
C2-C3	1.565	1.562	1.569
C2-C6	2.531	2.523	2.528
C1-C7	1.456	1.455	1.489
C2-C7	2.282	2.277	2.294

Figure 2. Selected internuclear distances of **1-*C₁***, **1-*C_S***, and **1-*C_2V*** at B3PW91, PBE1PBE, and CCSD(full). Values in square brackets obtained with the Carlier basis set (6-311++G(d,p)).

Table 1. Total and Relative Energies of Cations at B3PW91/6-311G(d,p)

				protonated alcohols	
cation	**1-*C*₁**	**1-*C*S**	**1-*C*2V**	**3**	**4**
E_{elec}[a]	−273.009458 (*0.00*)	−273.003910 (*3.48*) [e] (−233.4 cm⁻¹)[g]	−272.998026 (*7.17*) [e] (−321.2 cm⁻¹)[g]	−349.473743 (*0.00*)	−349.455802 (*11.26*) [f]
E_0[b]	−272.846302 (*0.00*)	−272.841025 (*3.32*)	−272.835234 (*6.65*)	−349.280161 (*0.00*)	−349.264575 (*9.78*)
E_{298}[c]	−272.840116 (*0.00*)	−272.835415 (*2.95*)	−272.829529 (*6.64*)	−349.272751 (*0.00*)	−349.256556 (*10.16*)
H_{298}[d]	−272.839172 (*0.00*)	−272.834471 (*2.95*)	−272.828585 (*6.64*)	−349.271806 (*0.00*)	−349.255612 (*10.16*)

[a] E_{elec} is the uncorrected total energy in hartrees. [b] $E_0 = E_{\text{elec}} + \text{ZPE}$. [c] $E = E_0 + E_{\text{vib}} + E_{\text{rot}} + E_{\text{trans}}$. [d] $H = E + RT$. [e] Values in brackets relative to **1-*C₁*** in kcal mol ⁻¹. [f] Relative to 3. [g] The imaginary frequency.

Table 2. Total and Relative Energies of Cations at PBE1PBE/6-311G(d,p)

cation	**1-*C*₁**	**1-*C*S**	**1-*C*2V**
E_{elec}[a]	−272.774310(*0.00*)	−272.768946(*3.36*)[e] (−244.4 cm⁻¹) [f]	−272.762413(*7.74*)[e] (−330.3 cm⁻¹) [f]
E_0[b]	−272.610299(*0.00*)	−272.605325(*2.79*)	−272.598889(*7.16*)
E_{298}[c]	−272.604191(*0.00*)	−272.599745(*2.79*)	−272.593200(*6.90*)
H_{298}[d]	−272.603247(*0.00*)	−272.598801(*2.79*)	−272.592256(*6.90*)

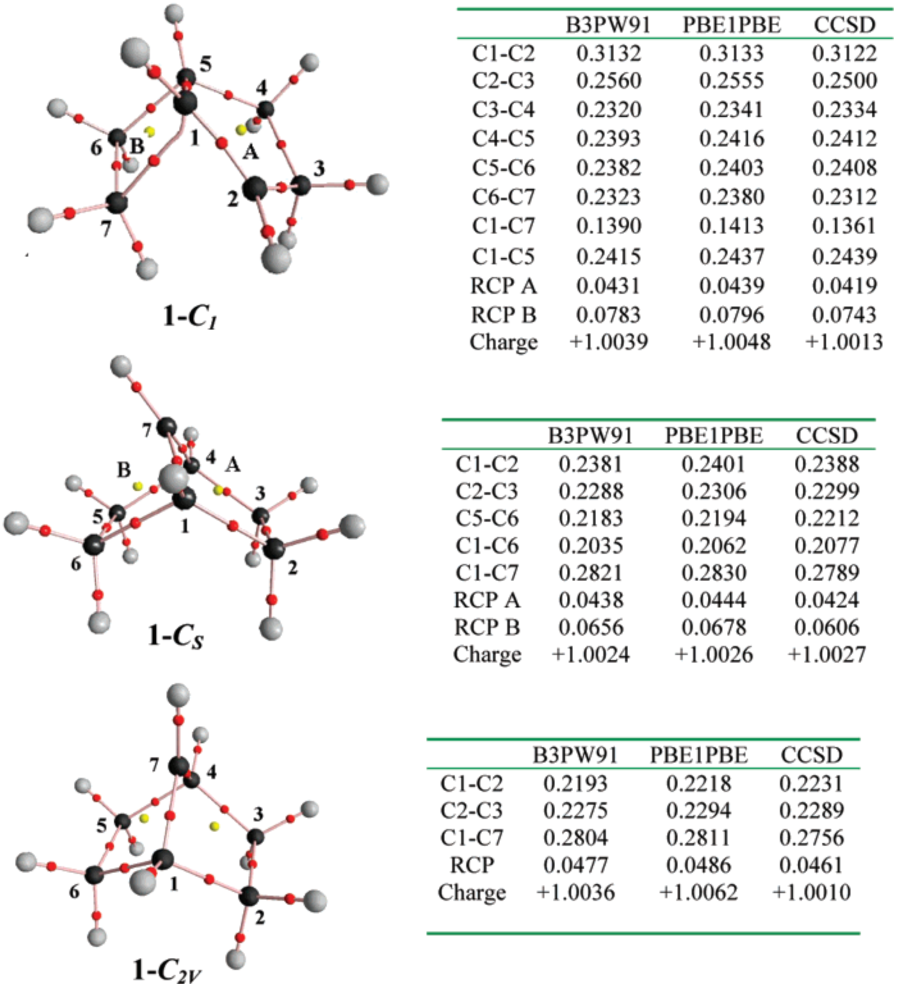
[a] E_{elec} is the uncorrected total energy in hartrees. [b] $E_0 = E_{\text{elec}} + \text{ZPE}$. [c] $E = E_0 + E_{\text{vib}} + E_{\text{rot}} + E_{\text{trans}}$. [d] $H = E + RT$. [e] Values in brackets relative to **1-*C₁*** in kcal mol⁻¹. [f] The imaginary frequency.

at ten intervals including geometries with the largest (2.118) and shortest distances (1.795 Å) between C2 and C7. Cartesian-coordinate files were written, and single-point calculations with SCF=TIGHT were carried out to obtain wave functions. The C2−C7 distances, uncorrected electronic energies (E_{elec}), and relative energies (ΔE_{elec}) of the ten geometries along with the equilibrium geometry of **1-*C₁*** are collected in Table 4. QTAIM analyses of the wave functions

Table 3. Total and Relative Energies of Cations at CCSD(full)/6-311G(d,p) and MP2(full)/6-311G(d,p)

cation	**1-C_1**	**1-C_S**	**1-C_{2V}**
CCSD(full)/6-311G(d,p)			
E_{elec}[a]	−272.440844 (0.00)	−272.436078 (+2.99)[f]	−272.431628 (+5.78)[f]
MP2(full)/6-311G(d,p)[b]			
E_{elec}[b]	−272.371813(0.00)	−272.363435 (5.26) (−315.0)[g]	−272.356284 (9.74) (−322.6)[g]
E_0[b,c]	−272.206471 (0.00)	−272.198947 (4.72)	−272.191814 (9.19)
E_{298}[b,d]	−272.200445 (0.00)	−272.193318 (4.47)	−272.186104 (9.00)
H_{298}[b,e]	−272.199501 (0.00)	−272.192372 (4.47)	−272.185160 (9.00)

[a] E_{elec} is the uncorrected CCSD(full) total energy in hartrees. [b] From a single point frequency calculation at MP2(full)/6-311G(d,p) on the CCSD(full)/6-311G(d,p) geometry. [c] $E_0 = E_{elec} + \text{ZPE}$. [d] $E = E_0 + E_{vib} + E_{rot} + E_{trans}$. [e] $H = E + RT$. [f] Values in brackets relative to **1-C_1** in kcal mol^{-1}. [g] The imaginary frequency.



	B3PW91	PBE1PBE	CCSD
C1-C2	0.3132	0.3133	0.3122
C2-C3	0.2560	0.2555	0.2500
C3-C4	0.2320	0.2341	0.2334
C4-C5	0.2393	0.2416	0.2412
C5-C6	0.2382	0.2403	0.2408
C6-C7	0.2323	0.2380	0.2312
C1-C7	0.1390	0.1413	0.1361
C1-C5	0.2415	0.2437	0.2439
RCP A	0.0431	0.0439	0.0419
RCP B	0.0783	0.0796	0.0743
Charge	+1.0039	+1.0048	+1.0013

	B3PW91	PBE1PBE	CCSD
C1-C2	0.2381	0.2401	0.2388
C2-C3	0.2288	0.2306	0.2299
C5-C6	0.2183	0.2194	0.2212
C1-C6	0.2035	0.2062	0.2077
C1-C7	0.2821	0.2830	0.2789
RCP A	0.0438	0.0444	0.0424
RCP B	0.0656	0.0678	0.0606
Charge	+1.0024	+1.0026	+1.0027

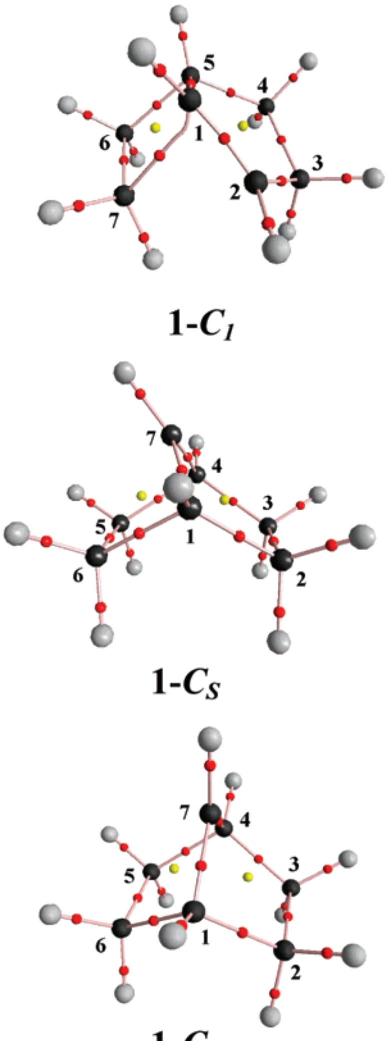
	B3PW91	PBE1PBE	CCSD
C1-C2	0.2193	0.2218	0.2231
C2-C3	0.2275	0.2294	0.2289
C1-C7	0.2804	0.2811	0.2756
RCP	0.0477	0.0486	0.0461
Charge	+1.0036	+1.0062	+1.0010

Figure 3. QTAIM molecular graphs of **1-C_1, 1-C_S**, and **1-C_{2V}** at B3PW91, values of $\rho(\mathbf{r})$ at bond and ring critical points of **1-C_1**, **1-C_S**, and **1-C_{2V}**, and total charges at B3PW91, PBE1PBE, and CCSD(full): black spheres carbons, gray spheres hydrogens, red spheres BCPs, and yellow spheres RCPs.

yielded 11 molecular graphs that were converted to JPEG files—the molecular graphs along with frame numbers are displayed in Figure 4S (Supporting Information). The JPEG files were combined in the sequence (also see Table 4 for assignments) F1, F2, F3, F4, F5, F6, F7, F8, F9, F8, F7, F6, F5, F4, F3,F2, F1, F10, F11, F10, and F1 to yield a 21-frame animation of the changes in the bicyclo[3.2.0]heptyl molecular graph during the nuclear motions associated with the 206 cm^{-1} mode. The resulting motion picture—viewable with common media players such as Windows Media Player—is included in the Supporting Information.

Results and Discussion

Thermochemistry. At all levels of theory **1-C_1** is a minimum on the PE surface, and **1-C_S** and **1-C_{2V}** are transition states. Based on the normal mode associated with its imaginary frequency, **1-C_S** is the transition state for the same-face rearrangement of **1-C_1** to its enantiomer. The barrier of this



1-C_1

	B3PW91	PBE1PBE	CCSD
C1-C2	1.2558	1.2517	1.0741
C4-C5	0.9649	0.9662	0.8139
C5-C6	0.9519	0.9533	0.8020
C6-C7	0.9511	0.9505	0.7953
C1-C7	0.6150	0.6069	0.4527
C1-C5	0.9346	0.9353	0.7895
C2-C7	0.4287	0.4570	0.3783

1-C_S

	B3PW91	PBE1PBE	CCSD
C1-C2	0.9590	0.9609	0.8080
C2-C3	0.9607	0.9622	0.8045
C1-C6	0.8513	0.8545	0.7243
C1-C7	1.0684	1.0665	0.9017
C2-C7	0.0621	0.0620	0.0483
C6-C7	0.2289	0.2328	0.1572
C2-C6	0.0422	0.0428	0.0321

1-C_{2V}

	B3PW91	PBE1PBE	CCSD
C1-C2	0.9019	0.9046	0.7683
C1-C7	1.0689	1.0657	0.8956
C2-C7	0.1224	0.1221	0.0851
C2-C6	0.0446	0.0451	0.0336

Figure 4. QTAIM molecular graphs of **1-C_1**, **1-C_S**, and **1-C_{2V}** at B3PW91 and delocalization indices of selected atom pairs of **1-C_1**, **1-C_S**, and **1-C_{2V}** at B3PW91, PBE1PBE, and CCSD(full).

degenerate rearrangement is very low—$\Delta H^{\ddagger}_{298}$ is in the range of 3 kcal mol^{-1} at B3PW91 (Table 1), PBE1PBE (Table 2), and CCSD(full) (Table 3) levels. There is little variation in going from $\Delta E(\Delta E^{\ddagger})$ to $\Delta H_{298}(\Delta H^{\ddagger}_{298})$. **1-$C_{2V}$** is higher in energy than **1-C_S**—the values of $\Delta H^{\ddagger}_{298}$ for the bridge flapping from **1-C_S** at the three levels of theory are 3.69 (B3PW91), 4.11 (PBE1PBE), and 4.53 (CCSD(full)/MP2(full)) kcal mol^{-1}. **1-C_{2V}** exhibits only one large negative eigenvalue and, as indicated by the nature of the 'vibrational mode' associated with the imaginary frequency, **1-C_{2V}** appears to be the transition state for bridge flapping between **1-C_S** ions ($\Delta H_{298} = \Delta H^{\ddagger}_{298}$). This result suggests that the '7-norbornyl cation' gas-phase PE surface is characterized by a bifurcation pathway as displayed in Figure 4(b) of a paper by Xantheas et al.[24] The values of $\Delta H^{\ddagger}_{298}$ relative to **1-C_1** are 6.64 (B3PW91), 6.90 (PBE1PBE), and 9.00 (CCSD/MP2) kcal mol^{-1}. O-protonated 7-norbornanol (**3**) and O-protonated exo-2-bicyclo[3.2.0]heptanol (**4**) were also studied at the B3PW91/6-311G(d,p) level. The molecular graphs of **3** and **4** are displayed in parts (a) and (b), respectively, of Figure 1S along with internuclear distances and values of $\rho(\mathbf{r}_c)$—in parentheses—of selected BCPs (Supporting Information). At this level, **3** is significantly lower in energy than **4**; ΔH_{298} is −10.16 kcal mol^{-1}. This result is in keeping with the fact that 7-norbornyl substituted products predominate in the solvolysis of 7-norbornyl and 2-bicyclo[3.2.0]heptyl substrates.[2−4]

Equilibrium Molecular and Geometrical Structures. ***1-C_1*.** The molecular graph of the equilibrium geometry of the '7-norbornyl' cation obtained at the B3PW91/6-311G(d,p) level is displayed in Figure 3 (**1**-C_1), Figure 4 (**1**-C_1), and Figure 5(a). It is clear that the so-called 7-norbornyl cation, in fact, exhibits the bicyclo[3.2.0]heptyl cation molecular graph at its equilibrium geometry! Identical molecular graphs (not displayed) were also obtained at the PBE1PBE/6-311G(d,p), CCSD(full)/6-311G(d,p), B3PW91/6-311++G(d,p), and PBE1PBE/6-311++G(d,p) levels. The plots of the density ($\rho(\mathbf{r})$) of the equilibrium geometry at contour values of 0.095 (Figure 5(b)) and 0.125 au (Figure 5(c)) and the Laplacian ($-\nabla^2\rho$) (Figure 5(d), contour value 0.005) are nicely in accord with the [3.2.0] molecular graph. Neither shows a 'bridge' that includes C2 and C7. The key point is that C7 does not have five bond paths terminating at the nucleus. Consequently, **1-C_1** is NOT a pentacoordinate

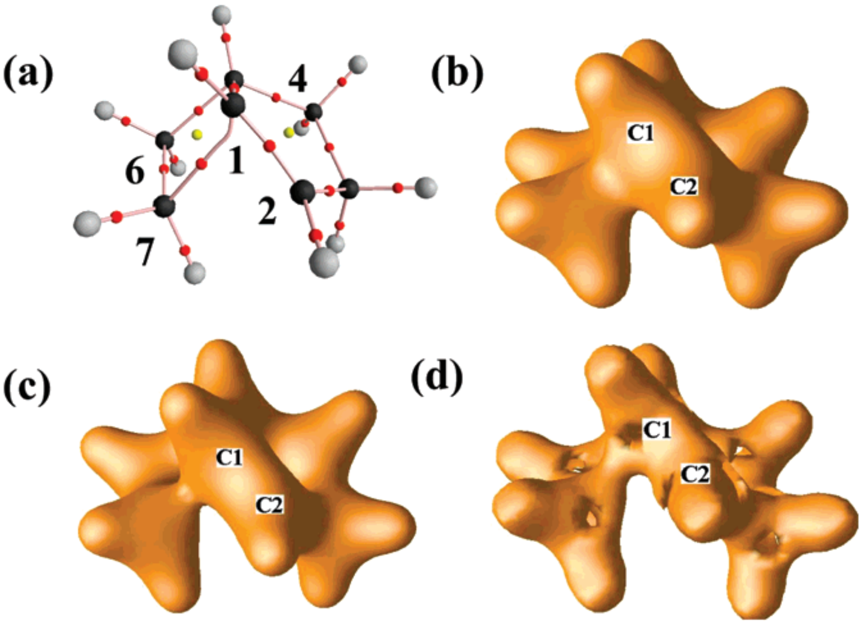


Figure 5. (a) Molecular graph of **1-*C₁*** at B3PW91/6-311G(d,p); (b) density (ρ) of **1-*C₁*** at a contour value of 0.095; (c) density (ρ) of 1-***C₁*** at a contour value of 0.125; and (d) Laplacian ($-\nabla^2\rho$) of **1-*C₁*** at a contour value of 0.005 at B3PW91/6-311G(d,p).

Table 4. Total and Relative Energies of Freeze-Frame Geometries of the 206 cm^{-1} Mode of **1-*C₁*** at B3PW91/6-311G(d,p)

C2−C7 distance/Å[a]	E_{elec}[b]	ΔE_{elec}/kcal mol^{-1} (relative to the equilibrium geometry)
2.118(max.)[**F11**][c]	−273.008135	0.830
1.994[**F10**]	−273.009382	0.048
1.954(equil geom)[**F1**]	−273.009458	0.000
1.906[**F2**]	−273.009338	0.075
1.858[**F3**]	−273.008939	0.326
1.821[**F4**]	−273.008412	0.656
1.816[**F5**]	−273.008327	0.710
1.814[**F6**]	−273.008291	0.733
1.812[**F7**]	−273.008256	0.754
1.807[**F8**]	−273.008142	0.826
1.795(min.)[**F9**]	−273.007949	0.947

[a] Freeze-frame analysis carried out with ChemCraft: the G03 206 cm^{-1} mode; displacement coordinates scaled by 0.35. [b] The uncorrected total energy E_{elec} in hartrees. [c] The frame numbers of the molecular graphs used in the molecular-graph motion picture. See Figure 3S (Supporting Information).

species in its equilibrium geometry. This is also the case at the PBE1PBE and CCSD levels—the CCSD molecular graph is displayed as Figure 2S(a) (Supporting Information)—even though the C2−C7 internuclear distances are significantly shorter in these cases relative to the B3PW91. At all levels, C1−C2 exhibits considerable double-bond character. The internuclear distances are all close to 1.39 Å, and the $\rho(r_c)$ values at the BCPs lie in the region of 0.31 au, considerably higher than the values (Figure 3) of 0.24 and 0.22 for the C1−C2 BCPs of **1-*C_S*** and **1-*C_{2V}***, respectively. C1−C7 of **1-*C₁*** is a weak bond—$\rho(\mathbf{r}_c)$ lies in the region of 0.13 to 0.14—relative to C1−C6 of **1-*C_S*** (0.2035) and **1-*C_{2V}*** (0.2193).

That vibrational frequencies were calculated prompted an examination of the normal modes of **1-*C₁***. In nuclear configuration space there may be an arrangement of nuclei where C2 and C7 are transiently connected by a BP due to molecular vibrations. This may be general in cases of this type where the density is flat and there is a high ellipticity of bonds with the soft axis laying in the 3-atom plane along the existing BPs. This appears to be the case in **1-*C₁***; the ellipticity is 1.790 at the C1−C7 BCP, significantly higher than the ellipticity (0.011) at BCP of the 'normal' single bond C4−C5 of **1-*C₁***. The simulated IR spectrum of **1-*C₁*** computed with G03 exhibited two strong bands at 206 and 536 cm^{-1}. When animated, only the 206 cm^{-1} mode of the bands below 600 cm^{-1} appeared to bring C2 and C7 closer together, possibly to a point where a BP and BCP transiently materialize between C2 and C7. To establish whether the nuclear motions of the 206 cm^{-1} mode resulted in the formation of a BCP/BP between C2 and C7, ChemCraft was used to freeze the nuclear motions—the G03 displacement coordinates were scaled by 0.35—at ten intervals to obtain snapshots of displacement geometries. The C2−C7 distances ranged from 2.118 to 1.795 Å (Table 4). It is seen that the C1−C7 BP switches to a highly curved BP between C2 and C7 when the C2−C7 distance was 1.814 Å (**F6**, Figure 4S (Supporting Information)) with the 1.814-Å geometry being 0.733 kcal mol^{-1} higher in energy than the equilibrium geometry. Given that the 206 cm^{-1} mode has a ZPE of 0.295 kcal mol^{-1} ((206 cm^{-1} × 2.86 cal cm^{-1})/2) it is unlikely that the 1.814-Å geometry is achieved at 0 K. Moreover, the average geometry of **1-*C₁*** at 0 K does not exhibit a BP between C2 and C7 (molecular graph not shown). The 1.816-Å geometry also exhibits a molecular graph that closely approaches a T-structure (**F5**, Figure 4S) we found for the equilibrium geometry of the 2-norbornyl cation.[10] It is important to note that Pendas et al.[25] have confirmed that QTAIM BCPs/BPs are valid probes of bonding in nonequilibrium structures. The point is that **1-*C₁*** exhibits a bicyclo-[3.2.0] molecular graph at its equilibrium geometry, and it does not exhibit a pentacoordinate C7 as it does not have

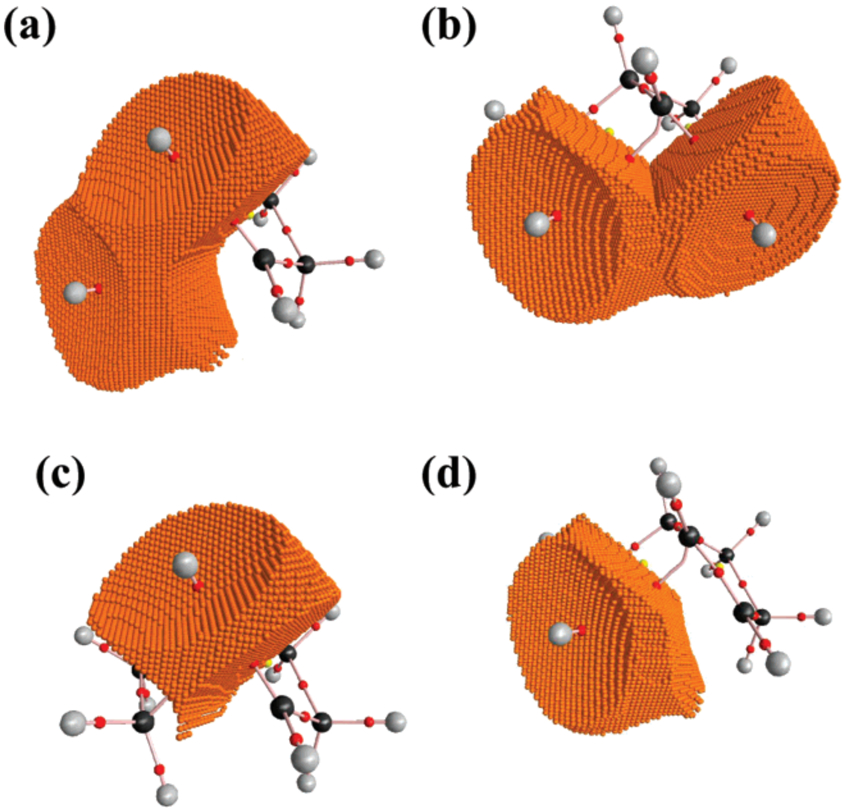


Figure 6. Atomic basins of **1-*C*$_1$** at B3PW91/6-311G(d,p): (a) C1, C7; (b) C2, C7; (c) C1; and (d) C7.

five bond paths terminating at the nucleus even during the 206 cm^{-1} vibration.

1-C$_S$. The molecular graph of **1-*C*$_S$** obtained at the B3PW91/6-311G(d,p) level is displayed in Figures 3 (**1-*C*$_S$**) and 7(a). Identical molecular graphs (not displayed) were obtained at the PBE1PBE/6-311G(d,p) and CCSD(full)/6-311G(d,p) levels as well. Even though C7 leans toward C5 and C6—the C7−C5(C7−C6) distance is 0.335 Å less than the C7−C2(C7−C3) distance—there is no bond path between C7 and C5 or C7 and C6. This was also the case even at PBE1PBE where the C7−C6(C7−C5) distance was 0.014 Å less than at the B3PW91 level. It is seen that the C5−C6 bond, toward which C7 leans, is longer (+0.026, +0.027, and +0.022 Å at B3PW91, PBE1PBE, and CCSD) and slightly weaker than the C2−C3 bond. This is supported by the fact that the $\rho(\mathbf{r}_c)$ values at the BCPs at B3PW91 are 0.2183 and 0.2387, respectively. C1−C7and C4−C7 exhibit double-bond character; the internuclear distances are 1.447 Å, and the $\rho(\mathbf{r}_c)$ values at the BCPs are 0.2821, 0.2830, and 0.2780 at the B3PW91, PBE1PBE, and CCSD levels, respectively. Based on the $\rho(\mathbf{r})$ values at the RCPs (0.0438 for ring A and 0.0656 for ring B), the bridge lean results in an increase in the density in ring B.

1-C$_{2V}$. The molecular graph of **1-*C*$_{2V}$** obtained at the B3PW91/6-311G(d,p) level is displayed as Figure 3 (**1-*C*$_{2V}$**) and as Figure 8(a). Identical molecular graphs (not displayed) were obtained at the B3PW91/6-311G(d,p) and CCSD/6-311G(d,p) levels as well. There are no bond paths between C7 and C2, C3, C5, and C6 at any level. It is seen that C1−C2 and C3−C4 bonds are shortened relative to the C1−C2 and C3−C4 bonds of **1-*C*$_S$** (0.0188 Å), and C4−C5, C6−C1 bonds are lengthened (0.0158 Å) relative to C4−C5, C6−C1 of **1-*C*$_S$**. The C2−C3 bond is marginally shorter than the C2−C3 bond of **1-*C*$_S$**, and C5−C6 is marginally longer than C5−C6 of **1-*C*$_S$**. The values of $\rho(\mathrm{r}_c)$ at the BCPs at the B3PW91 are 0.2183 and 0.2387.

O-Protonated Alcohols 3 and 4. The molecular graphs of O-protonated 7-norbornanol (**3**) and exo-2-bicyclo[3.2.0]-heptanol (**4**) along with selected internuclear distances and values of $\rho(\mathbf{r}_c)$ at the BCPs are displayed in Figure 1S. Protonation of the alcohols leads to a significant lengthening if the C−O bonds and several C−C bonds; the C−O distances of the parent alcohols 7-norbornanol and exo-2-bicyclo[3.2.0]heptanol are 1.410 and 1.426 Å, respectively. In terms of the internuclear distance and the value of $\rho(\mathbf{r}_c)$ at the BCP, the C−O bond of **4** is weaker than the C−O bond of **3**.

QTAIM-DI-VISAB Analyses. *1-C$_1$*. Selected atomic basins of **1-*C*$_1$** obtained at the B3PW91 level are displayed as Figures 6(a−d) and 2S(a−c). Figure 6(a) shows the C1 and C7 basins of the C1−C7 bond. That these basins share an atomic surface is clearly seen in this display. The DI is 0.6150, substantially lower than the DI (0.9519) of the ‘normal’ C5−C6 single bond across the ring; C1−C7 is a relatively weak covalent bond. Figure 6(b) shows the C2 and C7 basins that are in very close proximity to each other. From a visual standpoint these basins are similar to the pair of carbon atoms of the C1−C7 bond, yet no bond path connects them even though the DI (0.4287) is relatively large. The reason for this result is readily seen in the display of the C1 basin shown as Figure 6(c); a ‘wedge’ of density of this basin intervenes between the C2 and C7 basins and

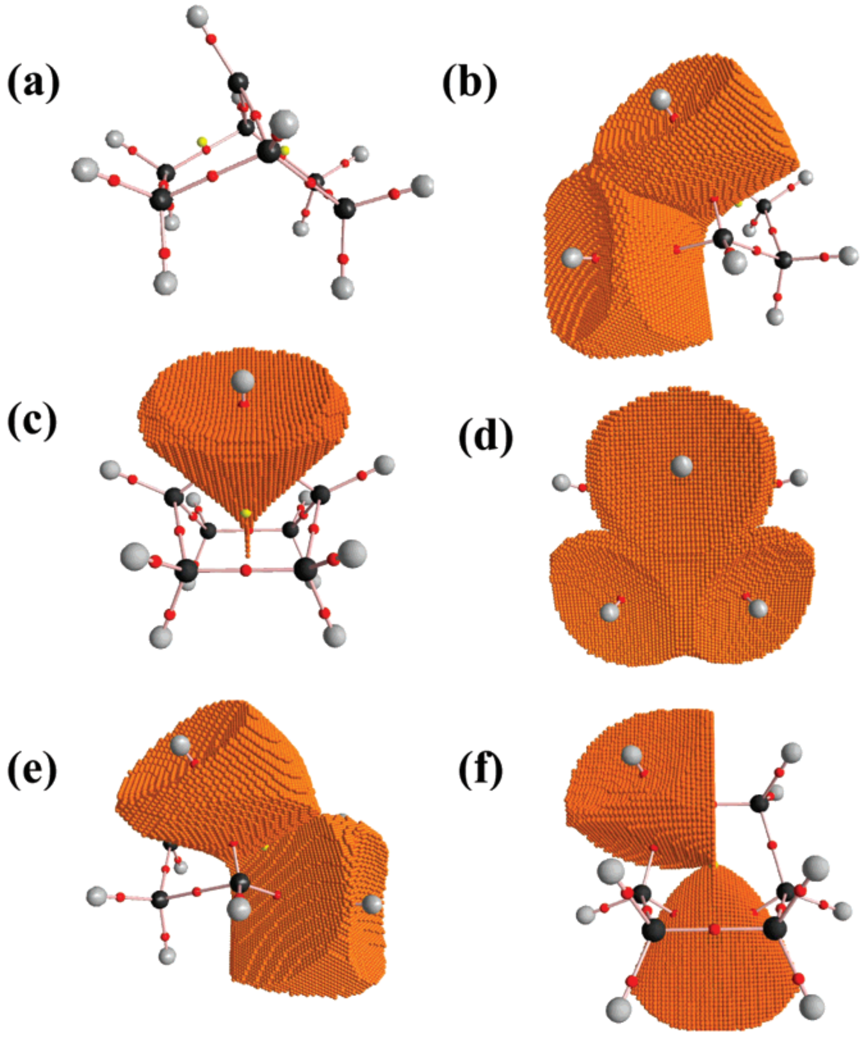


Figure 7. Molecular graph (a) and selected atomic basins of **1-*C_S*** at B3PW91/6- 311G(d,p): (b) C5, C7; (c) C7; (d) C5, C6, C7; (e) C2, C7; and (f) C2, C7 bottom view.

clearly precludes bond path formation. *Nevertheless there is a high degree of delocalization of electrons—the DI is 0.4287—between these basins in the absence of a bond path.* The flattening of the C7 atomic basin surface facing C2 is clearly seen in its display in Figure 6(d).

We observed this behavior—relatively large DIs but no BPs—previously in a number of cations[8−10] and in trimethylsilyl(carbene) and trimethylgermyl(carbene).[13] Farrugia et al. very recently observed this behavior in the case of the iron trimethylenemethane complex Fe(η^4-C{CH_3}$_3$)-$(CO)_3$ in an elegant high-resolution X-ray diffraction study that was coupled with B3LYP and QTAIM calculations.[26]

As a comparison, Figure 4S(a) shows the C4 and C6 basins on the other side of **1-*C_1*** that are not close to each other and exhibit a miniscule DI of 0.0455. It is interesting to note that there appears to be a weak interaction between $H7_{endo}$ and C2 as suggested by the proximity of these basins (Figure 4S(b)). The $H7_{endo}$ basin (Figure 4S(c)) shows some deformation, and the DI (0.0516) is nonzero. In keeping with our earlier results,[19,20] the CCSD DIs are somewhat lower than the DFT DIs. The molecular graph as well as the C7−C1, C7, C2, and C1 basins of **1-*C_1*** obtained at the CCSD(full)/6-311G(d,p) level (not shown) mirror the results obtained at B3PW91/6-311G(d,p).

1-C_S. The molecular graph and selected atomic basins of **1-*C_S*** obtained at the B3PW91 level are displayed as Figure 7(a)−(f). While there are no bond paths between C5, C6, and C7, there is a significant exchange of electrons between the basins with the DI being 0.2289 for each pair—at the CCSD level the value is 0.1565—in keeping with their proximity. Figure 7(b) shows the C6 and C7 basins, 7(c) shows C7, and 7(d) shows the 3-basin cluster of C5, C6, and C7. This exchange/delocalization of electrons between C5, C6, and C7 is undoubtedly one of the reasons for the pronounced lean of C7 toward C5 and C6. The DI for the C2(C3), C7 pair is much smaller (0.0621) in keeping with the fact that they are farther apart than the C5(C6), C7 pairs as seen in Figure 7(e) and a bottom view in Figure 7(f). In keeping with the values of $\rho(\mathbf{r}_c)$ for the DI of the C1−C6 and C4−C5 bonds is smaller (0.8513) than the DI (0.9590) of C1−C2 and C3−C4 pairs indicating that the C1−C6 and C4−C5 bonds are weaker than the C1−C6 and C4−C5 bonds. The molecular graph as well as selected atomic basins of **1-*C_S*** obtained at the CCSD(full) level (not shown) mirror the results obtained at B3PW91/6-311G(d,p).

1-C_{2V}*.** The molecular graph of **1-*C_{2V} and selected atomic basins obtained at the B3PW91 level are displayed as Figure 8(a)−(d). Figure 8(b) shows the C2 and C7 basins, 8(c)

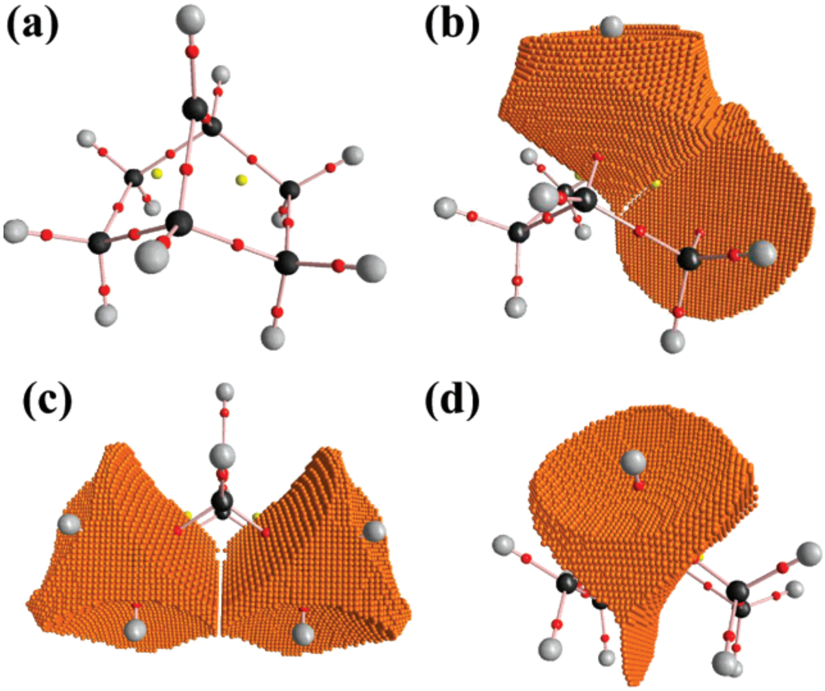


Figure 8. Molecular graph (a) and selected atomic basins of **1-*C_{2V}*** at B3PW91/6-311G(d,p): (b) C3, C7; (c) C2, C7; and (d) C1.

shows C6 and C2, and 8(d) shows the C1 basin. The C2 and C7 basins—this is also the case for the C3−C7, C5−C7, and C6−C7 pairs—do not have large proximate surface areas in keeping with the fact that the DI is 0.1224. While the DI is somewhat lower at the CCSD level (0.0849), these results clearly show that there is delocalization of electrons, although to a minor degree, between C7 and the ring carbons C2, C3, C5, and C6, contrary to the conclusions reached by Sunko et al.[27] on the basis of a simple molecular orbital analysis. Figure 8(c) is a display of the C2 and C6 basins that like the C2 and C7 basins do not have large proximate surfaces, and the DI (0.0446 at B3PW91 and 0.0334 at CCSD) is smaller than the DI for the C2−C7 pair. Figure 8(d) shows the C1 basin with the ‘wedge’ of density intervening between C2 and C6. In keeping with the relative values of $\rho(\mathbf{r}_c)$ (see Figure 3) the DI of C1−C2 (also C3−C4, C4−C5, and C1−C6) is smaller (0.9019) than the DI (0.9590) of the C1−C2(C3−C4) pairs of **1-*C_S*** indicating that these four ring C−C bonds are weaker than the C1−C2(C3−C4) bonds of **1-*C_S***. However, in keeping with the relative values of $\rho(\mathbf{r}_c)$ (see Figure 3) the DI of C1−C2 (also C3−C4, C4−C5, and C1−C6) is larger (0.9019) than the DI (0.8513) of the C1−C6(C4−C5) pairs of **1-*C_S*** indicating that these four ring C−C bonds are stronger than the C1−C6(C4−C5) bonds of **1-*C_S***. The molecular graph of **1-*C_{2V}*** and selected atomic basins obtained at the CCSD(full) level (not displayed) closely resemble the ones obtained at B3PW91.

Conclusions

This study shows that the so-called 7-norbornyl cation exhibits the molecular graph of the bicyclo[3.2.0]heptyl cation at its equilibrium geometry. It suggests that the QTAIM-DI-VISAB analysis is the method of choice for establishing the nature of the bonding in hypercoordinated so-called nonclassical carbocations. This approach obviates the need for dotted-line representations of bonding.

Acknowledgment. Financial support by the Natural Sciences and Engineering Research Council of Canada (NSERC) is gratefully acknowledged. Computing resources of the McMaster node of SHARCnet (Shared Hierarchical Academic Research Computing Network (of Ontario)) were used in this study.

Supporting Information Available: Cartesian coordinates of structures, selected figures, and a 21-frame animation (viewable with common media players such as Windows Media Player) of the changes in the bicyclo[3.2.0]-heptyl molecular graph during the nuclear motions associated with the 206 cm^{-1} mode. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mesić, M. M.; Sunko, D. E.; Vanĉik, H. *J. Chem. Soc., Perkin Trans.* **1994**, *2*, 11.

(2) Winstein, S.; Gadient, F.; Stafford, E. T.; Klinedinst, P. E. *J. Am. Chem. Soc.* **1958**, *80*, 5895.

(3) Miles, F. B. *J. Am. Chem. Soc.* **1968**, *90*, 1265.

(4) Krimse, W. W.; Streu, J. *J. Org. Chem.* **1985**, *50*, 4187.

(5) Sieber, S.; Schleyer, P. von R.; Vanik, H.; Mesi, M.; Sunko, D. E. *Angew. Chem. Int., Ed. Engl*. **1993**, *32*, 1604.

(6) Moss, R. A.; Fu, X. X.; Sauers, R. H. *Can. J. Chem.* **2005**, *83*, 1228.

(7) Bader, R. F. W. *Atoms in Molecules − A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.

(8) Werstiuk, N. H.; Muchall, H. M. *J. Mol. Struct. (THEOCHEM)* **1999**, *463*, 225.

(9) Werstiuk, N. H.; Muchall, H. *J. Phys. Chem. A* **1999**, *103*, 6599.

(10) Werstiuk, N. H.; Muchall, H. *J. Phys. Chem. A* **2000**, *104*, 2054.

(11) Werstiuk, H. H.; Muchall, H. M.; Noury, S. J. *Phys. Chem. A* **2000**, *104*, 11601.

(12) Bajorek, T.; Werstiuk, N. H. *Can. J. Chem*. **2005**, *83*, 1352.

(13) Poulsen, D. A.; Werstiuk, N. H. J. *Chem. Theory Comput.* **2006**, *2*, 77.

(14) Carlier, P. R.; Deora, N.; Crawford, T. D. *J. Org. Chem.* **2005**, *71*, 1592.

(15) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A., Jr. *Gaussian 03, Revision B.02 and C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(16) Dennington, R., II; Keith, T.; Millam, J.; Eppinnett, K.; Hovell, W. L.; Gilliland, R. *GaussView*, *Version 3.09*; Semichem, Inc.: Shawnee Mission, KS, 2003.

(17) Biegler-Konig, F. *AIM 2000*; Copyright 1998−2000, University of Applied Science: Bielefeld, Germany.

(18) Keith, T. A. *AIMALL97 Package (D2) for WINDOWS*. aim@tkgristmill.com (accessed September 7, 2007).

(19) Wang, Y.-G.; Matta, C.; Werstiuk, N. H. *J. Comput. Chem*. **2003**, *24*, 1720.

(20) Wang, Y.-G.; Werstiuk, N. H. *J. Comput. Chem*. **2003**, *24*, 379.

(21) *NABLA. Fortran Program for computing the density and Laplacian of the density on a 3D grid using G94, G98, and G03 wave functions*; Dr. Stephane Noury, Department of Chemistry, McMaster University: 2000.

(22) IBM(1999). *Open Visualization Data Explorer*. Available: http://www.research.ibm.com/dx/ (accessed September 7, 2007).

(23) *ChemCraft*, *Version 1.5*. http://www.chemcraftprog.com.

(24) Xantheas, X. S. X. S.; Elbert, T. S. T. S.; Ruedenberg, K. *Theor. Chim. Acta* **1991**, *78*, 365.

(25) Pendás, A. M.; Francisco, E.; Blanco, M. A.; Gatti, C. *Chem. Eur. J.* **2007**, DOI: 10.1002/chem.200700408.

(26) Farrugia, L. J.; Evans, C.; Tegel, M. *J. Phys. Chem. A* **2006**, *110*, 7952.

(27) Sunko, D. E.; Vančik, H.; Mihalić, Z.; Shiner, V. J.; Wigles, F. P. *J. Org. Chem.* **1994**, *59*, 7051.

CT700176D

JCTC Journal of Chemical Theory and Computation

Carbon−Hydrogen Bond Activation in Hydridotris(pyrazolyl)borate Platinum(IV) Complexes: Comparison of Density Functionals, Basis Sets, and Bonding Patterns

Benjamin Alan Vastine, Charles Edwin Webster,[†] and Michael B. Hall*

Department of Chemistry, Texas A&M University, P.O. Box 30012,
College Station, Texas 77841-3012

Received May 18, 2007

Abstract: The reaction mechanism for the cycle beginning with the reductive elimination (RE) of methane from κ^3-TpPt^{IV}(CH$_3$)$_2$H (**1**) (Tp = hydridotris(pyrazolyl)borate) and subsequent oxidative addition (OA) of benzene to form finally κ^3-TpPt^{IV}(Ph)$_2$H (**19**) was investigated by density functional theory (DFT). Two mechanistic steps are of particular interest, namely the barrier to C−H coupling (barrier 1 − Ba1) and the barrier to methane release (barrier 2 − Ba2). For 31 density functionals, the calculated values for Ba1 and Ba2 were benchmarked against the experimentally reported values of 26 (Ba1) and 35 (Ba2) kcal•mol^{-1}, respectively. Specifically, the values for Ba1 and Ba2, calculated at the B3LYP/double-ζ plus polarization level of theory, are 24.6 and 34.3 kcal•mol^{-1}, respectively. Overall, the best performing functional was BPW91 where the mae associated with the calculated values of the two barriers is 0.68 kcal•mol^{-1}. The calculated B3LYP values of Ba1 ranged between 20 and 26 kcal•mol^{-1} for 12 effective core potential basis sets for platinum and 29 all-electron basis sets for the first row elements. Polarization functions for the first row elements were important for accurate values, but the addition of diffuse functions to non-hydrogen (+) and hydrogen atoms (++) had little effect on the calculated values. Basis set saturation was achieved with APNO basis sets utilized for first-row atoms. Bader's "Atoms in Molecules" was used to analyze the electron density of several complexes, and the electron density at the Pt−N$_{ax}$ bond critical point (trans to the active site for C−H coupling) varied over a wider range than any of the other Pt−N bonds.

Introduction

The goal of facile conversion of saturated hydrocarbons into desirable organic materials motivates C−H bond activation research, and platinum is an important metal for these reactions.[1] Garnett and Hodges[2] were the first to report platinum mediated C−H bond activation, and they observed H/D exchange between deuterated water and aromatic substrates catalyzed by Pt^{II} salts in an acidic solution. Shilov and co-workers[3] investigated the catalytic oxidation of methane to methanol and chloromethane by PtCl$_4^{2-}$ and PtCl$_6^{4-}$ salts in acidic aqueous solution. The research into mechanistic aspects related to the Shilov chemistry is chronicled in two reviews,[4] and they include a discussion of the formation of 5-coordinate, coordinatively unsaturated Pt^{IV} complexes and their purported role in the reductive elimination (RE) step.

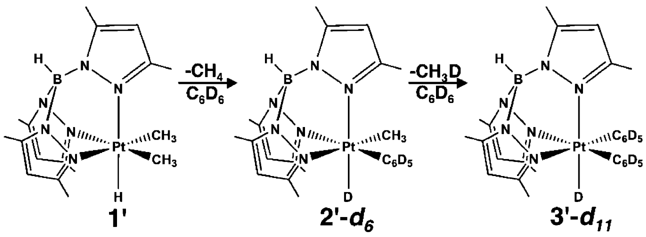
The isolation of 5-coordinate Pt^{IV} complexes is important because they are believed to be intermediates in platinum-mediated oxidative addition (OA) and RE chemistry. The first isolated 5-coordinate Pt^{IV} alkyl complex[5] was implicated in C−C bond-forming RE chemistry,[6] and Goldberg and co-workers[7,8] proposed 5-coordinate, coordinatively-unsaturated

* Corresponding author e-mail: hall@science.tamu.edu.

[†] Present address: Department of Chemistry, The University of Memphis, 213 Smith Chemistry Building, Memphis, TN 38152-3550.

10.1021/ct700120d CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/14/2007

Scheme 1



Pt^{IV} complexes as intermediates in C−H and C−C RE coupling reactions. Templeton and co-workers[9] isolated three different 5-coordinate Pt^{IV} complexes that were stabilized by silanes and proposed several 5-coordinate Pt^{IV} complexes as intermediates.[10,11]

In a theoretical study of Shilov chemistry, Siegbahn and Crabtree[12] argued that a σ-bond metathesis mechanism is preferred over the OA/RE mechanism; however, the possibility of the oxidative pathway could not be eliminated because of the similar energetics to that of metathesis. They also stated that the solvent was integral to the reaction. Bartlett et al. reported two studies of RE C−H coupling that used Pt^{II} and Pt^{IV} model complexes,[13] and both reports arrived at the same conclusion. For Pt^{II} complexes, direct elimination of methane was found to be favored energetically over phosphine loss prior to RE C−H coupling, but ligand loss prior to C−H coupling was preferred for the Pt^{IV} complexes.

Jensen et al.[14] reported the RE of methane and OA of benzene-d_6 to form κ^3-$Tp^{3,5-Me}Pt^{IV}(C_6D_5)_2D$ from κ^3-$Tp^{3,5-Me}$-$Pt^{IV}(CH_3)_2H$ (**1′**), where $Tp^{3,5-Me}$ (or Tp*) is the hydridotris-(3,5-dimethylpyrazolyl)borate ligand (Scheme 1).[15] From the kinetic studies, enthalpic barriers to methane formation (barrier 1 − Ba1) and methane release (barrier 2 − Ba2) from **1′** were measured and reported. The proposed mechanistic step for Ba1 is C−H coupling between a methyl ligand and the hydride of **1′**, and for Ba2, methane elimination from **1′**. The authors concluded that this elimination precedes benzene addition, which is consistent with a dissociative mechanism. Another recent report also concluded that the dissociative mechanism is the preferred pathway for methane elimination from Pt^{IV} complexes.[16] Suggestions have been made that the Tp* ring trans to the hydride could dechelate, bind in a κ^2-interaction to the platinum center, and provide an open coordination site. Zarić and Hall reported that loss of one degree of coordination of the Tp ligand ($\kappa^3 \rightarrow \kappa^2$) occurred prior to methane activation in a TpRh(CO) complex.[17]

Here, the results of a density functional theory[18] (DFT) study on the reaction in Scheme 1 are presented. Specifically, 31 density functionals and a variety of basis sets are benchmarked against the experimental values of Ba1 and Ba2 that were reported by Jensen et al.[14] For some of the reported results, the experimental Tp* ligand (**1′**, etc.) is replaced with the parent Tp ligand (**1**, etc.). The basic mechanism for the reaction studied is presented in section 1, and possible alternative pathways for the mechanism of C−H coupling and methane release are examined in section 2. The bonding schemes of several complexes are presented in section 3; studies in benchmarking DFT and various basis sets against the experimental values for Ba1 and Ba2 are presented in section 4.

Results and Discussion

1. Mechanism. In the following section, specific steps of the mechanism from the dimethyl reactant (**1**) to the methyl-phenyl intermediate (**10**) are studied. The mechanism and relative energies of the two barriers and the specific coordination modes of benzene in the methyl−benzene complexes (**6** and **8**) are presented and discussed. Then, the analogous reaction pathway for the release of the second methane and coordination of the second benzene to form the final diphenyl product (**19**) is presented.

Procedure. All calculations were performed by using the Gaussian 03 suite of programs.[19] Each complex reported in this section was fully optimized at the B3LYP/BS1 level of theory, and the analytical frequencies were calculated at this same level of theory for each complex to determine if the force constants were real (intermediate) or if one was imaginary (transition state). All optimizations were accomplished with the default convergence criteria, and each complex was optimized in C_1 symmetry. The B3LYP hybrid density functional is comprised of the Becke3 exchange functional and the Lee, Yang, and Parr correlation functional.[20] The basis set (BS1) that was used in the optimization and frequency calculations is as follows: platinum was assigned the Hay and Wadt small core Los Alamos National Laboratory effective core potential[21] (ECP = LANL2) and the valence double-ζ (341/341/21 = DZ) basis set (BS) as modified by Couty and Hall (ECP/BS = mLANL2DZ);[22] each nitrogen, boron, and the carbon and hydrogen atoms bound to the platinum were assigned Dunning's correlated consistent polarized valence double-ζ (cc-pVDZ) basis set;[23] all other atoms were assigned Dunning's full double-ζ D95 basis set.[24] Details for the density functionals and basis sets benchmarking studies will be given later. Unless noted otherwise, all energies are in kcal·mol^{-1} and relative to **1**. Most values discussed in the text are enthalpies ($\Delta H^{\circ/\ddagger}$) in the gas phase at standard conditions (298 K, 1 atm). The electronic energies (ΔE_{elec}), electronic energies with zero point corrections (ΔE_0), and free energies ($\Delta G^{\circ/\ddagger}$) are reported in tables. Three-dimensional molecular geometric representations were constructed with JIMP 2.[25]

κ^3-$TpPt^{IV}(CH_3)_2H$ (1) + C_6H_6 to κ^3-$TpPt^{IV}(CH_3)$-$(C_6H_5)H$ (10) + CH_4. The B3LYP/BS1 reaction energy profile for reductive elimination (C−H bond formation), methane release, benzene coordination, and oxidative addition of benzene is displayed in Figure 1. The orientations of the ligand atom positions in the complexes, as referenced in the text, are defined in Figure 2. The relative energy values (**1** + benzene = 0) for species **1**−**10** are tabulated in Table 1. The B3LYP/BS1 optimized geometries of complexes along the potential energy surface (PES), with relevant bond distances (Å), are shown in Figure 3.

C−H Coupling through Reductive Elimination of Methane (Ba1). In reactant **1**, the stronger trans influence of the hydride is noticeable in the slightly longer Pt−N_{ax} bond. The transition state for the C−H coupling mode (**2-TS**) has an enthalpic barrier of 24.3 kcal·mol^{-1} and can be

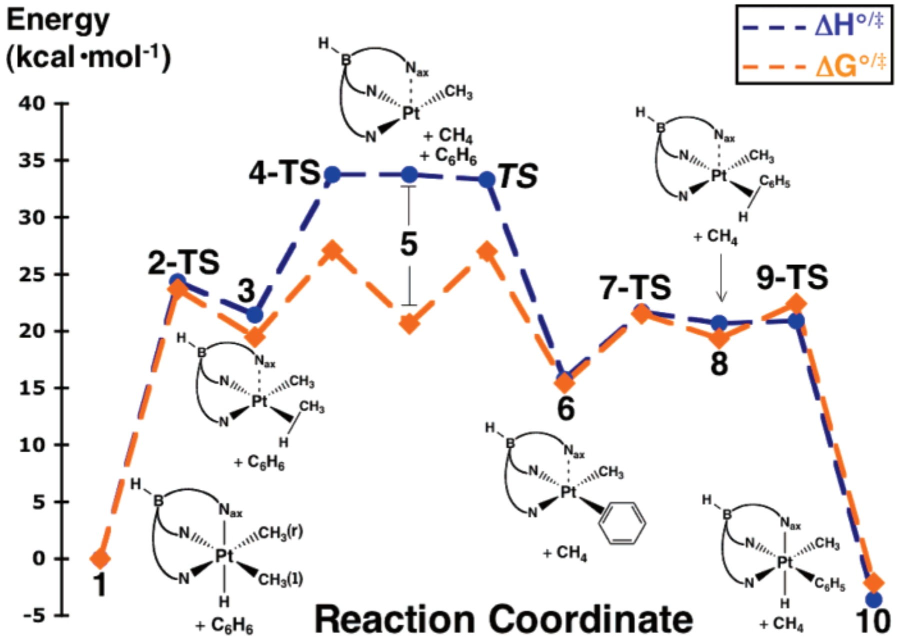


Figure 1. The B3LYP/BS1 relative enthalpies (blue) and free energies (orange) for complexes **1**−**10** (kcal·mol^{-1}). The complex designations correspond to the structures listed in Figure 3 and Table 1. The TS that connects **5** and **6** (***TS***) *was not calculated* and is only a qualitative representation.

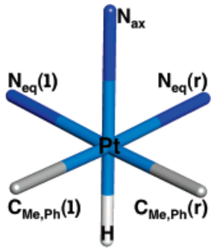


Figure 2. A generalized model that illustrates the orientations of the atoms within the ligands. These assignments are referenced in the text.

Table 1. Relative B3LYP/BS1 Energies for Complexes **1**−**10**

	energies			
complex	ΔE_{elec}	ΔE_0	$\Delta H^{\circ/\ddagger}$	$\Delta G^{\circ/\ddagger}$
1[a]	0	0	0	0
2-TS[a]	24.15	24.14	24.33	23.70
3[a]	20.82	20.82	21.43	19.49
4-TS[a]	32.28	32.28	33.76	26.90
5[a,b]	32.56	32.56	33.76	20.64
6[b]	14.46	14.46	15.76	15.42
7-TS[b]	20.58	20.58	21.68	21.52
8[b]	19.20	19.20	20.67	19.33
9-TS[b]	20.06	20.06	20.92	22.40
10[b]	−4.55	−4.55	−3.60	−2.15

[a] + C_6H_6. [b] + CH_4 energy values are given in kcal·mol^{-1} and are relative to **1**.

characterized as a late transition state in which the C_{Me}(l)−Pt−H angle has decreased by more than half its original value. This transition state leads to the formation of the relatively unstable intermediate **3** where the Pt−C_{Me}(l) bond has lengthened by 0.35 Å, and the C−H bond is 1.17 Å. During this process the Pt−N_{ax} progressively lengthens, and **3** is essentially a 4-coordinate square planar complex as expected for a d^8 metal (Pt^{II}). The Pt−N_{eq}(r), which is trans to the weakly bound CH_4 molecule, has shortened by 0.2 Å.

Methane Loss from 3 (Ba2). When the weakly bound CH_4 ligand of **3** is released, the Pt−N_{eq}(r) bond shortens to its minimum length (1.98 Å). The unimolecular dissociation transition state (**4-TS**) for this process is characterized by an enthalpic difference of 33.8 kcal·mol^{-1} (relative to **1**) and results in a coordinatively unsaturated (3-coordinate), 16e$^-$ Pt^{II} (d^8) species (**5**) where the Pt−N_{ax} distance shortens slightly from that of **3**. This intermediate, **5**, is nearly isoenthalpic with **4-TS**, and this result is explained below. The Pt−C_{Me}(r) and −N_{eq}(l) bond lengths are unaffected by methane release.

Common assumptions for the dissociation of a neutral dative ligand from a transition-metal complex that does not rearrange following a dissociation are as follows: (1) that entropy does not contribute until after the transition state is passed and (2) no enthalpic barrier exists for the recoordination.[26] With this assumption the free energy barrier ($\Delta G^{\ddagger}$) equals the enthalpic barrier ($\Delta H^{\ddagger}$). In Figure 4, the relative enthalpies and free energies versus length of the Pt−C_{Me}(l) coordinate starting from **3** are plotted for six points; the enthalpy curve plateaus at 4.42 Å and a value of 34 kcal·mol^{-1}, while the free energy curve plateaus at a length of 3.62 Å and a value of 26 kcal·mol^{-1}. The dissociation transition state, **4-TS**, is chosen to be located at 4.42 Å because both curves have plateaued by this point, and **4-TS** is isenthalpic with the relative enthalpy difference between **1** and **5**. Our primary purpose here is to consider the enthalpic barrier to methane release, and we have shown that the enthalpic difference between the separated products and the starting material is a good approximation for the experimental enthalpic barrier; therefore, *Ba2 is defined as the calculated relative enthalpic difference between* ***1*** *and* ***5*** *(+free methane)*. The relative free energy difference does not

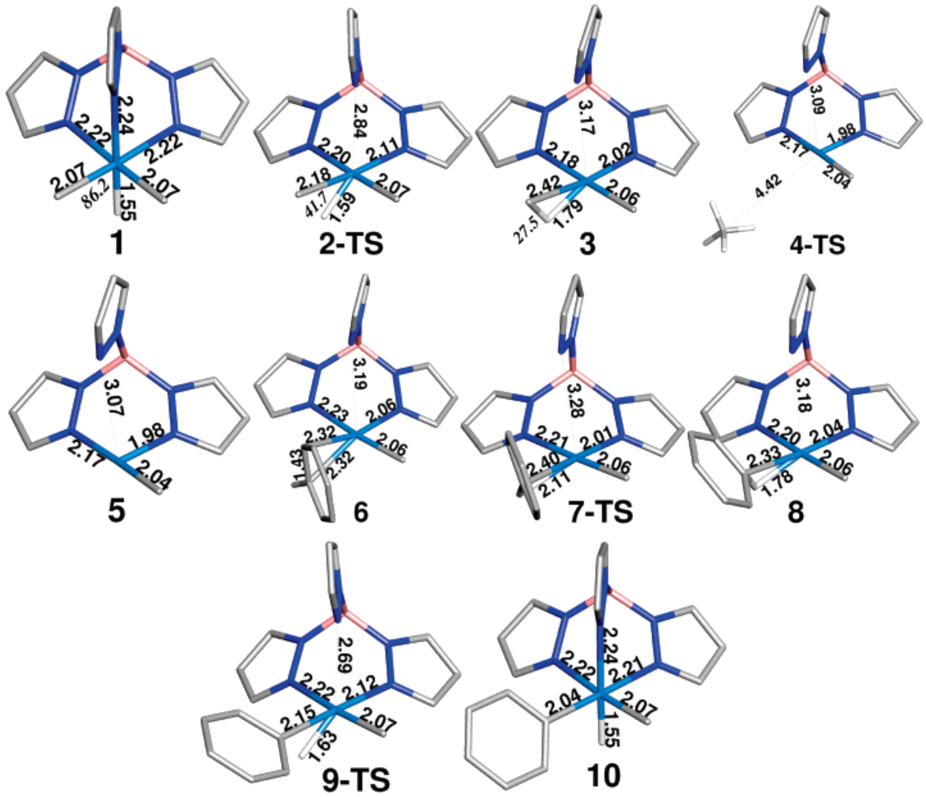


Figure 3. The optimized geometries for complexes **1**−**10**. Relevant bond lengths are included in the representations and are given in Å. The C_{Me}(l)−Pt−H angles (deg) are the numbers in italics. All nonessential hydrogen atoms have been removed for clarity.

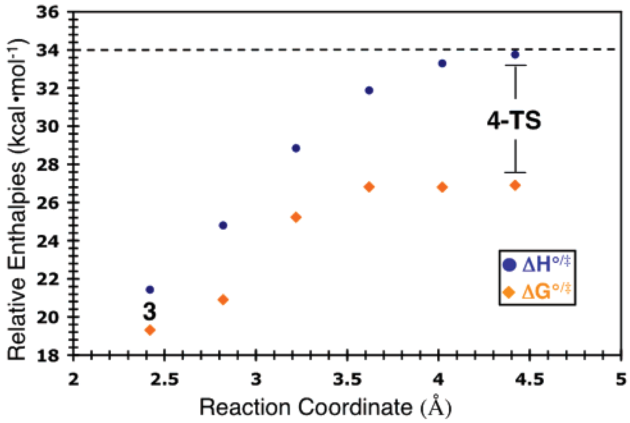


Figure 4. Relative enthalpy and free energy values for six select points along the Pt−$C_{\sigma-Me}$(l) coordinate. The dashed line represents the calculated enthalpic value for Ba2 (**4-TS**).

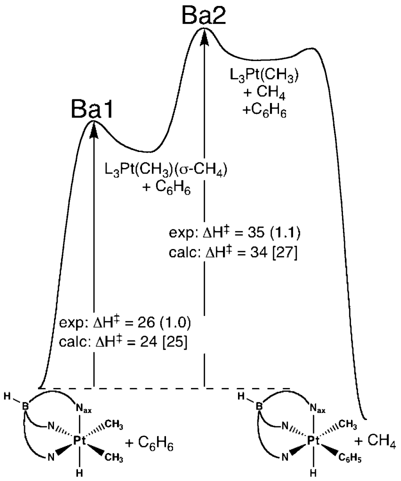


Figure 5. Comparison of the experimental and B3LYP/BS1 values for Ba1 and Ba2 (kcal•mol^{-1}). The Tp and Tp* ligands are denoted as "L_3". The reported uncertainties are given in parenthesis after the experimental values. The values in square brackets are the calculated values relative to **1′**.

increase at the same rate as the enthalpy, so there is a contribution of the entropy to the transition state. The difference between the common assumption and the free energy difference calculated for **4-TS** is 4.9 kcal•mol^{-1}, ca. 38% of the total entropy for the dissociation to **5** (+free methane).

Barriers 1 and 2. The experimental and calculated barriers (Ba1 and Ba2) are compared in Figure 5 where the energies are reported for Tp (**1** = 0.0) and for Tp* (**1′** = 0.0) in square brackets, and we observe agreement within two units of experimental uncertainty between the calculated and experimental value for both barriers; however, both calculated values are slightly less than the experimental value. For methane release from **1′**, the B3LYP/BS1 value for Ba1 is similar to that of **1**; however, the value for Ba2 is 8 kcal•mol^{-1} less than experiment. The Pt−N_{ax} distance of **5′** (2.15 Å) is 0.92 Å shorter than that of **5** (3.07 Å), and the

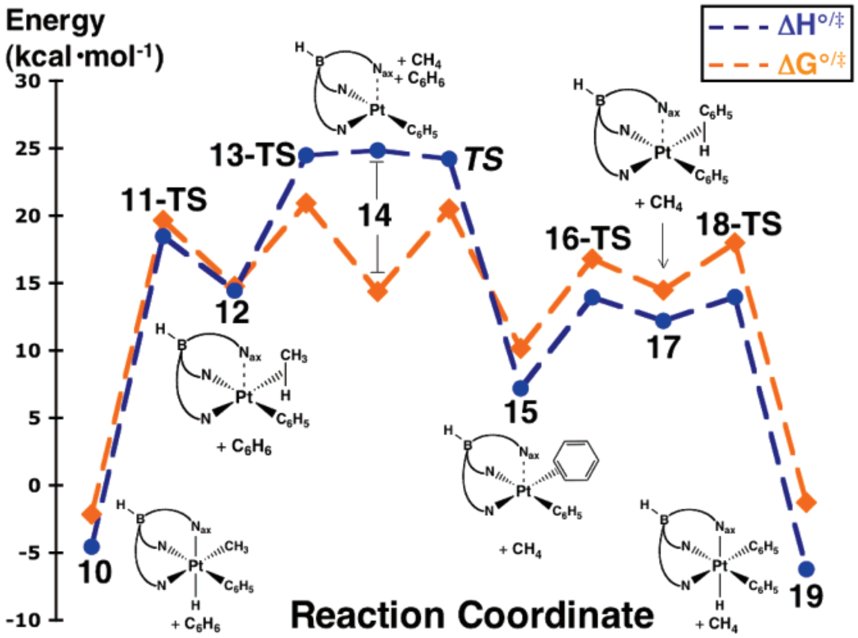


Figure 6. The B3LYP/BS1 calculated relative enthalpies (blue) and free energies (orange) in kcal·mol^{-1} for **10**−**19** (values relative to **1**). The species included in the figure are representative of those listed in Supporting Information Figure 1 and Table 2. The TS that connects **14** and **15** (***TS***) *was not calculated* and is only a qualitative representation.

result is the stabilization of the coordinatively unsaturated intermediate.

Benzene Coordination and OA To Form κ^3-TpPt^{IV}-(CH$_3$)(C$_6$H$_5$)H (10). The transition state for benzene coordinating to **5** (***TS*** − Figure 1) was not located on the B3LYP/BS1 PES, and its value was assumed to be similar to **4-TS**. There are two coordination modes for benzene to **5** that are in agreement with experimental observations:[27](1) an η^2-benzene bound through two carbons (π bond) forming complex **6** and (2) a σ-bound complex forming an η^2-benzene bound through a C−H bond (**8**). Species **6** is more stable than **8** by 4.91 kcal·mol^{-1}, and they are connected through **7-TS** with a barrier of 5.92 kcal·mol^{-1} (relative to **6**). Benzene acts as a π-donor/acceptor in **6** as the calculated carbon−carbon bond length of the two carbons π-bound to the platinum center (1.43 Å) is slightly longer than that calculated for the carbon−carbon bond length of free benzene (1.40 Å). The Pt−N$_{eq}$(r) bond length is slightly longer in **6**, which, coupled with the relative enthalpy difference, supports the view that benzene is in the π-bound form. Reinartz et al. reported[10] geometric parameters of an isolated η^2 benzene complex that is analogous to **6**, and the calculated parameters of **6** agree well with their complex; the experimentally determined Pt−C, Pt−N$_{eq}$(r), and C−C bond lengths are shorter compared to those in **6** by 0.08, 0.07, and 0.02 Å, respectively. The geometries of **6** and **8** are pseudo square planar (4-coordinate) at platinum, and the Pt−N$_{ax}$ distance is long for both. The facile OA splitting of the σ-C−H bond occurs (**9-TS**) to form the pseudo-octahedral complex, **10**. Overall, the exchange of phenyl for methyl is slightly exothermic.

From κ^3-TpPt^{IV}(CH$_3$)(C$_6$H$_5$)H (10) to κ^3-TpPt^{IV}(C$_6$H$_5$)$_2$H (19). The B3LYP/BS1 reaction profile for the elimination of the second methane and addition of the second benzene (**10**−**19**) is shown in Figure 6 and is analogous to Figure 1.

Table 2. Relative B3LYP/BS1 Energies for Complexes **10**−**19**

	energies			
complex	ΔE_{elec}	ΔE_0	$\Delta H^{\circ/\ddagger}$	$\Delta G^{\circ/\ddagger}$
10[a]	−4.55	−3.60	−3.60	−2.15
11-TS [a]	18.47	19.73	19.73	19.66
12 [a]	14.45	16.14	16.14	14.75
13-TS[a]	24.48	27.00	27.00	20.92
14 [a,b]	24.85	26.49	27.09	14.36
15[b]	7.20	9.64	9.64	10.15
16-TS[b]	13.95	16.12	16.12	16.80
17 [b]	12.20	14.69	14.69	14.48
18-TS[b]	13.96	16.04	16.04	18.00
19 [b]	−6.21	−4.36	−4.36	−1.26

[a] + C$_6$H$_6$. [b] + CH$_4$ energies are reported in kcal·mol^{-1} and relative to **1**.

The molecular geometries for complexes **11**−**19** are analogous to the complexes involved in the first methane elimination and benzene addition events, and these representations are included in Supporting Information Figure 1. Calculated relative energies for complexes **1**−**19** are reported in Table 2. The calculated bond lengths of **19** are in agreement with the bond lengths found in the crystal structure, and this result is shown in Figure 7. The overall reaction is calculated to be exothermic and exergonic by 4.36 and 1.26 kcal·mol^{-1}, respectively. To compare the two methane release events, the analogues of Ba1 and Ba2, in this second replacement, are 1 and 3 kcal·mol^{-1} less than the B3LYP/BS1 values of Ba1 and Ba2 for the first replacement. The analogous barriers are defined as **11-TS** and **14**, and both are relative to **10**. As with the addition of the first benzene, the transition state for benzene addition to **14** was not located; however, ***TS*** is estimated and included in Figure 6.

2. Alternative Pathways. In this section, alternative pathways are explored for the C−H coupling and methane

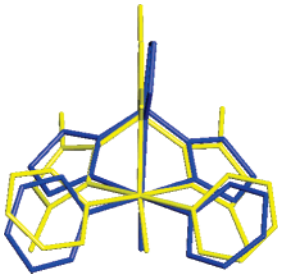


Figure 7. The crystal structure for κ^3-Tp*Pt^{IV}(Ph)$_2$H (yellow) and the B3LYP/BS1 equilibrium geometry for κ^3-TpPt^{IV}(Ph)$_2$H (blue) are overlaid. Bond lengths and angles are in general agreement between the two structures.

Table 3. Relative B3LYP/BS1 Energies (kcal·mol^{-1}) for *Rotation* and *Inversion* Mechanisms

	energies			
complex	ΔE_{elec}	ΔE_0	$\Delta H^{o/\ddagger}$	$\Delta G^{o/\ddagger}$
1[a]	0	0	0	0
TS$_{1-1a}$[a]	21.77	21.77	21.65	21.93
1a	21.28	21.28	21.60	20.92
2a-TS[a]	27.23	27.23	27.46	26.37
3a[a]	21.88	21.88	22.48	20.66
5a[a,b]	33.47	33.47	34.67	21.41
TS$_{1-1b}$[a]	32.16	32.16	32.06	31.27
1b[a]	22.82	22.82	23.32	21.78
2b-TS[a]	26.00	26.00	26.20	25.33
3b[a]	19.98	19.98	20.60	18.86
5b[a,b]	30.94	31.54	32.13	19.17

[a] + C_6H_6. [b] + CH_4 energies are reported in kcal·mol^{-1} and relative to **1**.

release. The possibility of an associative mechanism is examined where benzene coordinates prior to methane release. Two possible orientations of a pyrazolyl (pz) ring are also examined: (1) ring *rotation* about a B−N bond resulting in a side-on interaction of a pz ring with platinum and (2) *inversion* of the boron so that a pz ring is completely removed from the ligand sphere. Last, the possible formation of a dimer is examined. The relative energies are tabulated for the *rotation* and *inversion* pathways in Table 3, and these pathways are shown in Figure 8. In Figure 9, representative geometries are presented for complexes on the *rotation* and *inversion* pathways; and the difference is shown between the binding modes of the Tp ligand. The complexes along the *rotation* and *inversion* pathways not shown in Figure 9 are shown in Supporting Information Figure 2. All structures were calculated at the B3LYP/BS1 level of theory, as in section 1.

Though the experimental work by Jensen et al. supported a dissociative mechanism for methane loss, evidence for an associative mechanism for methane loss was reported by Johansson and Tilset where increased concentrations of solvent acetonitrile changed the ratio of CH_4/CH_3D released from protonated Pt^{II} complexes.[28] Therefore, several models were designed to investigate the possible associative mechanism where benzene and methane are both bound simultaneously to the platinum center; the benzene and methane are π- and σ-bound to the platinum center, respectively. All attempts to locate a transition state geometry for an associative complex were unsuccessful, and our data support the dissociative mechanism (Figure 2).

The next alternative pathway (*rotation*) is described by rotation of the pz ring *axial* to the hydride about the B−N$_{pz}$ bond (N$_{pz}$ is the nitrogen of the *axial* pz ring bonded to the boron) and formation of a complex where two pz rings are coordinated as usual and the third pz ring has "slipped" to form a κ^2-, κ'-Tp complex (**1a**). The barrier to pz ring rotation is 21.7 kcal·mol^{-1} (**TS$_{1-1a}$**). The N$_{pz}$ has a small amount of 4-coordinate character as the B−N$_{pz}$−Pt angle is 89.7° (Figure 9), and the Pt−N$_{pz}$ distance is 2.69 Å, which is ca. 0.5 Å longer than the Pt−N$_{ax}$ distance in **1**. The barrier to C−H coupling (**2a-TS**) is slightly greater than that of the

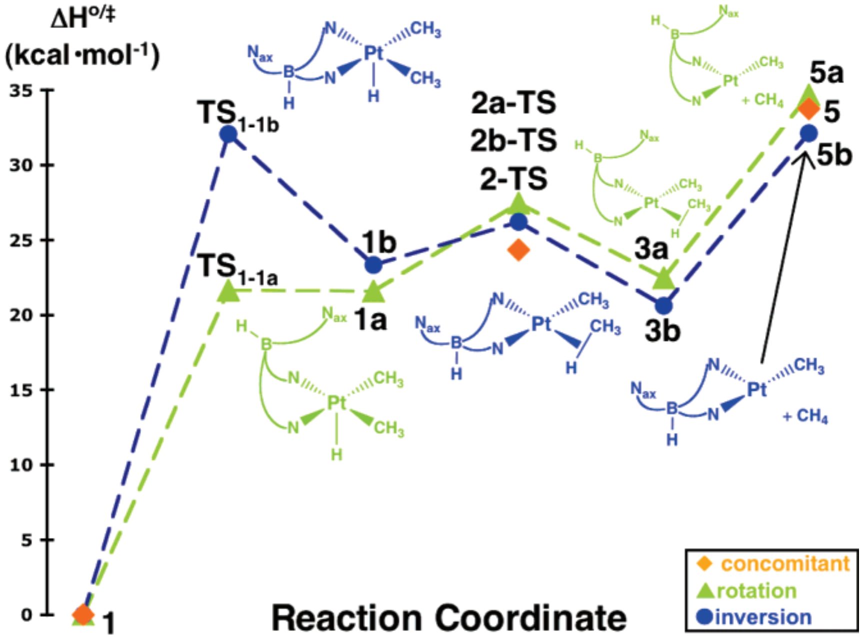


Figure 8. A comparison between the enthalpic PES for the concomitant (orange diamonds), *inversion* (blue dots), and *rotation* (green triangles) pathways leading to C−H bond formation (RE) and methane release. The energies, relative to **1**, are in kcal·mol^{-1}.

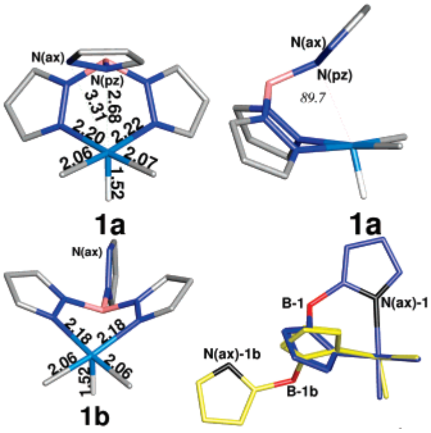


Figure 9. The B3LYP/BS1 optimized geometries for κ^2-, κ'-TpPt^{IV}(CH$_3$)$_2$H (**1a**) and κ^2-TpPt^{IV}(CH$_3$)$_2$H (**1b**) and the comparison between the starting material (**1**-blue) and the inverted form (**1b**-yellow). Bond lengths are reported in Å.

concomitant pathway in Figure 1 (*rotation*: 27.5 vs *con*: 24.1 kcal·mol^{-1}), and a pseudo-square planar (4-coordinate) complex (**3a**) is the result of C−H coupling. As with the *concomitant* pathway, the 16e$^-$, coordinatively-unsaturated Pt^{II} (d^8) complex that results from methane loss (**5a**) is stabilized by an increase in the interaction of the pz ring that was trans to the hydride but is now trans to the vacant coordination site.

In the *inversion* pathway, the *axial* pz ring is removed from the ligand sphere by inversion of the boron geometry, which results in a κ^2-Tp ligand. The barrier to inversion (**TS$_{1-1b}$**) is 32.1 kcal·mol^{-1}, which is the highest initial barrier of any of these pathways. A 5-coordinate Pt^{IV} species (**1b**) is formed where the *axial* pz ring is outside of the coordination sphere, and the boron is shown to reside below the *equatorial* pz rings (Figure 9). The C−H bond coupling transition state (**2b-TS**) is 26.6 kcal·mol^{-1} (relative to **1**), which is slightly greater than the *concomitant* pathway. A weakly bound methane complex is formed (**3b**), and loss of methane from this complex results in a 3-coordinate, Pt^{II} complex. This pathway has the lowest value for Ba2 of the three pathways at 32.1 kcal·mol^{-1}.

Summary of the Two Alternative Pathways. Facile C−H activation of benzene by [κ^2-[Ph$_2$B(pz)$_2$]Pt^{II}(Me)$_2$]$^+$ was reported by Thomas and Peters;[29] however, the *inversion* pathway is disfavored because the initial barrier (**TS$_{1-1b}$**) is higher in energy. Both barriers along the *rotation* pathway are similar to those of the *concomitant* pathway. The calculated values for Ba1 and Ba2 are not significantly altered when the interaction between the *axial* pz ring and the platinum is changed (*rotation*) or removed (*inversion*). A difference of 10.4 kcal·mol^{-1} is measured between the initial barriers to the *inversion* and *rotation* pathways ($\Delta\Delta H$: **TS$_{1-1b}$** − **TS$_{1-1a}$**); this difference is slightly greater than the difference of 6.4 kcal·mol^{-1} that was reported by Webster and Hall for the same barriers in the isomerization chemistry of TpRh(CO)$_2$.[30]

In a mixture of TpRu(PMe$_3$)$_2$OH and 1-methylpyrazole in C$_6$D$_6$, H/D exchange was reported by Gunnoe and coworkers at the four position of each pz ring, and this mechanism likely proceeds through a pathway where the pz ring coordinates to the ruthenium in a side-on interaction.[31] This experimental observation supports a competitive route via the *rotation* pathway. The *rotation* and *concomitant* pathways compete in the elimination of methane because of these similar relative energies.

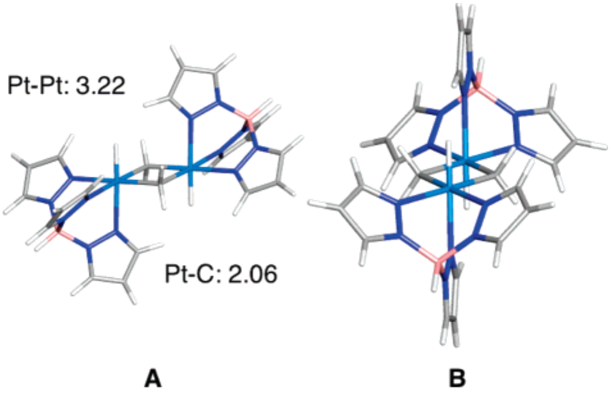


Figure 10. Two different views of the dimer complex. The view in **A** is down the bridging carbon−carbon atoms, while the view in **B** is down the Pt−Pt axis of the molecule. The opposing geometry of the Tp ligands is represented clearly in **A**.

Possible Formation of a [TpPt]$_2$ Dimer. A dimer was not observed in the kinetic studies of C−H coupling and methane release, but other studies have reported the formation and isolation of bridged binuclear complexes.[32] A common structural characteristic of the binuclear structures observed experimentally is opposing ligand geometries as seen in the calculated structure, Figure 10. Species **5** has an open coordination site available for dimer formation with a second molecule of **5**. In the optimized geometry of the calculated dimer, the two TpPt moieties are joined by a 4-center, 8e$^-$ bridge. In addition to reformation of the Pt−H bond, the Tp ligand returns to a tridentate interaction with the platinum. Dimer formation from **1** (2·**1** → dimer + 2·CH$_4$) is exergonic (−4.5 kcal·mol^{-1}) and endothermic (2.4 kcal·mol^{-1}); because of its instability, the dimer was not studied further.

3. Bonding Analysis. To investigate the bonding interactions that are involved in this chemistry, the B3LYP/BS1 electron densities of complexes **1**, **2-TS**, **3**, **5**, and **1a** were investigated with Bader's "Atoms in Molecules" (AIM) analysis.[33] Specific bond critical point (CP) densities that are relevant to the C−H coupling and methane release chemistry are tabulated in Table 4. AIM2000 was used to calculate the bond CPs.[34]

The electron density of **1** was analyzed with AIM, and six (3, −1) bond CPs were found between the platinum and the atoms listed in Table 4. The Pt−N$_{ax}$ bond CP has the least density, which results from the stronger trans influence of the hydride. The bond CP densities typically follow an inverse trend with respect to bond lengths; for example, the Pt−N$_{eq}$(r) bond CP density increases with C−H coupling and methane release, and the bond length shortens for this process. The Pt−C$_{Me}$(r) and −N$_{eq}$(l) bond densities are shown to be insensitive to the C−H coupling chemistry, and this correlates with geometric observations. Interestingly, a Pt−C$_{Me}$(l) bond CP was not located in the density of **2-TS** and

Table 4. Bond Critical Point (CP) Densities for Bonds Involved in C−H Coupling and Methane Release

	(3, −1) bond CP density (ρ(r)/e•bohr^{-3})				
bond: Pt−X	**1**	**2-TS**	**3**	**5**	**1a**
N_{ax}	0.0727040408	0.0241899527	0.0139997372	0.0167954419	NA
N_{eq}(l)	0.0757628782	0.0791088190	0.0811165097	0.0825009156	0.0795292241
N_{eq}(r)	0.0756648650	0.0979133965	0.1203747575	0.1334978956	0.0765416921
C_{Me}(l)	0.1329266488	NF[b]	NF	NA	0.1348946156
C_{Me}(r)	0.1330781962	0.1329927831	0.1337466884	0.1403112418	0.1336717972
H	0.1741331362	0.1485854220	0.0831396505	NA	0.1816352613
N_{pz}	NA[a]	NA	NA	NA	0.0304053343

[a] NA = not applicable. [b] NF = not found.

3; thus, the bond is manifested solely by a CP between the platinum and the hydrogen. For **1a** (one pz rotated), a bond CP was located along the Pt−N_{pz} coordinate with a density of 0.030405 e•bohr^{-3}, which is significantly less than the Pt−N_{ax} bond CP density value of **1**. The decrease in bond CP density is consistent with an increase in bond lengths, but the multiple CPs that are characteristic of a ligand π-bound to a metal (i.e., η^5-C_5H_5) are not observed for this rotated pz ring.

4. Density Functional and Basis Set Benchmarking. Benchmarking studies of density functionals and basis sets are presented in this section. Thirty-one functionals were benchmarked for the barriers. Basis set saturation was also studied, and the trends are presented. The procedure that was used for these studies is explained prior to each benchmarking study. The mean average error (mae) is reported for each study.

Functionals. For all but two of the functionals, the optimized geometry and analytical frequencies of **1**, **2-TS**, **5**, and methane were calculated at the functional/BS1 level of theory. Intermediates and transitions states were verified as having zero and one imaginary mode, respectively, as determined by frequency calculations. To calculate the barrier value at each level of theory, the functional/BSX//functional/BS1 (X = 2, 3) energies of **1**, **2-TS**, **5**, and methane were added to the function/BS1 correction to the enthalpy for each complex. For BS2 and BS3 basis sets, only the cc-pVDZ basis set of BS1 was replaced with cc-pVTZ and cc-pVQZ in BS2 and BS3, respectively, but the other basis sets remained as assigned in BS1. *All subsequent calculated values for the two barriers are presented at the functional/BS3 level of theory*. The procedure for the B2-PLYP[35] and mPW2-PLYP[36] functionals was slightly modified because of computational costs; the B2-PLYP/BSX// and mPW2-PLYP/BSX//B3LYP/BS1 (X = 1, 2, 3) energies of **1**, **2-TS**, **3**, **5**, and methane were added to the second-order correction and the B3LYP/BS1 correction to the enthalpy for each molecule to obtain the corrected enthalpy. The second-order perturbative correction was scaled by 0.27 and 0.25 for the B2-PLYP and mPW2-PLYP functionals, respectively.[37]

Pure density functionals, in which exact exchange is not incorporated, included in this study are BLYP,[38,20b] BPW91,[38,39b] BP86,[38,40] G96LYP,[41,20b] G96PW91,[41,39b] HCTH,[42] mPWPW91,[39] and PBE.[43] Hybrid density functionals (HDFT), which include a percentage of Hartree−Fock (exact) exchange, included in this study are the B3LYP,[20] B3PW91,[20a,39b] B3P86,[20a,40] B97-1,[44] mPW1PW91 (mPW0),[39b] PBE1PBE (PBE0),[43] MPW1K,[45] BH&HLYP,[46,20b] and MPWLYP1M.[47] Two newly developed hybrid functionals that include contributions from unoccupied virtual orbitals via perturbation theory are included in this report: B2-PLYP and mPW2-PLYP. Meta functionals (MDFT), which include the orbital kinetic energy component, included in this study are BB95,[38,48] mPWB95,[39a,48] mPWKCIS,[39a,49] PBEKCIS,[43,49] TPSS,[50] and VSXC.[51] Hybrid meta functionals (HMDFT), which includes exact exchange into meta functionals, employed in this study are B1B95,[48] MPWKCIS1K,[52] BB1K,[53] MPWB1K,[54] MPW1B95,[54] and TPSSh.[55]

Barrier 1. The values of Ba1, calculated with all the functionals previously mentioned, are shown in Figure 11. A value for Ba1 within 5 kcal•mol^{-1} of experiment, which is the typical margin of error for DFT in calculating barrier heights, was calculated for all but three of the functionals tested. However, a value within 1 kcal•mol^{-1} of experiment, which is the definition for “chemical accuracy” of a calculation, was calculated with the BPW91, G96LYP, G96PW91, B3P86, B97-1, mPW0, MPW1K, BH&HLYP, BB1K, and MPWB1K functionals. The error in these calculations is systematically below the experimental value; only the TPSS, TPSSh, BB1K, B2-PLYP, and mPW2-PLYP functionals calculated a value greater than the experimental value. Generally, the accuracy of the calculation does increase when exact exchange is included in the functional; for example, the MPW1K, BB1K, and MPWB1K return values that are more accurate than the mPWPW91, BB95, and mPWB95 parent functionals. The average value and standard deviation were calculated for each DFT category, and these numbers are included in Figure 11. A particularly poor value for Ba1 was calculated with the VSXC functional because the VSXC/BS1 optimized geometry of **2-TS** is similar to the structure of **2a-TS** where the *axial* pz ring has rotated to form the side-on interaction. For Ba1, a value of 24.3 kcal•mol^{-1} was calculated by using the HF method (HF/BS3 level of theory) and following the same procedure for this calculation as was performed with the density functionals.

Barrier 2. In Figure 12, the calculated value of Ba2 is presented for each functional and for the average values for each functional group. A value for Ba2 within 5 kcal•mol^{-1} was calculated for all but four of the functionals tested; however, a value within chemical accuracy was calculated for only the BPW91, MPWLYP1M, B3LYP, mPWKCIS, and PBEKCIS functionals. The accuracy and precision in calculating Ba2 is poorer for each functional category; the

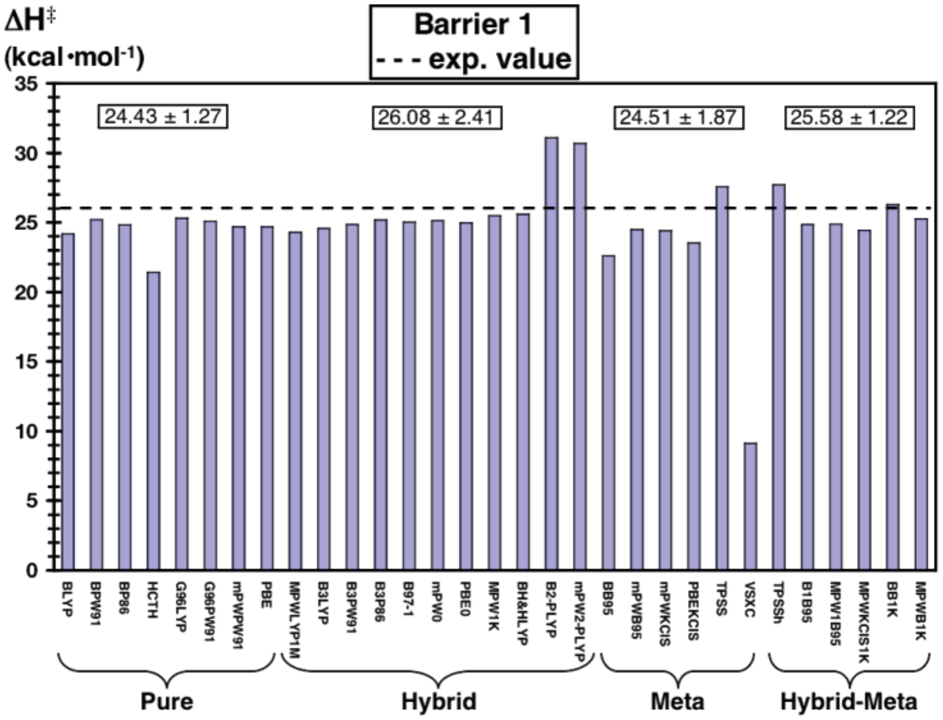


Figure 11. The calculated value for Ba1 for each functional. The dashed line represents the experimental value. In the boxes, the average values with standard deviations are presented for each group. The VSXC functional failed the Q-test (C.I. 90%) that was applied to the meta group and was not included in the statistics.

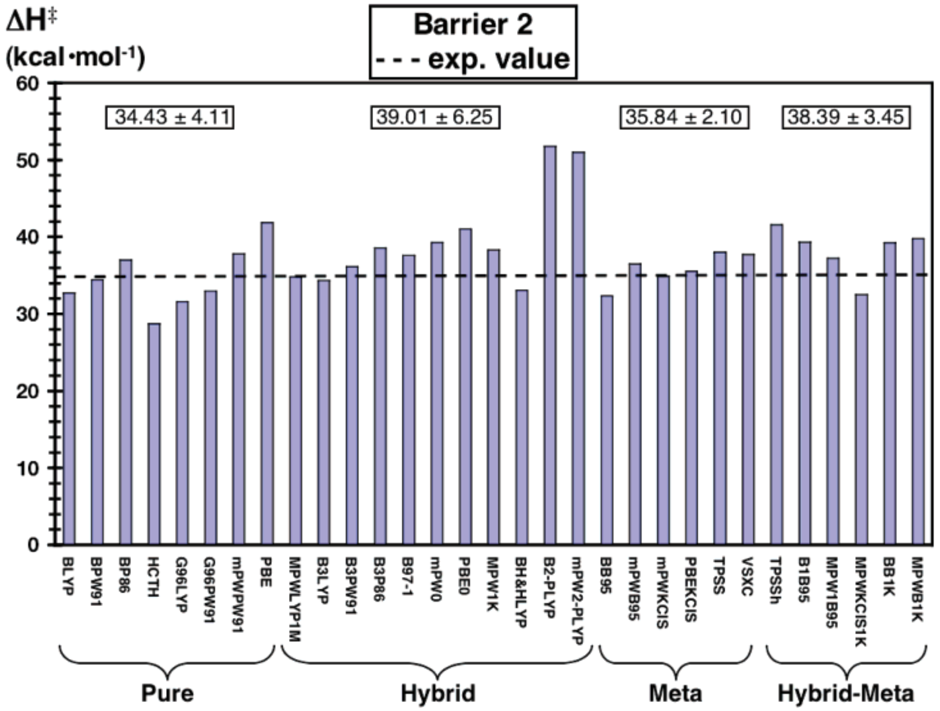


Figure 12. The calculated value of Ba2 for each functional. The dashed line represents the experimental value. The numbers in the boxes are the average values with standard deviations for each DFT category.

meta category is the most accurate and precise group. For Ba2, the calculated value does increase when exact exchange is included in the functional, but the accuracy generally decreases; for example, the meta group has a smaller average value and a lower deviation than the hybrid-meta group. At the HF/BS3 level of theory, a value of 11.2 kcal•mol^{-1} was calculated for Ba2.

Statistical Analysis. The mae for the functionals tested are listed in Table 5, and these values were determined for the results calculated at the functional/BS3 level of theory. From this error analysis, the best performing pure, hybrid, meta, and hybrid-meta density functionals are BPW91, MPWLYP1M, mPWKCIS, and MPW1B95, respectively; and the best overall performer is the BPW91 functional.

Table 5. mae for the Functionals Tested in This Report

pure		hybrid		meta		hybrid-meta	
functional	mae	functional	mae	functional	mae	functional	mae
BLYP	2.05	MPWLYP1M	*0.95*	BB95	3.04	TPSSh	4.15
BPW91	***0.68***	B3LYP	1.05	mPWB95	1.52	B1B95	2.74
BP86	1.58	B3PW91	1.15	mPWKCIS	*0.84*	MPW1B95	*1.68*
G96LYP	5.44	B3P86	2.18	PBEKCIS	1.51	MPWKCIS1K	2.03
G96PW91	1.48	B97−1	1.77	TPSS	2.28	BB1K	5.46
HCTH	2.04	mPW0	2.58	VSXC	9.81	MPWB1K	2.76
mPWPW91	2.06	PBE0	3.52				
PBE	4.07	MPW1K	1.93				
		BH&HLYP	1.17				
		B2-PLYP	10.92				
		mPW2-PLYP	10.33				

Summary of Density Functional Benchmarking Studies. Overall, the accuracy of the calculations is greater for Ba1 than for Ba2; the errors in the individual calculation of **5** and methane are summed, which decreases the accuracy of the calculations of Ba2. In previous studies, more accurate values for barriers were calculated with functionals where greater amounts of exact exchange were admixed into the functionals,[56] and this trend is supported with the data for Ba1. For example, the calculated value for Ba1 with the BLYP, B3LYP, and BH&HLYP functionals approaches the experimental value as the amount of exact exchange admixed into the functional increases. For both barriers, the average value increases when exact exchange is incorporated into the functionals; however, the deviation generally increases (Figures 11 and 12). In order to measure the effect of changing between common exchange and correlation functionals, the LYP, PW91, and P86 correlation functionals were paired with the B88 and B3 exchange functionals; and the general trend is that greater values (Ba1 and Ba2) were calculated in the order of LYP < PW91 < P86. The functionals with the B3 exchange functional calculated values that were greater than the corresponding functional with the B88 exchange functional. The only functional that calculated a value within 1 kcal·mol^{-1} for both barriers was BPW91. The B2-PLYP and mPW2-PLYP functionals, which include contributions from the virtual orbitals, are unsuitable for calculating these barrier heights as the values were much too high and diverged from experiment with basis set saturation.

Recently, Truhlar et al. performed a DFT benchmarking study[47] with a test set comprised predominantly of metal-containing compounds, and the G96LYP and MPWLYP1M functionals were shown to be suitable for these systems. In our study, more accurate values for Ba1 and Ba2 were returned with the MPWLYP1M functional. Quintal et al.[57] reported a benchmarking study of various functionals and found the kinetic functionals optimized for barrier heights (i.e., MPW1K) unsuitable for barriers of late row transition-metal reactions; in our study, these kinetic functionals performed well for Ba1 but not Ba2. The enthalpic values for Ba1 and Ba2 are tabulated for each functional in Supporting Information Table 1 at the BS1, BS2, and BS3 levels of theory.

Table 6. Results in Calculating Ba1 for Various ECP/BS That Were Assigned to Platinum[a]

no.	Pt: outermost 18e^{-}	ECP for Pt: inner 60e^{-}	$\Delta H^{\ddagger}$ Ba1 kcal·mol^{-1}
1	CRENBL	AREP	25.22
2	SBKJC	SBKJC	23.90
3	HW-VDZ (341/321/21)	LANL2	22.88
4	mLANL2DZ (341/341/21)	″	24.58
5	LANL2DZ(f) (341/341/21/1)	″	25.79
6	LANL2TZ (341/341/111)	″	24.95
7	SDD	Stuttgart RSC 1997	23.80
8	SDD(2f)	″	24.14
9	SV	″	21.80
10	TZVP	″	23.65
11	TZVPP	″	23.89
12	QZVP	″	20.38

[a] All other atoms were assigned the basis sets of BS3.

Basis Set Study. Only the first barrier (Ba1) was considered for the ECP/BS and all electron basis set benchmarking studies, and only the B3LYP functional was used in the large basis set study. Twelve ECP/BS were examined to measure the effect on the value of Ba1. The same procedure that was used to test the functionals was used here, but only the ECP/BS was replaced for each test. The geometries of **1** and **2-TS** were fully optimized with each ECP/BS (with the all-electron basis sets of BS1 for the first row elements), and single-point (SP) calculations were run on these optimized geometries with the ECP/BS and the all-electron basis sets of BS3 for the first row elements. These SCF energies were then added to the B3LYP/BS1 corrections to the enthalpy for **1** and **2-TS** to obtain the relative enthalpy difference. Four ECPs were used in this study for the inner 60e^{-} of platinum, and they are the Hay and Wadt LANL2,[21] the Stuttgart relativistic small core (RSC) 1997[58] ECP, the averaged relativistic (AREP) ECP of Ross et al.,[59] and the relativistic compact effective potential (RCEP) of Stevens et al. (SBKJC).[60] The basis sets coupled with the ECPs are the Hay and Wadt valence double-ζ BS[61] (HW-VDZ); the mLANL2DZ BS of Couty and Hall as previously mentioned;[22] the valence double-ζ SBKJC BS of Steven et al.;[60] the Stuttgart/Dresden double-ζ SDD[58] BS; the split valence (SV), triple-ζ with one (TZVP) and two (TZVPP) polarization functions, and quadruple-ζ with one polarization function (QZVP) of Weigend and Ahlrichs.[62]

In Table 6, the results of benchmark studies are shown for the platinum ECP/BS considered in this study. For each

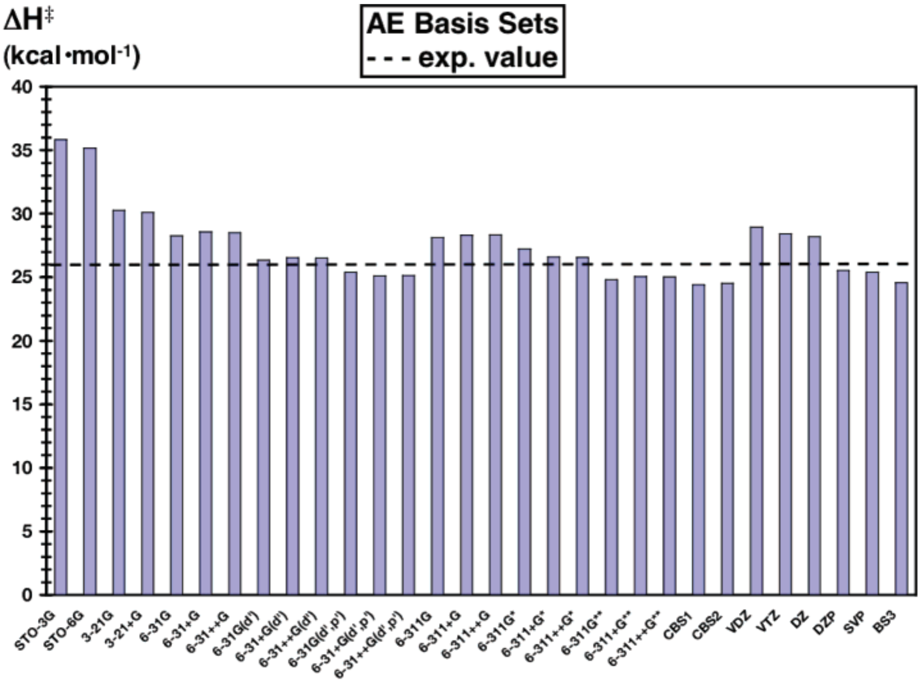


Figure 13. The effect of the basis set on the value of Ba1. The B3LYP/BS1 geometries were used in this study, and all non-platinum elements were assigned the basis set listed. The experimental value of Ba1 is represented by the dashed line.

BS used in this study, the addition of a polarization function resulted in an increased value calculated for Ba1. The modification of Couty and Hall to the HW-VDZ BS improved the value by nearly 2 kcal•mol^{-1}, while decontracting the d shell to form a triple-ζ quality BS returned a similar value to that of the mLANL2DZ. Similar values for Ba1 were calculated with the TZVP and TZVPP BS; however, the values that were calculated with the SV and QZVP BS are the lowest in this study. Of all the ECP/BS that were assigned to platinum, the SV and QZVP BS are the poorest for calculating the value of this barrier.

To benchmark the all-electron basis sets for the first row elements, platinum was assigned the ECP/BS of BS1, and the first row atoms were assigned the same basis sets from the list of Pople's n-Gaussian[63] (STO-nG, $n = 3,6$) basis sets; Pople-style split valence[64] from 3-21G to 6-311++G**; Dunning's full double-ζ basis set (DZ), double-ζ plus polarization basis set (DZP),[65] and split valence plus polarization (SVP) basis set;[66] and Ahlrich's valence double- and triple-ζ basis sets (VDZ, VTZ).[67] To measure basis set saturation, the large basis sets of the complete basis set atomic pair natural orbital (CBS-APNO) method of Petersson and co-workers were used,[68] and these basis sets are denoted CBS1 and CBS2.[69] To obtain the calculated value for Ba1, the SCF energies from these SP calculations were added to the B3LYP/BS1 correction to the enthalpy.

The results are shown in Figure 13 for the all-electron basis set benchmarking study. The most important factor for calculating accurate barrier values is the addition of polarization functions to the basis set, and this trend is seen for each family of basis sets. Diffuse functions, applied either to non-hydrogen atoms (+) or to all atoms (++), did not significantly alter the calculated value compared to the same basis sets without the diffuse functions. Increasing the size of the basis sets from double- to triple-ζ did not significantly alter the calculated value for the barrier. Basis set saturation was reached at the CBS1 level of theory as the addition of two f polarization functions to CBS1, producing CBS2, did not alter the calculated value of Ba1. The energies for each basis set is included in Supporting Information Table 2.

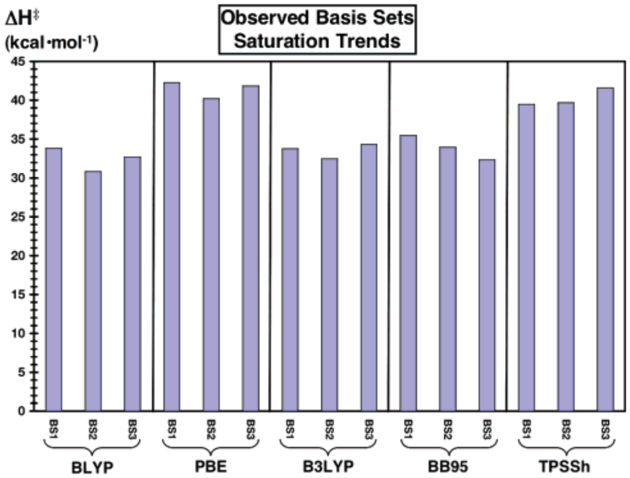


Figure 14. The three basis set saturation trends observed in this work. The trends represented by BLYP, PBE, and B3LYP are representative for most of the functionals tested. The exceptions are discussed in the text.

Basis Sets and Functionals. The trends in basis set saturation (BSS) are shown in Figure 14. For most of the functionals tested, the BSS trend is unexpected because the value calculated at the cc-pVTZ (BS2) level of theory is less than that of both the cc-pVDZ (BS1) and cc-pVQZ (BS3) levels of theory, and the data presented for the BLYP, PBE, and B3LYP functionals are representative for most of the functionals. However, there are exceptions; an expected BSS trend is observed for BB95 (Ba1 & Ba2) where the calculated value decreases with the increase in basis set size, while the

BSS trend for the TPSSh values increase and diverge from the experimental value (Ba2 only). The B2-PLYP and mPW2-PLYP functionals exhibit a similar trend as with TPSSh but for both barriers. For example, the values for Ba1 and Ba2, calculated with the B2-PLYP functional, increase from 27.7 to 31.1 and from 42.8 to 51.7 kcal•mol^{-1} for the BS1, BS2, and BS3 levels of theory, respectively.

Conclusion

We presented the reaction mechanism for the conversion of **1** into **19**, where the important mechanistic barriers to C−H coupling and methane release were analyzed. Against the experimental values of these barriers, 31 density functionals were benchmarked, and, within the definition of "chemical accuracy", 11 were found to be accurate for calculating the C−H coupling barrier, while only 5 were accurate for calculating the value of Ba2. In general, more accurate values for Ba1 were calculated with the functionals with higher values of exact exchange (ca. 40%) admixed into the functional, but those functionals did not perform well for calculating the dissociation barrier. Many of the common ECP/BS combinations available for platinum were found to be suitable for calculating reaction barriers; and polarization functions, added to each all electron basis set, were shown to be a requirement. In this study, DFT was shown to be a suitable method for including electron correlation, as it greatly outperformed the Hartree−Fock theory in calculating these two barriers.

Acknowledgment. This work was supported by grants from the NSF (CHE-0518074 and DMS-0216275) and The Welch Foundation (A0648). The authors thank Chad Beddie for valuable discussions.

Supporting Information Available: Every value for Ba1 and Ba2 (calculated with each functional and basis set), optimized geometries of complexes **11-TS−19**, and the remaining complexes for the *rotation* and *inversion* pathways. This material is available free of charge at http://pubs.acs.org.

References

(1) (a) Williams, T. J.; Labinger, J. A.; Bercaw, J. E. *Organometallics* **2007**, *26*, 281−287. (b) Zhang, F.; Kirby, C. W.; Hairsine, D. W.; Jennings, M. C.; Puddephatt, R. J. *J. Am. Chem. Soc.* **2005**, *127*, 14196−14197. (c) Heyduk, A. F.; Driver, T. G.; Labinger, J. A.; Bercaw, J. E. *J. Am. Chem. Soc.* **2004**, *126*, 15034−15035. (d) Owen, J. S.; Labinger, J. A.; Bercaw, J. E. *J. Am. Chem. Soc.* **2006**, *128*, 2005−2016. (e) Konze, W. V.; Scott, B. L.; Kubas, G. J. **2002**, *124*, 12550−12556. (f) *Activation and Functionalization of C−H Bonds*; Goldberg, K. I., Goldman, A. S., Eds.; Oxford University Press: Washington, DC, 2004; pp 1−440.

(2) Garnett, J. L.; Hodges, R. J. *J. Am. Chem. Soc.* **1967**, *89*, 4546−4547.

(3) (a) Gol'dshleger, N. F.; Tyabin, M. B.; Shilov, A. E.; Shteinman, A. A. *Russ. J. Phys. Chem.* **1969**, *43*, 1222−1223 (English translation). (b) Shilov, A. E.; Shul'pin, G. B. *Chem. Rev.* **1997**, *97*, 2879−2932. (c) Shilov, A. E. *Activation of Saturated Hydrocarbons by Transition Metal Complexes*; D. Riedel: Dordrecht, The Netherlands, 1984; pp 1−203. (d) Shilov, A. E.; Shul'pin, G. B. *Activation and Catalytic Reactions of Saturated Hydrocarbons in the Presence of Metal Complexes*; Kluwer: Dordrecht, The Netherlands, 2000; pp 1−534.

(4) (a) Stahl, S. S.; Labinger, J. A.; Bercaw, J. E. *Angew. Chem., Int. Ed.* **1998**, *37*, 2180−2192. (b) Lersch, M.; Tilset, M. *Chem. Rev.* **2005**, *105*, 2471−2526.

(5) Fekl, U.; Kaminsky, W.; Goldberg, K. I. *J. Am. Chem. Soc.* **2001**, *123*, 6423−6424.

(6) Fekl, U.; Goldberg, K. I. *J. Am. Chem. Soc.* **2002**, *124*, 6804−6805.

(7) Crumpton, D. M.; Goldberg, K. I. *J. Am. Chem. Soc.* **2000**, *122*, 962−963.

(8) Procelewska, J.; Zahl, A.; Liehr, G.; van Eldik, R.; Smythe, N. A.; Williams, B. S.; Goldberg, K. I. *Inorg. Chem.* **2005**, *44*, 7732−7742.

(9) Reinartz, S. R.; White, P. S.; Brookhart, M.; Templeton, J. L. *J. Am. Chem. Soc.* **2001**, *123*, 6425−6426.

(10) Reinartz, S. R.; White, P. S.; Brookhart, M.; Templeton, J. L. *J. Am. Chem. Soc.* **2001**, *123*, 12724−12725.

(11) Norris, C. M.; Templeton, J. L. *Organometallics* **2004**, *23*, 3101−3104.

(12) Siegbahn, P. E. M.; Crabtree, R. H. *J. Am. Chem. Soc.* **1996**, *118*, 4442−4450.

(13) (a) Bartlett, K. L.; Goldberg, K. I.; Borden, W. T. *J. Am. Chem. Soc.* **2000**, *122*, 1456−1465. (b) Barlett, K. L.; Goldberg, K. I.; Borden, W. T. *Organometallics* **2001**, *20*, 2669−2678.

(14) Jensen, M. P.; Wick, D. D.; Reinartz, S.; White, P. S.; Templeton, J. L.; Goldberg, K. I. *J. Am. Chem. Soc.* **2003**, *125*, 8614−8624.

(15) Trofimenko, S. *Chem. Rev.* **1993**, *93*, 943−980.

(16) Wik, B. J.; Ivanovic-Burmazovic, I.; Tilset, M.; van Eldik, R. *Inorg. Chem.* **2006**, *45*, 3613−3621.

(17) Zarić, S.; Hall, M. B. *J. Phys. Chem. A* **1998**, *102*, 1963−1964.

(18) Parr, R. G.; Yang, W. In *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989; pp 1−333.

(19) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(20) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(21) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270−283.

(22) Couty, M.; Hall, M. B. *J. Comput. Chem.* **1996**, *17*, 1359−1370.

(23) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(24) Dunning, T. H.; Hay, P. J. *Modern Theoretical Chemistry*; Schaefer, H. F., III, Ed.; Plenum: New York, 1976; pp 1−28.

(25) (a) Manson, J.; Webster, C. E.; Pérez, L. M.; Hall, M. B. *JIMP 2 Version 0.091 (built for Windows PC)*; Department of Chemistry, Texas A&M University: College Station, TX, 2006 (available @ http://www.chem.tamu.edu/jimp2/index.html). (b) Hall, M. B.; Fenske, R. F. *Inorg. Chem.* **1972**, *11*, 768−779.

(26) (a) Hall, M. B.; Fan, H. *Adv. Inorg. Chem.* **2003**, *54*, 321−349. (b) Hartwig, J. F.; Cook, K. S.; Hapke, M.; Incarvito, C. D.; Fan, Y.; Webster, C. E.; Hall, M. B. *J. Am. Chem. Soc.* **2005**, *127*, 2538−2552.

(27) Johansson, L.; Tilset, M.; Labinger, J. A.; Bercaw, J. E. *J. Am. Chem. Soc.* **2000**, *122*, 10846−10855.

(28) Johansson, L; Tilset, M. *J. Am. Chem. Soc.* **2001**, *123*, 739−740.

(29) Thomas, C. M.; Peters, J. C. *Organometallics* **2005**, *24*, 5858−5867.

(30) (a) Webster, C. E.; Hall, M. B. *Inorg. Chim. Acta* **2002**, *330*, 268−282. (b) Webster, C. E.; Hall, M. B. *Inorg. Chim. Acta* **2002**, *336*, 168.

(31) Feng, Y.; Lail, M.; Foley, N. A.; Gunnoe, T. B.; Barakat, K. A.; Cundari, T. R.; Petersen, J. L. *J. Am. Chem. Soc.* **2006**, *128*, 7982−7994.

(32) (a) Reinartz, S.; Baik, M. H.; White, P. S.; Brookhart, M.; Templeton, J. L. *Inorg. Chem.* **2001**, *40*, 4726−4732. (b) Davies, M. S.; Hambley, T. W. *Inorg. Chem.* **1998**, *37*, 5408−5409. (c) Schwartz, D. J.; Andersen, R. A. *J. Am. Chem. Soc.* **1995**, *117*, 4014−4025.

(33) Bader, R. F. W. *Atoms in Molecules, A Quantum Theory*; Oxford University Press: Ithaca, NY, 1990; pp 1−438.

(34) AIM2000 designed by Friedrich Biegler-König, University of Applied Sciences: Bielefeld, Germany. https://www.aim2000.de/ (accessed month year).

(35) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

(36) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398−4401.

(37) The values by which to scale the second-order correction were obtained by communication with the authors of refs 35 and 36.

(38) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(39) (a) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664−675. (b) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

(40) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822−8824.

(41) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433−445.

(42) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264−6271.

(43) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

(44) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264−6271.

(45) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Chem. Phys. A* **2000**, *104*, 4811−4815.

(46) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(47) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127−11143.

(48) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040−1046.

(49) Krieger, J. B.; Chen, J.; Iafrate, G. J.; Savin, A. In *Electron Correlations and Materials Properties*; Gonis, A., Kioussis, N., Eds.; Plenum: New York, 1999; p 463.

(50) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(51) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400−410.

(52) Zhao, Y.; González-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012−2018.

(53) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715−2719.

(54) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908−6918.

(55) Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2005**, *122*, 114102.

(56) (a) Truong, T. N.; Duncan, W. *J. Phys. Chem.* **1994**, *101*, 7408−7414. (b) Durant, J. L. *Chem. Phys. Lett.* **1996**, *256*, 595−602. (c) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405−3416. (d) Mori-Sáchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 201102. (e) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(57) Quintal, M. M.; Karton, A.; Iron, M. A.; Boese, A. D.; Martin, J. M. L. *J. Phys. Chem. A* **2006**, *110*, 709−716.

(58) Andrae, D.; Haussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123−141.

(59) Ross, R. B.; Powers, J. M.; Atashroo, T.; Ermler, W. C.; LaJohn, L. A.; Christiansen, P. A. *J. Chem. Phys.* **1990**, *93*, 6654−6670.

(60) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612−630.

(61) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299−310.

(62) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.

(63) (a) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657−2664. (b) Collins, J. B.; Schleyer, P. v. R.; Binkley, J. S.; Pople, J. A. *J. Chem. Phys.* **1976**, *64*, 5142−5151.

(64) 3-21G: Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939−947. 6−31G: Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257−2261. 6-311G: Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650−654. Diffuse functions (+ & ++): Clark, T.; Chandrasekhar, J.; Sptiznagel, G. W.; Schelyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294−301. Polarization functions: Foresman, J. B.; Frisch, Æ. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian, Inc.:

Pittsburgh, PA, 1996; p 110. The 6-31G(d′) basis set has the d polarization functions for C, N, O, and F taken from the 6-311G basis set instead of the original arbitrarily assigned value of 0.8 used in the 6-31G(d) basis set.

(65) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823−2833.

(66) Dunning, T. H.; Hay, P. J. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum Press: 1977; Vol. 2, pp 1−462.

(67) Schafer, A.; Horn, H.; Ahlrich, R. *J. Chem. Phys.* **1992**, *97*, 2571−2577.

(68) (a) Petersson, G. A.; Al-Laham, M. A. *J. Chem. Phys.* **1991**, *94*, 6081−6090. (b) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. *J. Chem. Phys.* **1988**, *89*, 2193−2218. (c) Montgomery, J. A., Jr.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1994**, *101*, 5900−5909.

(69) Note, in Gaussian03 C.02, the keyword used to retrieve the APNO basis sets appears similar to the keywords for Pople double- and triple-zeta basis sets; however, APNO basis sets are not Pople basis sets. The 6-311G(d′,p′) keyword, defined in this current article as CBS1, calls for a (14s9p4d,6s3p1d)/[6s6p3d,4s2p1d] APNO basis set. The 6-311G(d′) keyword, defined in this current article as CBS2, calls for an APNO basis set that adds two f polarization functions to first row elements, (14s9p4d2f,6s3p1d)/[6s6p3d2f,4s2p1d].

CT700120D

JCTC Journal of Chemical Theory and Computation

Mechanism of 5,5-Dimethylhydantoin Chlorination: Monochlorination through a Dichloro Intermediate

Akin Akdag, S. D. Worley, Orlando Acevedo, and Michael L. McKee*

Department of Chemistry and Biochemistry, Auburn University,
Auburn, Alabama 36849

Received July 23, 2007

Abstract: The hydantoin moiety is an important pharmacore, and when halogenated, hydantoin derivatives act as excellent biocides. However, there have been no computational studies concerning the chlorination mechanism for the hydantoin moiety reported. Herein we describe a computational mechanistic study of the chlorination of 5,5-dimethylhydantoin (**H**) at the B3LYP/6-311+G(2d,p) level. Under a 1:1 molar ratio of hydantoin and a chlorinating agent (HOCl), conproportionation is calculated to be favorable to give the N1 monochloro derivative as the major predicted product, which is in agreement with experiment. Initial direct chlorination at the N1 position is prevented by a high kinetic barrier. The first step involves the deprotonation of the hydantoin moiety (at the N3 position) which is followed by a S_N2 step transferring a chloronium ion (Cl^+) from HOCl to the ionized hydantoin anion. A mechanism is proposed where the N3 nitrogen is chlorinated first followed by the N1 position to form the dichloro derivative. When CPCM solvation free energies (ΔG(solv)) were added to the gas-phase free energies (ΔG(gas)) along the S_N2 reaction path, a sudden decrease in free energy was observed due to the incipient formation of the hydroxide ion. Explicit consideration of solvation within a box of 512 water molecules led to a much more gradual free energy change along the reaction path but a very similar free energy of activation.

Introduction

The hydantoin (2,4-imidazolidinone) moiety is an important medicinal core unit.[1] As can be seen from the structure, it can be derivatized at several positions. Substitution of the hydrogens on the ring with various organic groups has led to hydantoin based drugs,[2] e.g. 5,5-diphenylhydantoin and 5-ethyl-1-methyl-5-phenylhydantoin (Figure 1). The quest for hydantoin-based drugs remains in progress.[3]

Another hydantoin derivative is 5,5-dimethylhydantoin (**H**) whose chlorinated and brominated derivatives have been used both as biocides and organic reagents.[4] Halogenated **H** and similar heterocyclic ring compounds (see Figure 2 for the chlorinated **H** derivatives) have been employed as biocidal moieties in antimicrobial materials in these laboratories.[5] For example, 5,5-dimethyl-3-[3-(triethoxysilyl)propyl]hydantoin (Figure 2) has been shown to be a versatile biocide precursor and can easily be coated onto hard and soft surfaces,[6] and the halogenated derivates of polystyrene hydantoin beads are being employed in developing countries for disinfecting water.[5c]

In previous work, we have studied the stabilities and the mechanism of formation of the N−Cl bond in different heterocyclic moieties[7] and showed that the nature of the substitution around the nitrogen is important for N−Cl bond stability. Moreover, the stability order N−Cl(amine) > N−Cl(amide) > N−Cl(imide) was predicted, which was in accord with experimental observations.[8] Despite the usefulness of these compounds, no computational reports have been published concerning the mechanism of the halogenation of the hydantoin ring moiety. In very interesting early experimental studies concerning halogenation of hydantoin derivatives, it was shown that the thermodynamically controlled monohalogenation product was that containing halogen at the amide nitrogen N1 (Figure 2).[9] A mechanism was

* Corresponding author e-mail: mckee@chem.auburn.edu.

10.1021/ct7001804 CCC: $37.00
Published on Web 10/24/2007

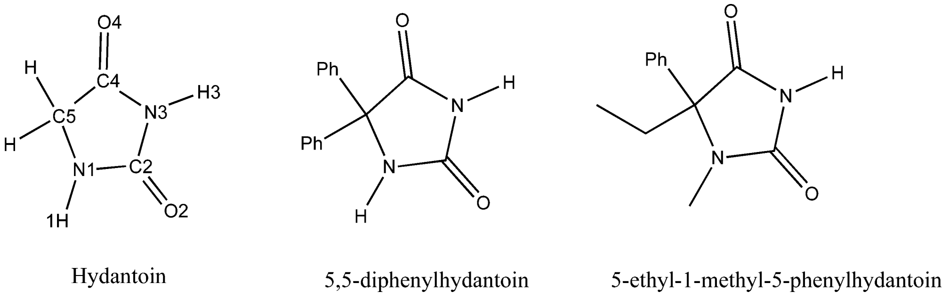


Figure 1. Structure of the hydantoin ring and its numbering system. Two examples of important medicinal hydantoin derivatives.

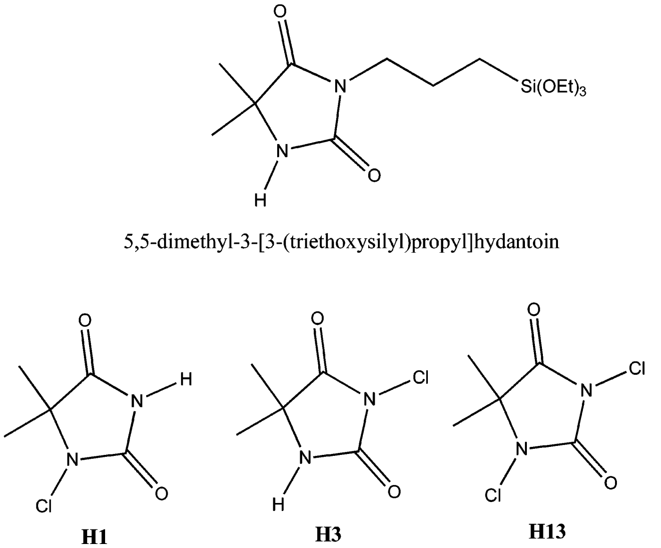


Figure 2. An example of a precursor of a biocidal hydantoin derivative and halogenated 5,5-dimethylhydantoin derivatives.

proposed in which the 1,3-dihalogenated intermediate transferred halogen from the imide nitrogen to the amide nitrogen on an unhalogenated hydantoin molecule.[9]

In this article chlorination of 5,5-dimethylhydantoin was investigated at the B3LYP/6-311+G(2d,p) level. The chlorination of the hydantoin required a two-step process: prechlorination (acid−base equilibrium) and chlorination. Each of these steps was studied computationally with solvation effects included.

Computational Methods

All electronic structure calculations were performed with Gaussian03.[10] The structures were optimized, and zero-point and thermal corrections were calculated at the B3LYP/6-311+G(2d,p) level. Solvation effects were included on the geometry obtained at the B3LYP/6-311+G(2d,p) level with CPCM and tesserae set to 0.05 Å.[2] In the conductor-like polarizable continuum model (CPCM),[12] the solute molecule is placed into a cavity surrounded by the solvent considered as a continuum medium with a dielectric constant of 78.39 (water). The charge distribution of the solute polarizes the dielectric continuum, which creates an electrostatic field that in turn polarizes the solute. In specifying the molecular cavity, the United Atom Topological Model was used with radii optimized for the PBE0/6-31G(d) level of theory (i.e., RADII=UAKS). The choice of UAKS radii was shown by Houk and co-workers[11] to give good results for anions. The aqueous free energies were computed using eq 1.

$$\Delta G(\mathrm{aq}) = \Delta G(\mathrm{gas}) + \Delta G(\mathrm{solvation}) \qquad (1)$$

A 1.9 kcal/mol correction was included in the calculation due to the fact that the molecules are changing in state from ideal gas to ideal solution. A correction shift of 2.4 kcal/mol was also applied to H_2O due to the fact that the water molarity is 55.56.[13] Experimental free energies of solvation were used for H_3O^+ (−110.2 kcal/mol) and OH^- (−104.6 kcal/mol).[14]

Explicit consideration of solvation[15] was made by using the BOSS program.[16] The solvent molecules were represented by the TIP4P water model[17] in a periodic box of 512 (minus the number of non-hydrogen atoms of the solute) water molecules at 25 °C and 1 atm in the NPT ensemble.[18] Each simulation consisted of 5 million configurations of equilibration and 10 million configurations of averaging. The solute energy and energy changes were treated quantum mechanically using PDDG/PM3[19] where the partial charges were obtained from the CM3 charge model,[20] unscaled for negatively charged solutes or scaled by 1.14 for neutral charged solutes[21] with solute−solvent and solvent−solvent intermolecular cutoff distances of 10 Å. This method is particularly well suited for the study of S_N2 reactions.[22]

The labeling used in this work indicates the location of the chlorine atom(s) and the location of the labile hydrogen atom(s). For example, **H** is the parent hydantoin (5,5-dimethylhydantoin) and **H3** is the 3-chloro derivative, **H-an1** is the anion formed by removing the proton attached to N1, and **H1-an3** is the 1-chloro derivative of **H** with a proton removed from N3. Likewise, **H-12t** is the tautomer of **H** with labile hydrogens at N1 and O2, and **H1-4t** is the tautomer of the 1-chloro derivative of **H** with hydrogens at N3 and O4. Occasional use is made of notations such as 1-Me-**H** and 1-Me-**H3** which are 1,5,5-trimethylhydantoin and 3-chloro-1,5,5-trimethylhydantoin, respectively.

Results and Discussion

When Corral and Orazi studied[9] the competitive chlorination of 3,5,5-trimethylhydantoin/1,5,5-trimethylhydantoin with HOCl (eq 2), they found the ratio of 3-chloro-1,5,5-trimethylhydantoin:1-chloro-3,5,5-trimethylhydantoin to be 97:0. When the chlorination agent was changed to be OCl^-, the ratio became 9:42 (eq 3).

$$\text{3-Me-}\mathbf{H} + \text{1-Me-}\mathbf{H} + \mathrm{HOCl} \rightarrow \text{1-Me-}\mathbf{H3} : \text{3-Me-}\mathbf{H1} + \mathrm{H_2O}\ (97{:}0) \qquad (2)$$

$$\text{3-Me-}\mathbf{H} + \text{1-Me-}\mathbf{H} + \text{OCl}^- \rightarrow \text{1-Me-}\mathbf{H3} : \text{3-Me-}\mathbf{H1} + \text{OH}^- \ (9{:}42) \quad (3)$$

This result indicates that two major products are possible. The work below will support the hypothesis that the **H1** product is the thermodynamic product, and **H3** is the kinetic product. In addition to the chlorinating agents HOCl and OCl^-, 1-Me-**H3** (3-chloro-1,5,5-trimethylhydantoin) could also act as a chlorinating agent (eq 4) at the N1 position. In acetone (eq 5), the dichloro derivative **H13** directs 100% chlorination of **H** at the N1 position (forming **H1**).

$$\text{1-Me-}\mathbf{H3} + \text{3-Me-}\mathbf{H} + 1\text{ eq OH}^- \rightarrow \text{3-Me-}\mathbf{H1} + \text{1-Me-}\mathbf{H} + \text{H}_2\text{O} \quad (4)$$

$$\mathbf{H} + \mathbf{H13} \rightarrow 2\ \mathbf{H1}\ (\text{in acetone}) \quad (5)$$

This conproportionation reaction suggests that the N1 position can be favored over the N3 position as the final site of chlorination. The main objective of this article is to account for the observed monochlorination of **H** (under equal hydantoin:HOCl molar ratio) at the N1 position to form the thermodynamically controlled product even though the N3 position provides the kinetically controlled product. A plausible chlorination mechanism begins with ionization, followed by an S_N2 transfer of a chloronium ion (Cl^+) to the nitrogen. Another possibility is the tautomerization of the hydantoin to another form which is more reactive than the hydantoin itself.

Tautomers and Ionization. For an illustration of the formation of various tautomers[23] see Schemes 1 and 2.

Relative energies of hydantoin tautomers are given in Table 1, while relative energies of species in the chlorination mechanism are given in Table 2. As seen in Table 2, the ionization of the amide (**H** → **H-an1**) requires 27.6 kcal/mol of free energy which is 5.3 kcal/mol higher than that of the corresponding tautomer (**H** → **H-32t**). Therefore, the ionized hydantoin **H-an1** is predicted to be in equilibrium with **H** and **H-32t**. It was found that if the hydantoin is chlorinated at N3 (**H3**), then the deprotonation (**H3** → **H3-an1**) is 8.0 kcal/mol more spontaneous than the corresponding unchlorinated derivative (**H** → **H-an1**, see Table 2).

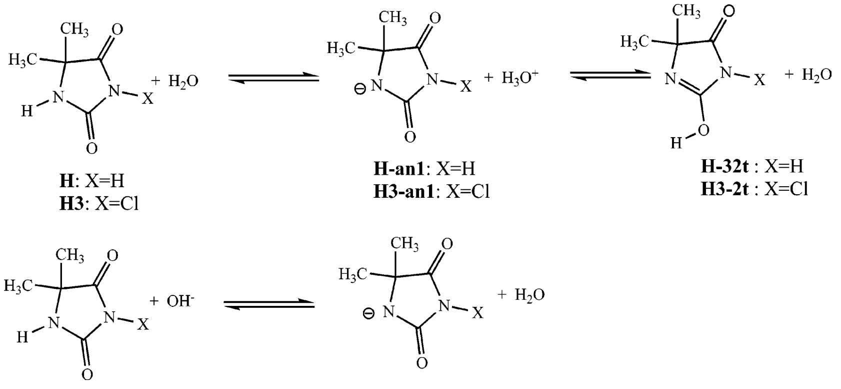
Table 1. Relative Enthalpies and Free Energies (kcal/mol) Tautomers of **H**, **H3**, and **H1**[a]

	ΔH (0K)	ΔH (g,298K)	ΔG (g,298K)	ΔG (aq,298K)
H	0.0	0.0	0.0	0.0
H-32t	17.3	17.7	20.4	22.3
H-12t	18.7	19.0	21.7	18.1
H-14t	18.0	18.4	20.8	18.0
H3	0.0	0.0	0.0	0.0
H3-2t	16.9	16.8	16.3	19.4
H1	−1.1	−1.2	−2.0	−3.9
H1-2t	17.4	17.2	15.6	15.2
H1-4t	15.3	16.3	13.5	11.8

[a] The labeling indicates the location of the chlorine atom and the location of the labile hydrogen atom(s) in the tautomer. For example, **H** is the parent 5,5-dimethylhydantoin, and **H3** is the 3-chloro derivative; **H-12t** is the tautomer of **H** with labile hydrogens at the N1 and O2 positions, while **H1-4t** is the tautomer of the 1-chloro derivative with hydrogen at the O4 position.

This can be attributed to the fact that chlorine is withdrawing electron density since it has partial positive charge (the natural charge, i.e., NPA charge on Cl is 0.14), i.e. the proton bonded to the N3 moiety is more acidic than is the proton bonded to N1. The pK_a and free energies changes of several relevant hydantions[1] are tabulated in Table 3. The pK_a of **H**

Scheme 1. Dissociation of **H** and **H3** under Neutral and Basic Conditions



Scheme 2. Dissociation of **H** and **H1** under Neutral and Basic Conditions

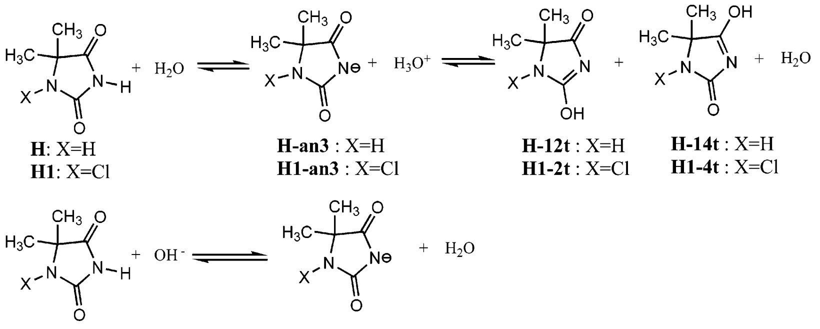


Table 2. Enthalpies and Free Energies for the Chlorination Step of 5,5-Dimethylhydantoin (**H**) by HOCl

		ΔH (0K)	ΔH (g,298K)	ΔG (g,298K)	ΔG (aq,298K)	adj. ΔG^a (aq,298K)
a	**H** + $2H_2O$ + 2HOCl	0.0	0.0	0.0	0.0	0.0
b	**H-an3** + H_2O + 2HOCl + H_3O^+	178.9	179.0	179.2	18.6	18.6
c	**H3** + H_2O + HOCl + H_3O^+ + OH^-	214.5	215.4	210.8	9.3	−9.8
d	**H3-an1** + HOCl + $2H_3O^+$ + OH^-	389.0	389.7	383.5	28.9	9.8
e	**H13** + $2H_3O^+$ + $2OH^-$	428.1	429.4	416.4	16.4	−21.8
f	**H-an1** + H_2O + 2HOCl + H_3O^+	183.8	184.0	185.4	27.6	27.6
g	**H1** + H_2O + HOCl + H_3O^+ + OH^-	213.4	214.2	208.8	5.4	−13.7
h	**H1-an3** + HOCl + $2H_3O^+$ + OH^-	382.2	382.8	374.6	20.8	1.7
i	**H-an3** + HOCl → **H3** + OH^- ^{b}	35.6	36.4	31.6	−9.3	−28.4
j	**H-an3** + **H13** → **H3** + **H1-an3**^{c}	−10.3	−10.2	−10.2	−4.9	−4.9
k	**H** + **H13** → 2**H1**^{d}	−1.3	−1.0	1.2	−5.6	−5.6
l	**H** + **H13** → 2**H3**^{e}	0.9	1.4	5.2	2.2	2.2

[a] The free energy has been reduced by 19.1 kcal/mol for each (H_3O^+/OH^-) pair which is the experimental free energy change for H_3O^+(aq) + OH^-(aq) → $2H_2O$(l). The calculated value for this process is 20.1 kcal/mol. [b] Reaction thermochemistry is equivalent to -**b**+**c**. [c] Reaction thermochemistry is equivalent to −**b**-**e**+**c**+**h**. [d] Reaction thermochemistry is equivalent to −**a**-**e**+2**g**. [e] Reaction thermochemistry is equivalent to −**a**-**e**+**2c**.

Table 3. Experimental and Calculated Free Energy Changes (kcal/mol) for Ionization in Hydantoins

hydantoin (ionization)	site of ionization	pK_a^{a}	exptl ΔG^b	calc ΔG	difference
5,5-dimethyl (**H** → **H-an3**)	N3	9.03	14.7	18.6	3.9
1,5,5-trimethyl (**H** → **H-an3**)^c	N3	9.02	14.7	18.6	3.9
3,5,5-trimethyl (**H** → **H-an1**)^c	N1	>14	>21.5	27.6	<6.1
1-chloro-5,5-dimethyl (**H1** → **H1-an3**)	N3	7.17	12.2	15.4	3.2

[a] Reference 9. [b] $\Delta G = -RT\ln K_a + 2.4$ kcal/mol (correction, see ref 24). [c] The effect of the methyl group at the N1 or the N3 position is assumed to be small relative to a hydrogen atom.

is 9.03[24] which is only slightly changed if the N1 position is methylated (9.02). On the other hand, if the N3 position is methylated, the pK_a is 14 or greater. Chlorination of **H** at the N1 position makes the hydantoin more acidic (pK_a=7.17). The experimental free energy changes computed from the pK_a values are in good agreement with the calculated free energy changes (Table 3).

S$_N$2 Chlorination. **H-an3** is thermodynamically favored over **H-an1** by 9.0 kcal/mol. The negative charge on nitrogen makes N3 nucleophilic such that it can react with hypochlorous acid in an S$_N$2 reaction to produce **H3** and hydroxide. Attempts to locate a transition state in the gas phase at the B3LYP/6-311+G(2d,p) level for the **H-an3** + HOCl → **H3** + OH^- S$_N$2 reaction failed. The problem is due to the poor representation of the aqueous free energy surface using gas-phase optimizations. At 298 K, the gas-phase enthalpy of reaction is +36.4 kcal/mol, while the aqueous phase free energy difference is −28.4 kcal/mol (Table 2), a 64.8 kcal/mol difference! The gas-phase potential energy surface is dominated by the ion−molecule complex **H-an3-complex**, 13.5 kcal/mol more stable than **H-an3** + HOCl, with N−Cl and O−Cl distances of 2.26 and 1.85 Å, respectively. When eq 1 is applied to **H-an3-complex**, the free energy is 6.8 kcal/mol higher than **H-an3** + HOCl. The aqueous-phase destablization is due to charge delocalization in the complex which has a free energy of solvation 9.8 kcal/mol smaller than **H-an3** + HOCl (Table S3).

While the ion−molecule complex **H-an3-complex** is a minimum in the gas phase, it may be close to the maximum along the aqueous S$_N$2 free energy reaction profile. To address this issue we have calculated a reaction path **H-an3** + HOCl → **H3** + OH^- by varying the Cl−OH distance from 1.85 (RC185) to 2.70 (RC270) Å in steps of 0.05 Å. The structures were fully optimized at the B3LYP/6-311+G(2d,p) level, and frequencies were computed after projecting out the reaction coordinate.[25] Solvation energies were computed using the CPCM solvation model and UAKS radii (see Tables S2 and S3). The results are plotted in Figure 3.

While the contribution of zero-point correction, heat capacity correction, and entropy to ΔG(aq) were constant to about 1 kcal/mol along the **H-an3** chlorination reaction path (Table S3), the interaction energies become more positive (less binding) systematically, from −13.5 to 9.8 kcal/mol. At the same time, the solvation free energy ΔG(solv) becomes more negative (−67.8 to −99.1 kcal/mol) which is due to the emergence of a strongly solvated hydroxide anion. The majority of the decrease occurs from RC215 to RC 220 where there is a 13.2 kcal/mol drop in ΔG(solv). Adding ΔG(g) and ΔG(solv) to give ΔG(aq) gives a maximum in the free energy curve at RC215 which is 9.9 kcal/mol above **H-an3** + HOCl.

The reaction path for **H-an1** + HOCl → **H1** + OH^-, also shown in Figure 3, has a free energy barrier of 5.8 kcal/mol. The plots of both reactions are relative to separated reactants, but we note that **H-an3** is 9.0 kcal/mol lower (more spontaneous) than **H-an1** (Table 2). Thus, even though chlorination of **H-an1** has a lower free energy barrier than **H-an3**, the overall process, including ionization, is higher (27.6 + 5.8 = 33.4 for **H-an1** versus 18.6 + 9.9 = 28.5 kcal/mol for **H-an3**).

To determine the cause of the discontinuity in free energy, we plotted the NPA charge of the OH group as a function of the **H-an3** + HOCl → **H3** + OH^- reaction coordinate (Figure 4). The charge of the OH group in the gas phase

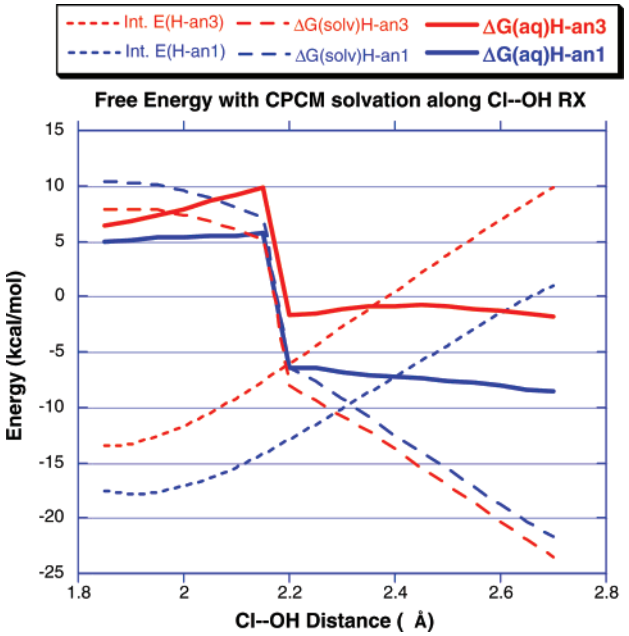


Figure 3. Plot of the electronic interaction energy (Int) between **H-an3**/**H-an1** and HOCl (red/blue upward dashed curves), the solvation free energy Δ*G*(solv) (red/blue downward dashed curves), and the aqueous free energy Δ*G*(aq) (red/blue solid curves) along the reaction coordinate in the reaction **H-an3** + HOCl → **H3** + OH⁻ (red lines) and **H-an1** + HOCl → **H1** + OH⁻ (blue lines).

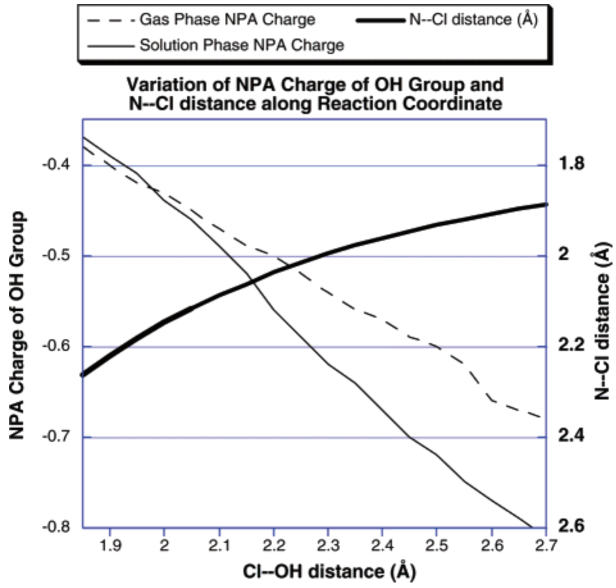


Figure 4. Plot of NPA charges for H group and N−Cl distance as a function of the Cl−OH reaction coordinate in the reaction **H-an3** + HOCl → **H3** + OH⁻.

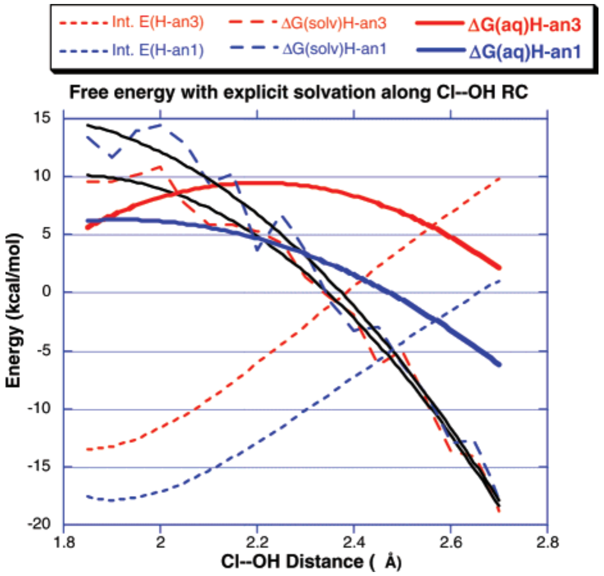


Figure 5. Plot of electronic interaction energy (Int) between **H-an3**/**H-an1** and HOCl (kcal/mol), the solvation free energy Δ*G*(solv), and the aqueous free energy Δ*G*(aq) along the reaction coordinate in the reaction **H-an3** + HOCl → **H3** + OH⁻ (red dashed/solid lines) and **H-an1** + HOCl → **H1** + OH⁻ (blue dashed/solid lines).

and the solution phase both increase as Cl−OH distances increase. The larger increase in solution phase reflects the greater polarizing ability of the environment. However, there is no abrupt change in the OH group charge. In addition, the N−Cl distance smoothly decreases as the Cl−OH distance increases. The discontinuity is caused by a change in the assigned radius of oxygen as the Cl−OH distances increases. Thus, the contribution of ΔG_{el}, the solvation free energy due to electrostatic interactions, to the CPCM solvation free energy shows a discontinuous change along the reaction coordinate between a Cl−OH distance of 2.15 and 2.20 Å. For a Cl−OH distance of 2.15 Å (and less), the program uses an atomic radius of 1.563 Å for the oxygen atom of the OH group, while for a Cl−OH distance of 2.20 Å (and greater), the program uses an atomic radius of 1.290 Å.

The very dramatic increase in solvation free energy between RC215 and RC220 is unusual and may be an artifact of using the CPCM method. Therefore, we considered an alternative way of computing solvation along the reaction path. Each structure along the reaction coordinate was equilibrated (5 M steps) and averaged (10 M steps) in a box of 512 water molecules. The variations of Δ*G*(solv) and Δ*G*(aq) from explicit solvation for the chlorination of **H-an3** and **H-an1** are given in Figure 5 and Tables S2 and S3.

The solvation free energy becomes more negative as the reaction proceeds. The variation of solvation free energy at the 18 individual points along the reaction path is superimposed on the fitted quadratic line. The reference energy is **H-an3** + HOCl (red lines) or **H-an1** + HOCl (blue lines).

The overall trend in the increase (more negative) of Δ*G*(solv) along the reaction path is the same as that obtained with the CPCM method. However, the position in the maximum in Δ*G*(aq) is displaced later ($\Delta G^{\ddagger}$≈9 kcal/mol at about 2.3 Å) for the chlorination of **H-an3** and earlier for the chlorination of **H-an1** (($\Delta G^{\ddagger}$≈6 kcal/mol at about 1.9 Å) relative to the maximum in Δ*G*(aq) from the CPCM method (Figures 3 and 5). In addition, the variation of the fitted Δ*G*(solv) and Δ*G*(aq) for explicit solvation is much smoother along the reaction path than for CPCM.

The solvation effect with the CPCM method is from the solute embedded in a cavity and the interaction between the

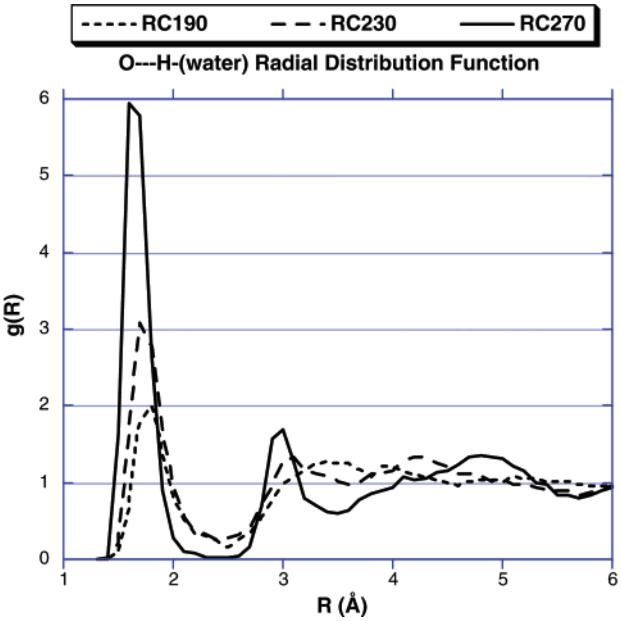


Figure 6. Radial distribution function g(R) for the O−H-(water) distances for three structures along the reaction coordinate in the reaction **H-an3** + ClOH → **H3** + OH$^-$.

solute and the solvent dielectric at the cavity surface. In explicit solvation, the accessibility of the solute to the solvent is more realistically modeled. As the hydroxide ion departs along the reaction path, the water molecules adopt a very tight solvation shell. This is illustrated in the radial distribution function (Figure 6) for the chlorination of **H-an3** at three points along the reaction path, RC190, RC230, and RC270 which represent points before, near, and after the maximum in ΔG(aq).

At the last point, RC270, a very sharp peak in the distribution between the oxygen of the developing OH and the hydrogens of water is apparent at about 1.7 Å with a strong secondary solvent shell at 3.0 Å. At earlier points (RC230 and RC190), there is a gradual decrease in the size of the peak and a movement to a larger O−H separation.

With these results in hand, a summary of relative free energies is given in Figure 7.

After monochlorination of the hydantoin occurs, the monochlorinated hydantoin can deprotonate. Therefore, the same argument is valid for the addition of second chlorine to the hydantoin in an S$_N$2-like reaction mechanism. The values of Δ$G^‡$ for the **H3** → **H13** step (28 kcal/mol) and the **H13** → **H1** step (31 kcal/mol) in Figure 7b are estimated by assuming that the chlorination step has a 8 kcal/mol barrier. For **H3** → **H13**, this is **H3** → **H3-an1** (19.6 kcal/mol) plus 8 kcal/mol. For **H13** → **H1**, the barrier for the reverse reaction **H1** → **H13** (i.e., **H1** → **H1-an3** (15.4 kcal/mol) + 8) is added to the free energy of the reaction (−13.7 + 21.8 = 8.1 kcal/mol) to give an estimated free energy barrier of (15.4 + 8 + 8.1) 31.5 kcal/mol for the forward reaction.

Table 2 shows that the formation of **H13** from **H1-an3** is less favored than that from **H3-an1** energetically. This can be attributed to fact that the thermodynamic stability of the amide N−Cl bond is higher than for the imide N−Cl bond. The high negative charge is localized on N3 relative to N1. **H1** is favored over **H3** by 1.2 kcal/mol at ΔH(g,298K). The stability is increased to 3.9 kcal/mol for ΔG(aq,298K). Hypochlorous acid is a much better chlorinating agent than **H13** as shown by reaction **i** (Table 2) which is 28.4 kcal/mol spontaneous compared to reaction **j** which is 4.9 kcal/mol spontaneous. However, under a 1:1 equal molar ratio of **H** to HOCl, HOCl will be exhausted after one-half of **H** is converted into **H13**. Thus, under conditions of limited chlorinating agent, the reaction will be thermodynamically driven to monochlorination at the N1 position **(**eq 6 and reaction **k**, Table 2). Under conditions of excess chlorination agent, the dichlorohydantoin derivative is expected.

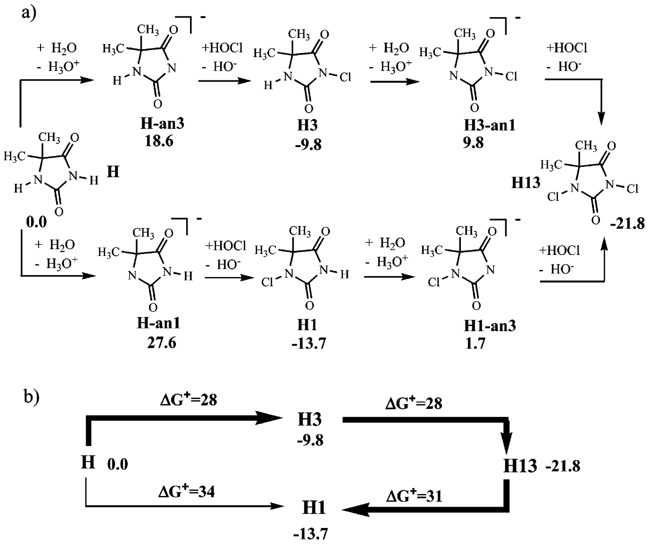


Figure 7. The most favorable path to **H1** is **H** → **H3** → **H13** → **H1**. (a) Aqueous free energies (kcal/mol) of species relative to **H** + 2H$_2$O + 2HOCl. (b) Aqueous free energies (kcal/mol) of species along the chlorination reaction path. The values Δ$G^‡$ are relative free energies with respect to the indicated reactant.

$$\mathbf{H} + \mathbf{H13} \rightleftarrows 2\mathbf{H1}\ \Delta G(\mathrm{aq}) = -5.6\ \mathrm{kcal/mol} \qquad (6)$$

Conclusions

In this study we have investigated a plausible mechanism for the chlorination of 5,5-dimethylhydantoin. The mechanism was broken into two parts: (1) prechlorination (acid−base equilibrium) and (2) chlorination. The dissociation of a proton from N3 is calculated to be more favorable than from N1, in agreement with experimental observations. When monochlorinated, the hydantoin become significantly more acidic.

The chlorination step was investigated in an S$_N$2-like mechanism, in which the anions act as nucleophiles and hydroxide as a leaving group. Since the N−Cl bond is stronger than the O−Cl bond thermodynamically, this reaction is favored. Moreover, the hydroxide anion has significantly more solvation free energy than the corresponding hydantoin anion derivatives. Based upon the prechlorination step, **H3** is favored kinetically. On the other hand, chlorination stabilizes **H1** over **H3**. The preferred route of chlorination is to form **H3** (monochlorination) and then **H13** (dichlorination). If limited chlorination agent is used, the conproportionation reaction **H** + **H13** ⇄ 2**H1** can take place to produce **H1** (monochlorination) as the thermodynamically favored product. Thus, the computations reported herein support the mechanism suggested for **H1** formation in the experimental paper by Corral and Orazi.[9] They also demonstrate the utility of such computations in the study of important complex organic reactions.

Acknowledgment. This work has been supported by the U.S. Air Force through contract F08637-02-C-7020 and the Vanson HaloSource Company. The computation time was provided by Auburn University and the Alabama supercomputer.

Supporting Information Available: Absolute energies, zero-point energies, enthalpy corrections, entropies, and solvation energies for various species (Table S1); absolute energies, zero-point energies, enthalpy corrections, entropies, and solvation energies (CPCM and explicit) for points along the S_N2 reaction path for the chlorination of **H-an3** and **H-an1** (Tables S2 and S3); the full citation for ref 10; and optimized Cartesian coordinates of all related species at the B3LYP/6-311+G(2d,p) level (Table S4). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Avendaño, C.; Menendez, J. C. *Hydantoin and Its Derivatives* In *Kirk-Othmer Encyclopedia Chemical Technology*, 4th ed.; 2000; pp 1−21. (DOI:10.1002/0471238961.0825040101220514.a01).

(2) Meusel, M.; Gütschow, M. *Org. Preparations Procedures Int.* **2004**, *36*, 391−443.

(3) (a) Kruger, H. G.; Mdluli, P. S. *Struct. Chem.* **2006**, *17*, 121−125. (b) Teng, X.; Degterev, A.; Jagtap, P.; Xing, X.; Choi, S.; Denu, R.; Yuan, J.; Cuny, G. D. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 5039−5044. (c) Riley, P.; Figary, P. C.; Entwisle, J. R.; Roe, A. L.; Thompson, G. A.; Ohashi, R.; Ohashi, N.; Moorehead, T. J. *J. Pharm. Sci.* **2005**, *94*, 2084−2095. (d) Wehner, V.; Stilz, H.-U.; Osipov, S. N.; Golubev, A. S.; Sieler, J.; Burger, K. *Tetrahedron* **2004**, *60*, 4295−4302. (e) Bakalova, A.; Buyukliev, R.; Momekov, G.; Ivanov, D.; Todorov, D.; Konstantinov, S.; Karaivanova, M. *Eur. J. Med. Chem.* **2005**, *40*, 590−596. (f) Krishnan, R. S. G.; Thennarasu, S.; Mandal, A. B. *Chem. Phys.* **2003**, *291*, 195−205.

(4) (a) Akdag, A.; Webb, T.; Worley, S. D. *Tetrahedron Lett.* **2006**, *47*, 3509−3510. (b) Walters, T. R.; Zajac, W. W.; Woods, J. M. *J. Org. Chem.* **1991**, *56*, 316−321. (c) Szumigala, R. H.; Devine, P. N.; Gauthier, D. R.; Volante, R. P. *J. Org. Chem.* **2004**, *69*, 566−569. (d) Rivera, N. R.; Balsells, J.; Hansen, K. B. *Tetrahedron Lett.* **2006**, *47*, 4889−4891. (e) Bartoli, G.; Bosco, M.; Carlone, A.; Locatelli, M.; Melchiorre, P.; Sambri, L. *Angew. Chem., Int. Ed.* **2005**, *44*, 6219−6222. (f) Soracco, R. J.; Wilde, E. W.; Mayack, L. A.; Pope, D. H. *Water Res.* **1985**, *19*, 763−766. (g) Koval, I. V. *Russ. J. Org. Chem.* **2001**, *37*, 297−317.

(5) (a) Tsao, T. C.; Williams, D. E.; Worley, C. G.; Worley, S. D. *Biotechnol. Prog.* **1991**, *7*, 60−66. (b) Chen, Y.; Worley, S. D.; Kim, J.; Wei, C.-I.; Chen, T.-Y.; Suess, J.; Kawai, H.; Williams, J. F. *Ind. Eng. Chem. Res.* **2003**, *42*, 5715−5720. (c) Worley, S. D.; Sun, G. *Trends Polym. Sci.* **1996**, *4*, 364−370.

(6) (a) Liang, J.; Owens, J. R.; Huang, T. S.; Worley, S. D. *J. Appl. Polym. Sci.* **2006**, *101*, 3448−3454. (b) Liang, J.; Wu, R.; Huang, T. S.; Worley, S. D. *J. Appl. Polym. Sci.* **2005**, *97*, 1161−1166.

(7) (a) Akdag, A.; Okur, S.; McKee, M. L.; Worley, S. D. *J. Chem. Theory Comput.* **2006**, *2*, 879−884. (b) Akdag, A.; McKee, M. L.; Worley, S. D. *J. Phys. Chem. A* **2006**, *110*, 7621−7627.

(8) Qian, L.; Sun, G. *J. Appl. Polym. Sci.* **2003**, *89*, 2418−2425, and references cited therein.

(9) (a) Corral, R. A.; Orazi, O. O. *J. Org. Chem.* **1963**, *28*, 1100−1104. (b) Petterson, R. C.; Grzeskowiak, U. *J. Org. Chem.* **1958**, *24*, 1414−1419.

(10) Frisch, M. J. et al. *Gaussian03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004 (for the full citation see the Supporting Information).

(11) Takano, Y.; Houk, K. N. *J. Chem. Theory Comput.* **2005**, *1*, 70−77.

(12) (a) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669−681. (b) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799−805. (c) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995−2001. (d) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1996**, *255*, 327−335.

(13) McKee, M. L. *J. Phys. Chem. A* **2003**, *107*, 6819−6827.

(14) Camaioni, D. M.; Schwerdtfeger, C. A. *J. Phys. Chem. A* **2005**, *109*, 10795−10797.

(15) For recent applications, see: (a) Acevedo, O.; Jorgensen, W. L. *Org. Lett.* **2004**, *6*, 2881−2884. (b) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2005**, *127*, 8829−8834. (c) Acevedo, O.; Jorgensen, W. L. *J. Org. Chem.* **2006**, *71*, 4896−4902. (d) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 6141−6144.

(16) (a) Jorgensen, W. L. *BOSS, version 4.6*; Yale University: New Haven, CT, 2004. (b) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689−1700.

(17) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(18) Jorgensen, W. L. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 3, pp 1754−1763.

(19) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601−1622.

(20) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2003**, *24*, 1291−1304.

(21) Udier-Blagović, M.; Morales De Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322−1332.

(22) (a) Tubert-Brohman, I.; Guimarães, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 138−150. (b) Vayner, G.; Houk, K. N.; Jorgensen, W. L.; Brauman, J. I. *J. Am. Chem. Soc.* **2004**, *126*, 9054−9058.

(23) For prior studies of hydantoin tautomers, see: (a) Kleinpeter, E. *Struct. Chem.* **1997**, *8*, 161−173. (b) Bausch, M. J.; David, B.; Dobrowolski, P.; Guadalupe-Fasano, C.; Gostowski, R.; Selmarten, D.; Prasad, V.; Vaughn, A.; Wang, L. H. *J. Org. Chem.* **1991**, *56*, 5643−5651. (c) Bagno, A.; Comuzzi, C. *Eur. J. Org. Chem.* **1999**, 287−295. (d) Kleinpeter, E.; Heydenreich, M.; Kalder, L.; Koch, A.; Henning, D.;

Kempter, G.; Benassi, R.; Taddei, F. *J. Mol. Struct.* **1997**, *403*, 111−122.

(24) For example, the equation $\Delta G = -RT\ln K_a$ and a pK_a value of 9.03 gives $\Delta G = 12.3$ kcal/mol. However, the free energy is increased by 2.4 kcal/mol to account for the concentration of liquid water. See: Da Silva, C. O.; Da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **1999**, *103*, 11194−11199.

(25) Baboul, A. G.; Schlegel, H. B. *J. Chem. Phys.* **1997**, *107*, 9413−9417.

CT7001804

JCTC Journal of Chemical Theory and Computation

First Hybrid Embedding Scheme for Polar Covalent Materials Using an Extended Border Region To Minimize Boundary Effects on the Quantum Region

Alexei M. Shor,[†] Elena A. Ivanova Shor,[†] Vladimir A. Nasluzov,*[,†] Georgi N. Vayssilov,[‡] and Notker Rösch*[,§]

Institute of Chemistry and Chemical Technology, Russian Academy of Sciences, 660049 Krasnoyarsk, Russian Federation, Faculty of Chemistry, University of Sofia, 1126 Sofia, Bulgaria, and Department Chemie, Technische Universität München, 85747 Garching, Germany

Received June 27, 2007

Abstract: We present an improved scheme for constructing the border region within a hybrid quantum mechanics/molecular mechanics (QM/MM) embedded cluster approach for zeolites and covalent oxides that ensures proper modeling of adsorption complexes with QM regions of moderate size. The procedure employs a flexible orbital basis set on monovalent oxygen pseudoatoms at the boundary of the QM cluster and introduces a pseudopotential description without explicit representation of valence electrons for their immediate Si neighbors in the MM region. This novel QM/MM border scheme, implemented in the elastic polarizable environment method for polar covalent materials (covEPE), provides an accurate description of the local structure of zeolites and other silica based materials. We assessed the performance of the novel border scheme by comparing calculated and experimental results for structures, vibrational frequencies, and binding energies of CO adsorption complexes at bridging OH groups in zeolites with FAU and MFI structures. In addition, when modeling zeolite-supported metal clusters, the new approach implies considerably reduced corrections due to the basis set superposition error, compared to our previous scheme for treating the border region of the QM partition [*J. Phys. Chem. B* **2003**, *107*, 2228].

1. Introduction

One of the most crucial features of a reliable hybrid quantum mechanics/molecular mechanics (QM/MM) scheme is an adequate construction of the border region between two subsystems.[1] With specific chemical bonding situations in mind, various schemes have been proposed to construct the border region between QM and MM partitions of a system.[2] In strongly ionic oxides such as MgO, the system can be partitioned into QM and MM regions without the QM/MM boundary cutting covalent bonds; in addition, atomic (ionic) centers at the border of the QM part are the same as inside the QM cluster. A characteristic challenge of QM/MM schemes for such systems is an artificial polarization of anionic centers at the border of the QM cluster due to neighboring cations of the MM region as the latter type of centers is represented by positive point charges. This artifact can be avoided by augmenting the representation of cations in the MM region immediately at the QM/MM boundary by total ion model potentials (TIMPs) which provide the proper repulsive contributions to electrons of anionic centers in the vicinity.[3] Hybrid schemes for systems with covalent bonding require a more complicated description of the border region because any partitioning of the system into QM and MM parts will cut covalent bonds. Various strategies have been

* Corresponding author e-mail: nv@icct.ru (V.A.N.), roesch@ch.tum.de (N.R.).

[†] Russian Academy of Sciences.

[‡] University of Sofia.

[§] Technische Universität München.

10.1021/ct700159k CCC: $37.00
Published on Web 10/17/2007

proposed to handle the resulting “dangling” bonds of the QM cluster; they are either saturated by hydrogen-like, artificial “link” atoms located between two centers on either side of the QM/MM boundary[1,4] or by invoking a special description for centers at the border of the QM region that compensates for the missing covalently bonded partner.[5]

Covalently bonded materials with some polar interactions are among the most complex systems for such modeling because both types of complications just mentioned can be encountered in a QM/MM scheme. For materials of this type, e.g., zeolites, we recently proposed the hybrid density functional/molecular mechanics (DF/MM) embedded cluster method covEPE.[6,7] In a preceding study[6] we described features and intrinsic problems of different types of cluster embedding schemes developed for modeling of zeolites[1,2,4] and silica.[8] Recently, those methods were applied to various problems, e.g., structural defects[9] and catalytic reactions.[10] A related new methodological development is the extension of the QM-pot scheme,[11] based on an “energy-subtraction” scheme; it combines periodic DF calculations as a low-level method with MP2 calculations of an embedded cluster model as a high-level method.

A key feature of our covEPE embedding scheme[6,7] is the representation of the oxygen centers at the border of the QM cluster by a specially parametrized pseudopotential O* which renders these species monovalent. To ensure consistency between the treatment of the QM and MM parts of the system, we developed a specific force field, parametrized on the basis of DF calculations. We found this approach suitable for modeling electronic and geometrical properties of isolated active sites of various zeolites (CHA, FAU, MFI), including pure silicalite as well as zeolites doped with Al and Ti atoms at framework positions.[6−12]

In that first covEPE parametrization we followed an earlier suggestion[6] and constructed the pseudopotentials for boundary centers with the rather small basis set (2s2p) for computational efficiency. As a consequence, one should avoid QM cluster models where this relatively poor representation of O* centers directly affects the description of the active center. Relatively close contacts with a larger adsorbate may induce an artificial interaction due to the fact that atoms in the QM region close to O* centers are usually described by very flexible basis sets, e.g., adsorbed transition metal species. As a result of the unbalanced description, one may encounter an artificial attraction between the adsorbate and border centers O*. Such artifacts are easily discernible via large counterpoise corrections when one estimates the basis set superposition error (BSSE) of the adsorption energy; below we will discuss this in more detail.

Unfortunately, the peculiar three-dimensional structure of zeolite frameworks renders it difficult to avoid such artificial interactions unless one opts for rather large QM models which, in turn, reduce the benefit of a hybrid QM/MM approach. In the present work, we opted for an alternative which relies on an improved, more flexible description of the pseudoatoms O*. In this augmented approach, the two-dimensional surface through the O* centers, which originally partitioned space into QM and MM regions, is extended to a formally three-dimensional boundary region which, besides the O* pseudoatoms of the QM cluster, also includes their immediate cationic neighbors in the MM region, Si^{PP}*, in the form of total ion model potentials (TIMPs). This choice formally combines the two previous variants of the EPE scheme;[3,6] it was motivated by the need to avoid (or, at least, to reduce strongly) the otherwise significant and unwanted polarization of O* centers. Hence, this improved covEPE method, to be presented here, is based on an *extended boundary region*.

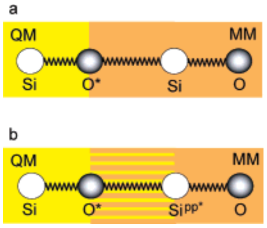


Figure 1. Structure of the border QM/MM region in the (a) previous and (b) the proposed novel extended termination scheme.

In the following, we will first discuss in detail the parametrization strategy for the new border centers O* and Si^{PP}*. Then we will evaluate the new variant of the covEPE embedding scheme by calculating (i) structural and spectral characteristics of two zeolite models, silicalite and Al-containing structures, based on a faujasite lattice as well as (ii) adsorption complexes of CO probe molecules and hexarhodium clusters in zeolite cavities.

2. The Novel covEPE Parametrization

2.1. Representation of the Border Region. A defining feature of the new covEPE parametrization is the representation of the border between the QM and MM partitions by pairs of atoms O*(QM)−Si^{PP}*(MM) that form a covalent bond of the original material (Figure 1). As in the original covEPE method, the charge of the whole QM/MM system is balanced by assigning incremental point charge Δq_{pp} to O* border pseudoatoms; these increments are half of the charge of oxygen centers in the MM region. As before, there are no dangling bonds at the QM boundary, because the O* centers are represented by adjusted pseudopotentials, carrying seven valence electrons, which renders these “pseudo”-oxygen centers monovalent. In the present implementation, we chose to describe O* centers with semilocal effective core potentials of the Stuttgart type:[13]

$$V(r) = -\frac{Q}{r} + \sum_{l} \sum_{k} A_{kl} \exp(-a_{kl} r^2) \sum_{m_l} |lm_l\rangle\langle lm_l| \quad (1)$$

Here, Q is the charge of the ionic core, while l and m_l are quantum numbers that designate eigenfunctions $|lm_l\rangle$ of the orbital angular momentum. The ten parameters of this pseudopotential were adjusted by starting with those of fluorine.[14] The previous implementation was based on a SBK pseudopotential[15] which (for fluorine) features only six adjustable parameters, hence it is less flexible. The valence orbital basis set of O* was derived from a (9s5p1d) all-electron basis set of F by removing the four largest *s* exponents.[16] The exponents of the resulting (5s5p1d) basis

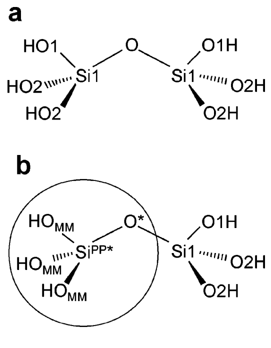


Figure 2. Sketch of systems used for fitting the pseudopotential parameters: (a) reference system calculated at the QM level and (b) target hybrid QM/MM system.

set were then adjusted as described below. This new basis set of O*, contracted to [4s3p1d], is considerably more flexible than the contracted [2s2p] basis set used previously with the SBK pseudopotential.[6] The new basis is of the same quality as the basis sets used for adsorbates and centers inside the QM region of the zeolite; therefore, counterpoise corrections of adsorption energies are significantly smaller (see sections 4.3 and 4.4).

However, neighboring bare positive point charges that represent Si cations of the MM region result in an artificial polarization of the new O* centers due to their flexible basis set. This problem is well-known from embedded cluster models of strongly ionic oxides.[3] To ensure an adequate polarization of O* centers, we followed the same strategy as in the EPE approach of ionic oxides:[3] we assigned a repulsive Si^{PP}* TIMPs (without any orbital basis set) to those cationic Si centers of the MM part that are located immediately at the QM/MM interface. Consistent with the O* pseudopotential, we chose to represent Si^{PP}* centers also by pseudopotentials of Stuttgart type. Parameter adjustment of Si^{PP}* centers was started with the values of sodium.[17] For Si^{PP}* centers at the QM/MM border we assumed the same effective charge, 1.2 e, as for their MM analogues of our silicate force field (FF) of shell-model type with potential derived charges (PDCs).[6,7]

2.2. Procedure of the Parametrization. We determined the parameters of O* and Si^{PP}* pseudoatoms in an iterative two-step procedure, similar to the strategy we used to establish the original covEPE scheme.[6] Some modifications were required because the new description of the QM/MM border is more sophisticated. In the first step of each iteration, we optimized the exponents of the O* basis set by minimizing the energy of an isolated pseudoatom O* that carries an incremental charge $\Delta q_{pp} = -0.3\ e$; that increment is derived from the potential derived charge, $-0.6\ e$, of oxygen centers in the MM region.[6] In a second step, the pseudopotential parameters of both types of border atoms, O* and Si^{PP}*, were adjusted to reproduce (i) selected electronic and structural characteristics of an isolated cluster which represents part of a zeolite framework and (ii) the electrostatic potential (ESP) produced by a periodic zeolite framework.

The reference data of type (i) had been produced by calculating the 2T model cluster $(HO)_3SiOSi(OH)_3$ at the QM level (Figure 2a). In the system to be trained, the O−Si(OH)$_3$ fragment of the reference cluster was replaced by O* and Si^{PP}* centers, and the remaining three OH groups were represented as point charges, $-0.7\ e$ for oxygen and $0.4\ e$ for hydrogen atoms (Figure 2b). These atomic charges for O and H reproduce best the electrostatic potential of the (finite) reference system. As before,[6] we derived the reference data of type (ii) with a periodic array of point charges located at crystallographic positions of a chabazite lattice which contains only Si atoms in tetrahedral positions (T-atoms). To mimic the electrostatic field of this zeolite, we used the same PDCs, 1.2 e for Si and $-0.6\ e$ for O, as reported earlier.[6] The system to be trained was constructed as the cluster (Si^{PP}*O*)$_3$SiOSi(O*Si^{PP}*)$_3$, embedded in a finite array of point charges which accurately mimics the periodic electrostatic potential of an extended chabazite environment.[3,6]

Using a least-squares approach and the simplex method,[18] we simultaneously optimized the pseudopotential parameters of O* and Si^{PP}*. The training set contained nine types of data, eight of which were derived from the 2T QM cluster: (i−iii) potential energy curves of the bonds Si1−O* and Si1−O1 and the bond angle O*−Si1−O1, each represented by a set of five points; (iv−vii) PDCs of the atoms Si1, O1, O2 and of the oxygen atom that was replaced by an O* pseudoatom; and (viii) the HOMO−LUMO gap.

Contribution (ix) to the least-squares sum was constructed from the ESP of the periodic array of point charges which was probed on a planar grid of 300 points near the center of the 8T ring of chabazite, covering an area of 1.5 Å × 2.0 Å.

After initial tests, the weighting factors of each squared deviation of data (i)−(iii) (in au) were set to 3, those of data (iv)−(viii) (in e and au, respectively) were kept at 1, while those of type (ix) (in au) were set to 0.025.

The potential energy curves calculated for the reference 2T cluster and the "trained" system with the final parameters for the O* and Si^{PP}* border atoms are provided as Supporting Information (Figure S1, Tables S1 and S2). The O*−Si1−O1 bending potential energy curves calculated for the pure QM reference cluster and the trained hybrid QM/MM system agree very well. The Si1−O* and Si1−O1 energy show some deviations, but the discrepancy never exceeds 0.6 kJ/mol. Thus, the results of the parametrization are certainly acceptable for these characteristics. Note that the previous covEPE parametrization scheme failed to reproduce the Si1−O* energy curve without a short-range FF correction term.[6]

While optimizing the parameters of the border centers O* and Si^{PP}*, we fixed the bond length between them in the system (HO)$_3$Si−O*−Si^{PP}* to be trained at the equilibrium distance of the reference cluster. If one used the optimized parameters for these centers and the FF parameters for the O*−Si^{PP}* interaction, then the potential energy curve of the O*−Si^{PP}* bond would fail to reproduce the reference curve; this is not really a surprise, given the purpose of these pseudopotentials. Therefore, we adjusted this interaction with a pair potential of Buckingham type, similarly to the Si(QM)−O* and O*−Si(MM) correction terms of the previous covEPE scheme.[6,19]

To assess the quality of hybrid QM/MM calculations, we optimized the structure of the (HO)$_3$Si−O*−Si^{PP}* moiety

Table 1. Characteristic Quantities Used in the Parameter Fitting of the Extended Border Scheme (New) and Corresponding Values Obtained with the Original, Simpler Scheme (Old)

		2T cluster [a]	new	Δ[b]	old	Δ[b]
PDC,[c] *e*	O*	−0.44	−0.49	−0.05	−0.38	0.06
	Si1	1.07	1.09	0.02	1.04	−0.03
	O1	−0.74	−0.72	0.02	−0.73	0.01
	O2	−0.67	−0.67	0.00	−0.68	−0.01
HOMO−LUMO, eV		6.313	6.286	−0.027	6.368	0.055
O*−Si^{PP}*, Å		1.6350	1.6351	0.0001		
O*−Si1, Å		1.6350	1.6382	0.0032		
Si1−O1, Å		1.6488	1.6523	0.0035		
O*−Si1−O1, deg		102.63	102.64	0.01		

[a] Values of the target system, see text. For the notation of the centers, see Figure 2a. [b] Deviations from the target values. [c] Potential derived charge.

and in addition compared the resulting values of the PDCs and the HOMO−LUMO gap of the trained system to those of reference 2T cluster (Table 1). The Si1−O* and Si1−O1 bonds reproduce the corresponding values of the reference cluster within 0.004 Å; the O*−Si^{PP}* bond deviates only 0.0001 Å from the reference. The O1−Si1−O* angle of the QM/MM cluster with optimum parameters and the corresponding O−Si−O angle of the reference QM system agree to 0.01°. Differences in the PDC of the trained system and the reference system deviate at most 0.02 *e* for the QM part of the hybrid system and 0.05 *e* for the border O* center. The HOMO−LUMO gap, 6.313 eV, is very well reproduced.

In Table 1, we also compare the results for PDCs and HOMO−LUMO gap obtained with the old and the new descriptions of the covEPE border. For calculations with the old approach[6] we used the same hybrid 2T cluster (Figure 2b), but the Si^{PP}* was replaced by a Si(MM) center with a positive charge of 1.182 *e* and for O* the previous parameters were used (smaller basis set, old pseudopotential parameters, compensating charge $\Delta q_{pp} = -0.282\ e$). Both schemes reproduce yield results of similar quality. The most significant discrepancy occurs for the PDCs of the centers Si1 and O*. With the previous model, these atomic charges in the hybrid QM/MM cluster are underestimated (in absolute values) with respect to the reference cluster, while the new extended QM/MM border scheme results in a slight overestimation, closer to values used in the silica FF to model the MM environment (Si: 1.2 *e*, O: −0.6 *e*). Therefore, with the new border region, one can expect a more consistent description of the ESP both in the QM and the MM regions.

In Figure 3, we present (a) a map of the ESP calculated for a periodic array of MM point charges of chabazite structure and deviations from that target quantity for the 2T QM cluster embedded in a corresponding infinite environment of point charges using old (b) and new (c) covEPE embedding. Obviously, the new covEPE parametrization allows one to reproduce the ESP of the periodic PC array more closely than the old parametrization—not only in the center of the chabazite 8T ring at about 5 Å from the QM centers but also close to the QM oxygen centers. The largest deviations occur close to the border pseudoatoms O*. Previously, in that region the ESP was 0.4 eV (or 25.0%) more positive than the ESP of the MM reference system, −1.6 eV. With the new covEPE parametrization, the ESP near O* centers is just 0.2 eV (or 12.5%) more positive (Figure 3c).

Finally, we turn to the set of FF parameters for the interaction of border atoms with atoms of the MM environment. With the previous description,[6] the characterization of O* centers (valence orbital basis set, effective core pseudopotential with a large positive point charge of 8.7 *e*) differed notably from that of MM oxygen centers (just a pair of point charges, 2.387 *e* and −2.987 *e*). To compensate for that difference and to correctly place the QM part within its MM framework, we originally had used a special pair-potential O*−Si(MM).[6] With the new scheme on the basis of an extended QM/MM border region, Si^{PP}* TIMP centers, substituting Si^{MM} of the old scheme, form an external coordination shell of the QM partition. From the MM side, these Si^{PP}* centers are treated on the same footing as the other Si centers of the MM part, e.g., carrying a charge of 1.2 *e*. Therefore, it is natural to describe the short-range interactions between these Si^{PP}* border centers and other atomic centers of the MM region with the standard parameters of the aluminosilicate FF that we had specifically derived to model the MM environment in the covEPE embedding approach.[7] Therefore, that FF was employed to describe Si^{PP}*−O(MM) pair interactions, all three-body interactions within the Si^{PP}*O_4 tetrahedron, including both O* and O(MM) centers as well as the three-body interactions Si^{PP}*−O(MM)−Si(MM) and Si^{PP}*−O(MM)−Si^{PP}*. Hence, in that regard, the new parametrization affords a notably simpler and more consistent description of the interactions of atomic centers at the boundary with the MM partition of the system.

3. Computational Details

The embedded cluster calculations were carried out with the covEPE scheme[6] as implemented in the parallel density functional program ParaGauss.[20,21] For the QM calculations, we employed the linear combination of Gaussian-type orbitals fitting-functions density functional method (LCGTO-FF-DF).[22] We used the gradient-corrected exchange-correlation functional suggested by Becke (exchange) and Perdew (correlation);[23] all calculations were performed in spin-restricted fashion. The Kohn−Sham orbitals were represented with the following Gaussian-type basis sets,

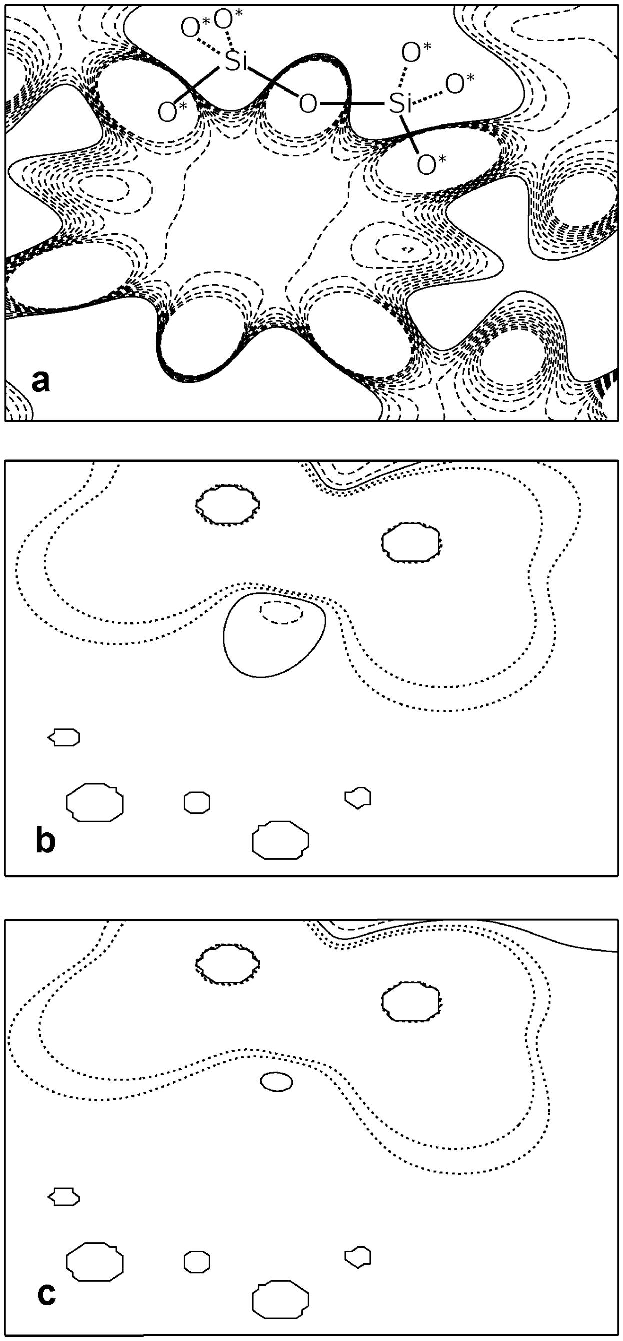


Figure 3. Electrostatic potential map calculated for a periodic array of potential-derived point charges (a) and differences between periodic ESP and ESPs calculated for a 2T QM cluster embedded in an array of point charges with (b) the old and (c) the new scheme for treating the border region. Solid, dashed, and dotted contours correspond to zero, negative, and positive ESP values, respectively; the contours represent equidistant values with an increment of 0.2 eV.

contracted in generalized fashion: (6s1p)→[4s1p] for H,[16,24] (9s5p1d)→[5s4p1d] for C and O,[16,24] and (12s9p2d)→[6s4p2d] for Al and Si.[16,24] The polarization d-exponents 0.50 and 2.05 for Si and 0.2881 and 1.0084 for Al atoms were taken from ref 25. For Rh the (19s15p10d)→[8s6p4d] basis set was constructed by adding to the (17s12p8d) basis set of Gropen[26] two s- (0.01303, 0.2253), three p- (0.03666, 0.09165, 0.2291), and two d-type exponents (0.04588, 0.1147). In the LCGTO-DF-FF method, the Hartree contribution of the electron−electron interaction is approximated by representing the electronic charge density with the help of an auxiliary Gaussian-type basis set.[22] The corresponding exponents were constructed by scaling the exponents of the orbital basis; in addition, "polarization" exponents, five each of p- and d-type, were added on each center, constructed as geometric series with a factor 2.5 starting at 0.1 au (p-exponents) or 0.2 au (d-exponents). For H centers, only p-type polarization exponents were added.

As done previously,[27,28] we applied C_3 symmetry restrictions when modeling faujasite-supported Rh_6 species to reduce the computational effort of the QM calculations (section 4.3). All other QM/MM calculations were carried out without any symmetry restrictions.

The MM part of all systems was described with the help of a FF of shell-model type[29] which we had developed for modeling silica and protonated aluminosilicates.[6,7] This force field is based on PDCs; hence, it is particularly suited for reproducing the ESP and the polarization of silica minerals and zeolite lattices.

The force constants for analyzing vibrational frequencies were calculated numerically, using finite differences of analytical energy gradients. To estimate the OH frequencies in adsorbate-free systems, we invoked the approximation of an independent harmonic oscillator, i.e., only the O−H internal coordinate was varied during the frequency calculation. Nine degrees of freedom—three for each of the centers H, C, and O (in a CO probe)—were varied during the calculations of OH and CO frequencies in the adsorption complexes of CO on zeolite OH groups.

Adsorbate−substrate binding energies, E_{ads}, were corrected for the basis set superposition error (BSSE) by applying the counterpoise method[30] in single-point fashion at the equilibrium geometry of the adsorption complexes.

4. Evaluation of the Embedding Scheme with an Extended Border Region

The quality of the new scheme for constructing an extended QM/MM border region and the corresponding parametrization of the border centers O* and Si^{PP}* were validated by calculations of structural characteristics of silicalite and alumosilicate frameworks of faujasite (FAU). In addition, we studied the OH frequencies of bridging hydroxyl groups of zeolites. Finally, we applied the new covEPE scheme to two adsorption complexes: CO probe molecules at bridging OH groups in FAU and MFI zeolites and Rh_6 clusters in the cavity of a FAU zeolite.

4.1. Silicalite of Faujasite Structure. As first check of the new parametrization of O* and Si^{PP}* centers (pseudopotentials, basis sets, correction term in the FF), we optimized the structures of QM clusters embedded in a faujasite framework containing only Si as T atoms (silicalite). The accuracy of the new scheme was evaluated from bond lengths and angles inside the embedded QM clusters and at the QM/MM border. For this comparison we used a series of five embedded QM clusters of increasing size, with two (2T) to six silicon centers (6T, Figure 4). Each QM cluster

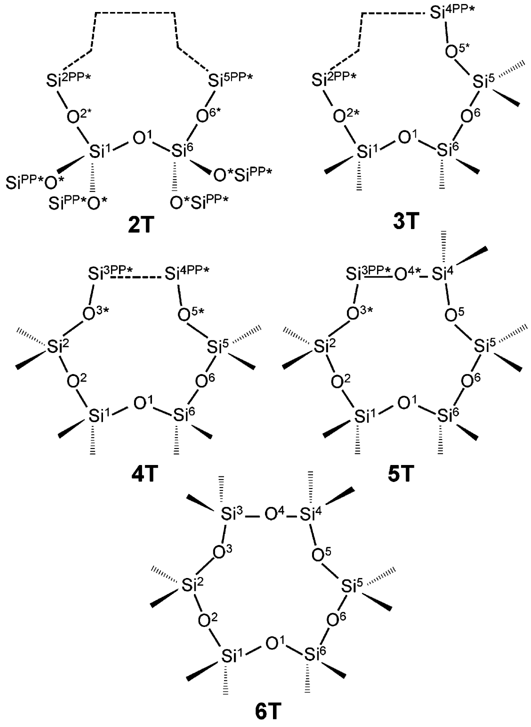


Figure 4. Pure silica QM clusters representing parts of a faujasite six-ring. The parts of the ring not included in the QM cluster are shown with dashed lines.

includes its smaller predecessors; this allows us to trace changes of structural parameters of the six-ring upon expansion of the region described at the QM level.

Inspection of Table 2 shows that structural parameters inside the QM region hardly deviate among the various models. Individual Si−O bond lengths are stable within 0.004 Å, O−Si−O angles within 1°, and Si−O−Si angles within 2°. With increasing cluster size, these structural parameters converge to the values of the largest 6T model in which a whole FAU six-ring is treated at QM level. Our calculations on the 6T model suggest that the crystallographically different oxygen centers O2 and O4 exhibit different properties in an all-silica FAU framework (Table 2, Figure 4): (i) for O2 centers, Si−O distances are 1.636 ± 0.002 Å and Si−O−Si angles are 150 ± 1° and (ii) for O4 centers, Si−O distances are 1.643 ± 0.001 Å and Si−O−Si angles are 136 ± 1°.

As our calculations do not impose structural restrictions, these differences should not be artifacts of the computational method but reflect features of the faujasite structure.

We observed slightly larger deviations, up to 0.01 Å for Si−O bonds and 4° for Si−O−Si angles, when comparing the structure of identical fragments of the FAU six-ring in different models 2T to 6T, that are treated at different computational levels, i.e., QM vs MM. Such larger deviations concern bonds and angles involving the O4 centers O^2, O^4, and O^6 which are oriented outside the six-ring. As just noted, Si−O distances of such centers are 1.643 ± 0.001 Å in the 6T model, while the same bonds described as Si(QM)−O* and O*−Si^{PP}* interactions at the border of the clusters 2T to 5T are parametrized to reproduce distances of 1.635 Å, very close to the Si−O bond lengths for O^1, O^3, and O^5 atoms in the six-ring of the 6T model. Tetrahedral O−Si−O* and O−Si−O bond angles deviate within 2° only.

The observed fluctuations of the structural parameters within the series of QM clusters, as obtained with the new covEPE description featuring an extended border region, are similar to those determined from our previous QM/MM scheme with a minimum basis set on O* border centers; calculated structure parameters, available only for 4T and 5T models,[6] agree within 0.01 Å and 3°.

4.2. Aluminosilicates. As a second test of the new method and the parametrization, we considered an aluminosilicate framework of the FAU structure with a Si/Al ratio of 47, a model system for which results are available with the previous border scheme.[7] We used 5T and 8T QM clusters with an OH group located at the O1 crystallographic position[31] and determined not only the structure of the embedded zeolite clusters but also OH frequencies and deprotonation energies (DE) of the bridging hydroxyl groups Al−O(H)−Si.

The average Al−Si distance in the 8T cluster embedded in the FAU lattice, 3.18 Å, optimized with the new scheme, is 0.03 Å shorter than the corresponding distance in the model optimized with the previous border scheme. In the same cluster the novel approach also yields shorter Al−O_b (by 0.04 Å) and Si−O_b bonds (by 0.015 Å), which agree better with recent EXAFS results.[32] In particular, with the new approach the distance Al−O_b, 1.915 Å, is quite close to corresponding experimental values, 1.89 ± 0.025 Å[32a] and 1.87 ± 0.01 Å;[32b] with the previous covEPE scheme, that distance was calculated notably longer, 1.954 Å.[7]

The observed structural changes around Al−O(H)−Si sites to some extent affect other properties of OH groups (Table 3). The OH frequency of the 8T FAU model, calculated with the extended border scheme, is 6 cm^{-1} lower than in the old approach; this result is in line with the elongation of the O_b−H bond by 0.001 Å in the new model. The deprotonation energy of faujasite in the novel scheme is reduced by 59 and 38 kJ/mol for the 5T and 8T QM clusters, respectively. The larger DE values of bridging OH groups, estimated in the previous approach, can be rationalized by a stabilization of the (neutral) initial-state structure which is caused by an artificial saturation of the small basis set of border O* centers with H basis functions of the hydroxyl group and the more positive ESP values around the border centers (see above).[7] The first factor is more pronounced in the smaller 5T QM model where the closest distance H−O* is 2.62 Å. The corresponding distance in the structure optimized with the novel border scheme is 0.11 Å longer, whereas the distance H−O_{Al} between the proton and the basic oxygen centers O_{Al} bound to Al is calculated 0.1−0.2 Å shorter (Table 3).

In summary, the novel scheme for representing the border region adequately describes the local structure of zeolites and the properties of acidic Al−O(H)−Si sites, notably improved compared to the previous border scheme of the covEPE approach.[6,7,12]

4.3. Adsorption of CO. To check the new QM/MM border scheme for interactions of guest species with a zeolite framework, we modeled the adsorption of carbon monoxide on Brønsted acidic Al−O(H)−Si centers of zeolites with

Table 2. Calculated Structural Parameters[a] of Various Silica QM Clusters, Embedded in a Faujasite Lattice, and Their Border: Bond Lengths (Å) and Bond Angles (deg)

	Si^3−O^3	O^3−Si^2	Si^2−O^2	O^2−Si^1	Si^1−O^1	O^1−Si^6	Si^6−O^6	O^6−Si^5	Si^5−O^5	O^5−Si^4	Si^4−O^4	O^4−Si^3
MM[b]	1.632	1.632	1.626	1.626	1.632	1.632	1.626	1.626	1.632	1.632	1.626	1.626
2T			1.634[d]	1.634[c]	1.639	1.640	1.636[c]	1.635[d]				
3T			1.634[d]	1.638[c]	1.638	1.638	1.644	1.644	1.635[c]	1.631[d]		
4T	1.631[d]	1.634[c]	1.643	1.643	1.636	1.637	1.645	1.644	1.634[c]	1.631[d]		
5T	1.632[d]	1.637[c]	1.647	1.642	1.636	1.637	1.644	1.642	1.637	1.636	1.637[c]	1.633[d]
6T	1.634	1.636	1.643	1.642	1.635	1.636	1.644	1.642	1.637	1.636	1.643	1.643
Δ[e]	0.003	0.002	0.009	0.008	0.004	0.004	0.008	0.007	0.003	0.005	0.006	0.010

	Si^3−O^3−Si^2	Si^2−O^2−Si^1	Si^1−O^1−Si^6	Si^6−O^6−Si^5	Si^5−O^5−Si^4	Si^4−O^4−Si^3
MM[b]	146	148	146	148	146	148
2T		140[c]	147	139[c]		
3T		139[c]	148	135	152[c]	
4T	151[c]	135	148	135	153[c]	142[f]
5T	151[c]	135	149	135	149	140[c]
6T	150	136	149	135	150	136
Δ[e]	1	4	2	4	3	4

	O^4−Si^3−O^3	O^3−Si^2−O^2	O^2−Si^1−O^1	O^1−Si^6−O^6	O^6−Si^5−O^5	O^5−Si^4−O^4
MM[b]	107	107	107	107	107	107
2T		108[d]	107	107	108[d]	
3T	108[d]	108[c]	107	108	110	110[d]
4T	109[d]	109	108	108	110	110[d]
5T	109[c]	109	108	108	109	108[c]
6T	109	108	108	108	109	109
Δ[e]	0	1	1	1	2	2

[a] For the notations of the models and various centers, see Figure 4. [b] Values of an infinite FAU lattice, described by the force field. [c] Si(QM)−O* and Si(QM)−O*−Si^{PP*} at the QM/MM border. [d] O*−Si^{PP*} and O*−Si^{PP*}−O(MM) at the QM/MM border. [e] Maximum deviation of QM results from the corresponding values of the largest QM model, 6T. [f] Si^{PP*}−O(MM)−Si^{PP*} at the QM/MM border.

Table 3. Selected Structural Parameters (in Å and deg), Harmonic OH Frequencies ν(OH) (in cm^{-1}), and Deprotonation Energies DE (in kJ/mol) for Acidic Al−O$_b$(H)−Si Sites from 5T and 8T QM Models Embedded in a Faujasite Lattice (Si/Al = 47) with Bridging Oxygen Centers O$_b$ Located at O1 Crystallographic Positions

	previous border scheme[a]			extended border scheme[b]		
QM cluster	5T	8T	Δ[c]	5T	8T	Δ[c]
Al−O$_b$	1.958	1.954	−0.004	1.929	1.915	−0.014
<Al−O>[d]	1.721	1.722		1.719	1.720	
Si−O$_b$	1.718	1.722	0.004	1.707	1.707	0.000
<Al−Si>[d]	3.21	3.21		3.18	3.18	
O$_b$−H	0.976	0.978	0.002	0.979	0.979	0.000
H−O*[e]	2.618	4.569		2.729	4.591	
H−O$_{Al}$[f]	2.80	2.70		2.59	2.61	
Al−O$_b$−Si	128.7	131.3	2.6	128.7	129.9	1.2
H−O$_b$−Al	114.0	111.2	−2.8	112.2	112.8	0.6
H−O$_b$−Si	117.1	115.8	−1.3	119.1	117.2	−1.9
ν(OH)	3754	3720	−34	3683	3714	31
DE	1285	1270	−15	1226	1232	6

[a] Reference 7. [b] Present work. [c] Changes in the results of the 8T QM model with respect to the results of the 5T QM model. [d] Average value. [e] Distance between the acidic H and the nearest pseudoatom O*. [f] Distance between the acidic H and the oxygen connected to the Al atom.

FAU and MFI structures as this molecule is often used as a probe for the acidity of zeolite OH groups.[33] This interaction is characterized with a relatively small adsorption energy,[34−37] and thus, minor inaccuracies in the description of the boundary region can be crucial, in particular for small embedded QM models. For FAU we modeled the interaction of CO with hydroxyl groups at O1 crystallographic positions, using a 8T QM cluster (Figure 5a). For HZSM-5 zeolite, the probe was assumed to adsorb at the Al7−O17(H)−Si4 site, modeled by a 9T QM cluster (Figure 5b).

In Table 4, we have collected calculated structural parameters, the adsorption energy of CO, E_{ads}, and shifts of the vibrational frequencies C−O and O−H, obtained with the original and the extended border schemes.

The adsorption energy of CO, corrected for the BSSE, determined with the novel scheme of the border region, is 18.6 kJ/mol, both in faujasite and HZSM-5. This value is close to the BSSE corrected adsorption energy of CO in chabazite, 16.0 kJ/mol, obtained in periodic B3LYP calculations.[37] With the previous border scheme, where a small basis set was used for O* centers, the uncorrected E_{ads} values were 28.8 kJ/mol for FAU and 44.8 kJ/mol for MFI, but only 11−12 kJ/mol remained after correction for the BSSE. Thus, 60−75% of these values were caused by basis functions of the adsorbate saturating the small O* basis set. With the novel border scheme, the fraction of the BSSE is considerably reduced, to 25%.

The geometry optimized with the novel approach suggests that CO molecules are almost linearly orientated relative to the hydroxyl group; the H−C−O angle is calculated at 173° in FAU and 176° in MFI. These values agree well with the

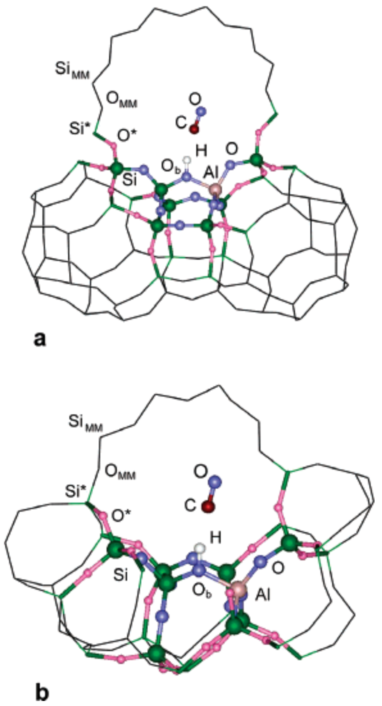


Figure 5. Optimized structures of models of CO adsorption on zeolite QM clusters embedded in lattices of (a) FAU and (b) MFI structure.

corresponding result, 176°, determined in a periodic B3LYP supercell calculation for CO adsorption in chabazite.[37] Our calculated distances, OC···H 1.90 Å for FAU, and 1.94 Å for MFI, are also close to the B3LYP result, 1.95 Å.[37] CO adsorption on acidic sites is accompanied by an elongation of the O−H bond, by 0.025 Å for FAU and 0.022 Å for MFI, while the C−O bond is contracted by 0.004 Å in both zeolite structures. These changes in bond lengths agree with calculated (Table 4) and experimentally observed[33] trends for the changes of the vibrational frequency shifts, a strong red shift of the OH band and a smaller blue shift of the CO frequency.

Our prediction of an elongated O−H bond and a related reduction of the O−H frequency upon adsorption of a CO probe on Al−O(H)−Si sites is in line with results of experimental[33] and previous theoretical studies.[34−37] However, the red shifts of the OH frequency, calculated at 425 cm^{-1} for MFI and 500 cm^{-1} for FAU (Table 4), are larger than the measured bathochromic shifts of the OH band, 307−320 cm^{-1} for MFI and 295−353 cm^{-1} for FAU.[33] This underestimation of the OH vibrational frequency of adsorption complexes of CO on OH groups can be considered as a feature of the density functional method used. For example, a large OH red shift of 375 cm^{-1} was recently reported from periodic B3LYP calculations on H-chabazite (Si/Al = 11).[37,38]

IR measurements predict an increase of the CO frequency by 32 ± 2 cm^{-1} after adsorption of CO probes at zeolite Brønsted sites, both for HZSM-5 and FAU.[33,37] With the extended border scheme, we calculated C−O red shifts of 37 cm^{-1} and 42 cm^{-1}, for CO adsorption on acidic OH group of zeolites with FAU and MFI structure, respectively, in good agreement with the IR data. However, the previous scheme for the QM/MM border region was less successful in describing CO adsorption at the OH groups; the interaction of CO molecules with the nearest border O* center was overestimated. This effect is particularly important for zeolites with narrower pores, e.g., MFI with 10-member ring of 5.5 Å diameter, where C and O centers of the adsorbate are located only 2.49 and 2.88 Å, respectively, from the nearest O* center (Table 4). As a consequence of this artificial interaction, adsorption complexes of CO species were notably bent, with an H−C−O angle of only 151°. A further consequence is an elongation of the C−O distance in adsorbed CO, by 0.002 Å compared to a free CO molecule, and a corresponding *reduction* of the C−O vibrational frequency upon adsorption (Table 4), i.e. the CO frequency was calculated to shift in the direction opposite to experiment. In the FAU model, adsorption complexes of CO at OH groups are less affected by border O* centers, because the cavity of this structure is wider (12-member ring, 12 Å diameter) and the active site is farther from the border O* centers. However, even in the FAU model, the C···O* and O···O* distances were calculated about 0.6 Å shorter with the previous approach than with the present extended border scheme.

Table 4. Selected Structural Parameters (in Å and deg), Frequency Shifts Δν(OH) and Δν(CO) (in cm^{-1}) of the OH and CO Vibrational Modes, Respectively, Adsorption Energy E_{ads} of CO Corrected for the Basis Set Superposition Error (BSSE) and BSSE Correction (in kJ/mol) of CO Complexes on Zeolites of FAU and MFI Structure[a]

	previous border scheme		extended border scheme	
	FAU	MFI	FAU	MFI
O$_b$−H	1.001	0.999	1.004	0.998
C−O	1.143	1.147	1.141	1.141
OC···H	1.919	1.970	1.896	1.941
O$_b$−H−C	158.7	167.3	168.3	170.9
H−C−O	174.1	150.9	172.9	175.7
C···O*	3.29	2.49	3.82	3.15
O···O*	2.92	2.88	3.65	3.14
C···Si(MM)	4.51	3.31	4.99	3.91
C···O(MM)	4.58	3.13	4.87	3.34
O···Si(MM)	3.82	3.01	4.52	3.43
O···O(MM)	3.60	2.29	4.07	2.46
Δr(O−H)[b]	0.023	0.022	0.025	0.022
Δν (OH)[c]	−459	−335	−502	−425
Δr(C−O)[d]	−0.002	+0.002	−0.004	−0.004
Δν(CO)[e]	31	−13	37	+42
E_{ads}	11.6	11.1	18.6	18.6
BSSE	17.2	33.7	5.9	6.7

[a] For the structure of the complexes and the notation of the centers, see Figure 5. [b] Change of the O−H distance with respect to an adsorbate-free zeolite. [c] Frequency shift with respect to ν(OH) in an adsorbate-free zeolite. [d] Change of the C−O distance with respect to a free CO molecule. [e] Frequency shift with respect to ν(CO) of a free CO molecule, 2091 cm^{-1}.

Thus, with the improved scheme for constructing the boundary region, one is able to predict structural, spectroscopic, and energetic characteristics of CO adsorption at acidic OH groups of zeolites in good agreement with

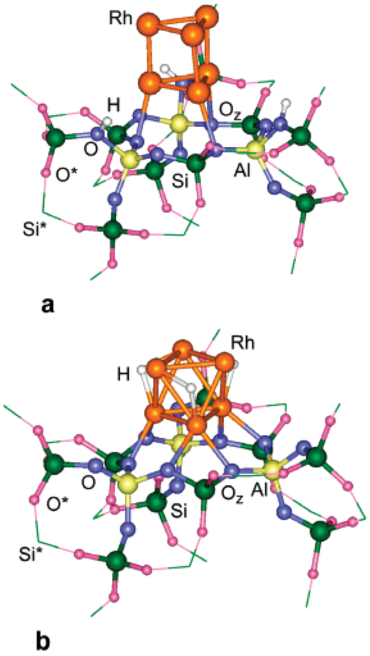


Figure 6. Optimized structures of the 12T QM embedded models representing adsorption complexes of (a) bare Rh_6 and (b) hydrogen-containing Rh_6H_3 clusters on a faujasite six-ring.

Table 5. Selected Interatomic Distances (in Å) in Faujasite-Supported Bare Rh_6 and Hydrogen-Covered Rh_6H_3 Clusters,[a] Energies E_{ads} of Adsorption of the Cluster Rh_6 on the Zeolite Support, Corrected for the Basis Set Superposition Error (BSSE), BSSE Values, and Energies E_{RS} of Reverse Proton Spillover onto the Supported Metal Clusters (in kJ/mol)

	previous border scheme		extended border scheme	
	Rh_6/zeo(3H)	Rh_6(3H)/zeo	Rh_6/zeo(3H)	Rh_6(3H)/zeo
Rh−Rh[b]	2.45−2.57	2.60−2.64	2.47−2.51	2.57−2.65
Rh−H		1.72; 1.77		1.72; 1.78
Rh−O_z[c]	2.29	2.14; 2.18	2.49	2.16; 2.21
Rh−O*[d]	3.01	3.34	3.28	3.35
E_{ads}	31		64	
BSSE	410		39	
E_{RS}	237		240	

[a] See Figure 6. [b] Experimental value 2.67−2.69 Å, ref 40. [c] Distance between nearest neighbors Rh and oxygen centers O_z of the zeolite support; experimental values: 2.10−2.17 Å, ref 40. [d] Distance between nearest neighbors Rh and O* centers.

available experimental and calculated data, thus avoiding artifacts that occurred with the previous, simpler QM/MM border approach.

4.4. Adsorption of Rh_6 Metal Cluster. Finally, we turn to an important benchmark system where bulky metal particles are adsorbed in zeolite cavities. Previously, we described Rh_6 in FAU with isolated cluster models of the zeolite support[27,39] to clarify earlier EXAFS studies.[40] With finite six-ring models of zeolite, we studied two forms of supported Rh_6 clusters: bare Rh_6, denoted as Rh_6/zeo(3H), and hydrogen-covered Rh_6H_3, denoted as Rh_6(3H)/zeo. The latter species, formally obtained by transfer of three protons from bridging OH groups of the zeolite to the metal cluster ("reverse hydrogen spillover"), were calculated to be preferred for a large variety of late transition metals, including Rh.[28] In the present context, we modeled both species (Figure 6) with either of the two border schemes (Table 5). For the Rh_6/zeo(3H) structure (Figure 6a) we considered bridging OH groups at O1 crystallographic positions of hexagonal prisms of the FAU structure, i.e., close to zeolite six-rings (Figure 6a).

With the new, extended border scheme, the BSSE-corrected adsorption energy, E_{ads}, of Rh_6 in Rh_6/zeo(3H) was calculated at 64 kJ/mol (Table 5). This value is by only 9 kJ/mol lower than $E_{ads} = 73$ kJ/mol, calculated with the finite model.[27] With the previous QM/MM border scheme, the BSSE-corrected value, $E_{ads} = 31$ kJ/mol, was underestimated as a consequence of a large BSSE, 399 kJ/mol. The main contribution to BSSE originates from saturating the basis set of the zeolite fragment (in particular, of O* centers with a small basis set) by the basis set of the adsorbate. With the extended border scheme, the BSSE is very significantly reduced, to only 39 kJ/mol, but the main contribution (79%) again is due to complementing the zeolite basis set.

In contrast to the strong influence on the adsorption energy of the Rh_6 cluster in Rh_6/zeo(3H), the border scheme has essentially no effect on the energy, E_{RS}, of reverse hydrogen spillover. It was calculated at 240 kJ/mol in the new scheme and 237 kJ/mol in the previous border scheme (Table 5). This can be rationalized by the fact that in this case one compares two structural isomers, Rh_6/zeo(3H) and Rh_6(3H)/zeo. Therefore, the influence of the BSSE on E_{RS} is, to a very large extent, eliminated as we are using the same basis sets when calculating the (formal) initial and final states of reverse spillover. With respect to isolated cluster calculations,[39] the energy E_{RS} is reduced by about 35%.

The calculations with both border schemes suggest that the bare adsorbed Rh_6 cluster is farther from its support and concomitantly features shorter Rh−Rh nearest-neighbor distances than the hydrogenated cluster Rh_6H_3, similarly to the results from the isolated cluster models.[27,39] In the previous border scheme, the artificial attraction between the adsorbate and the O* centers reduces the Rh−O* distance in the structure Rh_6/zeo(3H) by 0.27 Å compared to the structure optimized with the new approach. In addition, the Rh−O_z distance between the rhodium atoms at the "bottom" triangle of the Rh_6 species and the oxygen atoms of the zeolite six-ring increases notably, from 2.29 Å to 2.49 Å, when one switches from the previous to the new border schemes. The metal−metal distances in the Rh_6/zeo(3H) structure change much less between the two embedding schemes, at most 0.06 Å (Table 5).

Likely, due to the large Rh−O* separation, 3.35 Å, in the hydrogenated Rh_6(3H)/zeo complex, this structure remains essentially unaffected by the improved description of the border region. The distances Rh−O_z, Rh−Rh, and Rh−H of the two models agree within 0.03 Å, 0.03 Å, and 0.01 Å, respectively.

In a recent paper, we discussed in detail interatomic distances in the structures of the supported clusters, optimized with the novel border scheme.[41] There, we showed that available EXAFS data[40] for the metal−metal distances of Rh_6 species in Y zeolites, 2.67−2.69 Å, and distances between metal and oxygen centers of the support, 2.10−2.17 Å, are consistent with Rh−Rh (2.57−2.65 Å) and Rh−O_z distances (2.14−2.21 Å), calculated for the hydrogenated Rh_6(3H)/zeo model, while the corresponding values for the bare cluster Rh_6/zeo(3H) differ by about 0.2 Å. The same conclusion was drawn from results obtained with finite models of the zeolite support.[27,41]

5. Conclusions

The present work reported an improved scheme for constructing the border region within a hybrid embedded cluster approach covEPE for zeolites and covalent oxides. The new scheme assures proper modeling of adsorbates interacting with such types of support.[6,7] At variance with the original implementation of the covEPE method, where monovalent O* pseudoatoms at the QM border were described with a small basis set, the basis set on those centers in the present, an improved scheme is as flexible as that normally used for O centers of the QM cluster. These new O* centers are much more polarizable due to their large basis set. To avoid polarization artifacts, we extended the border region by a second "layer" of border centers, at Si centers of the MM region that are immediate neighbors of the QM O* centers. Both, O* and Si^{PP}* centers were modeled as pseudopotentials. The parameters of both types of pseudopotentials and the basis set of O* centers were optimized by targeting structural and electronic properties of model zeolite fragments.

The resulting improved hybrid QM/MM scheme affords a correct description of the local structure of silica and aluminosilicate zeolites and of the properties of bridging OH groups. These properties are reproduced at notably improved quality compared to results of covEPE models that were constructed with the previous border scheme.[7]

With the novel border scheme, we achieved good agreement with available experimental and reliable computational data for the adsorption of CO probe molecule on bridging OH groups of zeolites with MFI and FAU structures. Changes induced upon CO adsorption in the structure of zeolite acidic OH sites, the CO vibrational frequency shift, and the adsorption energy of CO, calculated with an 8T embedded QM cluster, are considerably improved as compared to analogous results obtained with the previous border scheme using a small basis set on the capping O* centers. The good performance of the new approach is accompanied with a substantial reduction of the BSSE, which in the previous construction of the border region originated from an implicit augmentation of the small basis set of the O* centers by basis functions of the adsorbates.

Models with the new border scheme are also well suited to describing large adsorbates, e.g., transition-metal clusters, with satisfactory accuracy. Here, too, the new scheme exhibits greatly reduced BSSE corrections compared to the previous method for constructing border O* centers. Recently, we successfully applied the new border scheme to calculate the energetics of reverse hydrogen spillover from zeolite hydroxyl groups to supported Ir_6 clusters and the lack of such spillover in the case of Au_6 clusters.[41]

Acknowledgment. E.A.I.S. gratefully acknowledges an individual grant of the Krasnoyarsk Regional Scientific Foundation (grant 15G127). This work was supported by Volkswagen-Stiftung (grant I/73653), Deutsche Forschungsgemeinschaft, Fonds der Chemischen Industrie, Bulgarian National Science Fund, and an integration project (no. 79) of the Siberian Branch of the Russian Academy of Sciences.

Supporting Information Available: Tables with the optimized parameters of the border centers and a figure with potential energy curves calculated for the reference 2T cluster and the trained system on the basis of the optimized parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrich, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struct. Theochem* **2003**, *632*, 1−28. (b) Sauer, J.; Sierka, M. *J. Comput. Chem.* **2000**, *21*, 1470−1493.

(2) (a) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*, *NIC Series Volume 1*; J. von Neumann Institute for Computing: Julich, 2000; pp 301−449. (b) Sokol, A. A.; Bromley, S. T.; French, S. A.; Catlow, C. R. A.; Sherwood, P. *Int. J. Quantum Chem.* **2004**, *99*, 695−712.

(3) (a) Nasluzov, V. A.; Rivanenkov, V. V.; Gordienko, A. B.; Neyman, K. M.; Birkenheuer, U.; Rösch, N. *J. Chem. Phys.* **2001**, *115*, 8157−8171. (b) Rösch, N.; Nasluzov, V. A.; Neyman, K. M.; Pacchioni, G.; Vayssilov, G. N. In *Computational Material Science*; Theoretical and Computational Chemistry Series, Leszczynski, J., Ed.; Elsevier: Amsterdam, 2004; Vol. 15, pp 365−448.

(4) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419−1432.

(5) (a) Gao, J. L.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714−4721. (b) Antes, I.; Thiel, W. *J. Phys. Chem. A* **1999**, *103*, 9290−9295.

(6) Nasluzov, V. A.; Ivanova, E. A.; Shor, A. M.; Vayssilov, G. N.; Birkenheuer, U.; Rösch, N. *J. Phys. Chem. B* **2003**, *107*, 2228−2241.

(7) Ivanova Shor, E. A.; Shor, A. M.; Nasluzov, V. A.; Vayssilov, G. N.; Rösch, N. *J. Chem. Theory Comput.* **2005**, *1*, 459−471.

(8) Sulimov, V. B.; Sushko, P. V.; Edwards, A. H.; Shluger, A. L.; Stoneham, A. M. *Phys. Rev. B* **2002**, *66*, 024108.

(9) Mukhopadhyay, S.; Sushko, P. V.; Stoneham, A. M.; Shluger, A. L. *Phys. Rev. B* **2005**, *71*, 235204.

(10) (a) Joshi, Y. V.; Thomson, K. T. *J. Catal.* **2005**, *230*, 440−463. (b) To, J.; Sherwood, P. Sokol, A. A.; Bush, I. J.; Catlow, C. R. A.; van Dam, H. J. J.; French, S. A.; Guest, M. F. *J. Mater. Chem.* **2006**, *16*, 1919−1926.

(11) Tuma, C.; Sauer, J. *Chem. Phys. Lett.* **2004**, *387*, 388−394.

(12) Deka, R. C.; Ivanova Shor, E. A.; Shor, A. M.; Nasluzov, V. A.; Vayssilov, G. N.; Rösch, N. *J. Phys. Chem. B* **2005**, *109*, 24304−24310.

(13) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866−872.

(14) Bergner, A.; Dolg, M.; Küchle, W.; Stoll, H.; Preuss, H. *Mol. Phys.* **1993**, *80*, 1431−1441.

(15) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026−6033.

(16) Van Duijnevelt, F. B. *IBM Research report No. RJ*; 1971; p 945.

(17) Fuentealba, P.; Preuss, H.; Stoll, H.; von Szentpály, L. *Chem. Phys. Lett.* **1982**, *89*, 418−422.

(18) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308−313.

(19) The parameters for the correction Buckingham type pair-potential for the O*−Si^{PP}* interaction are 31156.5 eV (A), 0.199145 Å (ρ), and 237.804 eV·Å^6 (C).

(20) Belling, T.; Grauschopf, T.; Krüger, S.; Nörtemann, F.; Staufer, M.; Mayer, M.; Nasluzov, V. A.; Birkenheuer, U.; Hu, A.; Matveev, A. V.; Shor, A. M.; Fuchs-Rohr, M. S. K.; Neyman, K. M.; Ganyushin, D. I.; Kerdcharoen, T.; Woiterski, A.; Gordienko, A. B.; Majumder, S.; Rösch, N. *ParaGauss version 3.0*; Technische Universität München: Garching, Germany, 2004.

(21) Belling, T.; Grauschopf, T.; Krüger, S.; Mayer, M.; Nörtemann, F.; Staufer, M.; Zenger, C.; Rösch, N. In *High Performance Scientific and Engineering Computing*; Lecture Notes in Computational Science and Engineering, Bungartz, H.-J., Durst, F., Zenger, C., Eds.; Springer: Heidelberg, 1999; Vol. 8, pp 441−455.

(22) Dunlap, B. I.; Rösch, N. *Adv. Quantum Chem.* **1990**, *21*, 317−339.

(23) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100. (b) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822−8824; **1986**, *34*, 7406.

(24) (a) *Gaussian Basis Sets for Molecular Calculations*; Huzinaga, S., Ed.; Elsevier: Amsterdam, 1984. (b) Veillard, A. *Theor. Chim. Acta* **1968**, *12*, 405−411.

(25) Bär, M. R.; Sauer, J. *Chem. Phys. Lett.* **1994**, *226*, 405−412.

(26) Gropen, O. *J. Comput. Chem.* **1987**, *8*, 982−1003.

(27) Vayssilov, G. N.; Gates, B. C.; Rösch, N. *Angew. Chem., Int. Ed.* **2003**, *42*, 1391−1394.

(28) Vayssilov, G. N.; Rösch, N. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4019−4026.

(29) Dick, B. G.; Overhauser, A. W. *Phys. Rev. B* **1958**, *112*, 90−103.

(30) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.

(31) The structures of the QM model clusters are shown in Figure 2a,b of ref 7.

(32) van Bokhoven, J. A.; van der Eerden, A. M. J.; Prins R. *J. Am. Chem. Soc.* **2004**, *126*, 4506−4507. (b) Joyner, R. W.; Smith, A. D.; Stockenhuber, M.; van den Berg, M. W. E. *Phys. Chem. Chem. Phys.* **2004**, *6*, 5435−5439.

(33) Hadjiivanov, K. I.; Vayssilov, G. N. *Adv. Catal.* **2002**, *47*, 307−511.

(34) Strodel, P.; Neyman, K. M.; Knözinger, H.; Rösch, N. *Chem. Phys. Lett.* **1995**, *240*, 547−552.

(35) Senchenya, I. N.; Garrone, E.; Ugliengo, P. *J. Mol. Struct. Theochem* **1996**, *368*, 93−110.

(36) Brand, H. V.; Redondo, A.; Hay, P. J. *J. Mol. Catal. A* **1997**, *121*, 45−62.

(37) Ugliengo, P.; Busco, C.; Civalleri, B.; Zicovich-Wilson, C. M. *Mol. Phys.* **2005**, *103*, 2559−2571.

(38) The vibrational shift seems to increase with the size of the basis set on hydrogen. For example, addition of a p-type polarization exponent to the basis set of hydrogen increases the shift of the OH frequency, from 300 cm^{-1} to 344 cm^{-1}, in BLYP calculations of CO adsorption at the 2T cluster model $H_3Al-O(H)-SiH_3$.[36] Therefore, results from a more flexible basis set should not necessarily be expected to agree better with experiment.

(39) Vayssilov, G. N.; Rösch, N. *J. Phys. Chem. B* **2004**, *108*, 180−197.

(40) Weber, W. A.; Gates, B. C. *J. Phys. Chem. B* **1997**, *101*, 10423−10434.

(41) Ivanova Shor, E. A.; Nasluzov, V. A.; Shor, A. M.; Vayssilov, G. N.; Rösch, N. *J. Phys. Chem. C* **2007**, *111*, 12340−12351.

CT700159K

JCTC Journal of Chemical Theory and Computation

Binding of Gold Nanoclusters with Size-Expanded DNA Bases: A Computational Study of Structural and Electronic Properties

Purshotam Sharma, Himanshu Singh, Sitansh Sharma, and Harjinder Singh*

Center for Computational Natural Sciences and Bioinformatics, International Institute of Information and Technology, Gachibowli, Hyderabad-500032, India

Received June 15, 2007

Abstract: Binding of gold nanoclusters with size-expanded DNA bases, xA, xC, xG, and xT, is studied using quantum chemical methods. Geometries of the neutral xA-Au_6, xC-Au_6, xG-Au_6, and xT-Au_6 complexes were fully optimized using the B3LYP density functional method (DFT). The gold clusters around xA and xT adopt triangular geometries, whereas irregular structures are obtained in the case of gold clusters complexed around xC and xG. The lengths of the bonds between atoms in the x-bases increase on gold complexation. The aromatic character of the x-bases also increases on gold complexation except for the five-member rings. A significant charge transfer from the x-base to gold atoms is seen in these complexes. Second-order interactions are observed in addition to direct covalent bonds between gold atoms and x-bases.

1. Introduction

Detailed understanding of the nature of interaction between metal particles and conjugated molecular systems in nanoparticle complexes is of fundamental importance in the development of potential miniature devices.[1a,b] Interest in the use of modified analogs of DNA as templates for growing nanoparticle complexes has increased significantly in recent years simultaneous with intensive investigations on whether alternative genetic systems could exist for therapeutic and biotechnological applications. Analogs such as peptide nucleic acids (PNAs),[2a,b] locked nucleic acids (LNAs),[3a] and threose nucleic acids (TNAs)[3b,c] have been synthesized by different research groups. In most cases, mainly the backbone of DNA has been subject to chemical modifications. Kool and co-workers have synthesized new modified DNA using size-expanded DNA bases called xDNA[4] and yDNA.[5] It is believed that size-expanded DNA could also have properties with potential nanotechnological applications, as they retain the recognition property of natural DNA to a certain extent. These size-expanded bases are formed by benzohomologation of the natural DNA bases. They pair with complementary normal DNA bases in size-expanded DNA. More recently xDNA Double Helix up to eight base pairs incorporating all four combinations of the x-bases and natural DNA bases has been prepared.[6]

The controlled assembly of metal nanoparticles into macroscopic materials using DNA oligonucleotides has opened new directions of research in nanosciences. The charge transport properties of DNA are of great importance in the development of nanotechnological devices.[7,8] It is known that metal bound DNA nanowires have enhanced conductivity.[8,9] A large number of experimental and theoretical studies are focused on gold-DNA interactions.[10a−j] Gold, known as a noble metal for its relatively inert chemistry, has turned out to be of remarkable use in a large number of investigations of nanobio systems. Molecular dynamics simulations have been carried out in order to understand the melting properties of DNA-linked gold nanoparticle assemblies.[10g] Ab initio calculations have been carried out on bare and thiolate passivated gold nanoclusters, gold nanowires, and fragments of DNA chains, in order to provide useful insights toward the complete understanding, design, and proper utilization of hybrid DNA-gold nanostructured materials.[10h] Theoretical studies on gold nanoparticles conjugated with small organic compounds such as acetone, acetaldehyde, and diethyl ketone have also been carried out recently.[10i]

* Corresponding author phone: +91 40 2300 1967 x277; fax: +91 40 2300 1413; e-mail: laltu@iiit.ac.in.

10.1021/ct700145e CCC: $37.00
Published on Web 10/16/2007

A few theoretical studies have been carried out regarding the structural and electronic properties of the x-bases.[11] The extra Π-electrons on benzene ring in the size-expanded DNAs induce stronger Π-Π coupling between stacked bases, that would facilitate band transport. It is established that metal bound natural DNA can be a good nanowire. Experimental and theoretical studies on the nature of binding between gold nanoclusters and natural DNA bases and base pairs have been reported.[12−14] Experiments showed that adenine (A), cytosine (C), guanine (G), and thymine (T) nucleobases interact specifically and in a sequence dependent manner with the Au surfaces.[12]

Using density functional theory techniques, we explore in this work the nature of binding between gold clusters and size-expanded bases and try to understand what similarities exist and what differences would arise in these structures as compared to natural bases. We are mainly interested in investigating the structural and electronic properties of metal bound size-expanded DNA bases and the nature of interaction between xDNA bases and metal atoms. The optimized structures of gold atoms bound to x-bases show features similar to earlier works on DNA bases.[15] Irregular structures as well as Au−Au distances seen by us are similar to those obtained in the study of a thiolate molecule anchored on a stepped gold surface leading to the formation of a monatomic gold nanowire, by Krüger et al.[15] The results from these analysis may lead to the possibility of newer families of nanowires and other technologically relevant devices.

2. Methods

The Gaussian03[16] suite of programs was used for all calculations. The initial geometries of individual bases were built from the coordinates extracted from NMR models of size-expanded DNA from Protein Data Bank (PDB) for xDNA (code: 2ICZ). After removal of phosphate and sugar backbones, the structures of xA, xC, xG, and xT were optimized using the HF-6-31G** basis set. The initial structures for the gold complexes were built by placing x-bases within two equilateral gold triangles using the HF-6-31G** optimized geometries of the x-bases as the initial starting point. These structures were optimized using B3LYP/LANL2MB basis sets. Vibrational analysis was carried out for all optimized structures, and real frequencies were obtained in all cases. Single point energy calculations followed by vibrational analysis were also carried out at the B3LYP/LANL2MB level on the HF-6-31G** optimized geometries of free x-bases, in order to allow meaningful comparisons of different characteristics of x-base structures before and after gold complexation at the same theoretical level.

A measure of changes in the chemical environment of atoms in aromatic molecules is the nucleus independent chemical shift[17] in NMR measurements. We have calculated these shifts for the free x-bases and those coupled to gold clusters. The nucleus independent chemical shifts (NICS) method allows the evaluation of aromaticity, antiaromaticity, and nonaromaticity of single ring systems and individual rings in polycyclic systems (local aromaticities). This method has been extensively used to assess the aromaticity and antiaromaticity of many organic and inorganic compounds, intermediates, and transition states.[18] Recently, total NICS values were used to assess the aromaticities of different polycyclic aromatic hydrocarbons with excellent agreement with other indices of aromaticity. A ghost atom placed in the center of the five- and six-member rings of these x-bases provides a measure of the shielding effect of ring current, which gives a measure of NICS.

Additionally, Natural Bond Orbital (NBO)[19] analysis was performed on the B3LYP/LanL2MB optimized structures, to find the second-order interactions among electrons in these molecular clusters.

3. Results

We describe below the results obtained followed by a discussion in the next section. Only select data are recorded in this article. Tables (labeled Sn) of data with complete details are given in the Supporting Information. The structures of optimized xDNA bases together with corresponding natural DNA bases (insets) are shown in Figure 1. It is known from earlier studies on natural DNA bases that the gold atoms act predominantly as acceptors of electronic charge from the DNA system.[12] The starting geometries of molecular clusters consisting of the x-bases and gold atoms for optimization were generated by placing two clusters consisting of three gold atoms each, near the atoms with relatively high electronegativity, O and N, of the x-bases in their optimized geometry (we use the generic term x-bases to indicate the size-expanded bases). The optimized structures of neutral x-DNA bases complexed with six gold atoms are shown in Figure 2 (coordinates in Table S1). The nonplanarity in the (xBase)-Au_6 complexes is measured in terms of relevant dihedral angles (Table 1a).

It is seen generally that on interacting with the gold atoms, all bonds tend to expand, and the electronic charges on atoms in the bases tend to decrease. The magnitudes of a few selected bond lengths in x-bases before and after complex formation with gold atoms are given in Table 1b and the same with several others in Table S2 in the Supporting Information. The changes in bond lengths on formation of complexes with the gold atoms have been compared for some selected common bonds in natural purines and x-purines in Figure 3(a) and in natural pyrimidines and x-pyrimidines in Figure 3(b), respectively. Mulliken population analysis was looked into. Extensive comparative histograms are plotted in Figure 4 to show the Mulliken charges over all the atoms in the bases before and after complexation, and the detailed data are given in Table S3.

The shapes and orientation of frontier molecular orbitals are a good indication of reactivity of chemical systems. Plots of highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) for the x-bases are given in Figure 5. The plots of HOMO, HOMO-1, and LUMO for gold complexed x-bases are given in Figure 6.

The changes in vibrational frequencies on complexation with gold for some selected bonds of the x-bases have been given in Table 2. We see a red shift in the stretching frequencies of amino and carbonyl groups indicating weakening of these bonds in synchrony with the increase of corresponding bond lengths.

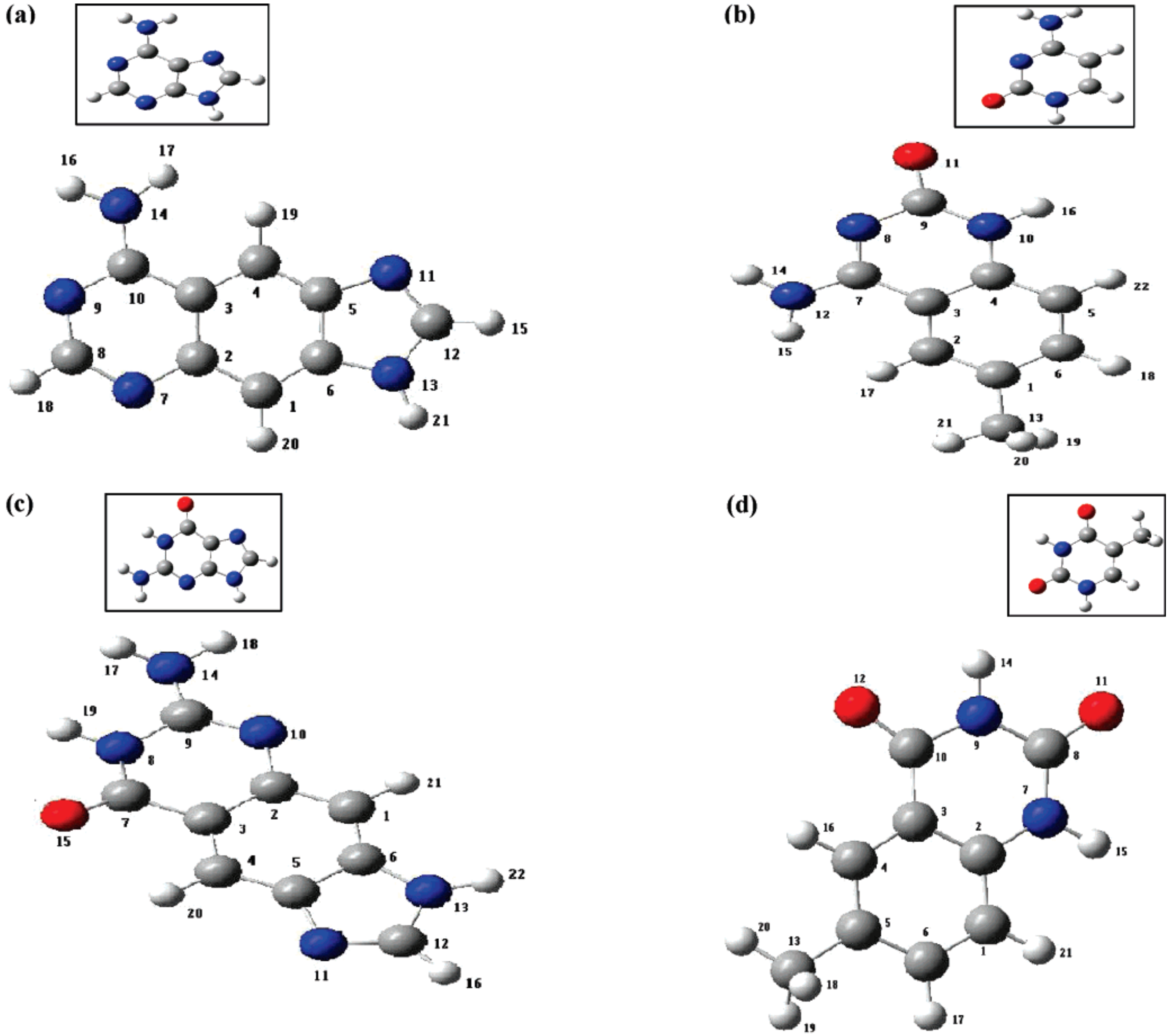


Figure 1. Size-expanded DNA bases (x-bases): (a) xA, (b) xC, (c) xG, and (d) xT. Insets show corresponding natural DNA bases. Color code: oxygen, red; carbon, dark gray; nitrogen, blue; and hydrogen, light gray.

The nucleus independent chemical shifts (NICS) for the aromatic rings of the x-bases before and after gold complexation are given in Table 3. It is found that there is a general increase in the NICS values on complexation. This is an indication of increase in the aromatic character after complex formation. In order to better understand the nature of binding between gold atoms and atoms of the x-bases, the second-order interactions present between electron density on gold atoms and on atoms of the x-bases were calculated using second-order perturbation theory analysis of gold complexed x-bases. A select cross-section of the NBO data is given in Table 4, and more detailed information is available in Table S4. The NBO analysis predicts certain second-order noncovalent interactions, in addition to direct covalent bonding between x-bases and gold clusters.

4. Discussion

Structures of Complexes. As mentioned earlier, we started the geometry optimization with placing three gold atoms near the relatively more electronegative elements, O and N, on two sides of the bases. The initial geometries of these complexes were built by placing two gold clusters, each of them forming an equilateral triangle on each side of the x-adenine near the electron rich sites, in order to model the first layer of the 111 face-centered cubic (FCC) bulk gold crystal. No change in the optimized structures was seen on using other structures in a wide neighborhood. We will generally refer to the complexes as xB-Au_6 complex except when it is imperative to indicate that they form two separate Au_3 clusters. It is found that structures obtained on optimization of x-adenine-Au_6 and x-cytosine-Au_6 complexes are nearly planar, with both the Au_6 cluster as well as the x-base separately adopting planar geometries. On the other hand, the optimized structures of x-guanine-Au_6 and x-thymine-Au_6 are significantly nonplanar. This absence of planarity in the xG-Au_6 and xT-Au_6 (the respective x-bases are planar in both cases) is noteworthy. It is possibly due to the anisotropy in electronic distribution around the x-DNA base.

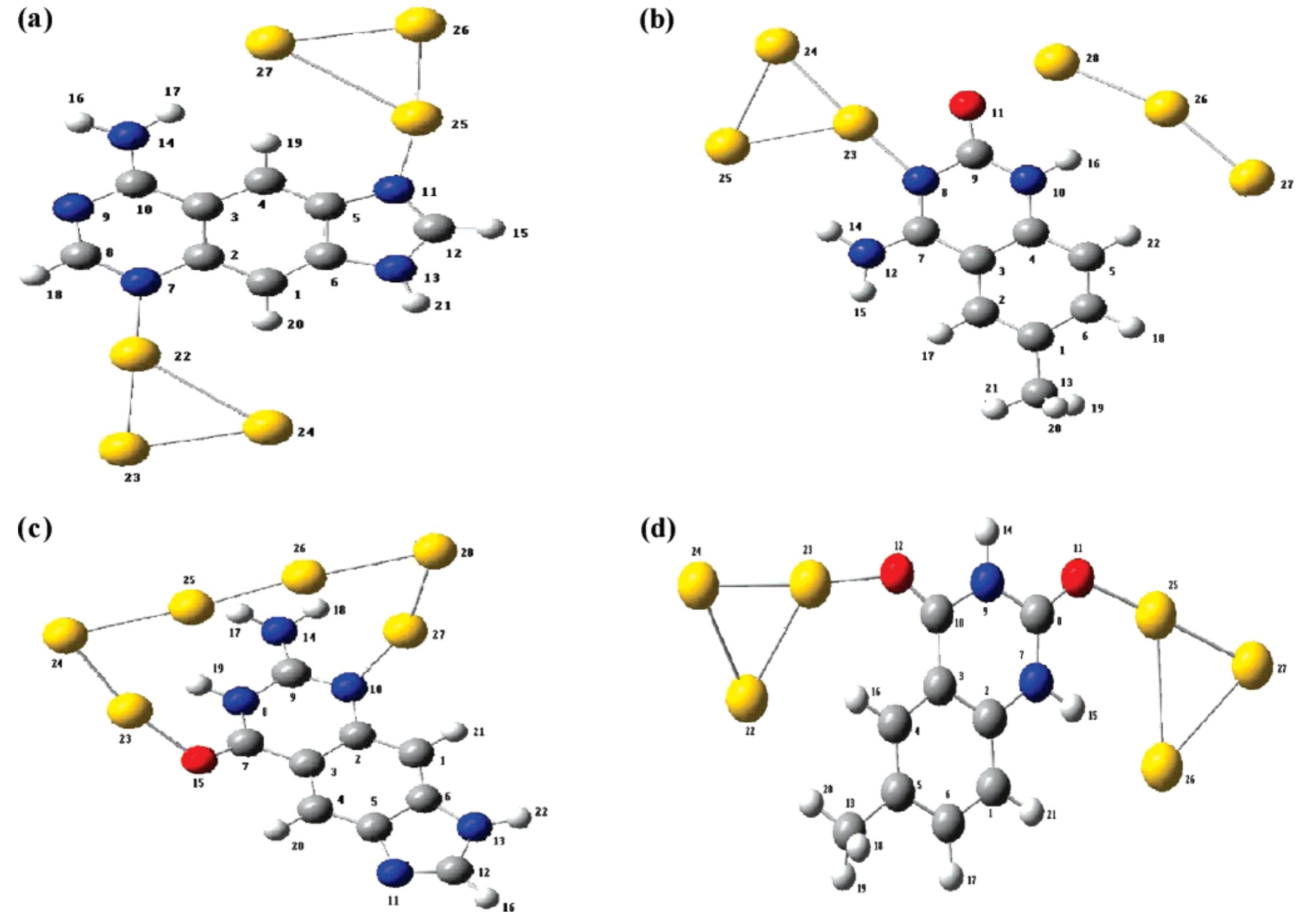


Figure 2. Optimized structures of (a) xA-Au_6 (b) xC-Au_6 (c) xG-Au_6, and (d) xT-Au_6, obtained at the B3LYP/LanL2MB level of theory.

In xG-Au_6, four of the six gold atoms are relatively closer to the base thus disrupting the consistency of the Au_3 cluster which remains preserved in the case of xA-Au_6 and xC-Au_6 complexes. In this sense, the general use of the word 'cluster' for the Au_3 units may be debatable.

A general trend seen is that the gold atoms bind to electron rich sites of the x-bases (N or O atoms). When bare nitrogen atoms (having no hydrogens attached) are available, at least one of the gold atoms (in some cases two) in the (x-base)-Au_6 complexes preferably bind to the bare nitrogen atoms forming anchor bonds, whereas the gold atoms which are near to the N−H nitrogens form unconventional N−H...Au type of hydrogen bonds. In these cases, the gold atoms optimize to geometries with minimal Au−N distances. In the case of x-thymine, where the bare nitrogen atoms are absent, the gold atoms bind to oxygen atoms.

In xA-Au_6, the gold atoms are distributed in two clusters of three atoms each, near the nitrogen atoms of xA. Although both the Au_3 clusters acquire a triangular geometry, the edges of these triangles are not equal in length. The N11−Au25 and N7−Au22 bond lengths (Table 1b) are comparable to the Au−N distance in coordination complexes of gold and nitrogen containing ligands[20] suggesting a substantial Au−N covalent binding. In xC-Au_6, two clusters of three gold atoms each are distributed near the heterocyclic ring of the xC base, where the electron rich N and O atoms are present. One of these is present near the N8 atom and is triangular in geometry. The other cluster is present near the N10 and O11 atoms and is almost linear in geometry (Table 1a). The closest interaction between the gold atoms and the ring atoms of the xC is between the N8 and Au23 for the triangular cluster and between O11 and Au28 for the other cluster (Table 1b). The gold atoms in the xG-Au_6 complex also acquire an irregular geometry, forming a Au_6 unit, in contrast to the other complexes. The gold atoms are distributed near the six-member heterocyclic ring of the xG base, and the cluster geometry deviates slightly from planarity (Table 1). The closest interaction of gold cluster and base is between O15−Au23 (2.17 Å) atoms (Table 1b). In xT-Au_6, the gold atoms arrange themselves in the form of two triangles on each side of the base xT. These Au_3 triangles are not in the plane of the x-base; they are placed on different sides of the base. A triangular geometry of three gold atoms is present near the O12 atom of the base. The other Au_3 cluster is present near the O11 and O7 atoms of the xT. The O11−Au25 and the O12−Au23 constitute the closest interactions (Table 1b). In all these studied complexes the major interaction between gold atoms and x-bases is seen to arise from covalent binding of either Au−N or Au−O type.

A major structural change observed in all the x-bases that are anchored to Au_6 clusters is the overall increase in all the bond lengths of the x-bases, leading to an expansion in their volume. In most cases, the change is small, less than 0.1 Å,

Table 1. (a) Selected Dihedral Angles (deg) of Gold Complexed x-Bases and (b) Selected Bond Lengths before and after Complex Formation with Gold Atoms

(a)

x-adenine-Au_6		x-cytosine-Au_6	
∠C2−N7−Au 22−Au 23	178.80	∠Au 24−Au 23−O8−C7	179.05
∠C5−N11−Au 25−Au 26	0.05	∠Au 26−Au 28−O11−C9	0.03
∠Au 22−N7−C2−C3	179.99	∠Au 25−Au 23−Au 28−Au 26	179.75
∠Au 23−Au 22−Au 2−Au27	−179.89		
x-guanine-Au_6		**x-thymine-Au_6**	
∠Au 24−Au 23−N8−C9	39.07	∠Au 24−Au 23−O12−C10	−154.54
∠Au 25−Au 28−C9−N8	−19.62	∠Au 27−Au 25−O11−C8	5.23
∠Au 26−Au27−N10−C9	33.61	∠O12−C10−C8−O11	0.94
∠Au 24−Au 23−Au 25−Au 28	178.01	∠Au 26−N7−C8−N9	179.50
∠Au 25−Au 28−Au 27−Au 26	−177.13	∠Au 23−C3−C10−N9	−149.50

(b)

bond	xA	xA-Au_6	bond	xC	xC-Au_6	bond	xG	xG-Au_6	bond	xT	xT-Au_6
C12−N11	1.28	1.36	C7−N12	1.32	1.37	C9−N10	1.28	1.37	C2−N7	1.38	1.43
N11−C5	1.39	1.44	C7−N8	1.31	1.39	C2−N10	1.38	1.45	N17−H15	0.99	1.06
N7−C2	1.37	1.44	N8−C9	1.36	1.42	C7−O15	1.21	1.30	N7−C8	1.37	1.40
N7−C8	1.28	1.39	C9−O11	1.21	1.34	C7−N8	1.37	1.43	C8−O11	1.20	1.29
C10−N14	1.34	1.37	C9−N10	1.37	1.40	N8−C9	1.37	1.42	C8−N9	1.37	1.43
C10−N9	1.31	1.40	N10−H16	0.99	1.06	C9−N14	1.34	1.40	N9−C10	1.37	1.43
N9−C8	1.35	1.38	N10−C4	1.37	1.43	N14−H18	0.99	1.05	C10−O12	1.20	1.29
C8−H18	1.08	1.11	N12−H15	0.99	1.04	C12−N11	1.27	1.35	O12−Au23		2.12
N14−H16	1.00	1.04	N12−H14	1.01	1.05	C5−N11	1.39	1.45			
N7−Au22		2.10	O11−Au28		1.35	Au27−N10		2.16			
N11−Au25		2.11	N8−Au23		2.15	O18−Au23		2.17			

and, in a few cases, the increase is more than 0.1 Å. Such a pervasive increase, seen even in bonds located farthest from the Au atoms, reflects the electronic density redistribution caused by the gold clusters (discussed in detail later). In all the cases, the C−N and C−O bonds show a greater expansion than the C−C bonds. The interaction with gold atoms enhances the polarity of the bonds between atoms of different electronegativity in the base, causing a significant redistribution of charge and a consequent increase in bond lengths. The optimized geometries of these base-Au complexes suggest predominant interaction of the gold atoms with electron rich regions in the bases involving mainly the hetero atoms like N and O. These atoms make their nonbonding electrons available for interaction via molecular orbitals of suitable energy. The C═C bonds are parts of too low-lying MOs and are not suitable for interaction with the gold atoms. The HOMO, LUMO diagrams suggest that the HOMO electrons which are crucial in determining the reactivity of complexes with Au atom are mainly concentrated on the polar bonds like C−N and C−O rather than C−C bonds, and they do not involve much of the gold atoms. This is a feature different from what is reported on similar complexes with natural DNA bases.[13b]

Small molecules such as H_20 and HF, when trapped inside spherical clusters such as fullerenes, are known to exhibit contraction in their volumes with blue shifts in stretching frequencies and shortening of bond lengths.[21] This is in contrast to the behavior seen with the Au atoms anchored to the natural bases in DNA as well as x-DNA. Some features in the optimized structures are similar to those seen by Krüger et al.,[15] for example, zigzag structures formed by Au atoms as well as similar Au−Au distances.

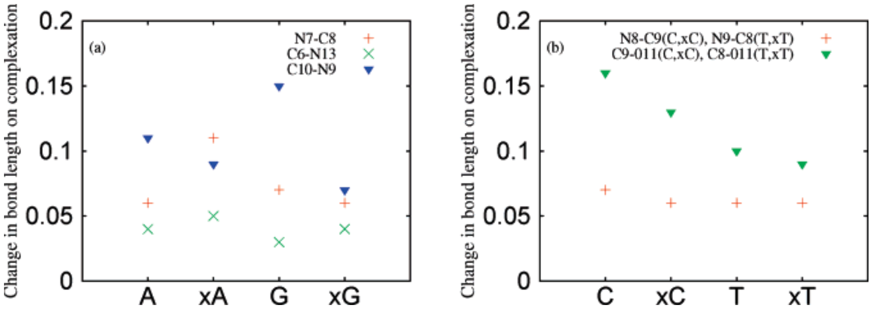


Figure 3. Change in bond length for selected atoms in natural and expanded purines (a) and in natural and expanded pyrimidines (b) on gold complexation.

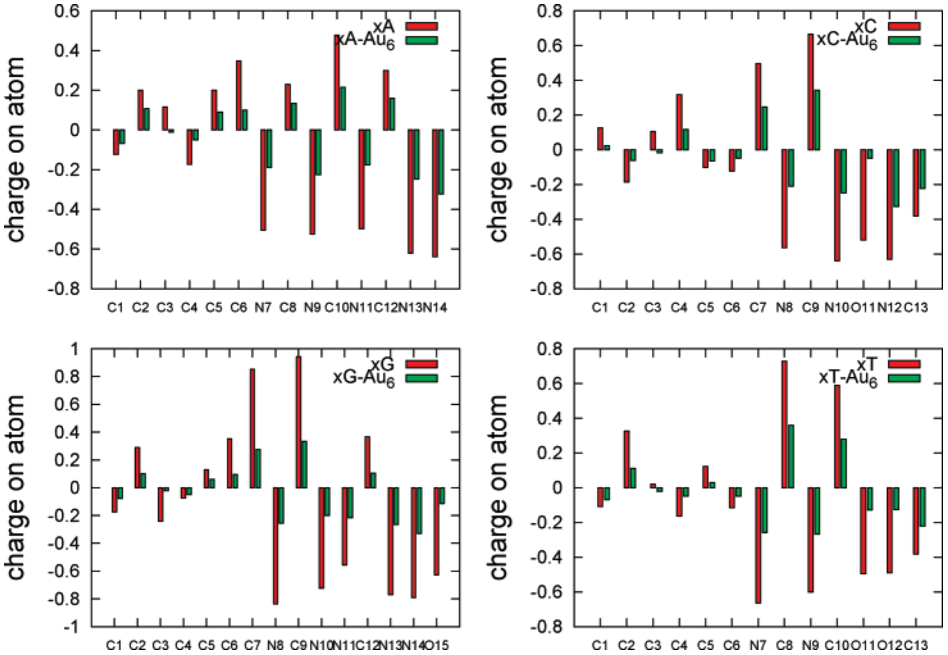


Figure 4. Mulliken charges on the atoms of the x-bases before and after gold complexation: (a) xA, (b) xC, (c) xG, and (d) xT.

The optimized complexes with gold atoms near natural DNA bases, using the same level of theory (see Figure S1) as above, show that the gold atoms form two equilateral triangles of three atoms on each side of the base in all cases, and the gold clusters are completely out of the plane of the bases. This is in contrast with the structures obtained in the case of xA-Au_6 and xC-Au_6 complexes, where the gold atoms are in the plane of the x-base. The presence of the aromatic benzene ring together with the heterocyclic rings (extension in the pyrimidines and in between the purines) provides additional stability due to delocalization of the electron density when the complex is planar. With xC-Au_6 and xG-Au_6, the triangular geometry of gold cluster seen in the case of the natural bases is not retained, and gold clusters acquire irregular structures. Expansion in bond lengths is observed in the case of natural DNA bases on complexing with gold atoms, as observed here in the case of the x-bases. To get a clear picture we plotted some of the changes in bond lengths. The increase in bond lengths for the selected common atoms of purines and x-purines are plotted in Figure 3(a), and the corresponding increase for selected common atoms for pyrimidines and x-pyrimidines are plotted in Figure 3(b). It is found that the C9−N10 bond of xG undergoes much less (nearly half) deviation than the corresponding bond in natural guanine, and the N7−C8 bond of xA undergoes greater expansion than natural adenine. The C9−N10 bond of natural cytosine undergoes much greater expansion on complex formation than the corresponding bond in the expanded bases xC and xT and natural thymine. However, the length of the bond between the N atom linked to a gold atom covalently and the adjacent carbon atom undergoes similar deviation in all cases.

Electronic Charge Distributions. A detailed analysis of the charge distributions in the uncomplexed and gold complexed x-bases was carried out using the Mulliken population analysis scheme.

In the case of xA-Au_6, we find that a substantial overall amount of charge (0.9 e, where e is the electronic charge) is transferred from x-adenine to the gold clusters. The carbon atoms of the x-adenine lose a lesser amount of electron density, in fact some of them show a gain, whereas the nitrogen atoms lose a greater charge (Table S3). This is due to the vicinity of the gold atoms to the N atoms. The $Au_{3(22-24)}$ cluster acquires a negative charge amounting to 0.37e units, whereas the other Au_3 cluster acquires 0.54e of charge. Generally, the gold atom bound to an electron rich site is found to withdraw electrons from the site and acquires negative charge. The N11 and N7 atoms bonded to gold atoms lose about 0.32e of charge to the nearest gold atoms, and the N13 and N9 atoms at the beta positions lose more than 0.3e of charge after redistribution of electron populations. These observations indicate the presence of charge-transfer interactions between gold clusters and x-adenine.

In the case of xC, there is an overall charge loss of 0.95e from the base to the gold clusters. The O atom loses charge amounting to half an electronic charge, as compared to the charge carried by it before gold complexation. The N8 bonded to the gold atom and the nitrogen atoms at the beta positions, N10 and N12, lose more than 0.3e charge each. The $Au_{3(23-25)}$ cluster gains 0.29e units of negative charge, whereas the other Au_3 cluster gains 0.66e units of electronic charge. The triangular structure of the former cluster does not permit a large amount of charge to accumulate as this will give rise to a high charge density as opposed to the open chain case of the latter.

In the case of xG, 0.87e of negative charge is transferred from xG to Au cluster. The N atoms 8, 10, 13, 14, and 015 atoms lose nearly half an electronic charge, and the N11 atom loses 0.34 e of charge. Interestingly, in this case, the C7 and C9 atoms in the vicinity of the covalent linkage between N and Au atoms also lose fairly high amounts of positive charge

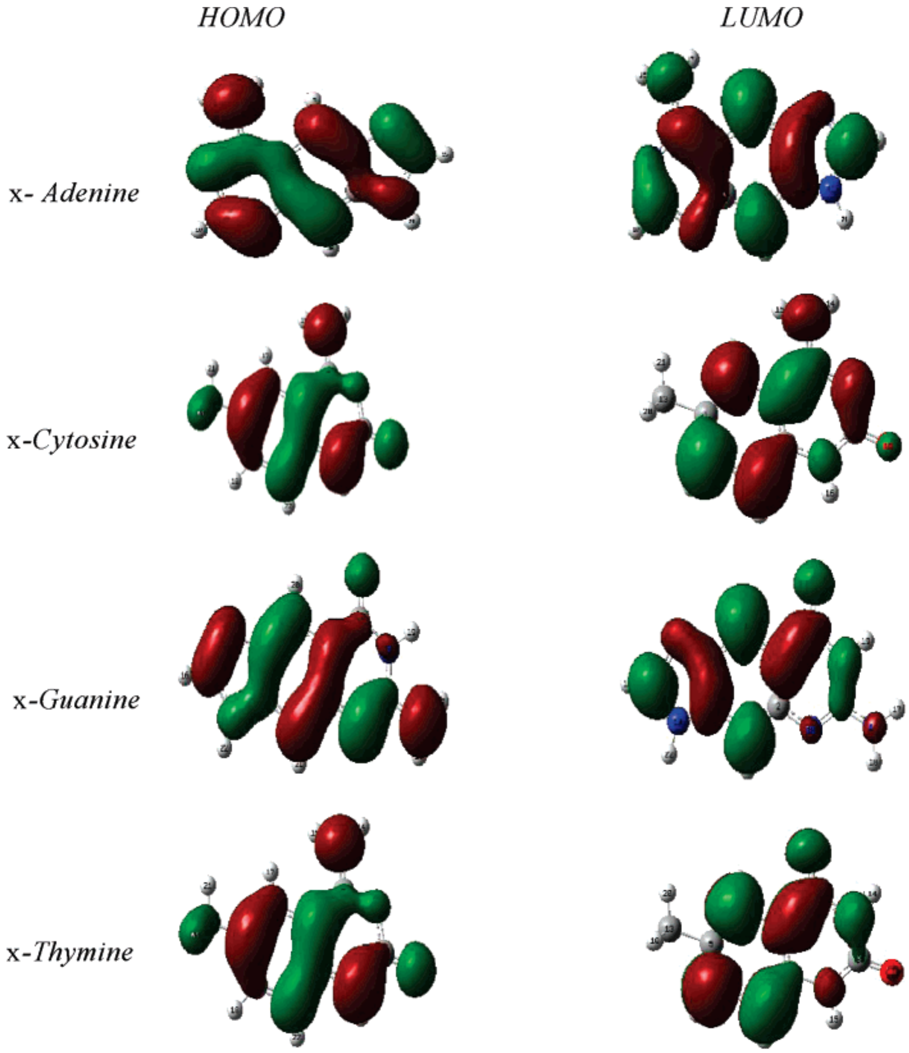


Figure 5. Plots of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) for expanded bases x-adenine, x-cytosine, x-guanine, and x-thymine.

(gaining electron density), namely 0.57e and 0.61e respectively, unlike the carbon atoms in the case of xA and xC.

The case of xT is distinct because here the binding with the gold atoms does not involve the N atoms; it is entirely with the O atoms. The N7 and N9 atoms lose 0.41e and 0.37e and C8 and C10 lose 0.33e and 0.31e positive charge (gaining electron density) on gold complexation. Both O11 and O12 atoms lose relatively less, namely 0.36 e charge on gold complexation, nearly the same as the analogous N atoms bound to a gold atoms in the case of xA and xC. A total of 0.82 e is transferred from xT to the gold cluster.

These observations suggest a massive redistribution of electronic charge when the x-bases come in contact with gold atoms. The charge distributions suggest that the gold clusters are stabilized around the x-bases due to electrostatic attractions between the gold atoms and x-base atoms. The Mulliken charges for some selected common atoms of purines have been plotted in Figure 3(c), whereas the corresponding values for selected common atoms for pyrimidines have been shown in Figure 3(d). It is seen that the N14 atom of xG loses a greater amount of charge on gold complexation than the corresponding atom in xA. On the other hand, the decrease in Mulliken charge on the O11 atom in the case of xC is greater than in the case of xT.

In all the expanded bases, the polarization in the six-member benzenoid ring is reduced significantly; this is consistent with the greater aromatic character found with NICS calculations discussed later.

Molecular Orbital Plots. Plots of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of the x-bases are given in Figure 5. It can be seen that there exists a strong intermixing between the atomic orbitals of the x-bases in the frontier molecular orbitals.

Plots of HOMO, HOMO-1, and LUMO for the gold complexed geometry are shown in Figure 5. The frontier orbitals reflect the reactive properties and sensitivity toward neighboring entities in larger assemblies. In contrast to findings on natural DNA bases,[13b] we find that the HOMO and LUMO of xA-Au_6 does not significantly involve the atomic orbitals of the Au atoms. The interaction of Au orbitals with those of the base is maximal in the HOMO of xG-Au_6, and it is found to be antibonding in nature, while for xT-Au_6 and xC-Au_6 it is marginal. The situation changes considerably on excitation, and a greater mixing of Au atomic

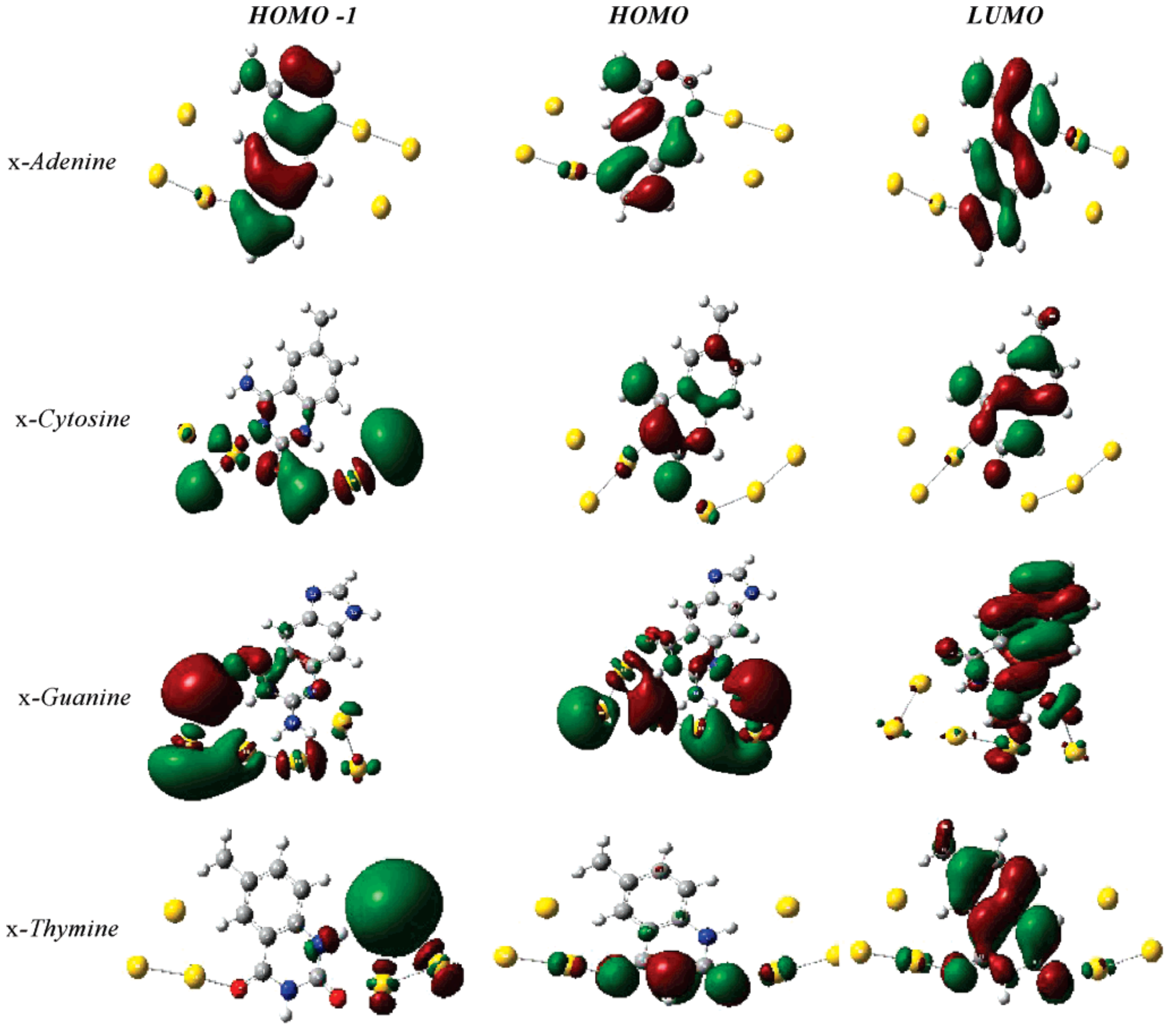


Figure 6. Plots of the frontier molecular orbitals HOMO-1, HOMO, and LUMO for the gold complexes of the expanded bases: x-adenine-Au_6, x-guanine-Au_6, x-cytosine-Au_6, and x-thymine-Au_6.

orbitals is seen in LUMO and LUMO+1 orbitals of the complex. For the xC-Au_6, the HOMO-1 and LUMO show some interaction between the gold atomic orbitals and those of the base. There is a significant intermixing of the atomic orbitals of gold as well as the x-base orbitals in the case of HOMO-1 and LUMO orbitals of the xG-Au_6 complex, but the HOMO is more localized on the gold atoms in this case. The HOMO and LUMO of xT show substantial bonding between gold atoms and xT atoms, but the HOMO-1 is more localized on the gold cluster. The HOMO−LUMO gap for xA-Au_6, xC-Au_6, xG-Au_6, and xT-Au_6 clusters are 0.27, 1.09, 2.92, and 0.27 eV, respectively, at the B3LYP/LanL2MB level. The HOMO−LUMO gaps for free x-bases are found to be 4.35, 4.62, 4.68, and 5.22 eV for xA, xC, xG, and xT, respectively, at the same level of theory. It may be noted that, as the HOMO−LUMO gap in these complexes is too small, this indicates that even a slight amount of thermal energy input can even excite the electrons to higher levels in these complexes. The smaller HOMO−LUMO gap in these complexes could facilitate band transport and charge migration in gold bound xDNA. This suggests the possible use of gold bound size-expanded DNA structures as nanowires.

Vibrational Analysis. Vibrational analysis has been carried out on the optimized structures of the x-bases, in order to examine the effect of gold complexation on stretching frequencies of certain functional groups present in x-bases (Table 2). It has been found that the stretching frequencies of amino as well as carbonyl groups get red-shifted on complexation of the x-bases with gold atoms. This is expected from the general lengthening of the bond distances observed and discussed earlier. The stretching frequencies of carbonyl groups in xC and xG get red-shifted by 226.08 cm^{-1} and 206.51 cm^{-1}, respectively, and the stretching frequencies of the two carbonyl groups of xT get red-shifted slightly (C8−O11 and C10−O12, respectively, by 174.76 and 184.43 cm^{-1}). Similarly, the stretching frequencies of the NH_2 groups of xA, xC, and xG also get shifted to lower

Table 2. Vibrational Frequencies of Some Selected Bonds in Free Bases and Gold Complexed x-Bases

bond	xA (cm^{-1})	xA-Au$_6$ (cm^{-1})	difference (cm^{-1})
NH2 antisymmetric stretch	3948.80	3842.63	106.17
NH2 symmetric stretch	3729.26	3599.63	129.63

bond	xC (cm^{-1})	xC-Au$_6$ (cm^{-1})	difference (cm^{-1})
NH2 antisymmetric stretch	3951.55	3878.48	73.07
NH2 symmetric stretch	3728.02	3629.85	98.17
N10−C16	3729.69	3510.89	218.8
C9−O11	1831.54	1605.46	226.08

bond	xG (cm^{-1})	xG-Au$_6$ (cm^{-1})	difference (cm^{-1})
NH2 antisymmetric stretch	3954.71	3779.30	175.41
NH2 symmetric stretch	3735.08	3592..99	142.09
N8−H19	3714.46	3611.73	102.73
C7−O11	1836.39	1629.88	206.51

bond	xT (cm^{-1})	xT-Au$_6$ (cm^{-1})	difference (cm^{-1})
N7−C15	3762.43	3459.37	303.06
C8−O11	1886.73	1711.97	174.76
C10−O12	1822.55	1638.12	184.43

wave numbers. These results provide information about the difference in behavior of the x-bases in the gold complexed and free state.

Aromatic Character of the x-Bases. To quantify the aromatic nature of the rings (at their ring centers) of the x-bases before and after complexation to the Au clusters, the nucleus independent chemical shifts (NICS)[17] are calculated at the centers of five- and six-member rings of the x-bases. NICS is a computational method that calculates the chemical shift of a hypothetical ghost atom positioned inside the ring. It is one of the methods of measurement of relative aromaticity of different rings with respect to the observed ring current. The more the negative value of NICS, the greater will be the aromaticity of the ring. The NICS values of five- and six-member rings are given in Table 3. It can be seen that on complexation with gold atoms, the aromatic character of the six-member carbon rings of the x-base increases, but the aromaticity of the five-member rings in the x-bases decreases in the case of purine x-bases. The aromaticity of the six-member heterocyclic ring increases in the case of xG, but it remains the same in the case of xA. On the other hand, the aromaticity of both the six-member carbon rings as well as the six-member heterocyclic rings increases on complexation with gold atoms. These calculations suggest a contribution of electronic effects, to the overall molecular expansion of the x-bases on gold complexation. We point out that the natural purine bases have been reported to become less aromatic, whereas natural pyrimidine bases become more aromatic on complexation with gold atoms.[13b]

Bonding and Interactions. We carried out NBO analysis of the xB-Au$_6$ complexes and recorded the data on interactions between gold clusters and x-bases. We look at the second-order perturbative estimates of the donor−acceptor (bond−antibond) interactions from NBO analysis to investigate the charge-transfer interactions between gold clusters and atoms of the x-bases. Since these interactions lead to donation of occupancy from the localized NBOs of the idealized Lewis structure into the empty non-Lewis orbitals (and thus, to departures from the idealized Lewis structure description), they are referred to as “delocalization” corrections to the zeroth-order natural Lewis structure. For each donor NBO (i) and acceptor NBO (j), the stabilization energy $E(2)$ associated with delocalization (“2e-stabilization”) $i \rightarrow j$ is estimated as[19]

$$E(2) = \Delta E_{i,j} = \frac{-q_i(F(i,j))^2}{E_j - E_i}$$

where q_i is the donor orbital occupancy; E_i and E_j are diagonal elements (orbital energies); and $F(i,j)$ is the off-diagonal NBO Fock matrix element.

The values of $E(2)$, $F(i,j)$, and $E_j - E_i$ for the predominant charge-transfer interactions between the x-bases and gold atoms are given in Table 4 with more details in Table S4 in the Supporting Information. We look at the charge-transfer interactions between the gold clusters and the atoms of the x-base in order to explore the possibility of hydrogen bonding and to understand the nature of interactions between them.

We observe that a significant charge transfer takes place from the lone pair of N7 in xA to antibonding orbitals of Au22−Au23. In addition, charge transfer from the lone pair of electrons in N11 to the antibonding orbital of Au25−Au27 is also seen. The hydrogen bonding interactions between C4−H1...Au26 as well as between N14−H17...Au26 also contribute to the stability of these clusters. In the case of xC, charge-transfer interactions from the lone pair on O11 to the antibonding orbital of Au23−Au24 and Au25−Au28 are the most significant. The hydrogen bonding interactions N10−H16...Au26 and N10−H16...Au28 are also seen in this case. In the case of xG, the charge-transfer interactions from the lone pair of O15 to the antibonding orbitals of Au23−Au24 is the most prominent. The charge transfers from the bonding orbital localized on C9−N10 to the antibonding orbitals of Au26−Au27 and from the lone pair of N10 to the antibonding orbital of Au26−Au27 are also significant. The hydrogen-bonding N8−H19...Au23 is also among the major interactions present in the xG-Au6 complex. In the case of xT, the major charge transfer takes place from the lone pair of O12 to the antibonding orbital of Au23−Au24 and from the lone pair of O11 to the antibonding orbital of Au25−Au27. Hydrogen bonding interactions N7−H15...Au26 and C4−H6...Au23 are also seen.

These observations suggest that the binding between x-DNA nucleobases and gold atom clusters is mostly due to the direct covalent bonds of Au−N or Au−O type. A variety of effects such as charge transfer as well as electrostatic effects and unconventional interactions such as N−H···Au hydrogen-bonding contribute to the stability of the complexes.

Table 3. NICS Calculations for x-Bases before and after Complexation with Gold Atoms

	xA		xG		xC		xT	
ring type	before	after	before	after	before	after	before	after
5-member	−9.57	−9.44	−9.78	−8.96				
6-member carbon ring	−9.68	−11.13	−9.27	−12.40	−7.81	−9.31	−8.03	−8.72
6-member heterocyclic	−4.22	−4.22	−1.08	−2.10	0.08	−3.33	−0.10	−3.87

Table 4. Selected Data from NBO Analysis for xA in Gold Complexed Geometry

	interaction (BD: bonding; LP: lone pair)		second-order interaction (kcal/mol)		
			$E^{(2)}$	$E_j - E_i$	F_{ij}
xA-Au$_6$	LP N 7	BD* Au22−Au23	82.88	0.40	0.165
	LP N11	BD* Au25−Au27	79.05	0.42	0.163
	LP N8	BD*Au23−Au24	69.83	0.39	0.148
xC-Au$_6$	LP O11	BD*Au23−Au24	11.05	0.12	0.033
	LP O11	BD*Au25−Au28	68.45	0.06	0.064
	LP O15	BD* Au23−Au24	12.61	0.60	0.083
	LP O15	BD* Au23−Au24	60.76	0.23	0.107
xG-Au$_6$	BD C9−N10	BD* Au26−Au27	17.87	0.24	0.061
	LP N10	BD* Au26−Au27	57.92	0.37	0.132
	LP O12	BD*Au23−Au24	16.38	0.56	0.091
xT-Au$_6$	LP O12	BD*Au23−Au24	65.81	0.27	0.120
	LP O11	BD*Au25−Au27	13.93	0.60	0.087
	LP O11	BD*Au25−Au27	73.47	0.25	0.122

5. Conclusion

In this paper, we have presented the first ab initio study of the binding with gold clusters, of size-expanded bases xA, xC, xG, and xT, where an additional benzenoid ring is inserted as compared to the bases of natural DNA. These bases have already been used experimentally to synthesize a new type of DNA called xDNA. There are some similarities and several clear differences in bonding in these complexes compared to bonding in complexes of gold atoms with the natural DNA bases. We find that the bond lengths of the x-bases expand on gold complexation. There is a significant intermixing of orbitals of gold and the x-base in these complexes in lower lying molecular orbitals. However, in clear contrast to the complexes with natural DNA bases, the frontier orbitals do not show a significant mixing of orbitals of the atoms on the expanded bases and those of Au atoms. It seems that most of the mixing must be occurring at lower energy levels, or excitation may result in greater mixing. The HOMO−LUMO gap of gold complexed x-bases is smaller than the free x-bases, thus opening up avenues for exploration of properties like enhanced conductivity in xDNA tagged by gold atoms. There is an appreciable amount of charge transfer from the x-base to gold atoms in all the complexes. The stretching frequencies of the amino and carbonyl groups of the x-bases get red-shifted on gold complexation. The aromatic character of rings in polycyclic x-bases increases on gold complexation. The stability of the complexes is best explained using results from natural bond orbital analysis. It is found that the binding between gold clusters and x-bases has substantial contribution from noncovalent interactions, in addition to the direct covalent bonding between N or O atoms and Au atoms in these complexes. Hydrogen bonding interactions like N−H...Au are likely to play a significant role in the stability of the complexes. We are hopeful that our findings will be of significant relevance to further the understanding of macromolecular assemblies. Further investigations exploring the robustness of these findings with respect to variation of the number of gold atoms in the clusters are desired.

Acknowledgment. H.S. wishes to thank Prof. Biman Bagchi of SSCU, IISc, Bangalore, for his valuable comments as well as his crucial role in getting us involved with nanocluster tagged DNA systems.[22] We thank the CDAC, Pune for computational support and CSIR, New Delhi for JRFs to P.S. and S.S. We also thank the referees for their wise and useful suggestions.

Supporting Information Available: Cartesian coordinates and NBO analysis data of gold complexed x-bases and comparison of the bond lengths and Mulliken charges of all the atoms of free and gold complexed x-bases (Tables S2 and S3). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Ratner, M.; Ratner, D. *Nanotechnology: A Gentle Introduction to the Next Big Idea*; Prentice Hall: Upper Saddle River, NJ, 2002. (b) Tour, J. M. *Molecular Electronics: Commercial Insights, Chemistry, Devices, Architecture and Programming*; World Scientific: River Edge, NJ, 2003.

(2) (a) Egholm, M.; Buchart, O.; Christensen, L.; Behrens, C.; Freier, S. M.; Driver, D. A.; Berg, R. H.; Kim, S. K.; Norden, B.; Nielsen, P. E. *Nature* **1993**, *365*, 566. (b) Nielsen, P. E.; Egholm, M.; Buchart, O. *Bioconjugate Chem.* **1994**, *5*, 3.

(3) (a) Petersen, M.; Wengel, J. *Trends Biotechnol*. **2003**, *21*, 74. (b)Schoning, K.; Scholz, P.; Guntha, S.; Wu, X.; Krishnamurthy, R.; Eschenmoser, A. *Science* **2000**, *290*, 134. (c) Eschenmoser, A. *Science* **1999**, *284*, 2118.

(4) (a) Liu, H.; Gao, J.; Maynard, L.; Saito, D. Y.; Kool, E. T. *J. Am. Chem. Soc.* **2004**, *126*, 1102. (b) Liu, H.; Gao, J.; Lynch, S. R.; Saito, Y. D.; Maynard, L.; Kool, E. T. *Science* **2003**, *302*, 868. (c) Liu, H.; Gao, J.; Lynch, S. R.; Kool, E. T. *J. Am. Chem. Soc.* **2004**, *126*, 6900. (d) Gao, J.; Liu, H.; Kool, E. T. *J. Am. Chem. Soc*. **2004**, *126*, 11826. (e) Liu, H.; Gao, J.; Kool, E. T. *J. Am. Chem. Soc.* **2005**, *127*, 1396. (f) Liu, H.; Gao, J.; Kool, E. T. *J. Org. Chem.* **2005**, *70*, 639. (g) Lee, A. H. F.; Kool, E. T. *J. Am. Chem. Soc.* **2005**, *127*, 3332. (h) Gao, J.; Liu, H.; Kool, E. T. *Angew. Chem., Int. Ed*. **2005**, *44*, 3118.

(5) Liu, H.; He, K.; Kool, E. T. *Angew. Chem., Int. Ed*. **2004**, *43*, 5834. (b) Lee, A. H. F.; Kool, E. T. *J. Org. Chem.* **2005**, *70*, 132.

(6) Lynch, S. R.; Liu, H.; Gao, J.; Kool, E. T. *J. Am. Chem. Soc.* **2006**, *128*, 14704.

(7) Braun, E.; Eishen, Y.; Sivan, U.; Ben-Yoseph, G. *Nature* **1998**, *391*, 775.

(8) Richter, J.; Mertig, M.; Pompe, W.; Monch, I.; Schackert, H. K. *Appl. Phys. Lett*. **2001**, *78*, 536.

(9) (a) Adessi, C. H.; Walch, S.; Anantram, M. P. *Phys. Rev. B* **2003**, *67*, 081405. (b) Health, J. R.; Ratner, M. A. *Phys. Today* **2003**, 43. (c) Long, Yi-T.; Li, C.-Z.; Kraatz, H.-B.; Lee, J. S. *Biophys. J*. **2003**, *84*, 3218. (d) Moreno-Herrero, F.; Herrero, P.; Moreno, F.; Colchero, J.; Gomez-Navarro, C.; Gomez-Herrero, J.; Baro, A. M. *Nanotechnology* **2003**, *14*, 128.

(10) (a) Hou, S.; Zhang, J.; Li, R.; Ning, J.; Han, R.; Shen, Z.; Zhao, X.; Xue, Z.; Wu, Q. *Nanotechnology* **2005**, *16*, 239. (b) Reed, M. A.; Zhou, C.; Miller, C. J.; Burgin, T. P.; Tour, J. M. *Science* **1997**, *278*, 252. (c) DiVenta, M.; Pantelides, S. T.; Lang, N. D. *Phys. Rev. Lett.* **2000**, *84*, 979. (d) Derosa, P. A.; Seminario, J. M. *J. Phys. Chem. B* **2001**, *105*, 471. (e) Xue, Y.; Ratner, M. A. *Phys. Rev. B* **2003**, *68*, 115407. (f) Kerman, K.; Morita, Y.; Takamura, Y.; Ozsoz, M.; Tamiya, E. *Anal. Chim. Acta* **2004**, *510*, 169. (g) Jin, R.; Wu, G.; Li, Z.; Mirkin, C. A.; Schatz, G. C. *J. Am. Chem Soc.* **2003**, *125*, 1643. (h) Garzon, I. L.; Artacho, E.; Beltran, M. R.; Garcia A.; Junquera, J.; Michaelian, K.; Ordejon, P., Rovira, C.; Portal, D. S.; Soler, J. M. *Nanotechnology* **2001**, *12*, 126. (i) Shafai, G. S.; Shetty, S.; Krishnamurty, S.; Shah, V.; Kanhere, D. G. *J. Chem. Phys.* **2007**, *126*, 014704. (j) West, J. L.; Halas, N. J. *Annu. Rev. Biomed. Eng.* **2003**, *5*, 285.

(11) Cabrera-Fuentes, M.; Sumpter, B. G.; Wells, J. C. *J. Phys. Chem. B* **2005**, *109*, 21135.

(12) (a) Demers, L. M.; Östblom, M.; Zhang, H.; Jang, N. H.; Liedberg, B.; Mirkin, C. A. *J. Am. Chem. Soc.* **2002**, *124*, 11248. (b) Storhoff, J. J.; Elghanian, R.; Mirkin, C. A.; Letsinger, R. L. *Langmuir* **2002**, *18*, 6666. (c) Kimura-Suda, H.; Petrovykh, D. Y.; Tarlov, M. J.; Whitman, L. J. *J. Am. Chem. Soc.* **2003**, *125*, 9014. (d) Petrovykh, D. Y.; Kimura-Suda, H.; Whitman, L. J.; Tarlov, M. J. *J. Am. Chem. Soc.* **2003**, *125*, 5219. (e) Chen, Q.; Frankel, D. J.; Richardson, N. V. *Langmuir* **2002**, *18*, 3219. (f) Giese, B.; McNaughton, D. *J. Phys. Chem. B* **2002**, *125*, 1112. (g) Rapino, S.; Zerbetto, F. *Langmuir* **2005**, *21*, 2512. (h) Otero, R.; Schöck, M.; Molina, L. M.; Lægsgaard, E.; Stensgaard, I.; Hammer, B.; Besenbacher, F. A*ngew. Chem., Int. Ed.* **2005**, *44*, 2270. (i) Östblom, M.; Liedberg, B.; Demers, L. M.; Mirkin, C. A. *J. Phys. Chem. B* **2005**, *109*, 15150. (j) Yonezawa, T.; Onoue, S.-Y.; Kimizuka, N. *Chem. Lett.* **2002**,1172.

(13) (a) Kumar, A.; Mishra, P. C.; Suhai, S. *J. Phys. Chem. A* **2006**, *110*, 7719. (b) Mohan, P. J.; Datta, A.; Mallajosyula, S. S.; Pati, S. K. *J. Phys. Chem. B* **2006**, *110*, 18661.

(14) Kryachko, E. S.; Remacle, F. *J. Phys. Chem. B* **2005**, *109*, 22746.

(15) Krüger, D.; Fuchs, H.; Rousseau, R.; Marx, D.; Parrinello, M. *Phys. Rev. Lett.* **2002**, *89*, 186402.

(16) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785. (c) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, J. A.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian03*, *Revision B.05*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(17) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; Hommes, N. J. R. v. E. *J. Am. Chem Soc***.** **1996**, *118*, 637.

(18) (a) Gomes, J. A. N. F.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349. (b) Lazzeretti, P. *Prog. Nucl. Magn. Reson. Spectrosc*. **2000**, *36*, 1. (c) Krygowski, T. M.; Cyranski, M. K. *Phys. Chem. Chem. Phys*. **2004**, *6*, 249. (d) Chen, Z.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R. *Chem. Rev.* **2005**, *105*, 3842.

(19) (a) Reed, A. E.; Weinstock R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735. (b) Reed, A. E.; Curtsiss L. A.; Weinhold, F. *Chem. Rev***.** **1988**, *88*, 899.

(20) Barranco, E. M.; Crespo, O.; Gimeno, M. C.; Jones, P. G.; Laguna, *Eur. J. Inorg. Chem.* **2004**, *2004*, 4820.

(21) (a) Shameena, O.; Ramachandran, C. N.; Satyamurthy, N. *J. Phys. Chem. A* **2006**, *110*, 2. (b) Sen, K. D. *J. Chem. Phys*. **2005**, *122*, 194324.

(22) (a) Singh, H.; Bagchi, B. *Curr. Sci*. **2005**, *89*, 1710. (b) Saini, S.; Singh, H.; Bagchi, B. *J. Chem. Sci.* **2006**, *118*, 23.

CT700145E

JCTC Journal of Chemical Theory and Computation

Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms

Jianyin Shao, Stephen W. Tanner,[†] Nephi Thompson,[‡] and Thomas E. Cheatham, III*

Departments of Medicinal Chemistry, Pharmaceutics and Pharmaceutical Chemistry,
and Bioengineering, College of Pharmacy, University of Utah, 2000 East 30 South,
Skaggs Hall 201, Salt Lake City, Utah 84112

Received May 17, 2007

Abstract: Molecular dynamics simulation methods produce trajectories of atomic positions (and optionally velocities and energies) as a function of time and provide a representation of the sampling of a given molecule's energetically accessible conformational ensemble. As simulations on the 10−100 ns time scale become routine, with sampled configurations stored on the picosecond time scale, such trajectories contain large amounts of data. Data-mining techniques, like clustering, provide one means to group and make sense of the information in the trajectory. In this work, several clustering algorithms were implemented, compared, and utilized to understand MD trajectory data. The development of the algorithms into a freely available C code library, and their application to a simple test example of random (or systematically placed) points in a 2D plane (where the pairwise metric is the distance between points) provide a means to understand the relative performance. Eleven different clustering algorithms were developed, ranging from top-down splitting (hierarchical) and bottom-up aggregating (including single-linkage edge joining, centroid-linkage, average-linkage, complete-linkage, centripetal, and centripetal-complete) to various refinement (means, Bayesian, and self-organizing maps) and tree (COBWEB) algorithms. Systematic testing in the context of MD simulation of various DNA systems (including DNA single strands and the interaction of a minor groove binding drug DB226 with a DNA hairpin) allows a more direct assessment of the relative merits of the distinct clustering algorithms. Additionally, means to assess the relative performance and differences between the algorithms, to dynamically select the initial cluster count, and to achieve faster data mining by "sieved clustering" were evaluated. Overall, it was found that there is no one perfect "one size fits all" algorithm for clustering MD trajectories and that the results strongly depend on the choice of atoms for the pairwise comparison. Some algorithms tend to produce homogeneously sized clusters, whereas others have a tendency to produce singleton clusters. Issues related to the choice of a pairwise metric, clustering metrics, which atom selection is used for the comparison, and about the relative performance are discussed. Overall, the best performance was observed with the average-linkage, means, and SOM algorithms. If the cluster count is not known in advance, the hierarchical or average-linkage clustering algorithms are recommended. Although these algorithms perform well, it is important to be aware of the limitations or weaknesses of each algorithm, specifically the high sensitivity to outliers with hierarchical, the tendency to generate homogenously sized clusters with means, and the tendency to produce small or singleton clusters with average-linkage.

Introduction

Molecular dynamics (MD) and free energy simulation methods provide valuable insight into the structure, dynamics, and interactions of biological macromolecules.[1−4] Over the past three decades, MD simulation methods have proven to be an accurate tool for probing the detailed atomistic dynamics of models of biological systems on the picosecond to microsecond time scales.[5−14] MD simulations give direct insight into protein folding,[8,15−29] drug-receptor interaction,[3,30−35] and fast time scale motions of biological molecules.[36−47] As computer power continues to increase, and

* Corresponding author e-mail: tec3@utah.edu.

[†] Current address: Bioinformatics Program, University of California San Diego, La Jolla, CA 92093.

[‡] Current address: Department of Physics, Wright State University, 248 Fawcett Hall, 3640 Colonel Glenn Hwy, Dayton, OH 45435.

10.1021/ct700119m CCC: $37.00
Published on Web 10/06/2007

simulations on the 10−100 ns time scale and beyond become routine, large amounts of data result. This sequence of data—the "MD trajectory"—fully specifies the history of the atomic motions in terms of a sequential time-dependent set of molecular configurations from the MD simulation and the larger set of derived properties calculated from the MD trajectory (such as energies, bond lengths, and angle distributions). These data not only provide insight into the structure, dynamics, and interactions of the biomolecules under study but also can be reused to score putative force field changes, as a set of "good" and "bad" representative structures sampled, and for the development of coarse-grained potentials. Although many of the properties derived from the MD trajectory are rather easy to extract, such as the time evolved root-mean-squared coordinate deviation (RMSd) to the initial structure or various distance and angle time series, some properties are more difficult to extract and may be significantly more time-consuming to evaluate (such as entropies and heat capacities). Further, even with elucidation of these properties, often the inherent relationships among the molecular configurations are hidden in the complexity of the data. One very useful way to expose some of these correlations is to group or cluster molecular configurations into subsets based on the similarity of their conformations (as measured by an appropriate metric).[48,49] Clustering is a general data-mining technique that can be applied to any collection of data elements (points) where a function measuring distance between pairs of points is available.[50,51] A clustering algorithm partitions the data points into a disjoint collection of sets called clusters. The points in one cluster are ideally closer, or more similar, to each other than to points from other clusters. In this work, we describe the implementation and application of a variety of well-known pairwise distance metric clustering algorithms into a general purpose (and freely available) C code library. To test and validate the implementations, a simple problem is the clustering of randomly (or systematically) placed points in the Euclidean plane where the pairwise metric is the distance between points. This provides an easy way, using the discrimination of our visual system, to *see* the results and to highlight bugs in the implementations. This contrived test system also nicely highlights the underlying limitations of each algorithm. After description of the algorithms and their relative performance, the clustering methods are then applied to a series of MD trajectories of various biomolecular systems.

The use of clustering algorithms to group together similar conformations visited during a MD simulation is not a novel concept.[48,49,52] A wide variety of algorithms has been applied in many studies to cluster molecular dynamics trajectories, group similar conformations, and otherwise search for similarities among structures. A subset of publications developing and applying clustering algorithms to analyze molecular dynamics trajectories spans the range from some of the earliest MD simulations to very recent studies.[48,49,52−75] In this work we build on the previous studies by comparing and contrasting the performance of various well-known clustering algorithms applied to the points in a plane example and multiple different sets of MD simulation data. The algorithms implemented include *top-down/divisive* (**hierarchical**), *bottom-up/agglomerative* (single-linkage/**edge**-joining, **centripetal**, **complete-linkage**, **centroid-linkage**, **average-linkage**, and **centripetal-complete**), *refinement* (**means**, **Bayesian**, and self-organizing maps or **SOM**), and *tree* clustering (**COBWEB**) algorithms. The choice of biomolecular systems to cluster includes MD simulation studies of a dynamic 10-mer polyadenine DNA single strand in aqueous solution, the interaction of the minor groove binding drug DB226 (the 3-pentyl derivative of 2,5-bis(4-guanylphenyl)furan) with a DNA hairpin loop in two different binding modes, and the conformational transition from an open to closed geometry of an drug-free cytochrome P450 2B4 structure (PDB: 1PO5).[76] In addition to the raw or production MD trajectory data, two artificial sets of data were constructed from independent trajectories of the polyA single strand to create trajectories containing 500 configurations at 1 ps intervals. The first represents five equally sized clusters created from 100 ps MD sampling around five distinct starting conformations, and the second is created from sampling around five distinct conformations to create clusters of different sizes, specifically containing 2, 15, 50, 100, or 333 configurations each.

When clustering the molecular configurations from a MD trajectory, ideally each clustering algorithm should group similar molecular configurations into distinct sets or groups. This gives a refined view of how a given molecule is sampling conformational space and allows direct characterization of the separate conformational substates visited by the MD.[77] As large-scale conformational change during the MD can lead to high variance for the calculation of time independent properties, such as MM-PBSA estimates of free energetics[3,78] or covariance estimates of the entropy,[79,80] it is expected that clustering of the trajectory into distinct substate populations can minimize this variance and provide more useful information about the ensemble of conformations sampled by MD. Clustering—no matter how valid in terms of its algorithmic success and ability to discern—is only useful if it can provide an unbiased means of exposing significant relationships and differences in the underlying properties. Ultimately, it is desired that an algorithm will naturally partition the data—with minimal user input—into representative clusters where each cluster may have different shapes, different variance, and different sizes. For example, structures sampled from a deep and narrow minimum energy well will typically have a smaller variance than those sampled from more flat and higher entropy wells. Clusters of configurations from MD simulation are also likely to have different sizes as sampling should ultimately progress according to a Boltzmann distribution, and, therefore, higher energy substates will be less populated than lower energy substates. In practice, except with artificially constructed and well-separated data, the performance of the underlying algorithm depends critically on the data, the pairwise comparison metric, the choice of atoms used in the comparison, and the choice of cluster count. With proper usage, we found that the clustering algorithms do seem to capture conformational substates of interest and that the clustering results highlight the similarity and differences among the structures. However, some of the clustering algorithms have key limitations and hence are not recommended for clustering MD trajectory data. Moreover, there appears to be no "one size fits all" clustering algorithm that always does an appropriate job of grouping the molecular configurations; in other words, the clustering algorithm ideally suited for clustering a particular data set will depend on the data.

To better characterize the relative performance, we implemented a range of different clustering algorithms. Assessment was made via visual inspection of the resulting clusters and also through the use of various clustering metrics. The algorithms chosen vary widely in their approaches, their computational complexity, their sensitivity to outliers, and their overall effectiveness. Our examination of several rather different clustering algorithms allowed us to quantitatively assess the quality of their output as well as their overall similarities and limitations. Surprisingly different behavior was observed in application of the different algorithms to the same MD simulation data. Whereas the fast and top-down divisive (or **hierarchical**) clustering algorithm tends to produce uniformly sized clusters or clusters with similar diameters, across the set of molecular configurations, various implementations of the bottom-up "merging" (or **linkage**) clustering algorithms tended to group most of the molecular configurations into a single large cluster with small singleton cluster outliers that contained only one or a few molecular configurations. Although the merging algorithms may produce singleton clusters, these algorithms can form clusters of any "shape" (such as elongated or concave clusters) in contrast to the **hierarchical** clustering algorithm. Depending on the data set, cluster count, and metric, differences in the relative performance of the various algorithms are clearly evident. The observation that clustering depends on the choice of algorithm strongly justifies the exploration of multiple clustering algorithms when initially characterizing the MD trajectory data. In addition to multiple algorithms, users need guidance on choosing the appropriate cluster count and atoms to use for the pairwise comparison. In general, the appropriate choices that will best partition the data are not known in advance. Strategies to assess the proper cluster count include dynamically choosing the number of clusters based on quantitative measures of clustering quality. Metrics investigated in this work include the pseudo F-statistic (pSF), the Davies-Bouldin index (DBI), the SSR/SST ratio, and the "critical distance". These indices and the detailed progression of the partitioning or merging can be cached, thereby allowing characterization of clustering performance across a range of cluster counts in a single clustering run. Further information about the relative performance and optimal choice of cluster count can come from visual examination of the tree of clusters. Finally, a significant concern when clustering based on pairwise distance evaluations is that the computational costs rapidly become excessive as the number of conformations to cluster grows. To partially mitigate this N^2 growth in computational costs and memory requirements, we implemented a two-pass "sieved" approach as a way to efficiently cluster many thousands of points. The MD trajectory is scanned first at a coarse level to do the initial clustering, with a second pass through the data to add skipped configurations to existing clusters. We examine the usefulness and limitations of these approaches.

Methods

Eleven different cluster algorithms[51] were implemented: **hierarchical**, single linkage (**edge**), average linkage (**average**), centroid linkage (**linkage**), complete linkage (**complete**), K-means (**means**), **centripetal**, **centripetal-complet**e, **COBWEB**,[81,82] **Bayesian**,[83] and self-organizing maps (**SOM**).[84] These were implemented in a library written in C, *libcluster*, which works abstractly on points and pairwise distances. It is not specific to MD simulations. By extending a few functions, such as the one that computes the centroid of a cluster, one can use this library to cluster arbitrary types of data. For instance, a separate program we developed, called *ClusterTest*, invokes *libcluster* to cluster collections of points in the plane. The *ClusterTest* utility can also measure distances between clustering outputs. In application to MD simulation, particular care needs be levied in calculation of the cluster centroid. Specifically, this relates to deciding the frame of reference for the averaging of conformations that form the centroid. In our initial development, the centroids produced were misleading as the molecules moved during the MD simulation and were not necessarily in the same reference frame as their centroid. To circumvent this problem, prior to construction of the centroid either the sampled configurations need to be placed into a common reference frame (such as by an RMSd fit to the first frame or a representative structure), or, better, separate reference frames should be created for each cluster where the frame of reference is the most representative configuration from that cluster. In the current implementation, all the structures in a given cluster are rms fit to the most representative structure before calculation of the centroid. The representative structure is the structure which has the minimal sum of the squared displacements between other structures in the cluster and itself. This is stored internally as the "bestrep" structure, and at present this is not necessarily equivalent to the representative structure output by **ptraj** (which currently writes out the structure closest to the centroid).

Code and Interface. All programs are written in portable C code and are available from the authors. For clustering MD trajectory data, this library was interfaced to the **ptraj** module of Amber.[85,86] We measured the distance between frames using mass-weighted, optimal-coordinate superposition root-mean-squared deviation (RMSd) or by using the distance measure D_{ab} (DME) defined by Torda and van Gunsteren.[52] Users can choose the subset of atoms to be used for pairwise comparison, specify the clustering algorithm and cluster count, and request to output new trajectories for each cluster, average structures for each cluster, and/or representative structures for each cluster. In this paper, each reference to distance indicates the RMSd between two simulation snapshots (i.e., two molecular configurations from different time points from the MD trajectory) unless otherwise specified.

Testing of the Implementation Using Points on a 2D Plane. To aid our analysis and algorithm development, we

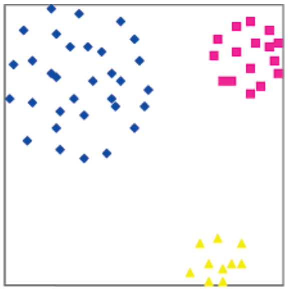


Figure 1. Clustering on a simple data set of points on a 2D plane. This data set is simple for several reasons: each cluster (represented by a different color) is convex, the clusters are clearly separated, nearby clusters do not differ greatly in size, and we requested the “right” number (three) of clusters when clustering. Changing any of these conditions will cause problems for some of the clustering algorithms.

utilized a very simple test set for evaluating the performance of the various clustering algorithms. This test uses points in the Euclidian plane where the pairwise metric is simply the distance between the points. Using either systematically placed (as shown in Figure 1) or randomly distributed points, it is very easy to construct and visualize the test set. As the data are somewhat contrived and not fully representative of the 3D configurations sampled in a MD trajectory, an algorithm’s good performance on the points in Euclidean space example does not guarantee it will classify simulation frames usefully. However, many properties—such as the relative memory requirements, the inability to generate concave clusters, or the sensitivity of **edge**-joining clustering to outliers—remain the same for any problem domain and are most easily discovered and visualized on a simple problem space. For simple test cases, where clusters are convex and clearly separated, all the algorithms perform equally well (see Figure 1). In other cases, the commonly applied algorithms (such as **hierarchical** clustering) break down.

In the following, we provide a general discussion of each of the clustering algorithms that were implemented. These common algorithms, or variants thereof, are classified as algorithms that are top-down (starting from a single cluster or divisive), bottom-up (starting from many clusters and merging or agglomerative), refinement (iteratively refining the membership of clusters starting from seed clusters), or tree based. A brief heuristic explanation of each is provided. A more technical description of each algorithm is provided in the Supporting Information.

Top-Down Clustering Algorithms. Top-down algorithms begin by assigning all points to one large cluster. They then proceed iteratively, splitting a large cluster into two subclusters at each stage. The cluster count increases by one at each step until the desired number of clusters is reached. **Hierarchical** clustering is the only top-down clustering algorithm we implemented. In our implementation, we defined the diameter of a cluster to be the maximal distance between any two points in the cluster. At each cycle, we find the cluster with greatest diameter. We split it around the two eccentric points that define the diameter, A and B: all points closest to A are assigned to one child cluster, and all points closest to B fall in the other.[52]

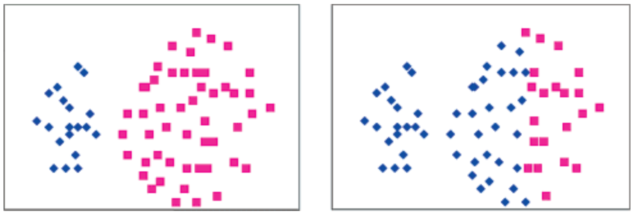


Figure 2. **Hierarchical** clustering (right) produces a nonintuitive clustering of the two distinct sets of points in the plane compared to the other algorithms (**centroid-linkage** clustering is shown on the left).

Hierarchical clustering tends to give clumsy results particularly near the boundary equidistant from the two eccentric points. Each “cut” made in hierarchical clustering may separate points near the boundary from their nearest neighbors. Hierarchical clustering can produce clusters of different population sizes (i.e., some with few points, some with many) but cannot produce clusters of greatly different diameters, such as might correspond to local energy minima of different depths (see Figure 2). This may or may not be mitigated by alteration of the refinement steps in the hierarchical algorithm to be cleverer about the “cut”. As implemented, in each of the refinement steps the cluster centroids are calculated, and the points are reassigned between the two new clusters. As will be seen later, this behavior differs from the refinement algorithms where reassignment of points can occur over all the clusters. This implies that the algorithm cannot overcome mistakes in partitioning made in previous steps. Hierarchical clustering is also sensitive to outliers since repositioning an extreme point changes the location of a boundary, and hierarchical clustering cannot produce concave clusters. Its main strengths are that it is the fastest of the clustering algorithms we examined at low cluster counts and changes in the performance metrics as a function of cluster count, such as the variance explained by the data or distance between split clusters, are easy to interpret.

Bottom-Up Clustering Algorithms. Bottom-up algorithms begin by assigning each point to its own cluster and proceed by iteratively merging clusters, one merge at each stage, into larger clusters until the desired number of clusters remains. Algorithmic differences relate to the specific choice of which pair of clusters to merge and the definition of the intercluster distance. **Edge or single-linkage**, **centroid-linkage**, **average-linkage**, **complete-linkage**, **centripetal**, and **centripetal-complete** clustering are the bottom-up algorithms implemented in this work. Bottom-up algorithms, like top-down algorithms, can produce a tree of clusters, where each “leaf” is a cluster, and the “root” is the cluster containing all points. An advantage of these methods is that the cluster merging information can be saved at each step to provide in a single run the set of distinct clusters that result across a range of cluster counts. Examination of these data in terms of the performance metrics can guide users to the appropriate cluster count for the data.

Edge. Under the single-linkage or “edge-joining” or **edge** algorithm, the distance from one cluster to another is defined as the shortest intercluster point-to-point distance. At each iteration step, the two closest clusters are merged. This merging continues until the desired number of clusters is obtained.[52] Centroid-linkage (or **linkage**) clustering is similar to single-linkage, except that the cluster-to-cluster distance is defined as the distance between the cluster centroids. **Average-linkage** (or **average**) clustering is also similar, except that the cluster-to-cluster distance is defined as the average of all distances between individual points of the two clusters. Complete-linkage (or **complete**) clustering defines the cluster-to-cluster distance as the maximal point-to-point intercluster distance between the two clusters. **Centripetal clustering** is derived from the CURE algorithm.[87] In centripetal clustering, we choose up to five "representatives" for each cluster. Representatives are taken as follows: choose up to five maximally distant points from the cluster and then move each point 1/4 of the way closer to the centroid to produce our representatives. This "centripetal" motion toward the centroid is intended to make the algorithm less sensitive to outliers. At each iteration step, the pair of clusters with the closest representatives is merged, and new representatives are chosen for the resulting larger cluster. The choice of five representatives and movement 1/4 of the way to the centroid is somewhat arbitrary. The centripetal clustering algorithms merging process is depicted graphically in Figure S0 of the Supporting Information.

Centripetal-complete is a variation on the centripetal algorithm that defines the distance between two clusters to be the largest distance ("complete") between the pairs of representative points from the two clusters (rather than the edge distance).

Refinement Clustering Algorithms. Refinement algorithms start with "seed" clusters. These seed clusters are refined, or "trained", over the course of one or more iterations through all the data points. After the clustering is determined to be good enough, or stable enough, the resulting clusters are saved. The number of clusters to form is set at the beginning and generally does not change during the refinement. These algorithms tend to depend on data presentation order and definition of the seed clusters. In our development, we evaluated the effect of the random (seed) and data order factors through multiple runs with different random seeds and comparison of chronological versus random ordering of the MD data. Means, Bayesian, and self-organizing maps are all refining algorithms. **Means** clustering starts by choosing a collection of seed points, each of which is assigned to its own cluster. We then iterate over all other data points. Each data point is assigned to the cluster whose centroid is closest; the centroid for this cluster is then recomputed.[88] To provide greater consistency between runs, we choose as our initial points a collection of maximally distant seed points, although random collections can also be used. **Bayesian clustering** starts with randomized seed clusters. A seed cluster has a random mean (and standard deviation) for each coordinate. Clusters are refined using an expectation-maximization (EM) algorithm. Points have probabilistic membership in each cluster. We first compute the odds that each point is in each cluster (the "expectation" step) and then alter the mean and standard deviation in the clusters to maximize the utility of each (the "maximization" step). This is based on the AUTOCLASS clustering algorithm.[88] In our experience, a large series of repetitive runs with different seeds need to be performed to get consistent results. **SOM**: Self-organizing maps are a form of artificial neural network. Each cluster is seeded with a random point, and the clusters are set up in a simple topology where each cluster has some "neighbor" clusters. The system is then run through several training cycles on the data. To process a data point, the most similar (closest) cluster is chosen. The coordinates of that cluster (and, to a lesser extent, its neighbors) are then shifted toward those of the training data point.[89]

Other Clustering Approaches. The **COBWEB**[81, 82] clustering system produces a tree describing the hierarchical relationships of members to their clusters. Each leaf node corresponds to a single point (or in the case of MD simulation, a single conformation or frame from the MD trajectory), and nonleaf nodes are clusters of all the descendant points. Points are placed in the tree by maximizing category utility (CU), a metric of cluster quality. Category utility is large for a cluster when the standard deviation of an attribute (over all points in the cluster) is smaller than the standard deviation of that same attribute in the cluster's parent.[88] Because of its unwieldy tree output, **COBWEB** results cannot be directly compared with those of other clustering algorithms. Although it is possible to "flatten" the tree into a standard partitioning of clusters, the straightforward flattening algorithm (choosing each merge in such a way as to maximize CU) may lead to terrible results, such as clusters consisting of disjoint batches of points. The thousand-node trees produced by **COBWEB** give a visualization of the relationships between MD configurations; however, they may be difficult to see and understand.

Clustering Metrics. To avoid bias, assessment via quantitative measures is desirable. Unfortunately, there is no universally accepted metric of "clustering quality". Despite this, metrics do provide a general indication of whether one clustering method is generally better than another.[49] In the current work we explored various distinct metrics, including the Davies-Bouldin index (DBI) and the pseudo F-statistic (pSF). DBI effectively measures the average over all clusters of the maximal values of the ratio that divides the pairwise sum of within-cluster scatter (where the scatter is the sum of the average distance of each point in the cluster from its centroid) by the intercluster separation.[90−93] It aims to identify clusters that are compact and well-separated. Low values of DBI indicate a better clustering. As the value of DBI is affected by cluster count, it makes sense to only compare DBI values for different clustering algorithms when the number of clusters is similar. Also, as the number of clusters decreases, the DBI value automatically tends toward smaller values. The pseudo-F statistic (pSF) is based on a comparison of intracluster variance to the residual variance over all points[94] and is determined from the classical regression model coefficients of SSR (sum of squares regression, or explained variation) and SSE (sum of squares error, or residual variation) through the ratio (for all points n and g clusters):

$$\mathrm{pSF} = \frac{\mathrm{SSR}/g - 1}{\mathrm{SSE}/(n - g)}$$

High values of pSF indicate better clustering. Since pSF sometimes rises with cluster count, one generally looks for a peak in pSF where the number of clusters is still manageably small. These metrics are imperfect. For instance, low DBI values result when the cluster algorithms produce several (likely uninformative) singleton clusters. On the other hand, pSF tends to give its highest scores when all clusters have approximately the same size, even if the clusters are badly formed. In our experience, using both metrics in conjunction with examination of the tree of clusters (see below) appears to be a promising way to assess clustering quality. Moreover, to assess cluster count, we can also use the "elbow criterion". This is a common evaluation tool that chooses the appropriate number of clusters by noting where adding in additional clusters does not add sufficient new information.[95] This can be seen by plotting the percentage of variance explained versus cluster count where a kink or angle in the graph (the elbow) illustrates the optimal cluster count. The percentage of variance explained by the data is the SSR/SST ratio where SSR is the sum of squares regression (or explained variation) from each cluster (summed over all clusters) and SST is the total sum of squares. The SSR/SST ratio is equivalent to the coefficient of determination or the R-squared value in classical regression. When this value is low, little variance is accounted for by the regression, and the clustering is likely poor. Another metric that provides insight into the proper cluster count is the critical distance. This is defined as the distance between the clusters that were just split or merged. The distance is different for each algorithm, as discussed previously; for example, the distance between clusters for the **centroid-linkage** algorithm is the distance between centroids, whereas for the **edge** algorithm it is the shortest point-to-point distance between clusters. Abrupt changes in the critical distance, as a function of cluster count, highlight optimal cluster counts. For example, if splitting a cluster leads to a significantly smaller critical distance than was seen previously, this suggests that the two new clusters are much closer together than clusters were in earlier splits and suggest that the split may have been unnecessary. The critical distance metric is not defined for the refinement algorithms.

One feature that emerged from the clustering of the real MD data is that the algorithms tended to group frames from a contiguous block of time together, even when sampling at 10−50 ps intervals. This is expected since with frequent sampling each simulation frame is necessarily close to its neighbors. However, the fact that clusters generally consist of frames from a single block of time shows that our sampling of conformational space may not be complete. Given a sufficiently long simulation, we would expect to see the system to revisit old clusters repeatedly (with sampling according to the Boltzmann distribution). As a rough quantification of this behavior, we define the "progress" of a cluster as $1 - S/E$, where S is the actual number of "switches" (i.e., the number of time points such that frames n and $n+1$ are in a different cluster), and E is the expected number of switches (based on the cluster's size and assuming random membership, $E = (n-1) * \Sigma(n_g/n * (n-n_g)/n)$ over clusters g). This number goes to zero as the actual number of switches approaches the expected number for a random distribution. The progress of large clusters for most of clustering is above 0.8, which means the continuous frames tend to be in the same cluster even at 10 ps sampling of the MD trajectory data. This observation can be used to guide choices of optimal cluster counts since the progress will likely decrease as the cluster count increases. Observation of progress values in the range of ~0.5 may suggest that the data are overpartitioned or poorly clustered. The subjective choice of a cutoff for progress values will depend on the sampling frequency and rate of conformational exchange and therefore cannot be considered in isolation; it is better to examine how the progress changes as a function of cluster count.

Molecular Dynamics Trajectories

A variety of different production-phase MD simulations were performed to provide the raw MD trajectory data used to test and validate the clustering algorithms. All of the MD simulations were performed with the Amber software suite.[86] For the simulations of nucleic acids in solution, a particle mesh Ewald treatment of the electrostatics (with less than 1 Å FFT grid spacing, cubic interpolation for the reciprocal space charge grid, a 9 Å direct space cutoff with the Ewald coefficient adjusted so that the direct space energy is less than 0.00001 kcal/mol at the cutoff, SHAKE[96] on all bonds with hydrogen, and constant temperature (300 K) and pressure (1 atm) with weak Berendsen scaling[97]) was applied. The all-atom Cornell et al. force field[98] for the DNA was applied with necessary supplemental parameters (for the bound drug) as outlined below and is available in the Supporting Information. Two distinct sets of MD data of nucleic acids in solution were investigated, specifically a rather dynamic trajectory of an "unfolded" polyA DNA strand (10-mer) sampled at broad (20 ps) intervals from more than 15 different trajectories (each starting from a different "unfolded" conformation) and also from an artificially created small trajectory that sampled around five different structures at 100 ps intervals (which should only produce "good" clustering with a cluster count of 5) and a dynamic trajectory of a DNA hairpin loop with the drug DB226 bound in the minor groove that shifts from one binding mode to another over the course of 36 ns of simulation. Additional biomolecular systems clustered include a ~75 ns simulation of a solvated mammalian cytochrome P450 with PDB entry 1PO5.[76] The relevant data for these systems are provided in the Supporting Information.

polyA Single Strand. Simulations were performed on a 10-mer polyadenine single strand of DNA. The initial model was built into an idealized B-DNA helix (of polyA-polyT deleting the polyT strand) using the Amber **nucgen** utility. The DNA was solvated with 2402 TIP3P[99] waters in a rectangular box (~53 Å × 42 Å × 35 Å with a 60 × 45 × 40 charge grid), and the charge was neutralized through the addition of nine Amber-adapted Aqvist sodium ions.[100] Simulations of the polyA single strand remain fully stacked and helical on a 5 ns time scale.[101] To investigate single strand structures more representative of the true ensemble

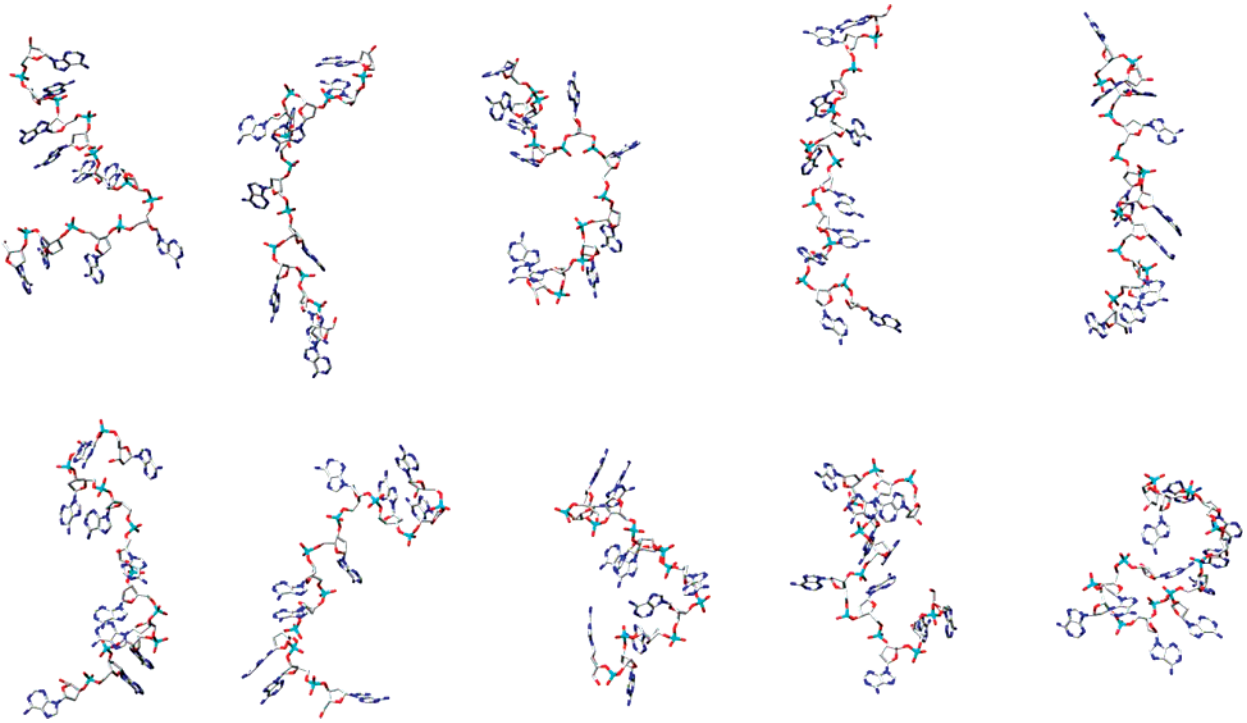


Figure 3. Snapshots from a 1050 ps self-guided MD simulation of a 10-mer polyA DNA single strand in explicit water (ions and water not shown) at 50 ps intervals (from left to right and top to bottom at 50 ps, 100 ps, 150 ps, 250 ps, 300 ps, 400 ps, 450 ps, 550 ps, 600 ps, and 650 ps) rendered with UCSF/Chimera.[106] Large guiding factors (0.5) and long (2 ps) local averaging times lead to considerable motion. When the SGMD is turned off, the single strands begin to "fold" into various stacked adenine structures.

and to generate a set of diverse conformations for clustering, the self-guided MD (SGMD) method was utilized with a guiding factor (0.5) and local averaging times (2 ps) significantly greater than are routinely applied (which are in the range of 0.1 and 0.2 ps, respectively).[102−105] When used in this manner, the SGMD rapidly moves the DNA and effectively samples a very wide range of "unfolded" conformations in short (1 ns) runs. Configurations from a 1050 ps SGMD simulation of this type, taken at 50 ps intervals (some of which are shown in Figure 3), were then run with standard MD protocols (Ewald and no SGMD) each on the 15−20 ns time scale. An aggregate trajectory for clustering was obtained by taking data from 15 of these trajectories at 20 ps intervals. As the starting configurations were widely different, this leads to a diverse set of single strand structures for clustering. In addition, an artificial trajectory of 500 frames was created from stable 100 ps regions of five of these independent trajectories. This creates a trajectory that should naturally split into five equally sized clusters. An additional 500 frame trajectory was created with unequally sized clusters of 2, 15, 50, 100, and 333 configurations, each sampling around distinct polyA geometries; this is a more difficult case to cluster as each resulting cluster has a different size. Moreover, the largest cluster samples multiple conformations and hence has relatively high variance compared to the smaller clusters. This is closer to what is expected for raw trajectory data; however, this trajectory is still easier to cluster than real MD trajectory data as there is no direct link or path between the clusters. During normal MD simulation and sampling on the picosecond time scale, the clusters are naturally linked due to the dynamics. As will be shown in the results, the contrived systems are easier to cluster. With real data, it is not obvious which algorithm is the best, and users likely have to explore multiple data clustering algorithms.

DNA Duplex-Drug Interactions. Simulations were performed on a model of the minor groove binding drug 2,5-bis[4-(N-alkylamidino)phenyl]furans (DB226)[107−109] bound to the ATTG region of a DNA hairpin loop. The DNA hairpin used has sequence 5′-CCA**ATTG**G-(TCTC)-C**CAAT**-TGG where the start binding site is indicated in bold and the loop is in parentheses. During the simulation the drug DB226 shifted back to the canonical **AATT** binding region. The hairpin DNA model was created by building an idealized B-DNA helix (for the full symmetric sequence d(CCAATTGGTC)$_2$) using the Amber **nucgen** utility followed by manually linking the two strands at one end. The model structure was relaxed with 1000 steps of steepest descent minimization (no cutoff) allowing only the six residues centered on the hairpin to move and 100 ps of dynamics with a generalized Born implicit solvent model (igb=1,[110,111] no cutoff, SHAKE on hydrogens,[96] 300 K with 1.0 ps coupling time with Berendsen temperature control[97]) allowing only the four loop residues to move. As this force field does not contain parameters for DB226, parameters (see the Supporting Information) were obtained using Antechamber[112] and the GAFF force field[113] using RESP charges[114] from a 6-31G* optimization with Gaussian 98.[115]

To build the initial model system, in analogy with the crystal structure of 2,5-bis(4-guanylphenyl)furan (furamidine) bound to the d(CGCGAATTCGCG)$_2$ dodecamer (PDB accession number 227D),[116] the 3-pentyl diamidine derivative

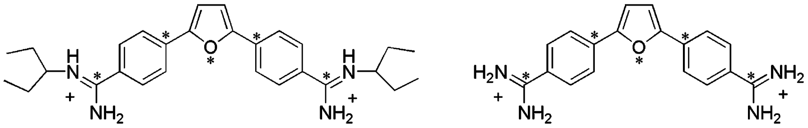


Figure 4. The molecular structure of DB226 (left) and furamidine (right). For full details on the parametrization, see the Supporting Information. The *'s denote atoms that were used for rms fits to the crystal structure during the initial docking, and the labels refer to atoms used for initial restraints to the DNA structure as discussed in the text.

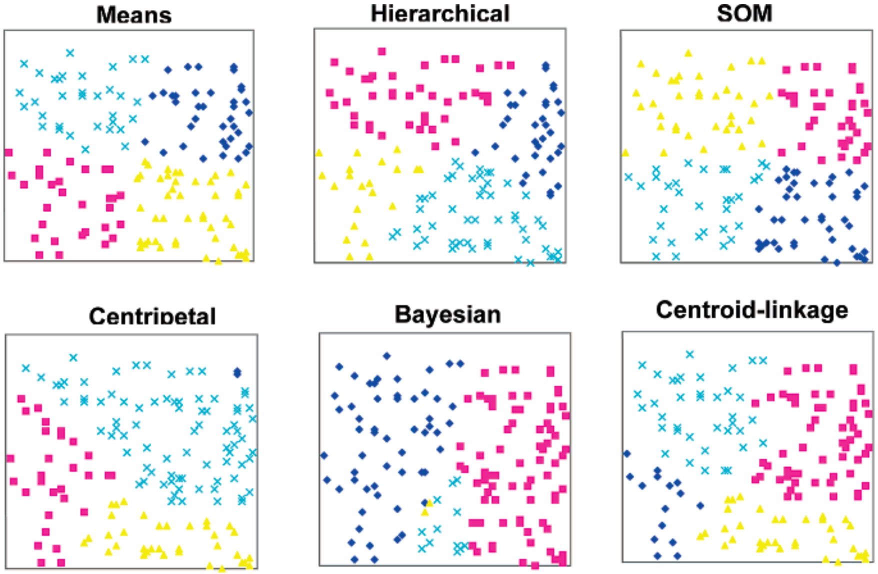


Figure 5. Clustering algorithms applied to the same set of random points in the 2D plane. The results highlight the features of six of the distinct clustering algorithms investigated, such as the uniform/linear cutting of the **hierarchical** clustering, the uniform sizes of the clusters created by the **means** clustering, and the ability of **centroid-linkage**, **centripetal**, and **Bayesian** algorithms to create clusters with distinct shapes and sizes.

of furamidine (DB226)[108,109] was hand-docked into the ATTG region. This was done by a rms fit of the Gaussian-optimized geometry of DB226 to the crystal structure of bound furamidine (using the five atoms denoted with an asterisk in Figure 4) and of four ATTG binding-site phosphates in the DNA hairpin to the crystal structure binding site. The system was then minimized for 500 steps using the steepest descent method in vacuo with no cutoff, followed by 100 ps of generalized Born implicit solvent simulations as above (except with a temperature coupling time of 10.0 ps). Distance restraints were applied (both to the heavy atoms and the hydrogens with a flat-well potential from 2.0 to 3.0 Å or 1.0 to 2.0 Å, respectively, with a 5.0 kcal/rad^2-mol lower bound force constant to 0.0 Å and a 15.0 kcal/rad^2-mol force constant beyond the upper bound) for DB226 atom N4 to O2 of base T17, N2 to N3 of base T7. Harmonic positional restraints with a force constant of 5.0 kcal/Å^2-mol were applied to the DNA duplex. Note that in early simulations of this system, where no restraints were applied during the in vacuo equilibration, the ligand either shifted to an alternate binding mode or escaped the groove entirely. This reinforces the need to be careful when initially setting such systems to avoid artifacts due to the equilibration and initial modeling procedure.

The system was then solvated with explicit TIP3P water[99] in a truncated octahedron periodic unit cell to a distance of 9 Å. Explicit net-neutralizing Na$^+$ and an additional 12 Na$^+$ and Cl$^-$ ions were added to bring the system to a salt concentration of ∼100−150 mM. The water and counterions were allowed to equilibrate via the same minimization and relaxation steps, with the DNA and ligand fixed. Finally, production MD was run for more than 36 ns.

Results

Clustering Points in the 2D Plane. To illustrate differences in the clustering algorithms, Figure 5 shows the performance of various clustering algorithms when applied to the same randomly selected collection of points in the 2D plane. Visualization of the data for each algorithm (run with a cluster count of four where each cluster is denoted by a different symbol and color) shows the significant variation in the cluster sizes and shapes. Each algorithm clusters the same data in very different ways. The properties of the various algorithms, such as the ability to handle cluster convexity and the preferences toward producing clusters of similar sizes, tend to carry over to other problem domains.

Unlike the data shown in Figure 1, the random distribution of points shown in Figure 5 does not have an obvious partitioning. How these data are clustered will depend on the details of the algorithm and how intra- and intercluster separation and variance are determined. As each algorithm is different, it is not surprising that rather different sets of clusters emerge with the different algorithms. The **means**, **hierarchical**, and **SOM** clustering algorithms tend to produce clusters of similar size with a linear partitioning of the data. The **centripetal** and **Bayesian** clustering algorithms are able

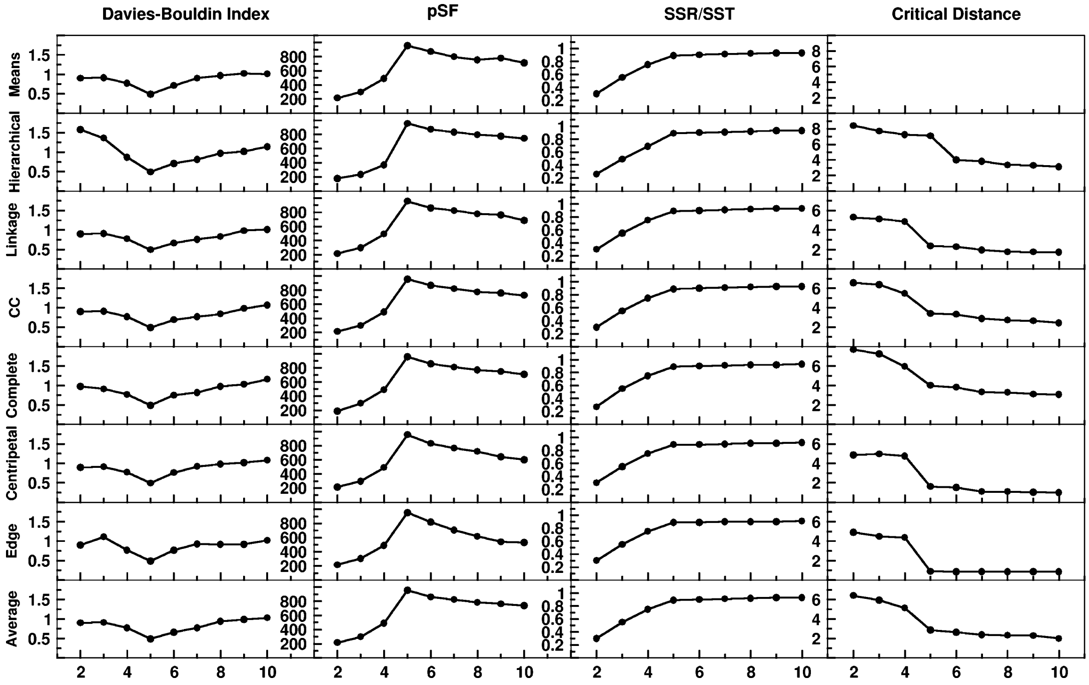


Figure 6. Cluster metrics for a subset of the algorithms investigated for the constructed polyA trajectory of five distinct equally sized clusters as a function of cluster count (*x*-axis). At the optimal cluster count of 5, DBI is at a minimum, pSF is at a maximum, the SSR/SST values plateau, and a transition occurs in the critical distance. "CC" is the centripetal complete clustering algorithm, and a critical distance is not shown for the **means** refinement clustering algorithm since it is ill-defined. Note that for the **means** refinement, five independent clustering runs (with random choices of configurations for the refinement steps) were performed, and the data shown are for the run with the highest pSF value.

to produce both small and large clusters. The **centroid-linkage** is able to produce clusters of very different shapes. The **centripetal** algorithm is able to associate distant points into a cluster despite having very few points near the "centroid" due to the use of representative points distant from the centroid. With the exception of the **Bayesian** algorithm, all shown tend to naturally partition the data.

Clustering Artificial MD Data: Five Equally Sized and Distinct Clusters. After development and testing of the algorithms on the points in the plane examples, clustering was performed on a series of MD trajectories using both the RMSd and the DME as a metric and a series of independent runs varying the cluster count and other variables. To demonstrate the results, two trajectories of 500 configurations from the polyA single strand MD simulations were created and then clustered. In each case, these trajectories were created from independent runs and sampling around five distinct conformations. The first test set has 100 configurations for each distinct conformation leading naturally to a partitioning into five equally sized clusters. Clustering metrics as a function of the cluster count are shown in Figure 6. The metrics show the expected (idealized) behavior including a minima in the DBI, maxima for pSF, a plateau in SSR/SST, and a sudden drop in the critical distance when a cluster count of 5 is reached. Beyond five clusters, little new information is gained from further partitioning of the data. The behavior of the critical distance at the transition point around the optimal cluster count is effectively opposite for the top-down compared to the bottom-up algorithms. In the case of the bottom-up algorithms and cluster merging, there is a sudden jump as the cluster count goes from 5 to 4; this indicates that the distance between the newly formed clusters is much larger than the distance (variance) between previous clusters. With the top-down or hierarchical algorithm, the change in the critical distance occurs as we split clusters from the optimal count of 5 to 6; this leads to a drop in the critical distance suggesting that the split leads to clusters that are significantly closer together than the previous clusters were. As an indicator of the proper cluster count, the drop in the critical distance occurs at the proper cluster count for the bottom-up algorithms and just after the proper cluster count for the top-down algorithms.

In general for this artificial data set most of the algorithms perform equally well. The exceptions are the **Bayesian** and **COBWEB** clustering algorithms which yield some of the expected 100-member clusters in some cases but incorrectly split other clusters; this contrasts with the **SOM** algorithm which correctly generates the five expected 100-configuration clusters. An additional limitation of the **SOM** and **Bayesian** algorithms that was uncovered is that both algorithms may fail to generate the expected cluster count (i.e., the algorithms can form clusters that contain no points). In some cases, when more than 5 clusters are requested, the **SOM** algorithm will yield only the 5 expected clusters. This property may be exploited to determine optimal cluster count. For more data on the **Bayesian**, **COBWEB**, and **SOM** clustering results refer to Table ST2 in the Supporting Information.

Clustering Artificial MD Data: Five Differentially Sized Clusters. The second set of artificial MD trajectory data was also constructed from sampling around five distinct conformations, but each cluster was constructed to be a different size, specifically with 2, 15, 50, 100, or 333

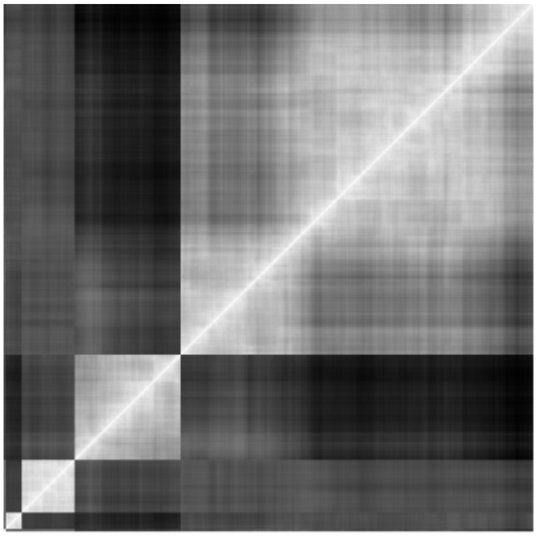


Figure 7. 2D RMSd plot (mass weighted) for all frames from the polyA single strand simulation with five differentially sized clusters. Note that only four clusters are readily visible since the first cluster is very small.

configurations in five separate clusters. This set is more difficult to cluster as it has both very small clusters with small variance and relatively large clusters with larger variance. The average distance of each conformation in the cluster to its centroid spans a large range. These average values for each cluster size : distance pair are as follows: 2 : 0.27 Å; 15 : 0.72 Å; 50 : 0.84 Å; 100 : 1.62 Å; and 333 : 2.34 Å. The maximal pairwise best fit RMSd between any two conformations is 9.8 Å. Although the intent was to create five clusters, the natural partitioning may be closer to six. This can be seen clearly from visualization of the 2D RMSd plot or effectively the visualization of the matrix of pairwise best fit RMSd values of every structure to every other structure for all the conformations (see Figure 7). The plot shows that each cluster is dissimilar from its neighbors and also that the largest cluster may best be represented by two similar clusters. The diagonal elements (white) represent self-comparison or zero RMSd and the black shows the largest pairwise RMSd of 9.8 Å. As the members of the smallest cluster are distinct from the others, partitioning ideally should create the small clusters before breaking the largest cluster. With a cluster count of 5, this does not happen with the **centripetal complete**, **COBWEB**, **complete**, **hierarchical**, **means**, and **SOM** algorithms. The cluster sizes for each of these algorithms, italicizing the incorrect cluster sizes, are **centripetal complete**: 15, *52*, 100, *106*, *227*, **COBWEB**: 50, *63*, *115*, *117*, *155*, **complete**: 15, *52*, 100, *137*, *196*, **hierarchical**: 15, *52*, 100, *112*, *221*, **means**: 15, *52*, 100, *113*, *220*, and **SOM**: *67*, *95*, 100, *114*, *124* algorithms. In most of these cases, the largest and smallest clusters are not found. **Edge**, **centripetal**, **average-linkage**, and **linkage** each partition the data as expected into five distinct clusters; when a cluster count of 6 is specified, the largest cluster is broken in two with each of these algorithms, although the sizes are different (**centripetal** 114, 219; **edge** 70, 263; **average-linkage** 102, 231; **linkage** 106, 227). The **centripetal complete**, **complete**, **means**, and **hierarchical** recover the natural partitioning when a cluster count of six is specified. This is not true of the **Bayesian**, **SOM**, or **COBWEB** algorithms, and the “small” cluster of 2 conformations is not found until a cluster count of 10 is reached with the **Bayesian** algorithm. The **SOM** and **COBWEB** algorithms appear to be unable to produce the smallest cluster. In the Supporting Information, trees outlining the partitioning of conformations by each of the algorithms are displayed. The **centripetal**, **edge**, **average-linkage**, and **linkage** algorithms show the best partitioning.

Shown in Figure 8 are the clustering metrics for some of the algorithms, omitting the poorly performing **Bayesian**, **SOM**, and **COBWEB** algorithms, on this artificial polyA MD trajectory with varying cluster sizes. It is clear from these data that when the configurations are not uniformly separated into similarly sized clusters, the metrics are less consistent across algorithms and also less informative. For many of the algorithms, a clear minimum in DBI or maximum in pSF is not readily evident. Similarly, rather than showing a clear elbow in the SSR/SST or critical distance plots as a function of cluster count, a smoother linear plot is often evident. These data can also be misleading. For example, the **hierarchical** clustering shows a clear minimum in DBI at a cluster count of 5 or 6, a maxima in pSF at a cluster count of 5, and a clear kink in the SSR/SST plot at a cluster count between 4 and 5, yet this algorithm does a poor job of clustering into five clusters. At a cluster count of 5, the **hierarchical** algorithm has already split the largest cluster but not split the 50 + 2 cluster into two separate clusters. Moreover, although the **centripetal** algorithm shows excellent clustering, the performance is not readily evident from the DBI, pSF, and SSR/SST metrics. The most definitive demonstration of metric success comes with the **edge** algorithm; this algorithm very naturally partitions the data.

Shown in the Supporting Information are schematics of the cluster trees or partitioning by various algorithms (Figures S2−S11) and the distinct performance metrics (Table ST3) for the **Bayesian**, **COBWEB**, and **SOM** algorithms.

As MD ideally samples according to a Boltzmann distribution, it is expected that the data will look more like the artificial trajectory with differentially sized clusters than data that partition into equally sized clusters. This is because the population of a given conformer is related to its free energy, with lower populations as the energy increases. This suggests that finding the ideal partitioning and clustering of the data will be messier with real data and that the various algorithms will each lead to distinct partitioning of the data.

Clustering Real MD Data: Simulation of Drug−DNA Interaction. A series of MD simulations were performed on a series of minor-groove binding agents binding into the minor grooves of various DNA hairpin sequences. The specific MD trajectory of DB226 binding to the ATTG sequence of a hairpin DNA was chosen for clustering and further analysis. This is an interesting case as the simulation revealed a major shift, by one base pair step, in the binding of the minor groove binding drug DB226 to the DNA hairpin. As this change in binding is easy to visualize, this is a good test of the clustering algorithms ability to discern and to naturally partition the data. To cluster the MD trajectory data,

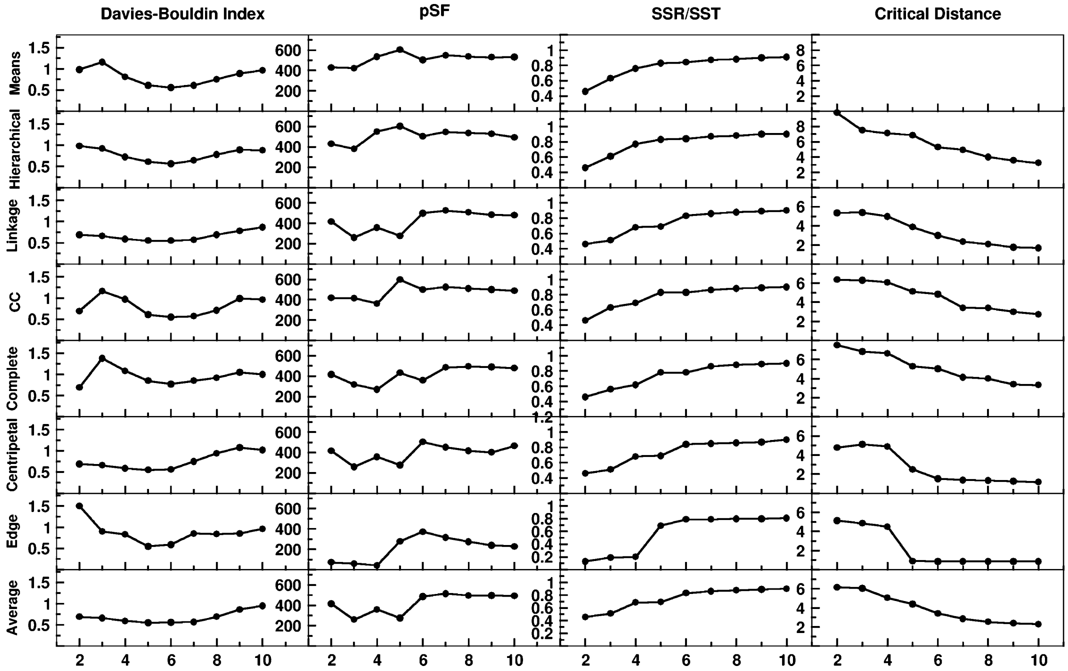


Figure 8. Cluster metrics for a subset of the algorithms investigated for the artificially constructed polyA trajectory representing five clusters of distinctly different sizes as a function of cluster count (*x*-axis). Note that for the **means** refinement, five independent clustering runs (with random choices of configurations for the refinement steps) were performed, and the data shown are for the run with the highest pSF value.

each algorithm was applied to the drug−DNA hairpin MD trajectory over 36 ns with configurations taken at 10 ps intervals, for a total of 3644 frames. For each of the algorithms investigated, a range of cluster counts from 2 to 20 was evaluated. To limit the structure comparisons to the binding region of the drug−DNA complex, the clustering metric used was the best-fit RMSd of the residues defining the drug and the binding region. Specifically, this included all the carbon, nitrogen, and oxygen atoms of the drug and from residues 3−8 and 14−19 of the DNA, i.e., the AATTGG binding region noting that the drug initially binds at the ATTG site. Figure 9 shows the shifting of DB226 to the AATT site that occurs during the MD simulation between 15 and 16 ns. The MD results suggest that multiple modes for DB226 binding to the DNA hairpin are thermally accessible. From the plot of the distances versus time shown, it appears that the drug attempts to shift down a base pair step at ∼6 ns, but moves back to the ATTG site, and then eventually successfully fully shifts by one base pair step by ∼16 ns. Additional data and discussion, including plots of the overall RMSd versus time (Figure S12) and molecular graphics of average structures before and after the change in drug binding (Figure S13) and the clustering data across the different cluster counts (Tables ST4−ST6), are shown in the Supporting Information.

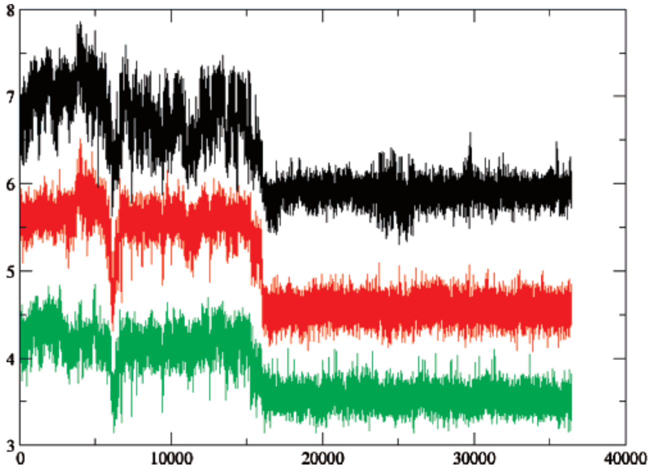


Figure 9. Graph of selected atom positions of the minor groove binding drug DB226 relative to the base pair step (*y*-axis, base pair step number) as a function of time (*x*-axis, in ps). In each frame, we calculated the least-squares fit plane for each base pair and averaged the normal vectors perpendicular to those planes. From this, we can interpolate the position of an arbitrary atom based on the midpoints of those least-square fit planes. The red (middle) represents the furan oxygen atom in the middle of DB226. The black and green represent the two guanyl nitrogen atoms at each side of the drug. From the plot, the shifting of DB226 from the ATTG region (base pair steps 4−7) to the AATT region (base pair steps 3−6) is evident.

Relative Performance of the Clustering Algorithms. The data in Table 1 provide a summary of the relative performance and properties of the various clustering runs as a function of cluster counts. Included are the runtimes, DBI, and pSF metrics, the SSR/SST ratio or R-squared value, the "progress" of the simulation, and the sizes of the resulting clusters. The full data, including cluster counts of 10 and 20, are in the Supporting Information (Table ST4−ST6). The runtime in the table is the time for running each clustering algorithm on an equivalent machine. The actual runtime will be increased by the time needed to precalculate the full pair wise distance matrix which amounted to 367 s for the

Table 1. Comparison of the Various Clustering Algorithms Applied to a 36 ns MD Trajectory of DB226 Bound to Hairpin DNA[a]

algorithm	cluster	runtime	DBI	pSF	SSR/SST	progress	cluster sizes
average	3	6197	1.10	1545.23	0.459	1.00	2061, 1581, 2
average	4	6279	1.59	1119.57	0.480	1.00	2061, 1454, 127, 2
average	5	6387	1.56	925.33	0.504	0.99	2061, 1340, 127, 114, 2
Bayesian	3	59	1.91	1820.10	0.500	0.89	2034, 929, 681
Bayesian	4	94	2.37	1395.03	0.535	0.73	1054, 996, 952, 642
Bayesian	5	120	2.17	1187.41	0.566	0.76	1366, 786, 689, 476, 327
centripetal	3	3272	1.14	108.29	0.056	0.98	3516, 127, 1
centripetal	4	3351	1.02	72.65	0.056	0.97	3516, 126, 1, 1
centripetal	5	3103	0.98	54.83	0.057	0.96	3515, 126, 1, 1, 1
CC	3	1516	1.54	1470.70	0.447	0.98	2148, 1492, 4
CC	4	1477	1.68	1066.73	0.468	0.98	2148, 1395, 97, 4
CC	5	1572	1.30	801.19	0.468	0.98	2148, 1395, 97, 3, 1
COBWEB	3	1854	1.87	1757.04	0.491	0.77	1594, 1109, 941
COBWEB	4	1236	2.12	1378.27	0.532	0.86	1804, 780, 764, 296
COBWEB	5	1221	2.88	1071.62	0.541	0.63	1025, 780, 764, 568, 507
complete	3	922	1.86	1585.13	0.465	0.85	1703, 1407, 534
complete	4	960	2.16	1163.30	0.490	0.83	1703, 1216, 534, 191
complete	5	1398	2.35	931.67	0.506	0.83	1703, 1060, 534, 191, 156
edge	3	923	0.54	3.43	0.0019	0.25	3642, 1, 1
edge	4	931	0.77	3.43	0.0028	0.37	3640, 2, 1, 1
edge	5	930	0.78	2.92	0.0032	0.30	3639, 2, 1, 1, 1
hierarchical	3	6	1.80	1898.54	0.510	0.91	2088, 960, 596
hierarchical	4	7	1.86	1362.83	0.529	0.90	2088, 960, 304, 292
hierarchical	5	9	2.13	1199.49	0.568	0.77	1350, 960, 738, 304, 292
linkage	3	2349	0.93	1544.21	0.459	0.99	2077, 1566, 1
linkage	4	2158	1.06	1035.76	0.460	0.99	2077, 1562, 4, 1
linkage	5	1782	1.07	805.17	0.469	0.99	2053, 1562, 24, 4, 1
means	3	1105	1.80	1899.67	0.511	0.91	2088, 965, 591
means	4	953	2.11	1490.24	0.551	0.79	1402, 967, 702, 573
means	5	909	2.02	1222.68	0.573	0.78	1322, 891, 756, 454, 221
SOM	3	663	1.70	1597.10	0.467	0.98	2066, 1546, 32
SOM	4	1391	1.96	1149.12	0.486	0.93	2035, 1396, 134, 79
SOM	5	1558	2.13	1059.96	0.538	0.74	1238, 1100, 956, 212, 138

[a] The RMSd of carbon, nitrogen, and oxygen atoms of the drug and residues 3−8 and 14−19 of the DNA hairpin were used as the pairwise distance between all configurations from the MD simulation at 10 ps intervals. The cluster count represents how many clusters were chosen. The DBI and psF values are metrics of clustering quality; low values of DBI and high values of pSF indicate better results. The R-squared (SSR/SST) value represents the percentage of variance explained by the data; plots of this as a function of cluster count can show where adding more clusters fails to add new information shown by the elbow criteria or a kink in the plot. The “progress” as discussed in the Methods section describes how often switching between the clusters is occurring. Larger values imply that most of the cluster members are sequential in time with values of 1.0 meaning all frames in a cluster are contiguous. **Edge** and **centripetal** clustering produced optimum values of DBI but generated pathological singleton clusters. **Centroid-linkage** clustering performs well under both metrics. **Hierarchical** clustering is the fastest of the algorithms applied. **CC** refers to **centripetal-complete** clustering. The **SOM**, **means**, **Bayesian**, and **COBWEB** algorithms were each run five times, and the results with the highest pSF are shown.

clustering data in Table 1. The time for calculating a DME pair wise distance matrix increases significantly as the number of atoms in the comparison increases. The DME matrix preparation time for the same atoms is 24 519 s. This also significantly increases the runtime for the algorithms which need to recalculate at each iteration the DME distance between the cluster centroid and every other clusters centroid, such as with the **centripetal** and **linkage** algorithms.

Clearly the fastest algorithm at low cluster counts is the **hierarchical** clustering and the most computationally demanding algorithms are **average linkage**, **centripetal**, **linkage**, the neural net refinement (**SOM**), and **COBWEB** algorithms. With the exception of the **Bayesian** and **SOM** refinement algorithms, the relative cost does not tend to increase dramatically as the number of clusters goes up.

In terms of the clustering quality metrics, algorithms that have high pSF and low DBI values at a given cluster count suggest better clustering. For all the algorithms applied to the MD trajectory of the drug bound to the DNA hairpin, where the similarity was ascertained by fitting to atoms in the binding region, this mix occurs at a cluster count of between 2 and 4. Based solely on pSF and DBI, a cluster count of 2 is suggested by the data; however, the SSR/SST ratio and critical distance plots suggests a count closer to 5

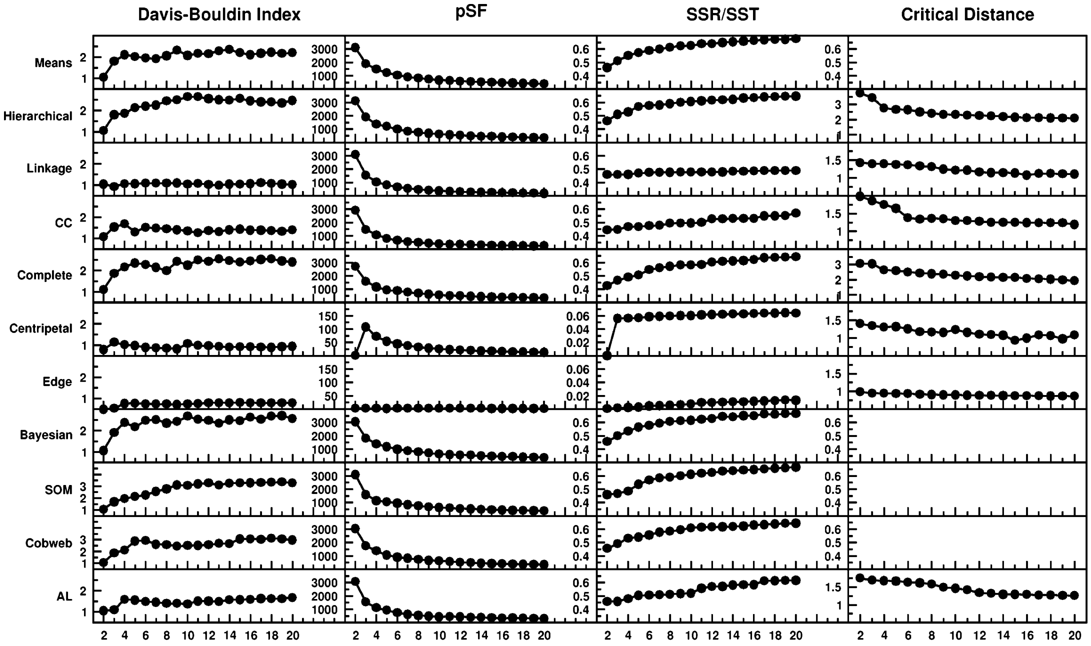


Figure 10. Cluster metrics for the MD trajectory of drug−DNA interaction as a function of cluster count (*x*-axis). Note that the scales of SSR/SST are different for the **centripetal** and **edge** clustering algorithms and that the scales for the critical distance are different for the **hierarchical** and **complete** clustering algorithms. The critical distance is not defined for the refinement algorithms and hence is not shown in these cases.

or 6. As the cluster count goes up, DBI is relatively constant, pSF gets smaller, and more information is added according to the SSR/SST. This is likely characteristic of MD simulation data as finer grained partitioning among the dynamic continuum of states is possible until all of the substates have been defined; as the potential energy surface is rough and there are many degrees of freedom, there are likely many different substates defining the path of the molecule in time. Although each algorithm suggests an optimal cluster count somewhere in the range of 2−6, the resulting cluster sizes vary considerably, and this impacts the relative performance of each. Most *inconsistent* with the natural partitioning of the data are the results from the **centripetal** and **edge** algorithms. These display a single large cluster with either small or singleton clusters outliers. In these cases, high pSF values are not obtainable, and the “progress” of 1.0 implies that each cluster has contiguous frames in time. Considering the data shown in Figure 9 and in the Supporting Information, and given our knowledge that two distinct binding modes for the drug were explored in the MD, we would expect the natural partitioning of this data to include the two distinct binding modes with drug binding to the ATTG and AATT binding sites. Shown in Figure 11 is a plot of the 2D RMSd values over the binding region during the ∼36 ns of simulation at 100 ps intervals. Two clear clusters are evident. The partial transition to the alternate binding mode at ∼6 ns is evident as the horizontal and vertical light lines which show agreement of frames from early in the trajectory (∼6 ns) with those from later in the trajectory. From the 2D rms plot, the next partitioning could split up either the smaller or larger cluster. Starting from the lower left (or the early part of the trajectory), the first cluster should have ∼1600 frames and the second ∼2040. This partitioning with a cluster count of three is seen with the **hierarchical** algorithm and cluster counts of 596, 960, and 2088. The highest pSF values are observed with **hierarchical** and **means** clustering algorithms. **Average-linkage** and **linkage** both obtain a low DBI and high pSF; however, the SSR/SST plot is essentially flat. The reason the plot is flat beyond a cluster count of 2 is that only clusters with very few configurations are newly formed.

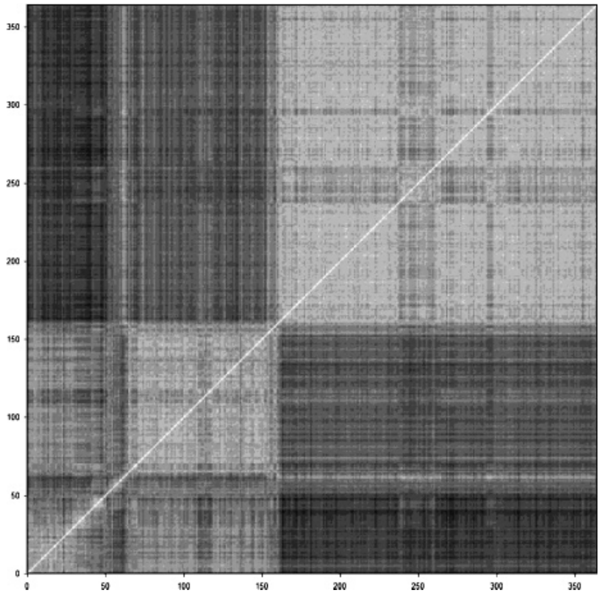


Figure 11. 2D RMSd plot (mass weighted) for frames at 100 ps intervals from the 36 ns simulation of DB226 binding to the DNA hairpin.

It is important to note that the performance of a given clustering algorithm is affected by the character of the data under consideration. In the case of the molecular dynamics trajectories of DNA interacting with DB226, the data did not include extreme outlier points. As a quick screen for outliers, we examined the rmsd from each simulation frame to its nearest neighbor. For the major heavy atoms in the binding region in the trajectory, these nearest-neighbor distances were distributed between 0.4 and 1.2 Å, mostly around 0.8 Å. If we call the rms deviation from one frame to the next the "velocity" of a simulation, it is reasonable to suspect that early equilibration stages will have a relatively high velocity and therefore account for the bulk of the variation between clusters. This is one reason why it is best, when clustering MD trajectories, to exclude the initial equilibration portion of the simulation. For our DB226 trajectory, the equilibration protocol was successful in starting the system in a reasonable state—the velocity is consistent over time.

Distances between the Various Clustering Algorithms. In addition to evaluating the performance of a single clustering algorithm, we can measure the distance between two sets of different clusters of the same data produced by different clustering algorithms. This provides a measure of the disagreement between different clustering algorithms. One reasonable approach to measure the distances between clusters is to compute the rms distance between cluster representatives from the different sets of different clusters. However, in practice this is tricky as it requires guesswork to set up the correspondence between the clusters in each distinct set. To avoid this problem, we devised the following metric, $\partial(A,B)$. To measure the distance between clustering A and B, we consider the set of all pairs of points being clustered. We say that A and B agree on a pair of points if both A and B assign the points to the same cluster. If one of A or B assigns the pair of points to the same cluster but the other does not, this is counted as a disagreement. We compare the actual number of disagreements to the number of disagreements we would expect to see if A and B were unrelated. To do this, we first note the odds that two randomly chosen points will fall in the same cluster in A:

$$P_A = \sum_k \frac{S_k(S_k - 1)}{n(n-1)}$$

Here n is the number of points, and S_k is the size of cluster k. Now the expected number of disagreements (ED) can be computed:

$$\mathrm{ED}(A,B) = C^*(P_A(1 - P_B) + P_B(1 - P_A))$$

Here C is the number of pairs of points, and P_A and P_B are the probability that a pair of points falls in the same cluster in A and B, respectively.

Similarly, we can compute the expected number of the true agreements (EA), where a pair of points is both in the same cluster in clustering A and clustering B:

$$\mathrm{EA}(A,B) = C^*P_A{}^*P_B$$

As a finer-grained metric, we define a function ∂ measuring the ratio of the actual and expected number of disagreements over the ratio of the actual and expected number of true agreements:

$$\partial(A,B) = \frac{\dfrac{\mathrm{AD}(A,B)}{\mathrm{ED}(A,B)}}{\dfrac{\mathrm{AA}(A,B)}{\mathrm{EA}(A,B)}}$$

This distance function has several intuitive properties: $\partial(A, B) = \partial(B, A)$, and $\partial(A, A) = 0$, as we would expect. For random clustering C and D, $\partial(C, D)$ is very near 1. In general, a small distance reading (<~0.2) indicates very similar clustering, such as is obtained when the same algorithm is used to generate a similar number of clusters. Readings less than ~0.6 indicate some similarity, and distances greater than ~0.6 indicate low levels of agreement.

The distance metrics, $\partial(A,B)$, for the various algorithms applied to clustering the DNA hairpin-DB226 trajectory are displayed in Table 2 and can be used to compare the output of the different algorithms. The data suggest that the **average-linkage**, **linkage**, and **centripetal-complete** algorithms generate very similar clusters. Similarly, the **Bayesian**, **SOM**, **hierarchical**, and **means** clustering all give related results, whereas the **complete** and **COBWEB** algorithms only generate somewhat similar sets of clusters when compared to the other algorithms. On the contrary, due to the production of singleton clusters, the resulting sets of clusters from the **edge** and **centripetal** algorithms are very dissimilar to the sets of clusters that result from the other algorithms under this distance metric. Direct comparisons of a given algorithm's clustering of the data using the two distance metrics, RMSd and DME, indicate that the resulting difference is fairly small—usually less than 0.2. However, again, for **edge** and **centripetal** algorithms, the differences go beyond 0.5. This indicates that both metrics (RMSd and DME) are capturing the conformational changes of the molecular system, though not in precisely the same way.

A useful application of this metric is to quantify the consistency between different runs of the stochastic clustering algorithms. The **SOM** clustering algorithm was run five times on the same system, and the distance between each resulting set of clusters was compared. The average distance was 0.03, indicating that **SOM** is very consistent. Similar runs for **means**, **Bayesian**, and **COBWEB** clustering yielded an average distance of 0.07, 0.15, and 0.25, respectively. Thus, it appears that the **SOM** and **means** algorithm provides more consistent (if not always better) results than those produced by the **Bayesian, COBWEB** clustering algorithm.

The Choice of Atoms for the Pairwise Comparisons. As might be expected, the clustering outcome is strongly influenced by which atoms are used to determine the similarity of the different molecular configurations. If too small a region is chosen, the fine partitioning that results may not be meaningful, similarly, choosing too many atoms may hide conformational substates visited by substructures during the MD. With the DB226-DNA hairpin trajectory, the distances between clusters are strongly influenced by the choice of atoms. For example, distances between clustering only the DB226 atoms compared to all solute atoms are generally above 0.9 for the bottom-up algorithms (i.e., **linkage**, **edge**, **complete**, **centripetal**) and ~0.3−0.4 for **SOM**, **hierarchical**, **means**, and **Bayesian** algorithms. If

Table 2. Distances between the Various Clustering Algorithms, in the Dimensionless Metric Defined in the Text That Compares the Ratio of the Actual to Expected Agreements and Disagreements[a]

RMS	average	linkage	CC	SOM	hierarchical	means	Bayesian	COBWEB	complete	centripetal	edge
average	**0.000**	**0.039**	**0.071**	0.169	0.182	0.205	0.240	0.245	0.209	0.887	0.994
linkage		**0.000**	**0.085**	0.169	0.196	0.224	0.255	0.259	0.234	0.951	0.996
CC			**0.000**	0.203	0.212	0.243	0.278	0.283	0.255	0.872	0.996
SOM				**0.000**	**0.137**	**0.154**	**0.158**	0.163	0.227	0.968	0.999
hierarchical					**0.000**	**0.064**	**0.130**	0.202	0.211	0.892	0.997
means						**0.000**	**0.100**	0.197	0.200	0.909	0.997
Bayesian							**0.000**	0.204	0.239	0.928	0.998
COBWEB								0.000	0.242	0.966	0.998
complete									0.000	0.914	0.997
centripetal										0.000	0.978
edge											0.000

[a] The trajectory being clustered was DB226 bound to hairpin DNA. Distances for cluster counts of 3, 4, 5, and 10 were computed and averaged. Clustering was performed using RMSd distances as the pairwise metric, the refinement algorithms were each run five times, and the data with the highest pSF values were used.

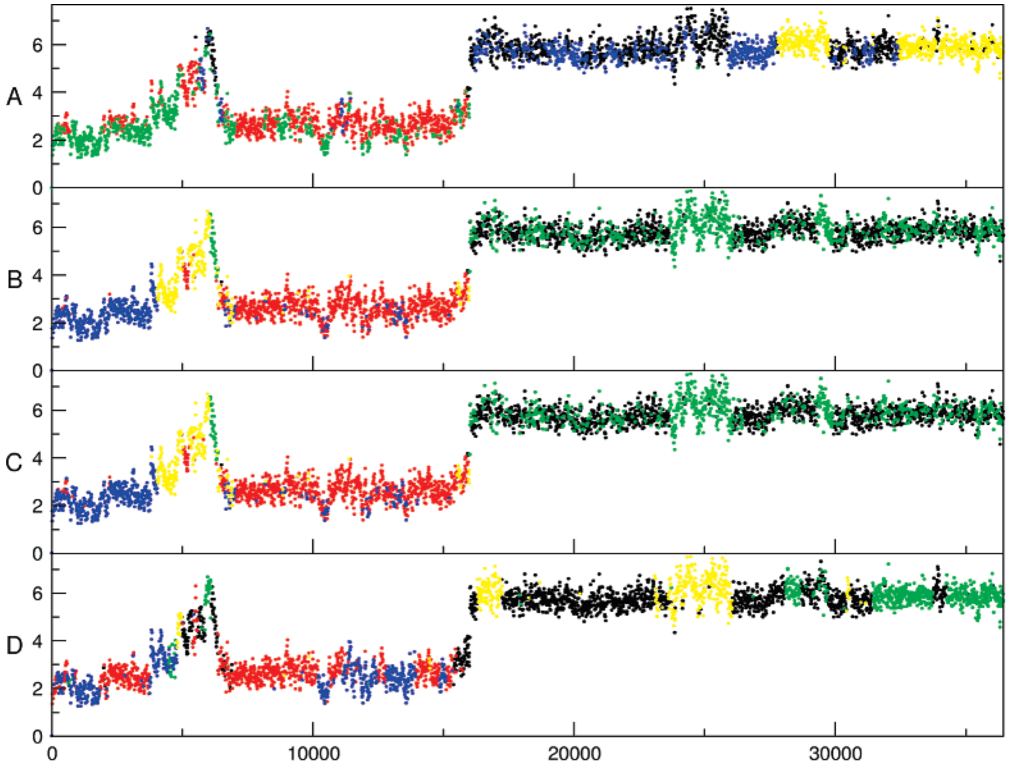


Figure 12. The effect of choosing different atoms for the pairwise comparisons when clustering. Based on the DNA hairpin-DB226 MD trajectory, shown is the RMSd (Å) as a function of time over the trajectory where individual points are colored based on their cluster identity. Four different sets of atoms for the pairwise comparison were chosen. (A) :1−21 represents all solute atoms, (B) :3−8:14−19:21 represents the atoms in binding region, (C) :3−8,14−19,21@C*,O*,N* represents the carbon, oxygen, and nitrogen atoms in binding region, and (D) :21 represents the atoms in the drug DB226 only. The trajectory data were taken every 10 ps and clustered using the **means** algorithm with a cluster count of 5.

similar atoms are chosen, the distances between the resulting sets of clusters are quite small, usually less than 0.01 for **linkage**, **hierarchical**, **means**, and **SOM** (i.e., comparing the major heavy atoms in the binding region to all atoms in the binding region). This suggests that it is best to narrow one's focus to the residues of interest before clustering a trajectory. Figure 12 shows the effects of clustering based on different choices of atoms for the pairwise comparison. Shown are the RMSd as a function of time for the atoms in the binding region with colors representing the distinct clusters that result based on different choices of the atoms for the pairwise comparison. The resulting sets of clusters based on pairwise comparisons of two sets of atoms describing the binding region, shown as the middle two plots in the figure, are almost identical as expected. The clustering based on the drug DB226 alone (residue 21, bottom), in comparison to the partitioning based on including all the DNA and drug atoms in the binding region, shows that although the drug conformation largely determines what cluster the configuration will adopt, clearly conformational substates of the DNA are also relevant. For example, compare the "green" cluster with and without the DNA binding region included. When the clustering is done on all of the DNA and drug atoms from the MD trajectory, some

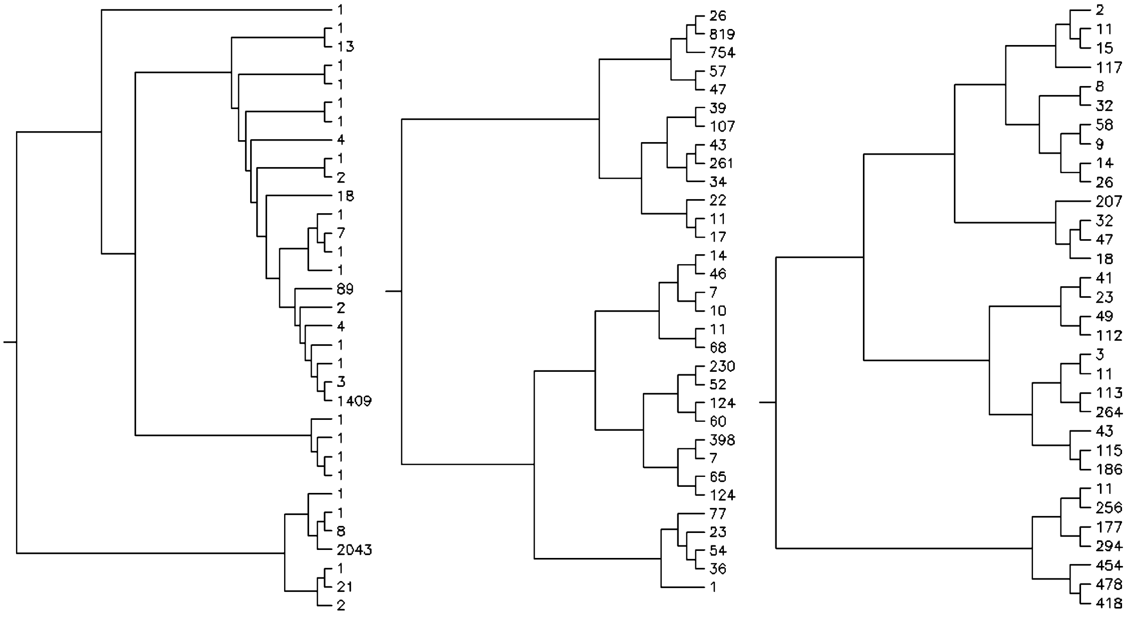


Figure 13. Cluster trees for **linkage** (left), **complete** (middle), and **hierarchical** (right) clustering algorithms applied to the DNA hairpin-DB226 trajectory. The last 30 steps of the algorithms are shown with the initial cluster size noted at each branch tip. Note that at this stage, the **linkage** algorithm has essentially already added most of the configurations to one of two large clusters. In contrast, the **complete** algorithm produces a well-balanced tree. The tree plots shown were generated using software available on the WWW from the Laboratory of Bioinformatics, Wageningen University and Research Center, The Netherlands; see http://www.bioinformatics.nl/tools/plottree.html.

of the important conformational differences or substates are concealed in the collective motion of the entire DNA structure. These data highlight the importance of narrowing the focus to relevant residues for the pairwise comparisons before clustering begins.

Critical Distance. Most clustering algorithms require the user to specify in advance the number of clusters to create. Doing this is difficult as the proper choice of the cluster count will depend on the underlying data. As an example of the difficulty, consider how poor the clustering would be for the trivial example of points in the plane shown in Figure 1 if a cluster count other than three was chosen. To provide users with more guidance on the proper choice of cluster count, we experimented with ways to dynamically choose a correct cluster count. For example, the bottom-up clustering algorithms can be instructed to stop when the intercluster distance or critical distance for the next merge is greater than some threshold ϵ rather than when a preselected number of clusters is reached. This approach has some promise, but it still requires the user to choose a value of ϵ. The appropriate value will be different from one algorithm to another—for instance, an ϵ value of 1.2 Å RMSd applied to **linkage** algorithm produces 12 clusters, while this same ϵ produces 74 clusters when the **centripetal** algorithm is applied, 181 clusters with the **complete** algorithm, and only 1 cluster with the **edge** algorithm. The appropriate value of ϵ also depends on the distance metrics used in the algorithm, the molecular system, and the choice of atoms used for the pairwise comparison.

Cluster Trees. For the bottom-up algorithms, a better approach may be to examine the tree of clusters that results (Figure 13). The cluster tree provides more information than does the individual clustering output at any particular stage. For instance, the cluster tree for a **linkage** clustering of the entire trajectory is not balanced; it consists mainly of two large clusters. For the **edge** and **centripetal** algorithms, the unbalanced tree that results will essentially be one big cluster. With the **complete** and **hierarchical** algorithms, the cluster trees are more balanced. In general, large clusters which remain unchanged for several iterations of the clustering algorithm seem more meaningful than clusters that shift at each stage, producing an unbalanced "bushy" tree.

Efficient Sieved Clustering. Speed and memory considerations make it difficult to cluster large trajectories using algorithms that compute the full pairwise comparison across all configurations. All of the algorithms investigated in this work precomputed an N × N symmetric matrix of the complete set of frame-to-frame distances, where N is the number of the frames in the trajectory. The matrix was precomputed since computing these distances on demand greatly increased the runtime. As discussed in the section comparing the relative performance of the different algorithms, computing this matrix is expensive and memory intensive. Even with precomputing the similarity matrix, the calculations quickly become intractable as more and more configurations are to be clustered. For the **SOM**, **COBWEB**, and **Bayesian** algorithms, the similarity matrix is only actually needed to calculate the DBI and pSF metrics; the pairwise comparisons could be calculated on the fly or loaded after the clustering is finished to improve performance. In spite of this, the refinement algorithms quickly become intractable as the trajectories to cluster become very long. To cluster very large trajectories, a better way to enable the efficient clustering is to cluster in a hierarchical fashion, specifically to initially cluster a subset of the data, such as that from a coarser-grained time sampling, with subsequent partitioning or clustering to put the skipped data into the existing clusters. To test the utility of this approach, we implemented a two-pass, or “sieved”, clustering method that initially clusters only part of the data and then on the second pass puts the missing data into existing clusters. With a sieve of *n*, initially clustering every “*n*” configurations, the pairwise-distance matrix is reduced in size by a considerable factor; specifically by n^2. The savings in computer time and memory more than compensate for the expense of making a second pass through the data. The one drawback of sieved clustering is that we may not sample all the conformations of the system during our first pass, especially if the sieve size is too large or if there are periodic components of the data with a period close to the sampling rate. In this case, the clustering output will not accurately partition the data. As a means to mitigate the problem with potential periodicity of the data, we can randomly select points for the first pass.

To assess the advantages and drawbacks of this scheme, we clustered the DB226-DNA hairpin trajectory data using sieves of various sizes. We compared the results and the runtime to the earlier results. Table 3 indicates the differences for the various clustering algorithms between the ordinary and the sieved results. In general, the sieving provides a dramatic decrease in runtime, particularly for the slower algorithms (bottom-up algorithms) where the time required decreases to about $1/n^2$. For the refinement and tree algorithms, including the **means**, **SOM**, **Bayesian**, and **COBWEB** algorithms, the time decreased to about $1/n$. Interestingly, with the top-down algorithm (**hierarchical**), the time decrease only occurs in the initial distance matrix calculation as the actual clustering algorithm is not very computationally demanding when small cluster counts are used. As the output, second pass through the data, and calculation of the statistics takes more time than the clustering, the time savings with this algorithm are modest.

Small sieve sizes (effectively less than 50 ps with this trajectory) produce negligible changes in DBI and pSF values with the **means**, **average-linkage**, **linkage**, and **SOM** algorithms. Additionally, for these algorithms the distances between the sieved clustering and unsieved clustering are small. This is likely a result of the second pass grouping procedure which assigns configurations to the closest centroid with each of these algorithms. Interestingly, the sieved clustering results are sometimes slightly better, with smaller DBI and larger pSF values, than the results obtained when clustering without sieving. This again suggests that the clustering depends on the data set. The data also suggest that the algorithms like **complete** and **COBWEB** seem more dependent on the choice of configurations for the first pass clustering. The small distance between the various distinct sets of clusters, as a function of sieve size, suggest that a sieve size of 5, with sampling every 50 ps, seems to be sufficient with this MD trajectory. The larger the desired number of clusters, the tighter the sieve should be, as rare conformational states (corresponding to smaller clusters) must still be adequately sampled in the first pass through the data. Interestingly, for larger sieve sizes (up to 500 ps), the **average-linkage**, **linkage**, and **SOM** algorithms still perform well, in contrast to the **means** algorithm which

Table 3. Performance of Sieved Clustering on the DNA Hairpin-DB226 MD Trajectory under Various Conditions[a]

algorithm	sieve size	sieve start	total time	DBI	pSF	clustering distance
means	no sieve		1376	2.02	1223	0.000
means	2	1	470	2.03	1223	0.014
means	2	2	386	1.99	1221	0.023
means	2	random1	421	2.00	1221	0.022
means	2	random2	671	2.05	1221	0.013
means	5	1	118	2.03	1220	0.013
means	5	2	140	2.04	1221	0.022
means	5	random1	117	2.02	1221	0.028
means	5	random2	127	2.05	1221	0.030
means	50	1	36	1.62	1007	0.181
means	50	2	33	2.10	1111	0.060
means	50	random1	34	1.99	1168	0.064
means	50	random2	37	2.03	1106	0.060
average	no sieve		6816	1.56	925	0.000
average	2	1	910	1.57	852	0.029
average	2	2	820	1.65	897	0.014
average	2	random1	897	1.66	842	0.069
average	2	random2	790	1.57	941	0.013
average	5	1	112	1.50	855	0.031
average	5	2	115	1.56	955	0.024
average	5	random1	203	1.76	882	0.030
average	5	random2	207	1.56	942	0.016
average	50	1	32	1.53	948	0.027
average	50	2	27	1.58	938	0.031
average	50	random1	33	1.58	973	0.041
average	50	random2	32	1.69	1031	0.081
linkage	no sieve		2149	1.04	805	0.000
linkage	2	1	259	1.49	795	0.019
linkage	2	2	305	1.76	830	0.015
linkage	2	random1	274	1.65	816	0.025
linkage	2	random2	336	1.64	856	0.026
linkage	5	1	70	1.70	798	0.020
linkage	5	2	64	1.88	842	0.041
linkage	5	random1	80	1.76	823	0.026
linkage	5	random2	75	1.65	865	0.030
linkage	50	1	17	1.61	870	0.035
linkage	50	2	18	1.38	926	0.042
linkage	50	random1	17	1.58	945	0.072
linkage	50	random2	18	1.48	856	0.026
SOM	no sieve		1925	2.13	1060	0.000
SOM	2	1	857	2.18	1092	0.019
SOM	2	2	587	2.18	1094	0.015
SOM	2	random1	730	2.17	1090	0.025
SOM	2	random2	546	2.17	1089	0.026
SOM	5	1	224	2.18	1121	0.020
SOM	5	2	281	2.16	1119	0.041
SOM	5	random1	293	2.16	1116	0.026
SOM	5	random2	212	2.16	1121	0.030
SOM	50	1	32	2.15	1132	0.076
SOM	50	2	32	2.09	1104	0.071
SOM	50	random1	31	2.05	1130	0.092
SOM	50	random2	31	2.40	1314	0.285

[a] In each case a cluster count of five was chosen. Various sieve sizes were chosen ranging from no sieve (10 ps sampling), to 2, 5, or 50 (representing 20, 50, or 500 ps sampling, respectively). For each algorithm, different choices of the starting configuration were investigated either with uniform sampling (sieve starting configuration of 1 or 2) or random sampling of the configurations to be clustered. Comparisons of the compute time, DBI, pSF, and clustering distance, relative to the unsieved clustering, show that the sieving process does not drastically alter the outcome.

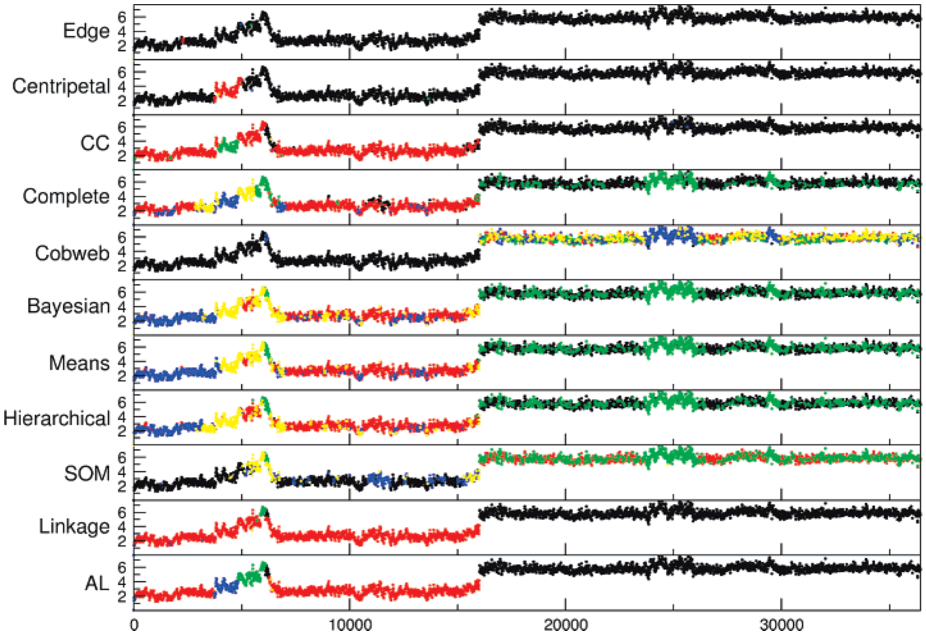


Figure 14. RMSd (Å) versus time (ps) for the DB226-DNA hairpin trajectory for different clustering algorithms. Trajectory data were taken every 10 ps, and a cluster count of 5 was used. The RMSd is the unfit distance (displacement) of the drug after fit the DNA structure to the first frame. The color scheme is based on the size of the cluster with black > red > green > blue > yellow.

sometimes breaks down. However, the performace of means may be improved by running multiple trials with different random selections of configurations for the refinement steps in each run. An additional problem of the **SOM** algorithm is that it may produce fewer clusters than are expected for a particular cluster count. The larger distance of **SOM** with a sieve size of 50 and a sieve start random2 (the last row in the table) is due to the fact that only 4 clusters have been formed during the **SOM** clustering. In addition to the applications to DNA shown, clustering has been applied to a relatively long trajectory of a dynamic protein system. Specifically, this involves a cytochrome P450 2B4 structure that converts from the “open” geometry seen in the crystal to a closed geometry over the course of a ~75 ns simulation. The cluster metrics for clustering over 7000 configurations at 10 ps intervals, a description of the simulation protocols, and RMSd plots and molecular graphics are shown in the Supporting Information.

Summary of the Relative Performance of the Various Clustering Algorithms. Using color to identify each distinct cluster, Figures 14 and 15 show the RMSd or MM-PBSA free energy of binding versus time for the DB226-DNA hairpin trajectories. This provides another means to visualize the relative performance of the various algorithms. From the figure, it is clear that similar RMSd or $\Delta G_{\text{binding}}$ values do not necessarily imply equivalent cluster membership. Also evident is that **edge**, **linkage**, and **centripetal** do not produce very meaningful sets of clusters as only one or two large clusters result. The **means**, **hierarchical**, **complete**, **Bayesian**, and **average-linkage** algorithms, on the other hand, all tend to produce meaningful sets of clusters.

Bottom-up algorithms iteratively merge small clusters into larger ones. With MD simulation data, the algorithms have a tendency to produce outlier or singleton clusters. If the algorithm generates 10 clusters, 9 of which are single points, little is learned about the underlying structure of the data (other than identifying the most extreme conformations). Careful choice of cluster count (see below) is one way to mitigate this sensitivity.

- The **single-linkage** algorithm is rather fragile in that the presence or absence of a single point can control the grouping of the rest. It is very sensitive to lines of closely spaced “breadcrumb” points, which can add arbitrarily long trails of data to one cluster.[52] Although it can handle clusters of differing sizes, the algorithm often does a poor job delineating clusters whose points are very close. Its results were passable on points in the plane but very poor on real MD trajectories.
- **Complete-linkage** and **centripetal-complete** clustering are the two bottom-up clustering algorithms that do not have the tendency to produce singleton clusters. In spite of this, the resulting clusters tend to be small.
- **Centripetal** clustering gives results that are similar, but inferior, to those of **centroid-linkage**. It tends to produce a larger minimum cluster size, since the representatives from a small cluster are not drawn away as far from the ‘frontier’ as those in a large cluster. The parameters of **centripetal** clustering (the number of representatives per cluster, and the distance they are drawn toward the centroid) may be amenable to further tuning to improve cluster quality.
- **Centroid-linkage** and **average-linkage** clustering gave consistently good results as quantified by the Davies-Bourdin Index (DBI) and pseudo-F statistic (pSF). They can produce clusters of varying sizes and possibly concave shapes. They are two of the most useful of the clustering algorithms we have examined for use with MD trajectories.

Refinement Clustering Algorithms. Because the refinement algorithms include a random factor, we ran the algorithms several times and kept the best (as measured by DBI and pSF) clustering results.

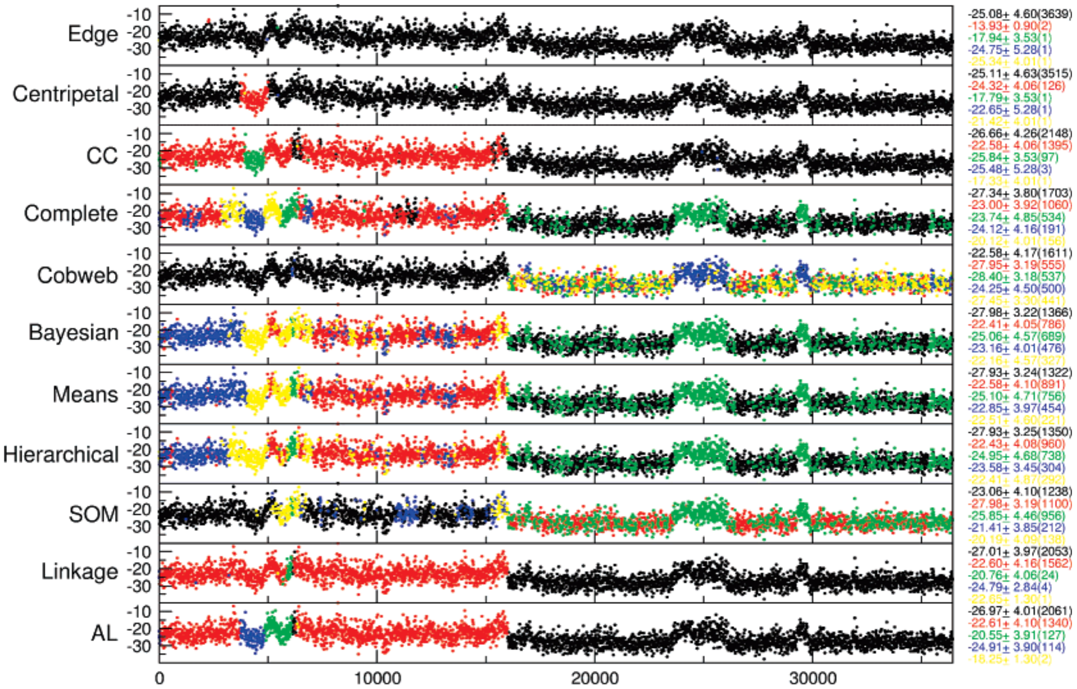


Figure 15. MM-PBSA binding energy versus time (ps) for the DB226-DNA hairpin trajectory for different clustering algorithms. Trajectory data were taken every 10 ps, and a cluster count of 5 was used. The color scheme is based on the size of the cluster with black > red > green > blue > yellow. The data on the right side of the figure show the approximate free energy and standard deviation (kcal/mol) of a cluster. The number in the parentheses is the number of snapshots in that cluster. Very different average free energies (all neglecting solute entropic components) are seen between the different clustering algorithms.

- **Means** clustering tends to produce "blocky" clusters of similar sizes. The seed cluster centroid positions start at the edges of the data set but move toward the eventual centroid over the course of the clustering run. This algorithm cannot produce concave clusters and does not generate clusters of different sizes, but in general it performs very well.
- **Bayesian** clustering produces decent results, but these results become poor for high cluster counts. It can produce clusters of different sizes. **Bayesian** clustering often has difficulty "recognizing" obvious clusters in simple test cases in the plane, even when the algorithm is reseeded and rerun many times. To give good results, the algorithm must be repeated many times with new random seeds. This is computationally expensive, particularly on MD trajectories where there are often hundreds of coordinates and thousands of configurations to consider.
- Self-organizing maps (**SOM**) produced the best results of the refinement algorithms. The performance was more consistent between runs than **Bayesian** clustering. However, self-organizing maps share some of the problems characteristic of the hierarchical clustering algorithm; specifically, the **SOM** algorithm cannot produce concave clusters, and it has difficulty producing clusters of varying sizes.

We also find that the **COBWEB** clustering algorithm is also promising. Visualization of the resulting tree structure, before flattening, can provide hints as to the reasonable number of clusters to specify the data. However, a severe limitation of the **COBWEB** algorithm is that it is highly dependent on the order of the points incorporated into the COBWEB tree. Thus, the variations between multiple **COBWEB** runs are relatively large.

Discussion

We described the development of a series of different clustering algorithms into a C program library, their application to the easy to visualize test case of clustering 2D points on the plane, integration of the clustering algorithms into the **ptraj** trajectory analysis program, and the subsequent application of the various algorithms to a series of contrived and real MD trajectories. Overall, we were rather surprised by the results which clearly show widely different behavior among the various algorithms. Moreover, the performance of a given algorithm is strongly dependent on the choice of cluster count and, less surprisingly, the choice of atoms for the pairwise comparison. On the other hand, the results appeared to only be weakly sensitive to the choice of the pairwise metric when comparing RMSd to DME measures of similarity. Evaluation of the relative performance was made possible through visualization of the results and also through the exploration of various metrics defining the performance. Specifically, low DBI values and high pSF values signal better clustering. Information on the appropriate cluster count comes from analysis of SSR/SST ratios and critical distance measures as a function of cluster count. In order to more efficiently handle very large data sets, a sieving approach was introduced where only a portion of the data is initially clustered, and then the remaining data are added to existing clusters. For the MD simulations investigated in this work, sieves up to 50 ps only moderately alter the outcome. Overall, the best performance was observed with the average-linkage, means, and SOM algorithms. If the cluster count is not known in advance, one of the other algorithms, such as hierarchical or average-linkage, are recommended. These two also can be used effectively with a distance threshold for separating clusters. In addition to performing reasonably well, it is important to be aware of the limitations or weaknesses of each algorithm, specifically the high sensitivity to outliers with hierarchical, the tendency to generate homogenously sized clusters with means, and the tendency to produce small or singleton clusters with average-linkage and linkage.

Acknowledgment. We would like to thank Carlos Simmerling, Adrian Roitberg, and Holger Gohlke and their group members for testing early versions of the code. We gratefully acknowledge the award of computational time and support from the Pittsburgh Supercomputer Center, the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the TeraGrid, and the University of Kentucky under NRAC and LRAC awards MCA01S027S. We also acknowledge computational time from the Center for High Performance Computing at the University of Utah including resources provided by the NIH (1S10RR17214-01) on the Arches Metacluster. Stephen Tanner would like to acknowledge support from the Department of Pharmaceutics and Pharmaceutical Chemistry's Summer Undergraduate Research Fellowship Program (2002).

Supporting Information Available: More technical descriptions of the cluster algorithms implemented, the parameters for the minor groove binding drug DB226, more detailed comparison of the relative performance and properties of various of the clustering algorithms, schematic cluster trees highlighting relative performance, RMSd plots, and molecular graphics. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) van Gunsteren, W. F.; Berendsen, H. J. Molecular dynamics: perspective for complex systems. *Biochem. Soc. Trans.* **1982**, *10*, 301−305.

(2) van Gunsteren, W. F.; Karplus, M. Protein dynamics in solution and in a crystalline environment: a molecular dynamics study. *Biochemistry* **1982**, *21*, 2259−2274.

(3) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.

(4) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glattli, A.; Hunenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064−4092.

(5) Levitt, M. Molecular dynamics of native protein: I. Computer simulation of trajectories. *J. Mol. Biol.* **1983**, *168*, 595−617.

(6) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(7) Cheatham, T. E., III; Kollman, P. A. Molecular dynamics simulation of nucleic acids. *Ann. Rev. Phys. Chem.* **2000**, *51*, 435−471.

(8) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740−744.

(9) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002**, *12*, 190−196.

(10) Tajkhorshid, E.; Aksimentiev, A.; Balabin, I.; Gao, M.; Israelwitz, B.; Phillips, J. C.; Zhu, F.; Schulten, K. Large scale simulation of protein mechanics and function. *Adv. Protein Chem.* **2003**, *66*, 195−247.

(11) Cheatham, T. E., III Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360−367.

(12) Feig, M.; Brooks, C. L., III Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.

(13) Wong, C. F.; McCammon, J. A. Protein simulation and drug design. *Adv. Protein Chem.* **2003**, *66*, 87−121.

(14) Rueda, D.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J. L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci.* **2007**, *104*, 796−801.

(15) Brooks, C. I. Protein and peptide folding explored with molecular simulations. *Acc. Chem. Res.* **2002**, *35*, 447−454.

(16) Daggett, V. Molecular dynamics simulations of the protein unfolding/folding reaction. *Acc. Chem. Res.* **2002**, *35*, 422−449.

(17) Simmerling, C.; Strockbine, B.; Roitberg, A. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.

(18) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* **2003**, *68*, 91−109.

(19) Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Raleigh, D. P.; Simmerling, C. L. The unfolded state of the villin headpiece helical subdomain: computational studies of the role of locally stabilized structure. *J. Mol. Biol.* **2006**, *360*, 1094−1107.

(20) Day, R.; Daggett, V. Direct observation of microscopic reversibility in single-molecule protein folding. *J. Mol. Biol.* **2006**, *366*, 677−686.

(21) Juraszek, J.; Bolhuis, P. G. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci.* **2006**, *103*, 15859−15864.

(22) Eleftheriou, M.; Germain, R. S.; Royyuru, A. K.; Zhou, R. Thermal denaturing of mutant lysozyme with both the OPLSAA and the CHARMM force fields. *J. Am. Chem. Soc.* **2006**, *128*, 13388−13395.

(23) Yoda, T.; Sugita, Y.; Okamoto, Y. Cooperative folding mechanism of a beta-hairpin peptide studied by a multicanonical replica-exchange molecular dynamics simulation. *Proteins* **2007**, *66*, 846−859.

(24) Baumketner, A.; Shea, J. E. The structure of the Alzheimer, amyloid beta 10−35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. *J. Mol. Biol.* **2007**, *366*, 275−285.

(25) Chen, H. F.; Luo, R. Binding induced folding in p53-MDM2 complex. *J. Am. Chem. Soc.* **2007**, *129*, 2930−2937.

(26) Paschek, D.; Nymeyer, H.; Garcia, A. E. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol.* **2007**, *157*, 524−533.

(27) Li, W.; Zhang, J.; Wang, W. Understanding the folding and stability of a zinc finger-based full sequence design protein with replica exchange molecular dynamics simulations. *Proteins* **2007**, *67*, 338−349.

(28) Periole, X.; Mark, A. E. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* **2007**, *126*, 014903.

(29) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-folding dynamics: overview of molecular simulation techniques. *Ann. Rev. Phys. Chem.* **2007**, *58*, 57−83.

(30) Spackova, N.; Cheatham, T. E., III; Ryjacek, F.; Lankas, F.; van Meervelt, L.; Hobza, P.; Sponer, J. Molecular dynamics simulations and thermodynamic analysis of DNA-drug complexes. Minor groove binding between 4′,6-diamidino-2-phenylindole (DAPI) and DNA duplexes in solution. *J. Am. Chem. Soc.* **2003**, *125*, 1759−1769.

(31) Bui, J. M.; McCammon, J. A. Protein complex formation by acetylcholinesterase and the neurotoxin fasciculin-2 appears to involve an induced-fit mechanism. *Proc. Natl. Acad. Sci.* **2006**, *103*, 15451−15456.

(32) Lu, Y.; Yang, C. Y.; Wang, S. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. *J. Am. Chem. Soc.* **2006**, *128*, 11830−11839.

(33) Xu, Y.; Wang, R. A computational analysis of the binding affinities of FKBP12 inhibitors using the MM-PB/SA method. *Proteins* **2006**, *64*, 1058−1068.

(34) de Jonge, M. R.; Koymans, L. H.; Guillemont, J. E.; Koul, A.; Andries, K. A computational model of the inhibition of Mycobacterium, tuberculosis ATPase by a new drug candidate R207910. *Proteins* **2007**, *67*, 971−980.

(35) Ode, H.; Matsuyama, S.; Hata, M.; Hoshino, T.; Kakizawa, J.; Sugiura, W. Mechanism of drug resistance due to N88S in CRF01_AE HIV-1 protease, analyzed by molecular dynamics simulations. *J. Med. Chem.* **2007**, *50*, 1768−1777.

(36) Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **2006**, *103*, 915−920.

(37) Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.* **2006**, *128*, 2812−2813.

(38) Lankas, F.; Lavery, R.; Maddocks, J. H. Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure* **2006**, *14*, 1527−1534.

(39) Noy, A.; Perez, A.; Laughton, C. A.; Orozco, M. Theoretical study of large conformational transitions in DNA: the B<−> A conformational change in water and ethanol/water. *Nucl. Acids Res.* **2007**, *35*, 3330−3338.

(40) van der Vaart, A.; Karplus, M. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 164106.

(41) Noe, F.; Horenko, I.; Schutte, C.; Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* **2007**, *126*, 155102.

(42) Li, D. W.; Han, L.; Huo, S. Structural and pathway complexity of beta-strand reorganization within aggregates of human transthyretin(105−115) peptide. *J. Phys. Chem. B* **2007**, *111*, 5425−5433.

(43) Patel, S.; Balaji, P. V.; Sasidhar, Y. U. The sequence TGAAKAVALVL from glyceraldehyde-3-phosphate dehydrogenase displays structural ambivalence and interconverts between alpha-helical and beta-hairpin conformations mediated by collapsed conformational states. *J. Pept. Sci.* **2007**, *13*, 314−326.

(44) Roccatano, D.; Barthel, A.; Zacharias, M. Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. *Biopolymers* **2007**, *85*, 407−421.

(45) Sefcikova, J.; Krasovska, M. V.; Sponer, J.; Walter, N. G. The genomic HDV ribozyme utilizes a previously unnoticed U-turn motif to accomplish fast site-specific catalysis. *Nucl. Acids Res.* **2007**, *35*, 1933−1946.

(46) Razga, F.; Zacharias, M.; Reblova, K.; Koca, J.; Sponer, J. RNA kink-turns as molecular elbows: hydration, cation binding, and large-scale dynamics. *Structure* **2006**, *14*, 825−835.

(47) Kormos, B. L.; Baranger, A. M.; Beveridge, D. L. A study of collective atomic fluctuations and cooperativity in the U1A-RNA complex based on molecular dynamics simulations. *J. Struct. Biol.* **2007**, *157*, 500−513.

(48) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of a 2.2-ns trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.

(49) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.

(50) Cormack, R. M. A review of classification. *J. R. Stat. Soc. A* **1971**, *134*, 321−367.

(51) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: A review. *ACM Comp. Surv.* **1999**, *31*, 264−323.

(52) Torda, A. E.; van Gunsteren, W. F. Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.* **1994**, *15*, 1331−1340.

(53) Marchionini, C.; Maigret, B.; Premilat, S. Models for the conformational behaviour of angiotensin-II in acidic aqueous solutions. *Biochem. Biophys. Res. Comm.* **1983**, *112*, 339−346.

(54) Willett, P. *Similarity and clustering in chemical information systems*; John Wiley & Sons, Inc.: New York, 1987; Vol 1, p 266.

(55) Kreissler, M.; Pesquer, M.; Maigret, B.; Fournie-Zaluski, M. C.; Roques, B. P. Computer simulation of the conformational behavior of cholecystokinin fragments: Conformational families of sulfated CCK8. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 85−94.

(56) Unger, R.; Harel, D.; Wherland, S.; Sussman, J. L. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **1989**, *5*, 355−373.

(57) Gordon, H. L.; Somorjai, R. L. Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins* **1992**, *14*, 249−264.

(58) Michel, A.; Jeandenans, C. Multiconformational, investigations of polypeptidic structures, using clustering methods and principal component analysis. *Comput. Chem.* **1993**, *17*, 49−59.

(59) Troyer, J. M.; Cohen, F. E. Protein conformational landscapes: energy minimization and clustering of a long molecular dynamics trajectory. *Proteins* **1995**, *23*, 97−110.

(60) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-unfolding thermodynamics of a b-heptapeptide from equilibrium simulations. *Proteins* **1999**, *34*, 269−280.

(61) Gabarro-Arpa, J.; Revilla, R. Clustering of a molecular dynamics trajectory with a Hamming distance. *Comput. Chem.* **2000**, *24*, 696−698.

(62) Watts, C. R.; Mezei, M.; Murphy, R. F.; Lovas, S. Conformational space comparison of GnRH and lGnRH-III using molecular dynamics, cluster analysis and Monte Carlo thermodynamic integration. *J. Biomol. Struct. Dyn.* **2001**, *18*, 733−748.

(63) Laboulais, C.; Ouali, M.; Le Bret, M.; Gabarro-Arpa, J. Hamming distance geometry of a protein conformational space: Application, to the clustering of a 4-ns molecular dynamics trajectory of the HIV-1 integrase catalytic core. *Proteins* **2002**, *47*, 169−179.

(64) Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810−818.

(65) Bystroff, C.; Garde, S. Helix propensities of short peptides: Molecular, dynamics versus Bioinformatics. *Proteins* **2003**, *50*, 552−562.

(66) Moraitakis, G.; Goodfellow, J. M. Simulations of human lysozyme: Probing, the conformations triggering amyloidosis. *Biophys. J.* **2003**, *84*, 2149−2158.

(67) Lee, M. C.; Deng, J.; Briggs, J. M.; Duan, Y. Large-scale conformational dynamics of the HIV-1 integrase core domain and its catalytic loop mutants. *Biophys. J.* **2005**, *88*, 3133−3146.

(68) Rao, F.; Settanni, G.; Guarnera, E.; Caflisch, A. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.* **2005**, *122*, 184901.

(69) Lyman, E.; Zuckerman, D. M. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.* **2006**, *91*, 164−172.

(70) Sullivan, D. C.; Lim, C. Quantifying polypeptide conformational space: sensitivity to conformation and ensemble definition. *J. Phys. Chem. B* **2006**, *110*, 16707−16717.

(71) Li, Y. Bayesian model based clustering analysis: application to a molecular dynamics trajectory of the HIV-1 integrase catalytic core. *J. Chem. Inf. Model.* **2006**, *46*, 1742−1750.

(72) Elmer, S. P.; Pande, V. S. Foldamer simulations: Novel, computational methods and applications to poly-phenylacetylene oligomers. *J. Chem. Phys.* **2004**, *121*, 12760−12771.

(73) Sorin, E. J.; Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **2005**, *88*, 2472−2493.

(74) Sims, G. E.; Choi, I.-G.; Kim, S.-H. Protein conformational space in higher order phi-psi maps. *Proc. Natl. Acad. Sci.* **2005**, *102*, 618−621.

(75) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Lett.* **2006**, *580*, 3422−3426.

(76) Scott, E. E.; He, Y. A.; Wester, M. R.; White, M. A.; Chin, C. C.; Halpert, J. R.; Johnson, E. F.; Stout, C. D. An open conformation of mammalian cytochrome P450 2B4 at 1.6-A resolution. *Proc. Natl. Acad. Sci.* **2003**, *100*, 13196−13201.

(77) Poncin, M.; Hartmann, B.; Lavery, R. Conformational substates in B-DNA. *J. Mol. Biol.* **1992**, *226*, 775−794.

(78) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of DNA, RNA and phosphoramidate helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.

(79) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.

(80) Harris, S. A.; Gavathiotis, E.; Searle, M. S.; Orozco, M.; Laughton, C. A. Cooperativity in drug-DNA recognition: a molecular dynamics study. *J. Am. Chem. Soc.* **2001**, *123*, 12658−12663.

(81) Fisher, D. H. In *Improving inference through conceptual clustering*; AAAI: Seattle, WA, 1987; pp 461−465.

(82) Fisher, D. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* **1987**, *2*, 139−172.

(83) Cheeseman, P.; Stutz, J. Bayesian classification (AutoClass): theory and results. In *Advances in knowledge discovery and data mining*; Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds.; American Association of Artificial Intelligence Press: Menlo Park, CA, 1996; pp 61−83.

(84) Kohonen, T. *Self-organizing maps*, 3rd ed.; Springer: Berlin-Heidelberg, 2001; Vol. 30, p 501.

(85) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules. *Comp. Phys. Comm.* **1995**, *91*, 1−41.

(86) Case, D. A.; Cheatham, T. E., III; Darden, T. A.; Gohlker, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A. V.; Simmerling, C.; Wang, B.; Woods, R. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(87) Guha, S.; Rastogi, R.; Shim, K. In *CURE: An efficient clustering algorithm for large databases*; Proceedings of the ACM SIGMOD International Conference on Management of Data: New York, 1998; pp 73−84.

(88) Witten, I. H.; Frank, E. *Data mining: Practical machine learning tools and techniques with Java implementations*; Morgan Kaufmann: 1999; p 525.

(89) Kohonen, T. *Self-organization and Associative Memory*; Springer-Verlag: Berlin, 2001; Vol. 30, p 501.

(90) Davies, D. L.; Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intelligence* **1979**, *1*, 224−227.

(91) Vesanto, J.; Alhoniemi, E. Clustering of the self-organizing map. *IEEE Trans. Neural Networks* **2000**, *11*, 586−600.

(92) Bolshakova, N.; Azuaje, F. *Cluster validation techniques for genome expression data*; University of Dublin, Trinity College: Dublin, 2002; p 13.

(93) Speer, N.; Spiet, C.; Zell, A. Biological cluster validity indices based on the gene ontology. In *Advances in intelligent data analysis VI*; Famili, A. F., Kok, J. N., Pena, J. M., Siebes, A., Feelders, A., Eds.; Springer: Berlin, Heidelberg, 2005; Vol. 3646, pp 429−439.

(94) Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Comm. Stat.* **1974**, *3*, 1−27.

(95) Mitchell, T. *Machine Learning*; McGraw-Hill: 1997; p 432.

(96) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* **1977**, *23*, 327−341.

(97) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Comp. Phys.* **1984**, *81*, 3684−3690.

(98) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(99) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparisons of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(100) Aqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021−8024.

(101) Cheatham, T. E., III; Srinivasan, J.; Case, D. A.; Kollman, P. A. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. *J. Biomol. Struct. Dyn.* **1998**, *16*, 265−280.

(102) Wu, X. W.; Wang, S. M. Self-guided molecular dynamics simulation for efficient conformational search. *J. Phys. Chem.* **1998**, *102*, 7238−7250.

(103) Wu, X.; Wang, S. Helix Folding of an Alanine-Based Peptide in Explicit Water. *J. Phys. Chem. B* **2001**, *105*, 2227−2235.

(104) Wu, X.; Brooks, B. R. Beta-hairpin folding mechanism of a nine-residue peptide revealed from molecular dynamics simulations in explicit water. *Biophys. J.* **2004**, *86*, 1946−1958.

(105) Wu, X.; Wang, S.; Brooks, B. R. Direct observation of the folding and unfolding of a beta-hairpin in explicit water through computer simulation. *J. Am. Chem. Soc.* **2002**, *124*, 5282−5283.

(106) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera−a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(107) Boykin, D. W.; Kumar, A.; Xiao, G.; Wilson, W. D.; Bender, B. C.; McCurdy, D. R.; Hall, J. E.; Tidwell, R. R. 2,5-bis-[4-(N-alkylamidino)phenyl]furans as anti-Pneumocystis carinii agents. *J. Med. Chem.* **1998**, *41*, 124−129.

(108) Wilson, W. D.; Tanious, F. A.; Ding, D.; Kumar, A.; Boykin, D. W.; Colson, P.; Houssier, C.; Bailly, C Nucleic acid interactions of unfused aromatic cations: Evaluation of proposed minor-groove, major-groove, and intercalation binding modes. *J. Am. Chem. Soc.* **1998**, *120*, 10310−10321.

(109) Mazur, S.; Tanious, F. A.; Ding, D.; Kumar, A.; Boykin, D. W.; Simpson, I. J.; Neidle, S.; Wilson, W. D. A thermodynamic and structural analysis of DNA minor-groove complex formation. *J. Mol. Biol.* **2000**, *300*, 321−337.

(110) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(111) Tsui, V.; Case, D. A. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.* **2000**, *122*, 2489−2498.

(112) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(113) Wang, J.; Kollman, P. A. Automatic parameterization of force field by systematic search and genetic algorithms. *J. Comput. Chem.* **2001**, *22*, 1219−1228.

(114) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges- the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(115) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; J. R. Cheeseman, V. G. Z.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98 (Revision A.10)*; Gaussian, Inc.: Pittsburgh, PA, 2001.

(116) Laughton, C. A.; Tanious, F. A.; Nunn, C. M.; Boykin, D. W.; Wilson, W. D.; Neidle, S. A crystallographic and spectroscopic study of the complex between d(CGCGAAT-TCGCG)2 and 2,5-bis(4-guanylphenyl)furan, an analogue of Berenil: structural origins of enhanced DNA-binding affinity. *Biochemistry* **1996**, *35*, 5655−5661.

CT700119M

JCTC Journal of Chemical Theory and Computation

Folding Simulations of the Transmembrane Helix of Virus Protein U in an Implicit Membrane Model

Jakob P. Ulmschneider*,† and Martin B. Ulmschneider‡

Department of Chemistry, University of Rome "La Sapienza", Rome, Italy, and Department of Biochemistry, University of Oxford, Oxford, U.K.

Received May 2, 2007

Abstract: Vpu is an 81-amino-acid auxiliary membrane protein encoded by human immunodeficiency virus type 1 (HIV-1). One of its roles is to amplify viral release by self-assembling in homo-oligomers to form functional water-filled pores enabling the flux of ions across the membrane. Various NMR and CD studies have shown that the transmembrane domain of Vpu has a helical conformation. With a recently developed implicit membrane model and an efficient Monte Carlo (MC) algorithm using concerted backbone rotations, we simulate the folding of the transmembrane domain of Vpu at atomic resolution. The implicit membrane environment is based on the generalized Born theory and enables very long time scale events, such as folding to be observed using detailed all-atom representation of the protein. Such studies are currently computationally unfeasible with fully explicit lipid bilayer molecular dynamics simulations. The correct helical transmembrane structure of Vpu is predicted from extended conformations and remains stably inserted. Tilt and kink angles agree well with experimental estimates from NMR measurements. The experimentally observed change in tilt angle in membranes of varying hydrophobic width is accurately reproduced. The extensive simulation of a pentamer of the Vpu transmembrane domain in the implicit membrane gives results similar to the ones reported previously for fully explicit bilayer simulations.

Introduction

One of the most interesting challenges of theoretical biophysics is the direct computational prediction of membrane protein structure from sequence information. Unfortunately, molecular mechanics simulations using explicit lipid-bilayer membranes[1−4] are usually limited to the 1−100 ns time scale due to the large number of nonbonded interactions that need to be evaluated for such complex systems. While this allows for the study of protein stability in a lipid bilayer[1,2] or even self-assembly of protein/detergent micelles for various proteins,[5,6] it is unfortunately inadequate to study protein folding, which requires time scales in the micro- to millisecond range. In principle folding can be simulated for tiny systems in explicit lipid bilayer membranes when very large computational resources are available, e.g., 64 CPUs for 2.6 ns replica exchange molecular dynamics of a 16 residue peptide in a 36 lipid bilayer solvated by 1048 water molecules.[4] Nevertheless, even this approach is currently unfeasible for larger systems or for studies of protein function. Simulations in the multi-μs range for molecular dynamics (MD) or in the billions of Monte Carlo (MC) steps are needed to study folding and to obtain converged averages of experimentally measurable macroscopic properties. A further overview on the large number of present and anticipated future applications of implicit membrane methods is given in recent reviews.[7−9]

Implicit solvation models generally treat the solvent as a polarizable continuum. For spherical ions in a homogeneous isotropic dielectric the solvation energy can be determined analytically as demonstrated by Born.[10] The generalized Born solvation model extends this equation to macromolecules, which are approximated as an assortment of charged

[*] Corresponding author e-mail: Jakob@ulmschneider.com.
[†] University of Rome "La Sapienza".
[‡] University of Oxford.

10.1021/ct700103k CCC: $37.00
Published on Web 10/03/2007

spheres.[11] The immense success of this method in globular protein and peptide folding simulations[12−16] has encouraged attempts to apply the generalized Born formalism to represent the membrane environment implicitly.[17−27] These models describe the membrane environment as a uniform hydrophobic slab and have been used successfully to fold and assemble small helical membrane peptides.[18,23,24] The combination with sophisticated Monte Carlo methods has enabled us to successfully study the folding and orienting of membrane associated peptides into their experimentally observed native conformations.[21,27]

In this study, we report folding simulations of Virus protein U (Vpu), a 81-residue membrane protein of the human immunodeficiency virus type 1 (HIV-1).[28,29] It consists of one N-terminal hydrophobic membrane helix and two shorter amphipathic helices that remain in the plane of the membrane on the cytoplasmic side.[30] Two main functions of Vpu are observed: The first, which involves the cytoplasmic domain in the C-terminal half of the protein, is to accelerate the degradation of the CD4 receptor in the endoplasmic reticulum (ER) of infected cells.[31,32] Second, Vpu has been shown to amplify the release of virus particles from infected cells, a process that involves the transmembrane (TM) domain.[33,34] Vpu and its isolated TM part oligomerize in lipid membranes[35] and show channel activity.[36−40] In this work, we focus on the TM α-helix of Vpu: Its structure has been determined experimentally,[41] and its orientation relative to the plane of the lipid bilayer has been estimated from both NMR spectroscopy[41−44] and Fourier Transform Infrared Dichroism (FTIR) spectroscopy.[45]

Several previous MD simulation studies have been performed on Vpu in explicit bilayers, with either the complete peptide,[46] part of the peptide,[47,48] or only the N-terminal TM helix as monomer or as oligomer.[49−55] However, the short time scale (∼1−5 ns) of these simulations was not sufficient to study folding or function. Longer simulations (200 ns) of Vpu have been performed using a coarse-grain method.[56] In this work, we use our implicit membrane model together with a MC scheme to simulate the folding of the TM helix of Vpu as well as study its oligomeric structure in the membrane.

Simulation Methods

The Generalized Born Membrane. The development of the present generalized Born (GB) membrane has been described in detail in a previous publication.[21] The GB equation[11] is left unchanged, and only the method to calculate the Born radii is modified. The total effective free energy of solvation in the membrane is given by $\Delta G_{\mathrm{sol}} = \Delta G_{\mathrm{pol}} + \Delta G_{\mathrm{np}}$, where ΔG_{pol} is the electrostatic contribution (GB equation)

$$\Delta G_{\mathrm{pol}} = -166\left(\frac{1}{\epsilon_{\mathrm{m}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right)\sum_{i}^{n}\sum_{j}^{n}\frac{q_i q_j}{\sqrt{{r_{ij}}^2 + \alpha_j\alpha_j \exp(-{r_{ij}}^2/4\alpha_i\alpha_j)}} \qquad (1)$$

and ΔG_{np} is the nonpolar hydrophobic contribution. The membrane is treated as a planar hydrophobic region in a uniform polar solvent with a dielectric constant $\epsilon_{\mathrm{w}} = 80$, that becomes increasingly inaccessible to the solvent toward its center.

Both the protein interior and the membrane are assumed to have the same interior dielectric constant of $\epsilon_{\mathrm{m}} = 2$. The Born radii are calculated using the fast asymptotic pairwise summation of Qiu and Still,[57] where the integral of $1/r^4$ over the solute interior is approximated as a sum

$$G'_{\mathrm{pol},i} = \Gamma(z_i, R_i, L) + \underbrace{\sum_{j}^{1\text{-}2}\frac{P_2 V_j(z_j)}{r_{ij}^{4}} + \sum_{j}^{1\text{-}3}\frac{P_3 V_j(z_j)}{r_{ij}^{4}} + \sum_{j}^{1\geq 4}\frac{P_4 V_j(z_j)\ ccf}{r_{ij}^{4}}}_{\text{sums only involving atoms with } |z| > \mathrm{L}}, \qquad (2)$$

where L is the membrane half width, z_i is the z-position of the atom i, P_1−P_4 are the parameters determined by Qiu et al.,[57] the sums are over 1−2, 1−3, and 1 ≥ 4 neighbors and ccf is a close contact function, and $V_i(z)$ is the volume of atom i. The main advantage of the asymptotic approach over other methods to obtain Born radii is speed: Pairwise evaluation of the costly $1/r_{ij}$ terms already occurs for the nonbonded Coulomb and van der Waals interactions. In our program, the calculation of the Born radii $1/{r_{ij}}^4$ terms in eq 2 is combined with the other nonbonded calculations. As a result, the evaluation of the Born radii (through eq 2) takes no additional computational time, i.e., is obtained 'for free'. The increase in computation is entirely due to the evaluation of eq 1 and results in a slowdown of ∼2.0−2.2 compared to vacuum simulations. This is at the lower end of the values reported for other GB models, which usually are in the range ∼4−5.[58,59] In addition to the good performance, the method has been demonstrated to yield excellent results in predicting experimental free energies of solvation as well as hydration effects on conformational equilibria.[60]

By modifying the pairwise summation to solute atoms, the self-solvation terms $\Gamma(z_i, L)$ as well as the atomic volumes $V(z_i)$ were made to vary smoothly between full solvation and a limiting value for burial at the center of the membrane. We use a Gaussian shape

$$\Gamma(z_i) = g_{\mathrm{bulk}} + (g_{\mathrm{center}} - g_{\mathrm{bulk}})e^{\gamma({z_i}^2/L^2)} \qquad (3)$$

where g_{bulk} is the limiting value of Γ at a large distance from the membrane (i.e., $z \gg L$) corresponding to the self-solvation term of the unmodified generalized Born method $g_{\mathrm{bulk}} = -166/(R_i + \text{offset} + P_1)$, while g_{center} is the value of Γ at the membrane center. We used a Gaussian with $\gamma = -2.0$ and a membrane half width of $L = 15$ Å, while $g_{\mathrm{center}} = -7.67$ kcal/mol, as reported previously.[17,21,61]

The nonpolar part of the solvation free energy ΔG_{np} is modeled using an effective surface tension associated with the solvent accessible surface area.[57] Instead of a costly calculation of the accurate surface area, a mimic based on the Born radii is used, which has been shown to be very accurate but much faster.[62] As it is moved toward the center of the membrane the surface energy contribution of each atom is scaled down by a Gaussian function of the same width as Γ. For distances far from the membrane (i.e., $z \gg$

L) the nonpolar contribution is included with the positive surface tension of solvation in water, while in the center of the membrane the surface tension is negative (i.e., energy is gained by moving into this phase from the gas phase) as determined experimentally.[63] The surface tension contribution of each atom was varied using a Gaussian function with $\gamma = -1.5$, interpolating between the limiting values of 12 cal/mol·Å^2 in bulk solvent and −19 cal/mol·Å^2 at the membrane center.

The present membrane model neglects any effects due differences in lipid composition, density, and charge distribution of the two bilayer leaflets as well as effects due to the transmembrane voltage. However, it is in principle possible to include these properties by replacing the Gaussians with an equivalent nonsymmetric function. The nonpolar part of the implicit membrane model was previously parametrized against experimental transfer free energies of hydrophobic side-chain analogs,[63] and no parameters were optimized for the present simulations.

Monte Carlo Sampling. The implicit membrane model has been implemented as part of an all-atom Monte Carlo program. An efficient concerted rotation sampling technique[64] is used to move the protein backbone; in addition there are single rapid side-chain moves, with a ratio of 3 side-chain moves per backbone move. The potential energy is evaluated with the OPLS all-atom force field.[65] All nonbonded interactions as well as the GB energy are truncated using a residue-based cutoff of 14 Å, but no cutoffs are used in the pentamer simulation. In addition, one folding simulation is run without cutoff to compare to the run using cutoffs. The Born radii are recomputed for every configuration. The described setup has been shown to perform well in sampling DNA[66] and protein folding simulations.[16] We have recently demonstrated that this method is equivalent to molecular dynamics sampling, with both methods able to find the native state of several polypeptides with comparable computational effort.[67]

Replica Exchange MC (REMC). The replica exchange method has recently been reviewed in detail.[68,69] Ten replicas of each system were set up with identical fully extended initial configuration and exponentially spaced temperatures in the range 300−500 K. Every 10^4 Monte Carlo moves a replica swap with transition probability

$$p_{1\rightarrow 2} = \exp(-\Delta) \qquad (4)$$

where

$$\Delta = \left(\frac{1}{kT_1} - \frac{1}{kT_2}\right)(E_1 - E_2) \qquad (5)$$

is attempted. E_1 and E_2 are the total energies of two conformers at temperatures T_1 and T_2, respectively. High-temperature replicas facilitate the crossing of energy barriers, while low-temperature replicas extensively sample low-energy conformations. This enables the efficient and increased sampling of the entire system by frequent crossing of high-energy barriers. The exponential temperature spacing ensures a constant acceptance rate of all adjacent replica swaps.[68]

Vpu. The 30-residue TM domain of Vpu was set up identical to the NMR experiments (PDB code 1pje)[41] and has the sequence MQPIQIAIVALVVAIIIAIVVWSIVI-IEGR. An additional six-residue “solubility tag” at the C terminus used in the experiments is omitted in the simulations. The experiments did not locate all residues present in the peptide. The missing residues (1−6, 26−30) are the polar and charged residues at the helix termini, which are important for the correct orientation of the helix in the membrane. To be able to compare the current analysis with the experiment, the missing residues were added with optimized geometry in an α-helical secondary structure for the pentamer simulations. The Vpu TM monomer folding simulations were started from completely extended conformation arranged so that they span the membrane.

Free Energy Analysis. The free-energy was calculated as a function of the helix tilt and center-of-mass position along the membrane normal. For a system in thermodynamic equilibrium, the change in free energy on going from one state to another is given by

$$\Delta G = -RT \ln \frac{p_1}{p_2} \qquad (6)$$

where R is the ideal gas constant, T is the temperature, and p_i is the probability of finding the system in state i. The free energy is plotted on a two-dimensional grid, and the values are shifted so that the lowest bin is zero.

Rigid Body Energy Scan. The minimal energy conformation was calculated by exploring the entire translational and rotational space of a completely helical rigid structure of VPU in the membrane. The principal axis of the protein was determined through diagonalization of the inertia tensor using only the heavy backbone atoms. The tilt angle was defined as the angle of the principal axis with respect to the membrane normal, while the rotation angle was defined as the angle of rotation around the principal axis.

The helix was translated from −50 Å to +50 Å along the membrane normal (membrane center = 0 Å) in 1 Å steps. At each step the protein was rotated through all space to find the orientation of minimum energy by first tilting it with respect to the membrane normal and subsequent rotation around its principal axis until all tilt and rotational states have been sampled with a step size of 5°. The lowest energy conformation encountered was then subjected to a rigid body minimization in order to locate the precise location of the global energy minimum.

Results

Insertion Energy Landscape. In order to investigate the insertion-energy landscape for the local minimum energy orientations the implicit membrane potential was plotted as a function of position along the membrane normal and tilt angle, while the rotation angle was optimized (i.e., the rotation angle for each position and tilt angle is such that the energy is minimal). Figure 1 panel C shows the resulting insertion energy landscape for a completely helical structure of Vpu. The zero point of the potential was chosen at an infinite distance from the membrane. Vpu has four distinct

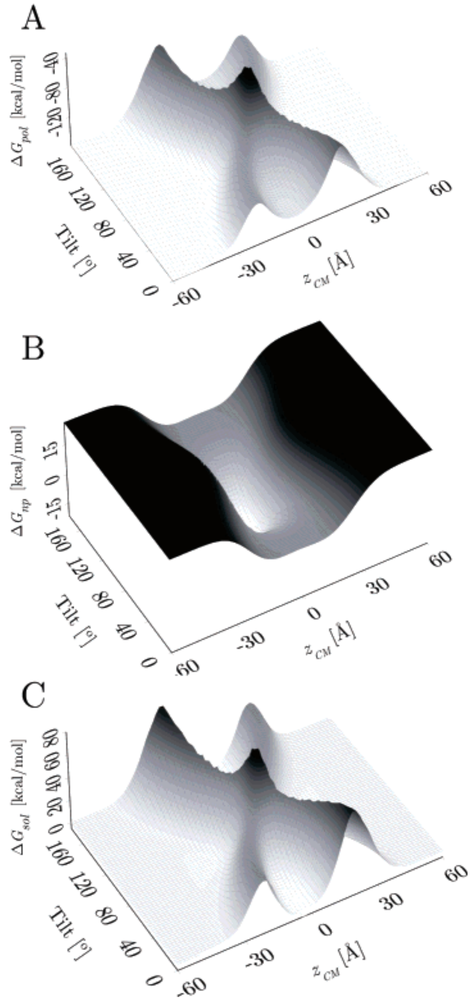


Figure 1. Insertion energy profiles. The figure shows the insertion energy of the Vpu helix with charged termini as a function of the helix tilt and center-of-mass position along the membrane normal for the optimized rotation angle (around the long axis of the helix). Panel A shows the polarization energy, panel B nonpolar energy, and panel C the total solvation energy, shifted such that it is zero at an infinite distance from the membrane.

minima, the two deepest corresponding to inserted configurations with the helices approximately parallel to the membrane normal. The other two minima are surface bound configurations with the helix axis parallel to the plane of the membrane. It should be noted that due to the symmetry of the membrane model, the cytoplasmic and intracellular minima have identical insertion energies, as do the two inserted minima.

Generally the inserted TM configuration corresponds to the global energy minimum. The insertion energy is −5.5 kcal/mol, with a tilt angle of 5.6° as well as position close to the center of the membrane—slightly shifted to 3.7 Å. Adsorption of the peptide onto the membrane surface is also favorable but to a significant lesser extent, with an energy minimum of −0.8 kcal/mol at 20 Å, and a parallel orientation with tilt angle of 80°.

To investigate the relative roles of the polar and nonpolar part of the implicit membrane energy, their contributions to the total insertion potential was also calculated. Figure 1 shows the contributions of ΔG_{pol} (panel A) and ΔG_{np} (panel B) to the overall insertion-energy landscape (panel C). Burial of charged and polar residues in the membrane interior is highly unfavorable, and the characteristic 'X' shape of ΔG_{pol} is caused by the position of such residues at the helix termini.[70,71] The hydrophobic effects are the main contributors to helix insertion and, as expected, give the lowest contribution for a completely buried helix parallel and in the center of the membrane (panel B). It is generally recognized that overall hydrophobicity is the main driving force for the integration of TM helices into the lipid bilayer.[72] Indeed the vast majority of residues in TM helices are hydrophobic.[73] Nevertheless, polar, charged, and aromatic residues are known to be important for anchoring the helix termini into the lipid headgroup environment at the membrane interfaces.[74−76] The overall potential favors TM orientations since hydrophobic residues strongly prefer an inserted to a surface-bound configuration. For the burial of a typical TM peptide of about 20 residues in the membrane, White et al. roughly estimate a hydrophobic contribution of ~40 kcal/mol, offsetting a unfavorable dehydration of the α-helical peptide backbone of about ~30 kcal/mol. This results in a net favorable free energy of about ~10 kcal/mol.[77] From the orientational scan of Vpu, we estimate a hydrophobic contribution of −22.5 kcal/mol and a polarization penalty of +17 kcal/mol, resulting in the −5.5 kcal/mol insertion of the TM helix with respect to a helix in solution.

Role of Terminal Charges. When calculating the properties of membrane proteins, it is important to take into account the charge state of the amino acid side chains. The implicit membrane model is very sensitive to changes in charge since burial of charged groups is highly unfavorable. For Vpu, both Glu 28 and Arg 30 are modeled in their charged state (pH = 7). In addition, the chain ends can be modeled as either charged (NH_3^+/COO^-) or capped with methyl groups, neutralizing the termini. The lack or presence of this additional dipole has a strong effect on the outcome of the simulations: in the capped case, there is no strongly charged group on the N-terminal part of the peptide. This has a significant effect on the insertion energy landscape discussed above. Figure 3 shows ΔG_{pol} recalculated for the capped system, revealing a markedly different shape than the uncapped system given in Figure 1 panel C. The lack of a charged residue on the N-terminal side of the helix results in a considerable lowering of the barrier (~6 kcal/mol) on the N-terminal side of the transmembrane inserted minimum. The characteristic 'X' shape is lost. While the peptide remains in a TM inserted conformation, the system can ultimately overcome the barrier and exit the membrane.

The experimental setup does not use an N-terminal cap but adds an additional six-residue "solubility tag" at the C terminus that facilitates the isolation, purification, and sample preparation of the peptide.[41] We model Vpu without the tag but both with caps and without to reveal the differences and role the charge state has on the implicit membrane model.

Helix Tilting. It has been experimentally observed that membrane proteins avoid unfavorable exposure of their hydrophobic surface to a hydrophilic solvent by matching

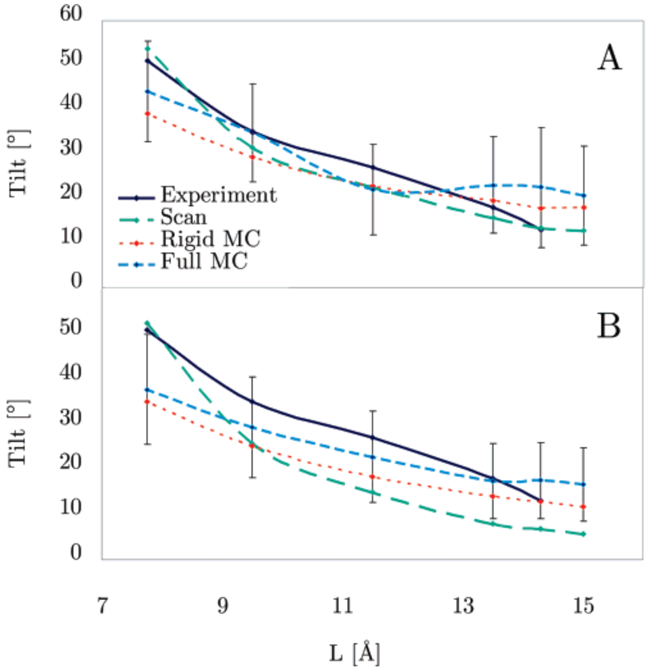


Figure 2. Tilt angle of Vpu as a function of the hydrophobic membrane half-width *L*, calculated for a α-helical TM conformation. Panel A gives the results for the capped peptide, panel B for the system with uncapped chain ends. The tilt was calculated in three ways: The dashed line gives the tilt angle of the minima encountered in the rigid body scan. The dotted line shows the average tilt angles determined in a rigid body MC simulation (10^6 MC steps) and the fine dashed line for a fully flexible MC simulation (10^9 MC steps). In the MC simulations, the tilt angle fluctuates up to ±30°, resulting in a large standard deviation of ~10°, shown here as error bars for the full MC runs. There is little difference in the results of the capped and uncapped peptide, and in both cases the tilt angle obtained with the fully flexible simulations best matches the experimental values.

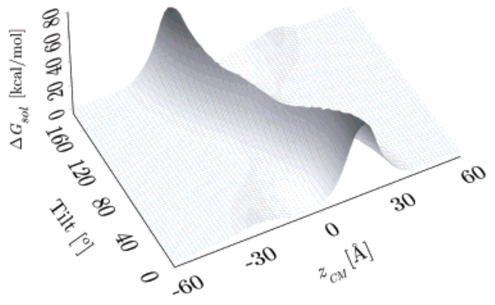


Figure 3. Insertion energy profile of the Vpu helix with capped uncharged termini as a function of the helix tilt and center-of-mass position along the membrane normal for optimized rotation angle (around the long axis of the helix). The lack of a charged residue on the N-terminal side of the helix results in a weak barrier (~6 kcal/mol) on one side of the transmembrane inserted minimum. This barrier is much larger in the uncapped helix (panel C of Figure 1).

the length of their hydrophobic helical segments to the thickness of the lipid bilayers (‘hydrophobic mismatch’).[43,78−80] Most easily, this occurs by structural adaption such as changes in helix tilt and kink. Recently, the tilt angle of the transmembrane segment of Vpu was determined experimentally in lipid bilayers of various thickness using solid-state NMR experiments of aligned samples[43] as well as bicelles.[42] These studies demonstrated that changes in tilt angle appear to be the principal mechanism for compensating the mismatch, with an increase from 18° for a hydrophobic width of $2L = 27$ Å to a much larger 51° for $2L = 15.5$ Å. In order to investigate this behavior with the implicit membrane model, a series of simulations was performed by adjusting the width of the hydrophobic segment, $2L$, reproducing the effect of the various lipid environments used in the experiments. For each membrane thickness, a complete translational and rotational scan (see method section) was performed with the perfectly α-helical conformation found in the NMR measurements, in order to determine the tilt angle and *z*-position of the energy minimum. In the second stage, the helix was run in the membrane using a rigid-body MC simulation of 1×10^6 MC steps length, giving the average values of the tilt and *z*. Finally, a third simulation was performed using a full MC run of the helix with complete flexibility for 1×10^9 MC steps. Although computationally demanding, the full MC run corresponds most closely to the experimental setup, as the system can freely move, breathe, and reveal helix kinking or even unfolding. Thus, the averages here will be the most indicative of the quality of the model.

Due to the strong dependence of the results on the charge state of the terminal residues, all simulations were performed twice, for the capped peptide and the uncapped peptide. Table 1 shows the results for the case of the uncapped peptide. The experimental estimates of the membrane thickness and tilt angles were taken from the solid-state NMR measurements of Park et al. on phospholipids bilayers[41,43] and on bicelles.[42] For all simulations, the helix remains firmly inserted in the TM state, with a slight off-center position toward the N-terminus of 2.9−4.5 Å and an insertion energy of ~5.5−8 kcal/mol. The tilt angles are also plotted in Figure 2 panel A. There is overall good agreement with the experimental results, with the observed increase in tilt as the membrane width decreases. However, in a different experimental study of Vpu using infrared dichroism, the tilt angle was determined to be 6.5° ± 1.7°.[45] This is significantly lower than what we report here, and our values better fit the NMR data. There is a progressively better match as the simulation methodology becomes more thorough: the best results are obtained with the fully flexible MC simulations, revealing the importance of conformational flexibility in determining the configurational averages. The tilt angle fluctuates up to ±30°, resulting in the large standard deviation of ~10°, which could be due to the implicit nature of the membrane model. Thus the tilt cannot be calculated more accurately than ±10°. The results for the capped Vpu are shown in Table 2 and plotted in Figure 2 panel B. Due to the low barrier, some of peptides in the long MC simulations exit the membrane after ~500 × 10^6 MC steps. For these data points, the averages are only over the TM part of the run.

The simulations all show strong kinking with angles of 25−40° at the center of the helix. However, no persistently

Table 1. Helix Tilt and *z*-Position of Uncapped Vpu[a]

		rigid scan (min)			rigid MC (10^6 steps)		full MC (10^9 steps)		
L [Å]	exp. tilt [deg]	*z* [Å]	tilt [deg]	Δ*G* [kcal/mol]	*z* [Å]	tilt [deg]	*z* [Å]	tilt [deg]	kink [deg]
15		3.8	5.7	−5.5	3.7 ± 1.5	11.8 ± 5.9	4.5 ± 2.1	16.6 ± 8.1	29.4 ± 12.6
14.3	13	3.8	6.8	−6.3	3.6 ± 1.5	12.9 ± 6.3	6.1 ± 2.0	17.5 ± 8.4	27.7 ± 11.4
13.5	18 (21)	3.7	7.9	−7.1	3.6 ± 1.5	14.0 ± 6.8	2.9 ± 2.1	17.5 ± 8.3	39.9 ± 17.1
11.5	27 (30)	3.9	14.9	−8.1	3.5 ± 1.6	18.3 ± 8.4	4.0 ± 2.5	22.8 ± 10.1	43.4 ± 15.3
9.5	35	3.5	25.8	−8.2	3.3 ± 2.1	25.1 ± 10.6	4.9 ± 2.7	29.4 ± 11.1	55.1 ± 16.4
7.75	51	4.3	52.5	−8.8	3.0 ± 2.7	34.9 ± 12.9	3.6 ± 3.3	37.7 ± 12.2	48.6 ± 16.4

[a] *L* is the membrane half width. The experimental helix tilt is taken from the solid-state NMR measurements of Park et al. on phospholipids bilayers[43] and on bicelles (brackets).[42] For the rigid body scan, the *z*-position and tilt of the minimum and its insertion energy with respect to infinite separation from the membrane is given. For the MC simulations, the averages of the *z*-position, tilt, and kink are given.

Table 2. Helix Tilt and *z*-Position of Capped Vpu[a]

		rigid scan (min)			rigid MC (10^6 steps)		full MC (10^9 steps)		
L [Å]	exp. tilt [deg]	*z* [Å]	tilt [deg]	Δ*G* [kcal/mol]	*z* [Å]	tilt [deg]	*z* [Å]	tilt [deg]	kink [deg]
15		5.0	12.8	−4.3	6.3 ± 2.3	18.1 ± 8.4	6.7 ± 2.3	20.8 ± 11.1	37.2 ± 10.6
14.3	13	4.4	13.4	−5.0	5.4 ± 2.2	17.8 ± 8.0	5.1 ± 3.4	22.5 ± 13.4	29.0 ± 12.0
13.5	18 (21)	4.3	15.6	−5.6	5.1 ± 2.6	19.5 ± 10.3	5.9 ± 2.4	23.1 ± 10.8	30.8 ± 12.6
11.5	27 (30)	3.9	22.6	−6.8	3.9 ± 2.0	22.8 ± 9.5	6.5 ± 2.3	22.0 ± 10.1	25.0 ± 11.2
9.5	35	3.5	31.5	−7.4	3.2 ± 2.3	29.3 ± 11.2	3.2 ± 2.5	34.8 ± 11.1	42.8 ± 12.2
7.75	51	4.5	53.7	−8.4	2.9 ± 2.6	39.1 ± 12.6	2.2 ± 2.7	44.1 ± 11.4	38.3 ± 14.5

[a] *L* is the membrane half width. The experimental helix tilt is taken from the solid-state NMR measurements of Park et al. on phospholipids bilayers[43] and on bicelles (brackets).[42] For the rigid body scan, the *z*-position and tilt of the minimum and its insertion energy with respect to infinite separation from the membrane is given. For the MC simulations, the averages of the *z*-position, tilt, and kink are given.

kinked structures are observed, with very fast fluctuation of the kink angle during the simulations. Interestingly, the kinking behavior is little influenced by either the membrane width or the charge state of the termini (see Tables 1 and 2). Experimentally, only a slight kink of 1−5° is observed at Ile 17 in both micelle and lipid mixture (9:1, DOPC: DOPG) bilayer environments,[41] but none is found in thinner bilayers[43] or in bicelles.[42] Simulations of Vpu oligomers with explicit lipid and solvent molecules have reported higher kink angles of 12.7−19.9°,[52] and in a recent similar simulation of the monomer kink values of 3.7−10° were found.[46] The stronger kinking in this study is almost certainly due to the implicit nature of the membrane model. In the absence of an explicitly represented strongly ordered lipid phase, the helix can flex and bend more easily.

Insertion Energy Profile from Simulations. The fully flexible MC simulations of 1 × 10^9 MC steps are sufficiently long to yield converged insertion free energy landscapes from a direct population analysis. For both the capped and uncapped simulations with $2L = 30$ Å, a two-dimensional population histogram was calculated as a function of the center-of-mass position along the membrane normal and the tilt angle. The negative logarithm of the histogram bins gives the overall solvation free energy profile of the system and is plotted in Figure 4. The free energies are relative to the lowest bin, which has been set to zero. A close similarity to the profiles shown in Figures 1 and 3 is evident and expected. Note that the profiles in Figure 4 can only extend over the conformational space that was physically sampled, while the rigid-body scan results above can plot the entire landscape—albeit for a fixed conformation. Figure 4 panel A reveals the uncapped peptide has thoroughly explored the TM inserted minimum at $z = 4.5$ Å, tilt = 16.6° and remains strongly contained by large barriers, as already visible in Figure 1. The results for the capped peptide shown in Figure 4 panel B are very different. After spending considerable time sampling the TM bound state at $z = 6.3$ Å, tilt = 18.1°, the peptide overcomes the weak barrier (see also Figure 3) after ∼500 × 10^6 MC steps and exits to the surface of the membrane. The small barrier height—caused by the lack of charged groups on the N-terminal side of the Vpu peptide—is only ∼2 kcal/mol, even smaller than the estimate of ∼6 kcal/mol from the rigid body scan.

Folding Simulations. The next step is to demonstrate that the implicit membrane model can predict the experimentally determined native state of Vpu in an ab initio protein folding simulation. For this, REMC simulations were run with 10 replicas for 1 × 10^9 MC steps each (see Methods), starting from completely extended conformations perpendicular to the membrane plane. The simulations were performed both for capped Vpu and for the uncapped system. Figure 5 shows the folding progress of the transmembrane system over the course of the simulations. Only the 318 K replica, the temperature closest to the NMR experiments, is shown. Both capped and uncapped Vpu fold into stable membrane spanning helices within the first ∼400 × 10^6 MC steps. Replicas with higher temperatures contain a large amount of helical secondary structure but do not form stable helices. No beta structure is observed in any of the simulations. Once formed, the helix shows strong tilting and kinking. To quantify the similarity to the native state—the completely helical structure found in the NMR measurements,[41] we calculated the overall system helicity as it increases over the course of the simulation, and the results are shown in Figure 6. After a steady buildup of helical content, a plateau is reached after ∼400 × 10^6 MC steps. The chain ends are

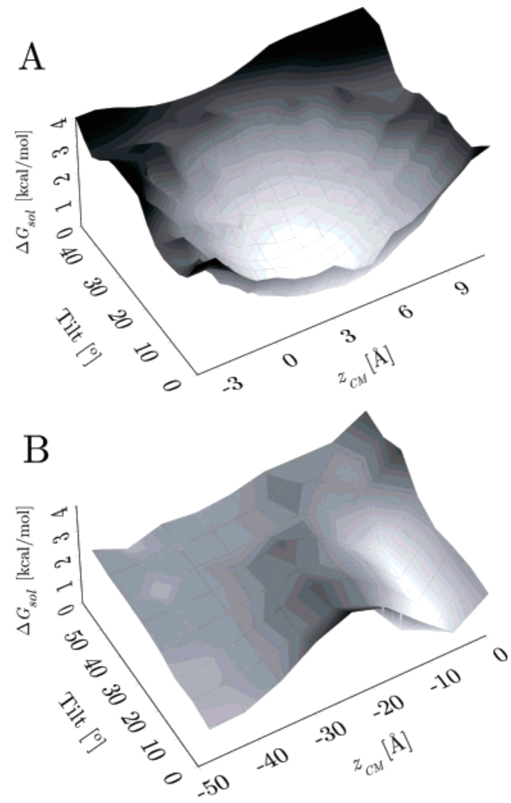


Figure 4. Free energy profile of Vpu as calculated from a population analysis for the fully flexible MC simulations in TM bound conformation. ΔG is plotted as a function of the helix tilt and center-of-mass position along the membrane normal, and is the free energy relative to the lowest bin that has been set to zero. The uncapped system is shown in panel A, revealing a stable TM inserted minimum at $z = 4.5$ Å, tilt = 16.6°, as found in the rigid-body scan (Figure 1). The capped system plotted in panel B shows the same TM minimum but is not stable and exits the TM state after 500 × 10^6 MC steps. The weak barrier (see also Figure 3) is caused by the lack of charged groups on the N-terminal side of the Vpu peptide.

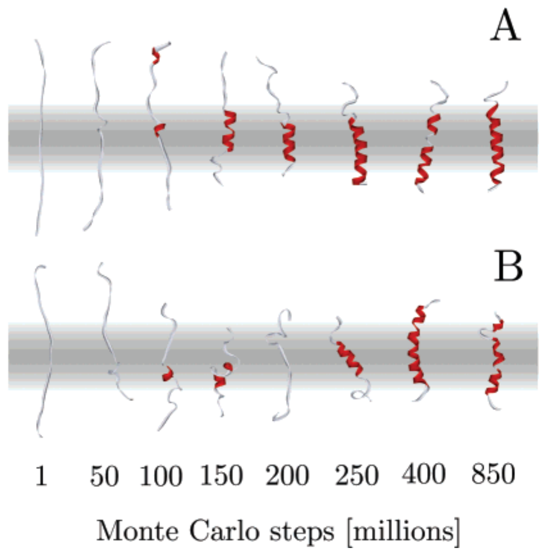


Figure 5. Transmembrane folding of Vpu for the 318 K replica of the REMC simulations, showing the capped system (panel A) and the uncapped system (panel B). Vpu folds into a stable membrane spanning helices within the first ~400 × 10^6 MC steps. Higher temperature replicas retain largely extended or coiled conformations (data not shown). It should be noted that the implicit membrane does not represent a hydrophobic slab but rather a Gaussian shaped hydrophobic zone, thus the 30 Å slabs shown are for reference only.

found to be flexible and mostly do not sample helical conformations.

The folding results are directly compared to the single fully flexible 1 × 10^9 MC step simulations of Vpu starting from the helical TM conformation, and the helicity is plotted in the same Figure 6. For both the capped and uncapped peptide strong tilting and kinking is observed (see Tables 1 and 2), but the completely helical conformation remains intact during the runs, with 78.5% ± 6.6% helicity for the uncapped peptide and 66.5% ± 2.3% for the capped system. The lower helicity of the capped peptide is due to the more flexible chain termini. While the REMC folding run for the capped peptide reaches the same plateau in helicity as the reference native run, the helicity observed for the REMC folding run with the uncapped peptide is lower. This indicates a sampling problem of the more highly polar system, where partly helical structures present in the various replicas persist much longer due to charge−charge interactions than in the case of the capped system, where the complete helix quickly dominates. Such partly helical structures are swapped into the 318 K replica and thus contribute to the overall helicity. Contrary, in the native TM state simulation of the uncapped peptide, partly helical conformations are not sampled at all. Overall, the results prove that the native state of Vpu can be accurately predicted after a relatively short MC simulation from a completely random conformation. The choice of the chain termini seems to not influence these results. This matches experimental observations that found the effect of terminal caps on the helicity of a designed TM bound peptide to be marginal.[81]

In order to investigate the role of the cutoff of the nonbonded interactions, an additional REMC run was performed with a complete evaluation of all nonbonded interactions (no cutoff), including the GB terms. In general, the GB model is ideal when truncating electrostatic interactions: The Coulomb term (E_{coul}) plus the GB polarization term (ΔG_{pol}) for two largely separated atoms is simply a screened Coulomb interaction, weakened by $\epsilon_{\mathrm{water}}$. Thus, the contribution of these far terms to the total nonbonded energy is much smaller than in vacuum electrostatics. However, for deeply buried atoms (e.g., atoms in the membrane interior), the resulting large Born radii will reduce their contribution to ΔG_{pol} so much that cutoff artifacts of E_{coul} may still be significant.

The simulation was carried out identically to the folding simulations above. No significant deviation to the folding runs using a cutoff were detected. The buildup of helicity during the simulation is shown in Figure 5 and is almost the same as in the other runs. Thus we conclude that cutoff effects do not significantly influence the folding results.

Vpu Channel Simulation. Vpu polypeptides that contain the membrane spanning segment have ion-channel activity, and since Vpu forms single TM helices, this is achieved by

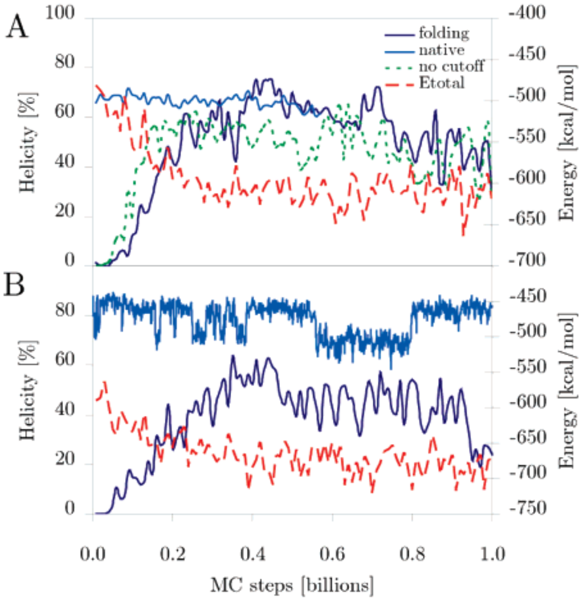


Figure 6. This graph shows the buildup of the helicity of Vpu for the 318 K replica during the folding runs of 1 × 10^9 MC steps (left axis − solid lines) and the change of the total system energy (internal energy plus solvation free energy, right axis − dotted line). The helicity of the single simulations starting from the native helical TM inserted state is also shown in the same plot for comparison. Panel **A**: capped system. The two folding trajectories are with and without cutoffs, respectively. Panel **B**: uncapped peptide. In both cases, the complete TM helix forms in ~400 × 10^6 MC steps. Due to frequent replica swaps to only partly helical structures that contribute to the average, the helicity is lower in the REMC folding runs and fluctuates much more, indicating longer sampling is still required (i.e., not all replicas have folded). In the native MC runs, the helix remains completely stable, but the capped peptide exits the membrane after ~500 × 10^6 MC steps.

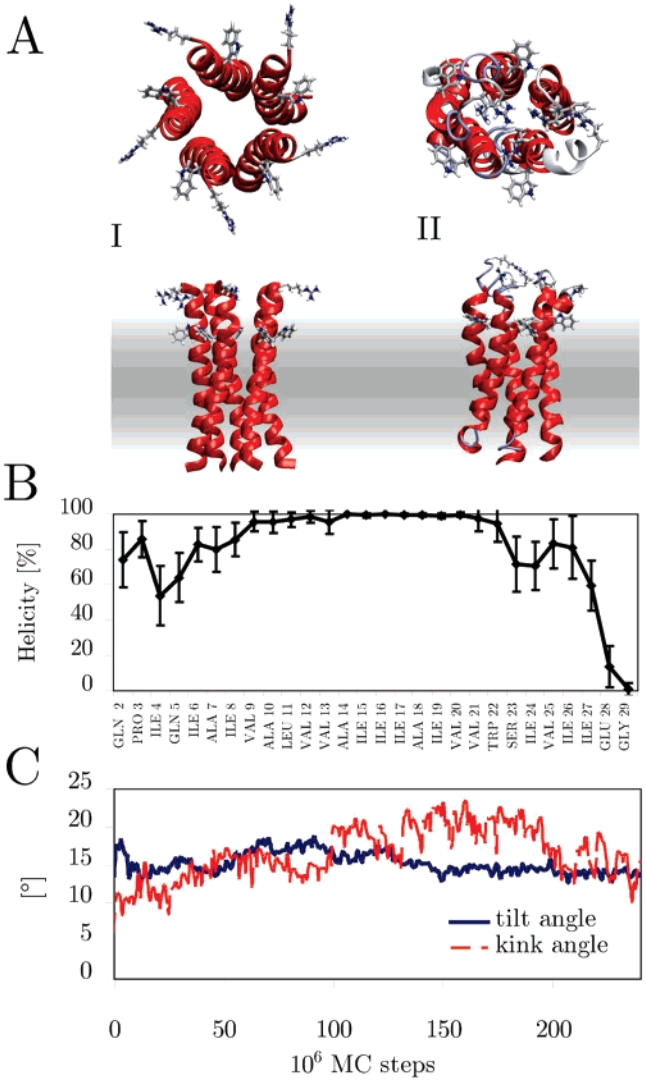


Figure 7. Vpu pentamer simulation. **A**: View from the top (C-terminal to N-terminal) and the side of the first (I) and the final structure (II) of the 240 × 10^6 MC step simulation. **B**: average helicity per residue over the course of the simulation. **C**: average tilt and kink angle calculated over the center segment of the 5 monomers during the simulations.

their oligomerization in lipid bilayers. The N-terminal TM domain of Vpu has been shown to form homo-oligomers both in vivo and in vitro.[35] Oligomeric Vpu has a cation-specific channel activity for which only the TM sequence is required.[36−40] The structure of the oligomer is not known, but a pentamer is thought most likely, as suggested by several simulation studies[49−55] and estimated from single-channel conductance measurements.[52] We set up the pentamer in the implicit membrane, using the structure proposed by Park et al.[42] (PDB code 1PI7), with the interfacial tryptophan residues facing the lipid environment. The missing residues (1−6, 26−30) of each chain were added with an α-helical secondary structure. The MC simulation was run for 240 × 10^6 steps without any constraints and without any cutoffs for the nonbonded interactions.

The pentamer remains stable throughout the simulation, as shown in Figure 7, with all five helices firmly interlocked at their hydrophobic segments. No displacement normal to the membrane plane is observed, and the pentamer remains in its inserted state, anchoring the tryptophan residues at the interface, the preferred location for this amino acid.[73,74,82] There is no loss of helical structure except at the chain ends, where a slight unwinding is observed (panel B). Panel C shows the development of the tilt and kink angles averaged over the individual helices as a function of simulation time. The average tilt of 15° ± 1.4° is similar to the 14.5° reported by Cordes et al. in explicit lipid and solvent simulation,[52] while the average kink angle of 16° ± 3.5° is slightly smaller than the 19.9° obtained in that study. These values are smaller than those of the Vpu monomer, which can be explained by the thicker membrane environment and the fact that the helices are firmly bound to each other, preventing stronger tilting. For the same reason, there is less fluctuation in the tilt and kink angle throughout the run.

The loss of helical structure at the N-terminus, involving a proline residue, is only minor. Indeed it was observed experimentally that the tendency of proline to disrupt helical structures on membrane interfaces is weak.[81] Structural deviation from the helix is much more pronounced at the C-terminus and could be due to a limitation of the all-implicit model: In a fully explicit bilayer, this region corresponds to the very dense lipid headgroup layer that would prevent significant helix perturbation. At present the implicit membrane model does not account for the polar nature of the lipid headgroup region, but it is in principle possible to add such a contribution. Interestingly, perturbation of helical structure was only sporadically observed in the monomer folding simulations, since the single helix can tilt more strongly to bury completely in the hydrophobic zone. In previously reported explicit MD pentamer simulations, significant structural change was not seen.[49−55] However, the limited sampling time (1−5 ns) of these simulations is much shorter than the extensive sampling achieved with the implicit membrane model. Cordes et al. speculate that the destabilization of the Vpu bundles at the C-terminal end are due to a EYR-motif, with the arginines covering the pore acting as a selectivity filter.[52] In our simulations, the even lower stability of the C-terminus is almost certainly caused by using a different mutant, with a glycine (EGR instead of EYR), but similar to the previous results, the arginines, with their flexible side chains, are found to point to the inside covering the pore throughout the simulations (Figure 7 panel A).

Conclusion and Discussion

The secondary structure of the N-terminal transmembrane helix of Vpu is predicted in protein folding simulations using an implicit membrane model and all-atom representation of the protein. It forms a stable helix firmly inserted in the membrane, and the observed average tilt and kink angles closely match experimental results from NMR measurements. In addition, the experimentally observed increase of the helix tilt in membranes of decreasing hydrophobic thickness (‘hydrophobic mismatch’) is accurately reproduced. The results reveal the strength of the generalized Born implicit membrane model in capturing the essential membrane energetics through a polarization term and a hydrophobic burial term. The lack of explicit lipid and solvent molecules enables greatly accelerated sampling currently not achievable in explicit bilayer simulations. A simulation of a pentamer of the transmembrane Vpu helix reveals a stable channel, in agreement with previous MD simulation efforts.

Simulating the folding of small membrane bound polypeptides and oligomeric TM bundles in a completely implicit membrane model is challenging. To be useful for studying a wide range of peptides, such methods must essentially fulfill several requirements: (a) single TM or surface bound helices as well as integral membrane proteins must retain their experimentally observed structure (e.g., NMR data) despite being surrounded only by a continuum environment, (b) such stability must not be caused by the use of models and parameter sets that overly bias helical structures or by the use of artificial constraining potentials, (c) in order to justify the substantial simplifications entailed by an implicit representation of the membrane the model must be significantly faster than equivalent fully atomistic membrane simulations, enabling extensive conformational sampling that goes beyond simple rigid orientational scans, and (d) a wide range of experimental data must be reproduced, especially the experimental determined partitioning free energy of polypeptides into both the membrane interfaces and the membrane interior (biophysical and statistical hydrophobicity scales)[70,83] as well as the recent biological hydrophobicity scale determined by translocon mediated insertion.[84]

While the presented implicit membrane model performs well on these points, we have identified two key deficiencies that have yet to be overcome: The neglect of effects due to the complex lipid headgroup environment and the improper treatment of charged residues. In practical terms, this signifies that the interfacial regions of the membrane are poorly described, and some loss of defined secondary structure is observed in the segments of membrane bound peptides in this region. This will be especially problematic for simulating surface bound peptides (e.g., antimicrobials) and matching experimental partitioning free energy of unfolded polypeptides into membrane interfaces.[85] For charged residues, the GB model predicts a large desolvation penalty on moving into the hydrophobic region. In the biological scale of Hessa et al., the effect of the additional charge is almost nonexistent: For example, the apparent free energy of insertion $\Delta G_{\mathrm{app}}^{\mathrm{aa}}$ of an amino acid located at the center of a 19 residue TM helix is roughly equally unfavorable for glutamine (2.36) and glutamic acid (2.68),[84] whereas the burial penalty due to the additional full charge is large in the GB model. It would therefore be more appropriate to use variable protonation state models, where residues can be neutralized upon entering the membrane. Alternatively, White at al. suggests that the strong positional dependence of charged residues in the biological scale could be due to distorted bilayer states, where the headgroups are in contact with buried peptide charges, and the hydrophobic thickness is significantly reduced.[86] This is currently beyond the limits of the implicit membrane model.

The subtle energetic and entropic effects that can be neglected when representing the complex lipid bilayer environment implicitly are illustrated by a recently reported study of TM bundles in an implicit GB membrane by Bu et al.,[22] where predicting the correct native oligomerization state of several homo-oligomers was only partially achieved. A high population of non-native (as compared to the NMR structure) equilibrium structures were encountered at experimental temperatures. The authors used a cylindrical harmonic restraining potential to prevent the oligomers from disintegrating into individual helices drifting away from each other, enabling stability at the elevated temperatures used in the replica exchange runs. In our simulations, no restraining potential is used, with the Vpu pentamer remaining tightly packed throughout the simulation.

Ultimately, further improvement is required for a more accurate modeling of the polar lipid headgroup region of the membrane, which will involve additions to the implicit membrane that go beyond simple dielectric treatment. In addition, the inclusion of variable protonation state models is probably a good idea if sequences with many charged residues are studied. Such efforts are currently underway.

Acknowledgment. We are grateful to A. Di Nola for helpful discussions. J.P.U. is funded by an Emmy Noether fellowship of the German Research Foundation. M.B.U. is funded by an international fellowship of the Wellcome Trust.

References

(1) Biggin, P. C.; Sansom, M. S. Interactions of alpha-helices with lipid bilayers: a review of simulation studies. *Biophys. Chem.* **1999**, *76*, 161−183.

(2) Forrest, L. R.; Sansom, M. S. P. Membrane simulations: bigger and better? *Curr. Opin. Struct. Biol.* **2000**, *10*, 174−181.

(3) Domene, C.; Bond, P. J.; Sansom, M. S. P. Membrane protein simulations: Ion channels and bacterial outer membrane proteins. In *Protein Simulations*; Daggett, V., Ed.; Academic Press Inc.: San Diego, CA, 2003; Vol. 66, p 159+.

(4) Nymeyer, H.; Woolf, T. B.; Garcia, A. E. Folding is not required for bilayer insertion: Replica exchange simulations of an alpha-helical peptide with an explicit lipid bilayer. *Proteins* **2005**, *59*, 783−790.

(5) Braun, R.; Engelman, D. M.; Schulten, K. Molecular Dynamics Simulations of Micelle Formation around Dimeric Glycophorin A Transmembrane Helices. *Biophys. J.* **2004**, *87*, 754−63.

(6) Bond, P. J.; Cuthbertson, J. M.; Deol, S. S.; Sansom, M. S. MD simulations of spontaneous membrane protein/detergent micelle formation. *J. Am. Chem. Soc.* **2004**, *126*, 15948−9.

(7) Woolf, T. B.; Zuckerman, D. M.; Lu, N. D.; Jang, H. B. Tools for channels: moving towards molecular calculations of gating and permeation in ion channel biophysics. *J. Mol. Graph.* **2004**, *22*, 359−368.

(8) Efremov, R. G.; Nolde, D. E.; Konshina, A. G.; Syrtcev, N. P.; Arseniev, A. S. Peptides and proteins in membranes: What can we learn via computer simulations? *Curr. Med. Chem.* **2004**, *11*, 2421−2442.

(9) Feig, M.; Chocholoušová, J.; Tanizaki, S. Extending the horizon: towards the efficient modeling of large biomolecular complexes in atomic detail. *Theor. Chem. Acc.* **2006**, *116*, 194−205.

(10) Born, M. Volumen und Hydratationswärme der Ionen. *Z. Phys.* **1920**, *1*, 45−48.

(11) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(12) Chowdhury, S.; Zhang, W.; Wu, C.; Xiong, G. M.; Duan, Y. Breaking non-native hydrophobic clusters is the rate-limiting step in the folding of an alanine-based peptide. *Biopolymers* **2003**, *68*, 63−75.

(13) Jang, S.; Shin, S.; Pak, Y. Molecular dynamics study of peptides in implicit water: Ab initio folding of beta-hairpin, beta-sheet, and beta beta alpha- motif. *J. Am. Chem. Soc.* **2002**, *124*, 4976−4977.

(14) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.

(15) Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **2002**, *420*, 102−106.

(16) Ulmschneider, J. P.; Jorgensen, W. L. Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation. *J. Am. Chem. Soc.* **2004**, *126*, 1849−1857.

(17) Spassov, V. Z.; Yan, L.; Szalma, S. Introducing an implicit membrane in generalized Born/solvent accessibility continuum solvent models. *J. Phys. Chem. B* **2002**, *106*, 8726−8738.

(18) Im, W.; Feig, M.; Brooks, C. L., III. An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys. J.* **2003**, *85*, 2900−2918.

(19) Tanizaki, S.; Feig, M. A generalized Born formalism for heterogeneous dielectric environments: application to the implicit modeling of biological membranes. *J. Chem. Phys.* **2005**, *122*, 124706.

(20) Tanizaki, S.; Feig, M. Molecular Dynamics Simulations of Large Integral Membrane Proteins with an Implicit Membrane Model. *J. Phys. Chem. B* **2006**, *110*, 548−556.

(21) Ulmschneider, M. B.; Ulmschneider, J. P.; Sansom, M. S. P.; Di Nola, A. A generalized Born implicit membrane representation compared to experimental insertion free energies. *Biophys. J.* **2007**, *92*, 2338−2349.

(22) Bu, L.; Im, W.; Brooks, C. L., III. Membrane Assembly of Simple Helix Homo-Oligomers Studied via Molecular Dynamics Simulations. *Biophys. J.* **2007**, *92*, 854−863.

(23) Im, W.; Brooks, C. L. Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6771−6776.

(24) Im, W.; Brooks, C. L., III. De novo Folding of Membrane Proteins: An Exploration of the Structure and NMR Properties of the fd Coat Protein. *J. Mol. Biol.* **2004**, *337*, 513−519.

(25) Lee, J.; Im, W. Implementation and application of helix-helix distance and crossing angle restraint potentials. *J. Comput. Chem.* **2007**, *28*, 669−680.

(26) Im, W.; Chen, J. H.; Brooks, C. L. Peptide and protein folding and conformational equilibria: Theoretical treatment of electrostatics and hydrogen bonding with implicit solvent models. In *Peptide Solvation and H-Bonds*; Baldwin, R., Baker, D., Eds.; Elsevier Academic Press Inc.: San Diego, CA, 2006; Vol. 72, pp 173−198.

(27) Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. Monte Carlo folding of trans-membrane helical peptides in an implicit generalized Born membrane. *Proteins* **2007**, in press.

(28) Cohen, E. A.; Terwilliger, E. F.; Sodroski, J. G.; Haseltine, W. A. Identification of a Protein Encoded by the Vpu Gene of Hiv-1. *Nature* **1988**, *334*, 532−534.

(29) Strebel, K.; Klimkait, T.; Martin, M. A. A Novel Gene of Hiv-1, Vpu, and Its 16-Kilodalton Product. *Science* **1988**, *241*, 1221−1223.

(30) Marassi, F. M.; Ma, C.; Gratkowski, H.; Straus, S. K.; Strebel, K.; Oblatt-Montal, M.; Montal, M.; Opella, S. J. Correlation of the structural and functional domains in the membrane protein Vpu from HIV-1. *Proc. Natl. Acad. Sci.* **1999**, *96*, 14336−14341.

(31) Willey, R. L.; Maldarelli, F.; Martin, M. A.; Strebel, K. Human-immunodeficiency-virus type-1 Vpu protein regulates the formation of intracellular Gp160-Cd4 complexes. *J. Virol.* **1992**, *66*, 226−234.

(32) Levesque, K.; Finzi, A.; Binette, J.; Cohen, E. A. Role of CD4 receptor down-regulation during HIV-1 infection. *Curr. HIV Res.* **2004**, *2*, 51−59.

(33) Klimkait, T.; Strebel, K.; Hoggan, M. D.; Martin, M. A.; Orenstein, J. M. The human immunodeficiency virus type 1-specific protein Vpu is required for efficient virus maturation and release. *J. Virol.* **1990**, *64*, 621−629.

(34) Strebel, K.; Klimkait, T.; Maldarelli, F.; Martin, M. A. Molecular and biochemical analyses of human immunodeficiency virus type-1 Vpu protein. *J. Virol.* **1989**, *63*, 3784−3791.

(35) Maldarelli, F.; Chen, M. Y.; Willey, R. L.; Strebel, K. Human-immunodeficiency-virus type-1 Vpu protein is an oligomeric type-I integral membrane-protein. *J. Virol.* **1993**, *67*, 5056−5061.

(36) Ma, C.; Marassi, F. M.; Jones, D. H.; Straus, S. K.; Bour, S.; Strebel, K.; Schubert, U.; Oblatt-Montal, M.; Montal, M.; Opella, S. J. Expression, purification, and activities of full-length and truncated versions of the integral membrane protein Vpu, from HIV-1. *Protein Sci.* **2002**, *11*, 546−557.

(37) Kochendoerfer, G. G.; Jones, D. H.; Lee, S.; Oblatt-Montal, M.; Opella, S. J.; Montal, M. Functional characterization and NMR spectroscopy on full-length Vpu from HIV-1 prepared by total chemical synthesis. *J. Am. Chem. Soc.* **2004**, *126*, 2439−46.

(38) Schubert, U.; FerrerMontiel, A. V.; OblattMontal, M.; Henklein, P.; Strebel, K.; Montal, M. Identification of an ion channel activity of the Vpu transmembrane domain and its involvement in the regulation of virus release from HIV-1-infected cells. *FEBS Lett.* **1996**, *398*, 12−18.

(39) Ewart, G. D.; Sutherland, T.; Gage, P. W.; Cox, G. B. The Vpu protein of human immunodeficiency virus type 1 forms cation-selective ion channels. *J. Virol.* **1996**, *70*, 7108−7115.

(40) Romer, W.; Lam, Y. H.; Fischer, D.; Watts, A.; Fischer, W. B.; Goring, P.; Wehrspohn, R. B.; Gosele, U.; Steinem, C. Channel activity of a viral transmembrane peptide in micro-BLMs: Vpu(1-32) from HIV-1. *J. Am. Chem. Soc.* **2004**, *126*, 16267−16274.

(41) Park, S. H.; Mrse, A. A.; Nevzorov, A. A.; Mesleh, M. F.; Oblatt-Montal, M.; Montal, M.; Opella, S. J. Three-dimensional structure of the channel-forming trans-membrane domain of virus protein "u" (Vpu) from HIV-1. *J. Mol. Biol.* **2003**, *333*, 409−24.

(42) Park, S. H.; De Angelis, A. A.; Nevzorov, A. A.; Wu, C. H.; Opella, S. J. Three-dimensional Structure of the Trans-membrane Domain of Vpu from HIV-1 in Aligned Phospholipid Bicelles. *Biophys. J.* **2006**, *91*, 3032−3042.

(43) Park, S. H.; Opella, S. J. Tilt angle of a trans-membrane helix is determined by hydrophobic mismatch. *J. Mol. Biol.* **2005**, *350*, 310−318.

(44) Sharpe, S.; Yau, W. M.; Tycko, R. Structure and dynamics of the HIV-1 Vpu transmembrane domain revealed by solid-state NMR with magic-angle spinning. *Biochemistry* **2006**, *45*, 918−33.

(45) Kukol, A.; Arkin, I. T. Vpu transmembrane peptide structure obtained by site-specific fourier transform infrared dichroism and global molecular dynamics searching. *Biophys. J.* **1999**, *77*, 1594−601.

(46) Lemaitre, V.; Willbold, D.; Watts, A.; Fischer, W. B. Full length Vpu from HIV-1: Combining molecular dynamics simulations with NMR spectroscopy. *J. Biomol. Struct. Dyn.* **2006**, *23*, 485−496.

(47) Sramala, I.; Lemaitre, V.; Faraldo-Gomez, J. D.; Vincent, S.; Watts, A.; Fischer, W. B. Molecular dynamics simulations on the first two helices of Vpu from HIV-1. *Biophys. J.* **2003**, *84*, 3276−3284.

(48) Sun, F. Molecular dynamics simulation of human immunodeficiency virus protein U (Vpu) in lipid/water Langmuir monolayer. *J. Mol. Model.* **2003**, *9*, 114−123.

(49) Grice, A. L.; Kerr, I. D.; Sansom, M. S. P. Ion channels formed by HIV-1 Vpu: A modelling and simulation study. *FEBS Lett.* **1997**, *405*, 299−304.

(50) Moore, P. B.; Zhong, Q. F.; Husslein, T.; Klein, M. L. Simulation of the HIV-1 Vpu transmembrane domain as a pentameric bundle. *FEBS Lett.* **1998**, *431*, 143−148.

(51) Cordes, F. S.; Kukol, A.; Forrest, L. R.; Arkin, I. T.; Sansom, M. S. P.; Fischer, W. B. The structure of the HIV-1 Vpu ion channel: modelling and simulation studies. *Biochim. Biophys. Acta* **2001**, *1512*, 291−298.

(52) Cordes, F. S.; Tustian, A. D.; Sansom, M. S. P.; Watts, A.; Fischer, W. B. Bundles consisting of extended transmembrane segments of Vpu from HIV-1: Computer simulations and conductance measurements. *Biochemistry* **2002**, *41*, 7359−7365.

(53) Lopez, C. F.; Montal, M.; Blasie, J. K.; Klein, M. L.; Moore, P. B. Molecular dynamics investigation of membrane-bound bundles of the channel-forming transmembrane domain of viral protein U from the human immunodeficiency virus HIV-1. *Biophys. J.* **2002**, *83*, 1259−1267.

(54) Kim, C. G.; Lemaitre, V.; Watts, A.; Fischer, W. B. Drug-protein interaction with Vpu from HIV-1: proposing binding sites for amiloride and one of its derivatives. *Anal. Bioanal. Chem.* **2006**, *386*, 2213−2217.

(55) Lemaitre, V.; Ali, R.; Kim, C. G.; Watts, A.; Fischer, W. B. Interaction of amiloride and one of its derivatives with Vpu from HIV-1: a molecular dynamics simulation. *FEBS Lett.* **2004**, *563*, 75−81.

(56) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.* **2007**, *157*, 593−605.

(57) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

(58) Gallicchio, E.; Levy, R. M. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* **2004**, *25*, 479−499.

(59) Im, W.; Lee, M. S.; Charles, L.; Brooks, I. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691−1702.

(60) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. Free energies of hydration from a generalized Born model and an ALL-atom force field. *J. Phys. Chem. B* **2004**, *108*, 16264−16270.

(61) Parsegian, A. Energy of an Ion crossing a Low Dielectric Membrane: Solutions to Four Relevant Electrostatic Problems. *Nature* **1969**, *221*, 844−846.

(62) Schaefer, M.; Bartels, C.; Karplus, M. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.* **1998**, *284*, 835−48.

(63) Radzicka, A.; Wolfenden, R. Comparing the Polarities of the Amino-Acids - Side-Chain Distribution Coefficients between the Vapor-Phase, Cyclohexane, 1-Octanol, and Neutral Aqueous-Solution. *Biochemistry* **1988**, *27*, 1664−1670.

(64) Ulmschneider, J. P.; Jorgensen, W. L. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J. Chem. Phys.* **2003**, *118*, 4261−4271.

(65) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(66) Ulmschneider, J. P.; Jorgensen, W. L. Monte Carlo backbone sampling for nucleic acids using concerted rotations including variable bond angles. *J. Phys. Chem. B* **2004**, *108*, 16883−16892.

(67) Ulmschneider, J. P.; Ulmschneider, M. B.; Di, Nola, A. Monte Carlo vs Molecular Dynamics for All-Atom Polypeptide Folding Simulations. *J. Phys. Chem. B* **2006**, *110*, 16733−42.

(68) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(69) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910−3916.

(70) Ulmschneider, M. B.; Sansom, M. S.; Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **2005**, *59*, 252−65.

(71) Ulmschneider, M. B.; Sansom, M. S.; Di Nola, A. Evaluating tilt angles of membrane-associated helices: comparison of computational and NMR techniques. *Biophys. J.* **2006**, *90*, 1650−60.

(72) von Heijne, G. Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* **1997**, *66*, 113−139.

(73) Ulmschneider, M. B.; Sansom, M. S. P. Amino acid distributions in integral membrane protein structures. *Biochim. Biophys. Acta-Biomembr.* **2001**, *1512*, 1−14.

(74) Yau, W. M.; Wimley, W. C.; Gawrisch, K.; White, S. H. The preference of tryptophan for membrane interfaces. *Biochemistry* **1998**, *37*, 14713−14718.

(75) Strandberg, E.; Morein, S.; Rijkers, D. T.; Liskamp, R. M.; van der Wel, P. C.; Killian, J. A. Lipid dependence of membrane anchoring properties and snorkeling behavior of aromatic and charged residues in transmembrane peptides. *Biochemistry* **2002**, *41*, 7190−7198.

(76) de Planque, M. R.; Kruijtzer, J. A.; Liskamp, R. M.; Marsh, D.; Greathouse, D. V.; Koeppe, R. E., II; de Kruijff, B.; Killian, J. A. Different membrane anchoring positions of tryptophan and lysine in synthetic transmembrane alpha-helical peptides. *J. Biol. Chem.* **1999**, *274*, 20839−20846.

(77) White, S. H.; von Heijne, G. Transmembrane helices before, during, and after insertion. *Curr. Opin. Struct. Biol.* **2005**, *15*, 378−386.

(78) Killian, J. A.; Salemink, I.; de Planque, M. R.; Lindblom, G.; Koeppe, R. E., II; Greathouse, D. V. Induction of nonbilayer structures in diacylphosphatidylcholine model membranes by transmembrane alpha-helical peptides: importance of hydrophobic mismatch and proposed role of tryptophans. *Biochemistry* **1996**, *35*, 1037−1045.

(79) White, S. H.; Wimley, W. C. Hydrophobic interactions of peptides with membrane interfaces. *Biochim. Biophys. Acta* ***1998***, 1376, 339−*52.*

(80) Duong-Ly, K. C.; Nanda, V.; Degrad, W. F.; Howard, K. P. The conformation of the pore region of the M2 proton channel depends on lipid bilayer environment. *Protein Sci.* **2005**, *14*, 856−861.

(81) Ladokhin, A. S.; White, S. H. Interfacial Folding and Membrane Insertion of a Designed Helical Peptide. *Biochemistry* **2004**, *43*, 5782−5791.

(82) de Planque, M. R.; Bonev, B. B.; Demmers, J. A.; Greathouse, D. V.; Koeppe, R. E., II; Separovic, F.; Watts, A.; Killian, J. A. Interfacial anchor properties of tryptophan residues in transmembrane peptides can dominate over hydrophobic matching effects in peptide-lipid interactions. *Biochemistry* **2003**, *42*, 5341−5348.

(83) Wimley, W. C.; Creamer, T. P.; White, S. H. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* **1996**, *35*, 5109−24.

(84) Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **2005**, *433*, 377−81.

(85) Hristova, K.; White, S. H. An Experiment-Based Algorithm for Predicting the Partitioning of Unfolded Peptides into Phosphatidylcholine Bilayer Interfaces. *Biochemistry* **2005**, *44*, 12614−12619.

(86) White, S. H.; von Heijne, G. Do protein-lipid interactions determine the recognition of transmembrane helices at the ER translocon? *Biochem. Soc. Trans.* **2005**, *33*, 1012−*5.*

CT700103K

JCTC Journal of Chemical Theory and Computation

Prediction of the Structure of Complexes Comprised of Proteins and Glycosaminoglycans Using Docking Simulation and Cluster Analysis

Tsubasa Takaoka,[†] Kenichi Mori,[†] Noriaki Okimoto,[‡] Saburo Neya,[†] and Tyuji Hoshino*,[†,§]

Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan, Bioinformatics Group, GSC, RIKEN, Yokohama, Kanagawa 230-0046, Japan, and PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan

Received February 5, 2007

Abstract: A typical docking simulation provides information on the structure of ligand−receptor complexes and their binding affinity in terms of a docking energy. We have developed a potent method combining a docking simulation with cluster analysis to extract adequate docking structures from the many possible output structures of the simulation. First, we tried to predict the structure of basic fibroblast growth factor (bFGF) bound to heparin, using the docking simulation program AutoDock 3.0. Two X-ray crystal structures had already been obtained for bFGF. One was a complex of the protein and heparin, a kind of glycosaminoglycan, and the other, only the protein itself, hereafter called a simplex. We docked a heparin molecule onto the protein simplex and generated many trial structures for the bFGF−heparin complex. The structures of those docked complexes were optimized through energy minimization by AMBER8. Although neither the docking energy calculated by AMBER8 nor that calculated by AutoDock 3.0 could be used satisfactorily by themselves to select a proper heparin-binding complex from the output structures, the majority of the structures generated by AutoDock 3.0 were fairly close to each other in atom geometry, and the averaged geometry over these structures was also close to that of the crystal. Hence, we utilized only the atom geometry for evaluation and carried out cluster analysis with the collection of geometries. This procedure enabled selection of a structure considerably close to the crystal's. We applied this approach to two other heparin-binding proteins: antithrombin and annexin V. Two crystal structures, a complex and a simplex, had been elucidated for these proteins as well as for bFGF. Our trials gave an exact prediction of the heparin-binding structures of these proteins, showing the approach in this study is effective in studying the docking of ligands that have a variety of docking conformations due to the presence of multiple rotatable bonds and charged chemical groups.

Introduction

In silico screening is a powerful and indispensable computational tool in drug discovery and development because it enables analysis of the intermolecular interactions between proteins (receptors) and chemical compounds (ligands) and prediction of their interaction energy. A key process in in silico screening is modeling of complexes of lead compounds and target proteins. The structure for the target protein bound to a chemical compound is usually not available even if structural information on the unliganded protein has been disclosed by X-ray crystal analysis or nuclear magnetic resonance (NMR). In this case, the complex must be modeled from the individual structures of the unliganded protein

* Corresponding author e-mail: hoshino@faculty.chiba-u.jp.

[†] Chiba University.

[‡] RIKEN.

[§] PRESTO, Japan Science and Technology Agency.

10.1021/ct700029q CCC: $37.00
Published on Web 10/10/2007

simplex and the chemical compound. A docking simulation is a computational technique for enabling such modeling. In the docking simulation, a small molecule, such as a peptide or compound, is to be bound to a macromolecule, such as a protein or enzyme. Numerous conformations of the small molecule and the corresponding energies when bound to the macromolecule are calculated in one simulation. A lower binding energy indicates a higher probability of formation of a complex. The calculated binding energy, however, is poorly correlated with the closeness of the structure of the complex to that of the crystal, and it is difficult to select an optimal structure from the numerous ligand conformations generated by the docking simulation. This problem is particularly serious in the docking of glycosaminoglycans (GAGs) and GAG-binding proteins because the intended ligands, glycosaminoglycans, have many conformational variations.

The aim of this study is to establish a procedure for finding an adequate conformation of heparins bound to a target protein. Most of the currently available software programs for docking simulations are based on the assumption that a ligand molecule is held inside the binding pocket of the target protein, a pocket often composed of many hydrophobic amino acid residues. Paul and Rognan attempted to reproduce 100 crystal structures of ligand−protein complexes using several kinds of docking software programs.[1] They evaluated the ability of the software programs to correctly predict the ligand-binding structures and found that the rates for correct prediction were 39% for DOCK,[2] 51% for FlexX,[3] and 56% for GOLD 3.0.[4] In addition, they incorporated cluster analysis into the three docking programs and succeeded in improving the accuracy of the docking output. Success in the docking of glycosaminoglycans (GAGs) to proteins, however, has not been reported yet. The aim of this work is to provide a promising approach for the docking of GAGs to proteins. GAG−protein docking is very challenging because of the highly flexible nature of the GAG chain, high charge-density of the GAG binding site, and weak surface complementarity at the GAG−protein interface.

The interaction of GAGs with proteins plays a significant role in the regulation of many physiological processes, such as homeostasis, growth factor activity, anticoagulation, cell adhesion, and enzyme regulation.[5−8] For example, heparin is now used as a coagulator in surgery. However, little is known about the mechanism of the interaction of GAGs with proteins. Since it is difficult to crystallize a GAG−protein complex, few crystal structures of GAG−protein complexes have so far been obtained. Consequently, modeling software for GAG−protein complexes would be a useful tool for analysis of the interaction of GAGs with proteins.

AutoDock 3.0, a docking program provided by Garrett M. Morris, can explore an extensive conformational space.[9] Morris et al. demonstrated the accuracy of AutoDock 3.0 using seven protein−ligand complexes whose tertiary structures and binding constants were known. They classified the protein−ligand complexes into three groups. The first group contained complexes that have small and rigid ligands. This group was utilized as the simplest docking test case. The second group contained moderately flexible ligands, providing a typical test set of intermediate difficulty. The third group contained ligands having many rotatable bonds and diverse chemical characteristics, the most difficult test cases. They compared the performances of the Monte Carlo simulated annealing algorithm (SA) used in earlier versions of AutoDock,[10,11] a genetic algorithm (GA),[12] and a Lamarckian genetic algorithm (LGA) newly employed in AutoDock 3.0.[9] In their study, there was little difference in computational results among the three methods for the first test group. In the test cases with the intermediate and highest levels of difficulty, a structure close to the crystal one was rarely generated by using the SA or GA method. On the other hand, many structures generated by using the LGA method were very close to that of the crystal, even for the third test group.

Goodford and his co-worker reported success in prediction of the binding for two GAGs—monosaccharide and disaccharide—using the docking programs GRID (version 15, Molecular Discovery Ltd.),[13−16] AutoDock 2.4,[17] and DOCK.[2] However, the closeness in structure of the predicted complex to the crystal's was not sufficient when they carried out the docking for hexasaccharide.[18] The results of our preliminary trial employing AutoDock 3.0 for docking of GAGs to proteins were also unsatisfactory. This failure is essentially due to the fact that GAGs have a large electric charge and many rotatable bonds.

In this paper we propose a method for reliable modeling of GAG−protein complexes using a docking simulation and cluster analysis together. We have executed a procedure comprised of the generation of ligand binding conformations by AutoDock 3.0, energy minimization with AMBER8,[19] and cluster analysis for selecting a reasonable ligand structure. We focused on basic fibroblast growth factor (bFGF),[20,21] antithrombin,[22,23] and annexin V,[24,25] whose structures had already been experimentally determined for both the simplex and heparin-bound complex. The effectiveness of our methodology for predicting the GAG−protein structure was evaluated by assessing the similarity between the experimentally determined crystal structures and the computationally derived structures.

Method

Docking Simulation by AutoDock 3.0. The following Brookhaven database entries were used for the docking simulations: (A) bFGF, 1bfc[20] and 1bfg;[21] (B) antithrombin III (ATIII), the L-chain of 1e03[22] and 1e04;[23] (C) and (D) annexin V, 1a8a[24] and 1g5n.[25] All of these test cases ((A)−(D)) satisfy the condition that tertiary structures are available both for the GAG-bound conformation and the unbound one. All of them are heparin-binding proteins. That is, the GAG-bound crystal in each case is a complex of heparin and protein.

Annexin V, a calcium-binding protein, has two heparin-binding sites. These two distinct GAG-binding sites are positioned on the protein surfaces opposite to each other, so annexin V provides two test cases: (C) and (D). One site, (C), holds Ca^{2+} ions that influence the interaction with heparin. This site is formed by two calcium-binding loops, I_{AB} and I_{DE}, termed from the numbering of domains and

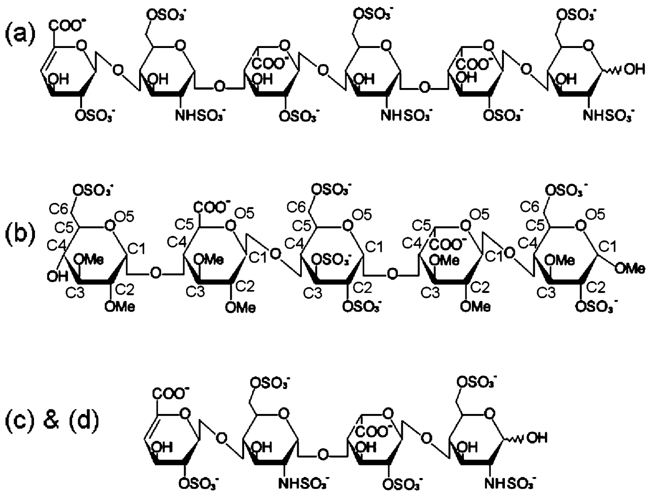


Figure 1. Chemical structures of heparins used in the docking simulation and the subsequent cluster analysis. RMSDs are measured with respect to the atoms composing the sugar ring, which are usually labeled as C1, C2, C3, C4, C5, C6, and O5 as shown in (b).

helices of annexin V. No sulfate groups of heparin directly interact with the Ca^{2+} ions. Instead, the sulfate oxygen atoms make hydrogen bonds with the backbone nitrogen atoms in the I_{AB} calcium-binding loop. Additional hydrogen bonds are formed with water molecules coordinated to the Ca^{2+} ions in both loops and with the side chain of a serine residue in the I_{DE} loop. In contrast, the second heparin-binding site (D) is located on the concave surface of the protein and is not associated with calcium binding.[25] We carried out docking simulations, targeting these two heparin-binding sites separately.

The protocol for the docking simulations is as follows. First, the initial heparin structure was deduced from each crystal structure of the complex. To detach the heparin from the heparin-binding site, only the coordinate of the heparin was translated outward by 8−10 Å. The translated heparin was assumed to be a ligand for the docking simulation (Figure 1). The unliganded protein simplex was regarded as a receptor, which is called a macromolecule in AutoDock 3.0. Docking simulations were carried out using standard AutoDock 3.0 parameters, with 256 runs, the maximum value for AutoDock 3.0, performed for each protein. Grids with a spacing of 0.375 Å were generated around the geometric center of the original ligand position, so that the grid dimensions were (A) 82 × 54 × 54; (B) 100 × 80 × 80, (C) 100 × 80 × 80, and (D) 100 × 80 × 80. The rotatable bonds of the glycosidic link of the heparin were set rigid. If they are flexible or partially rigid, the heparins often have an unrealistic structure. Other rotatable bonds, for example, hydroxyl groups (−OH), sulfate groups ($-OSO_3$), and methyl groups ($-CH_3$), were set flexible. A genetic algorithm and local search procedure were employed. The calculation of internal electrostatic energy in a docking run was activated because heparin has a large negative charge. In cases (C) and (D), the charges of calcium ions were set to −2.0.

Minimization with AMBER8. The output file of AutoDock 3.0 contains the binding ligand structures and their binding energies but not information on the protein structure, because a receptor is regarded as a rigid body in AutoDock 3.0. To obtain the structures of the ligand−protein complex, it is necessary to combine the protein structure with the ligand structures that are extracted from the output file. Hence, in each test case, 256 structures generated by AutoDock 3.0 for the ligand were saved as PDB format files. The protein coordinate was added to each file to prepare 256 protein−heparin complexes. Energy minimization was executed for the complexes using the AMBER8 sander module[19] with the parm99 all-atom force field for proteins[26] and with the glycam04 parameters for heparin.[27,28] Since parameters for heparin with sulfate groups are not provided in glycam04, they were prepared by ourselves. The parameters for the bonded terms were assigned in accordance with the parm99 force field. In order to determine atom charges for sulfate groups, the structures of glucosamine and iduronic acid extracted from the respective crystal structure were optimized at the HF/6-31g(d, p) level using the Gaussian03 program.[29] The charges of the atoms of these glycans were calculated by the two-stage RESP method[30] using the electrostatic potential computed at the rb3lyp/cc-pvtz level and with an ether solvation condition in a manner similar to that used in a previous study.[31,32] This procedure is the same as that used for the development of ff03.[33] To relax the strain in the complexes, the complexes were energetically minimized for 5000 steps by the generalized Born method.[34]

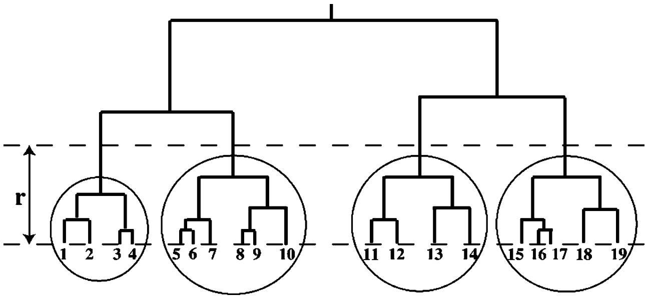


Figure 2. Example of a family tree in the cluster analysis. The RMSD between any two structures in a cluster is less than *r*. The structures, with sequential numbers in a circle, belong to the same cluster.

After energy minimization of the docking complexes, the calculated structures for each complex were superimposed on the crystal structure with respect to the main chain atoms of protein, and the coordinates of the heparins were saved. Simultaneously, the similarities between the docking structures of heparin and the crystal structure were measured by examining the root-mean-square deviations (RMSDs) of atom coordinates for the C1, C2, C3, C4, C5, C6, and O5 atoms. The VMD package[35] was used for superposition, RMSD measurement, and visualization.

Cluster Analysis. We carried out the hierarchical cluster analysis using the RMSDs on the C1, C2, C3, C4, C5, C6, and O5 atoms of the docking structures (Figure 2).

The 256 ligand structures are labeled $s_1, s_2, ..., s_{256}$. Initially, corresponding 256 clusters are designated C_1, C_2, ..., C_{256} with each cluster containing only a single structure. The group containing these 256 clusters is labeled A_{256}. The RMSD between s_i and s_j is represented as $d(s_i,s_j)$, and the distance between two clusters C_m and C_n is defined as

$$D(C_m,C_n) = \max_{si \in Cm, sj \in Cn} d(s_i,s_j) \qquad (1)$$

First, the distances of all pairs of C_m and C_n, $D(C_m, C_n)$, in A_{256} are measured. The pair that has the smallest distance is coupled and registered as a new cluster labeled C_{257}. As a result, the number of clusters becomes 255. The new group containing 255 clusters is labeled A_{255}. Next, the distances among the 255 clusters are measured, and the pair that has the smallest distance in A_{255} is coupled and registered as a new cluster. By iterating this procedure until A_1 is obtained, a family tree for the 256 structures is derived.

The cluster analysis suggests how the structures are distributed and where the generated ligand structures are concentrated. First, we examine all of the ligand structures, setting the distance *r* at 1.5 Å. Then an additional cluster analysis is carried out with the largest cluster in the first step set as a parent group and the distance *r* set at 1.2 Å. The structure closest in the RMSD sense to the average of the atom coordinates of all the ligand structures composing the largest cluster in the second cluster analysis is concluded to be our solution and is called a “representative model”.

In our calculations of the number of hydrogen bonds, a combination of donor, hydrogen, and acceptor atoms is regarded as forming a hydrogen bond when the donor−acceptor distance is within 3.5 Å and the hydrogen-donor−acceptor angle is within 60°. This generous criterion for hydrogen bonding is applied so as not to miss even weak interactions and has been adopted in our previous studies to closely survey intermolecular interactions.[36−38]

Docking Simulation by GOLD 3.1. In order to examine the performance of our approach when other docking software is used, we have executed the same cluster analysis using GOLD 3.1. The protocol for the docking simulation is the same as that for AutoDock 3.0. The PDB files used in AutoDock 3.0 were converted into mol2 files by the BABEL[39] program in all the cases (A), (B), (C), and (D). For each test case, 256 runs were performed using standard GOLD 3.1 parameters. In order to obtain 256 docking ligand structures, the calculation was not terminated even if the top solutions in ranking were close to each other in RMSD. The binding site for the docking search was set to within 20 Å from the position of the grid center in AutoDock 3.0. The rotatable glycosidic bonds are rigid, while the other rotatable bonds are flexible.

Results

AutoDock 3.0 predicts the binding of small ligand molecules to receptors. In this study, 256 docking ligand structures were generated, and their docked energies were computed. In the docking results for bFGF, no significant correlation is observed between the RMSDs of the ligand structures, relative to the crystal structure, and their docked energies (Figure 3). For example, the lowest docked-energy structure shows a quite large RMSD value. The ligand is bound to the heparin-binding site in this structure, but the direction of the glycan chain is opposite to that of the crystal structure (Figure S1 in the Supporting Information).

In spite of poor correlation between the docked energy and the RMSDs from the crystal structure, a certain number of the ligand structures are fairly close to the crystal structure. Accordingly, we speculated that many of the structures generated by AutoDock 3.0 are distributed around the crystal structure and plotted the geometrical centers of the structures generated by AutoDock 3.0 to obtain scatter diagrams. The diagrams suggest that the majority of structures are concentrated in a specific area (Figure 4). It is reasonable to assume that the structure at the center of this area is considerably similar to the crystal. Hence, a representative model can be extracted from the docked ligand structures, and a reasonable structure for the GAG-bound complex is obtained without information on the crystal structure. The two-step cluster analysis is a promising method for extracting a good representative model because carrying out the cluster analysis twice is effective in removing structures that are localized to an area but not the major area.

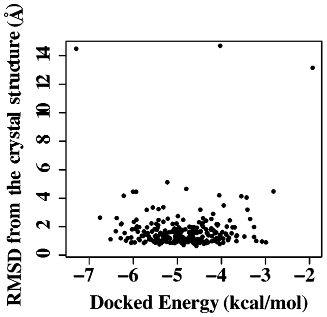


Figure 3. Comparison between the docked energies calculated by AutoDock 3.0 and the RMSDs from the crystal structure for the 256 structures for bFGF generated by AutoDock 3.0. Docked energy represents the stability of FGF−heparin complexes. No significant correlation is observed between RMSD and docked energy.

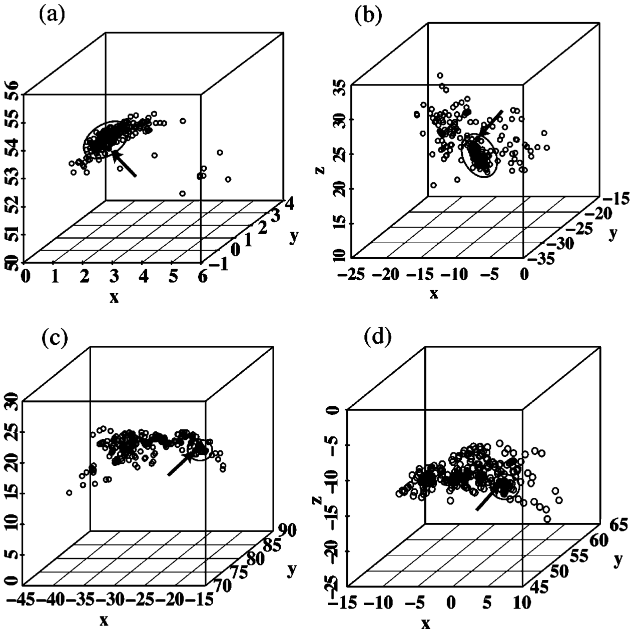


Figure 4. Scatter diagrams of geometrical centers of the structures generated by AutoDock 3.0. The units for the *x*, *y*, and *z* axes are Å. The densest concentration of structures is marked by an oval. The crystal structure is marked by an arrow. (a) bFGF, (b) antithrombin, (c) annexin V-Ca(+), (d) annexin V-Ca(−).

Table 1. Comparison of RMSDs in the Cluster Analysis

protein		*N*[a]	max RMSD[b] (Å)	average RMSD[c] (Å)
bFGF	all[d]	256	14.57	1.53
	first[e]	73	1.34	0.98
	second[f]	30	0.84	0.85
antithrombin	all	256	16.32	7.05
	first	38	1.32	2.63
	second	17	0.72	2.59
annexin V−Ca(+)[g]	all	256	16.54	8.77
	first	13	1.39	1.56
	second	7	1.09	1.28
annexin V−Ca(−)[g]	all	256	14.08	8.52
	first	16	0.68	0.62

[a] Number of structures in each cluster. [b] Largest RMSD measured from the averaged geometry over the structures in the cluster. [c] Average RMSD of all structures relative to the crystal structure.[d] A cluster containing all 256 structures. These structures are generated by AutoDock 3.0 and are minimized by AMBER8. [e] A cluster of the structures categorized into the largest group when performing a cluster analysis using “all” as a parent set. [f] A cluster of the structures categorized into the largest group when performing a cluster analysis using “first” as a parent set. [g] Ca(+) and Ca(−) indicate whether Ca^{2+} ions are present or not to interact with heparin.

Table 2. RMSDs between the Experimental Crystal Structure and the Model Selected by Cluster Analysis or the Model Selected from the Lowest Binding Energy of AutoDock 3.0

	cluster analysis		AutoDock 3.0	
protein	rmsd (Å)	rank[a]	rmsd (Å)	rank[a]
bFGF	0.66	3	14.25	255
antithrombin	2.45	7	6.95	150
annexin V−Ca(+)	0.83	4	10.89	175
annexin V−Ca(-)	0.57	6	8.05	105

[a] The rank of the model is the order among all 256 models, determined by closeness to the crystal structure.

As shown in the average RMSD of Table 1, the number of structures close to the crystal’s contained in the 1st_cluster is larger than that of the whole group in every case. The max RMSD of the 1st_cluster in (A), (B), and (C) is larger than 1.3 Å. Accordingly a more concentrated area of docked structures was extracted with the 1st_cluster set as a parent group. The results of the second cluster analysis show that the number of structures close to the crystal’s contained in the 2nd_cluster is larger than that in the 1st_cluster in every test case (Table 1).

The max RMSD of the 1st_cluster in (D) is very small (0.68 Å). The structures in this cluster are especially close to one another. Hence, further classification was not necessary for (D), and we did not carry out a second cluster analysis. Consequently, the final representative models in (A), (B), and (C) are the closest to the averaged geometry of all structures in the 2nd_cluster, while that of (D) is closest to the averaged geometry of all structures in the 1st_cluster. The RMSDs between the representative model and the crystal structure are shown in Table 2. The accuracy is moderate in (B) and good in (A), (C), and (D). The reason for this difference is that the structures generated by AutoDock 3.0 in (B) contain very few structures that are close to the crystal’s. A comparison of the representative model and the crystal structure shows that our approach has a high level of accuracy, especially for (A), (C), and (D) (Figure 5).

The rank of the representative model among all 256 structures with respect to closeness to the crystal structure is shown in Table 2. The representative model selected from the cluster analysis has a rank within the top ten in every case. On the other hand, the model selected from the binding energy of AutoDock 3.0 is not good, with all of their ranks over 100. Hence, the energy analysis using AutoDock 3.0 does not discriminate an adequate docking structure for the binding of heparin from the generated structures. Although even the cluster analysis could not extract the closest model, i.e., rank 1, in this study, the ranks of the representative models demonstrate that this method is a useful approach for predicting an adequate heparin-binding structure.

Furthermore, when the first cluster analysis was carried out, there were some structures, “singular models” not comprising a cluster with any others, that provided useful information on GAG binding in docking simulations. The average RMSDs of the singular models in Table 3 are larger than those of the whole group in Table 1, implying they are very different from the crystal structure in cases (A)−(D). By inspecting what residues interact with the singular models, we found several residues positioned far from the heparin-binding site of the protein. Details of these residues are shown in Figures S2 and S3 in the Supporting Information.

Additionally, we examined the applicapability of our method to GOLD 3.1, which is one of the most widely used docking simulation tools. The entire procedure of generation of 256 structures with GOLD 3.1, energy minimization with AMBER8, and cluster analysis were performed in a manner similar to that for AutoDock 3.0. The energy minimization could not be completed for antithrombin because GOLD 3.1 generates many inadequate structures in which heparin is bound to the inside of antithrombin, and the forces on atoms are too large to execute molecular mechanical calculation in AMBER8. The clustering results are shown in Table S1 in the Supporting Information. The average RMSDs from the crystal structure of all 256 structures generated by GOLD 3.1 are equal or larger than those of the structures generated by AutoDock 3.0. Table S1 clearly indicates that the level of accuracy for predicting the heparin-binding structure was low compared with the results from AutoDock 3.0. A comparison of the models selected by the cluster analysis and those selected from the lowest binding energy from GOLD 3.1 shows that the selected models differ considerably from the crystal; i.e., their rankings among the 256 structures are not good with respect to closeness to the crystal structure. No notable improvement is observed except for the case of annexin V−Ca(+) (Table S2 in the Supporting Information). Furthermore, even the structure closest to the crystal’s is inferior to the representative model selected by the cluster analysis combined with AutoDock 3.0 in every test case (Table S3 in the Supporting Information). Accordingly, GOLD 3.1 is deemed inappropriate for GAG−protein docking.

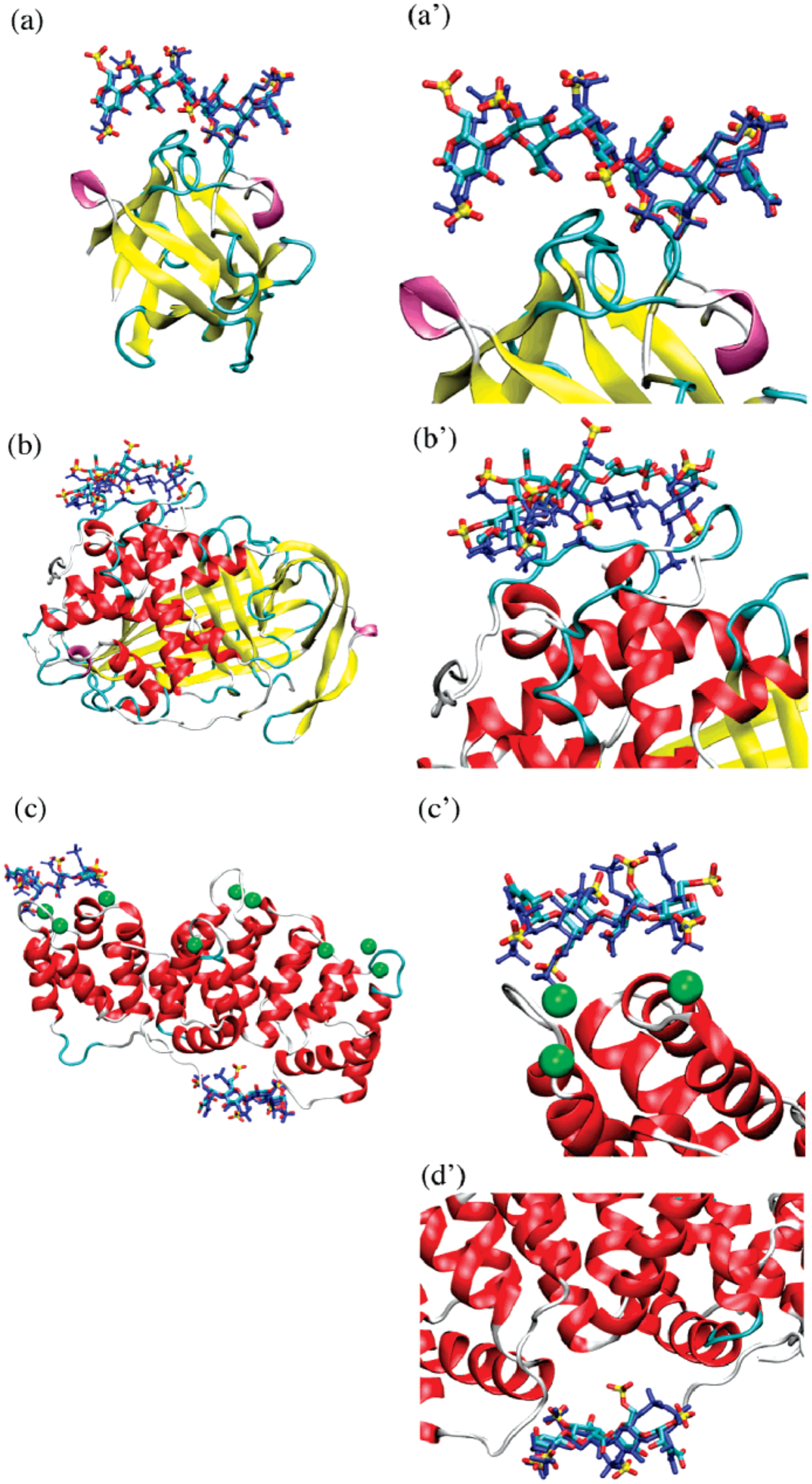


Figure 5. Comparison of the representative model to the crystal structure. Proteins and heparins are shown as cartoon and ball-and-stick representations. Green spheres represent Ca^{2+} ions. Heparins colored blue are crystal structures, and those colored cyan, yellow, and red are representative models. (a) bFGF, (a′) magnification of (a), (b) antithrombin, (b′) magnification of (b), (c) annexin V, (c′) magnification of Ca(+) area of (c), (d′) magnification of Ca(−) area of (c).

Discussion

Cluster analysis seems to be effective in predicting the structure of GAG−protein complexes. AutoDock 3.0 is able to generate many structures close to the crystal structure, and, as shown in the average RMSDs in Table 1, it is plausible that the largest cluster contains the structure most

Table 3. Average RMSD of All Singular Models Measured from the Crystal Structure

protein	average RMSD (Å)	number of singular models
bFGF	4.21	18
antithrombin	10.03	88
annexin V−Ca(+)	9.50	81
annexin V−Ca(-)	9.72	136

adequately reproducing the crystal structure. In this study, the representative model is determined by selecting the structure closest to the averaged atom geometry of the structures in the 2nd_cluster (exceptionally in the 1st_cluster for (D)). Although the top-ranked structure could not be extracted in each case, the average RMSD of the 2nd_cluster in Table 1 and the RMSD in Table 2 demonstrate that the representative model is acceptable and that its rank is satisfactory.

Cluster analysis has already shown a substantial degree of success in predicting protein folding structures. Based on the supposition that there are a greater number of conformations surrounding the correct folding structure than the incorrect folding one, Shortle et al. performed cluster analysis for small proteins with the 1000 lowest-energy conformations produced by random structure generation and subsequent energy minimizaion.[40] They clearly suggested that the analysis can identify conformations considerably closer to the native structure than the conformation with the lowest energy. This finding is the basis for our trial prediction of heparin-binding structure by cluster analysis. Zagrovic and co-workers closely examined the average structure of the ensemble generated by molecular dynamics simulations for small polypeptides both in folded and unfolded states.[41] They found that none of the conformations of the unfolded state exhibited a nativelike structure but that the mean structure obtained by averaging over the entire set of unfolded conformations showed a nativelike geometry. This approach is quite useful because information on the native heparin structure is usually not available when performing binding predictions. Zagrovic et al. further suggested the advantage of evaluation with the distance-based RMSD and the preference of the average structure over the unfolded ensemble of small protein structures for predicting the native geometry.[42] Their reports provided good justification for our present trial.

The criteria for selecting the representative model from 256 structures should be determined carefully because the averaged atom geometry is greatly influenced by the selection of clusters. The structures in (C) and (D) are dispersed compared with those in (A) and (B) (Figure 4). As a result of the first cluster analysis in (C) and (D), the number of structures is less than 10 for all the clusters except the 1st_cluster, which contains structures fairly close to the crystal's (Table 1). This result suggests that proper selection of the 1st_cluster is very important for prediction of the structure of the GAG−protein complex. In the first clustering shown in Table 1, the distance *r* was set to 1.5 Å. To examine the dependency of the clustering results on the distance criteria, the first clustering was executed with *r* set at 2.0, 1.0, and then 0.5 Å (Table 4). For 0.5 Å, clusters became too small, and a cluster far from the crystal structure occasionally became the largest cluster. This caused a decrease in the level of accuracy of prediction, as seen for annexin V−Ca(+) in Table 4. For *r* equal to 2.0 Å, the largest cluster was likely to contain structures too different from the crystal structure, thus lowering the prediction accuracy. Judging from these findings, an *r* from 1.0 Å to 1.5 Å is most suitable for the first clustering.

In the geometry search step, AutoDock 3.0 computes the interaction energy between a receptor and a ligand by intermolecular van der Waals, hydrogen bond, and Coulomb potentials and evaluates the stability of the ligand after generating a large number of ligand binding conformations by the Lamarckian genetic algorithm. The scoring function for the geometry search is[43]

$$\Delta E = \sum\left(\frac{A_{ij}}{{r_{ij}}^{12}} - \frac{B_{ij}}{{r_{ij}}^{6}}\right) + \sum\left(\frac{C_{ij}}{{r_{ij}}^{12}} - \frac{D_{ij}}{{r_{ij}}^{6}}\right) + \sum\left(\frac{q_i \times q_j}{\epsilon(r_{ij}) \times r_{ij}}\right) + \sum K_\phi(1 + \cos(n\phi) - \delta) + \Delta H_{\mathrm{vdW}}^{\mathrm{ligand}} + \Delta H_{\mathrm{eleq}}^{\mathrm{ligand}} + \Delta H_{\mathrm{hbond}}^{\mathrm{ligand}} \tag{2}$$

where *A* and *B* are the van der Waals parameters for atoms *i* and *j*, *C* and *D* are hydrogen bond parameters, r_{ij} is the interatomic distance between atom *i* and atom *j*, *q* is the Coulomb charge of each atom, and $\epsilon(r)$ is the distance-dependent dielectric function. The first three terms are the receptor−ligand interaction energy terms. The next four are the internal energies of the ligand—the torsion potential, van der Waals force, electrostatic force, and intramolecular hydrogen bonding, respectively. In the LGA, the structure with the highest ΔE is deleted, the structure with the lowest ΔE is always retained, and the other structures are merged using a crossover technique or random mutation technique.[9] The estimation of ΔE will highly influence the accuracy of our approach using cluster analysis because this energy dominates the structures generated in the docking simulation.

The most prominent difference between (B) and the other cases is in the functional groups of the different species of heparin. Uronic acids or glucosamine of natural heparin has a hydroxyl group at C3. Tetrasaccharide and hexasaccharide fragments of porcine mucosal heparin, the heparin of the crystal structure in (A), (C), and (D), were experimentally prepared by partial digestion with heparin-lyase I, followed by collection with strong anion exchange liquid chromatography.[21,25,44] On the other hand, the heparin of the crystal structure in (B) is a synthetic heparin analog.[45,46] This pentasaccharide contains O-alkyl ethers in place of hydroxyls that are usually sulfated at the early stage of the synthesis. As in the representative model (B) of Figure 5, carbon (red) and oxygen (cyan) atoms are seen at the C3 site of uronic acids or glucosamine. Consequently, the ligand in the docking simulation of (B) contains O-alkyl ethers instead of hydroxyls. AutoDock 3.0 does not consider any nonpolar atoms of a ligand. Accordingly, protein atoms directly interact with the carbon of heparin. Those atoms would interact with hydroxyls if the ligand were a natural heparin. Hence, the energy evaluation of van der Waals force,

Table 4. Dependency of Cluster Analysis Accuracy on Distance Criterion *r*

	number of structures[a]			average RMSD[b] (Å)			best RMSD[c] (Å)			RMSD of selected model[d] (Å)		
protein	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å
bFGF	73	29	13	0.98	0.85	1.14	0.66	0.66	0.96	0.94	0.66	1.12
antithrombin	37	24	9	2.60	2.62	2.59	2.35	2.37	2.45	2.60	2.61	2.59
annexin V−Ca(+)	14	10	5	1.65	1.33	15.23	0.72	0.73	15.03	1.76	1.51	15.23
annexin V−Ca(−)	19	16	9	0.77	0.62	0.61	0.4	0.4	0.4	0.56	0.57	0.57

[a] Number of structures categorized into the largest cluster by the first clustering. [b] Average RMSD of all structures in the cluster, relative to the crystal structure. [c] RMSD between the crystal structure and the closest one in the cluster. [d] RMSD between the crystal structure and the model selected by the cluster analysis.

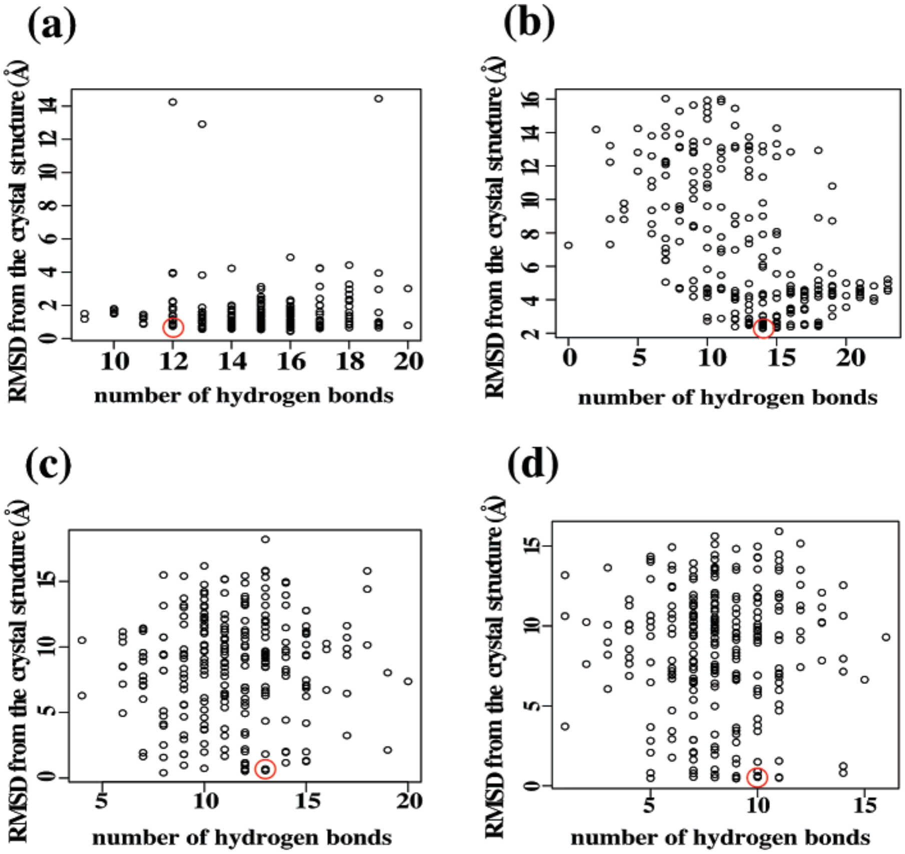


Figure 6. Comparison between the number of hydrogen bonds in protein−heparin complexes and the RMSD from the crystal structure. Each circle corresponds to one docking structure, and 256 circles appear in the respective graphs of (a)−(d). Data on the representative models are indicated by red circles. No significant correlation is observed between RMSD and number of hydrogen bonds. (a) bFGF, (b) antithrombin, (c) annexin V−Ca(+), (d) annexin V−Ca(−).

hydrogen bonding, and Coulomb force in the geometry search step is considerably different from that in other cases.

A study of the energetics of the interaction of bFGF with GAG by Thompson et al. showed that the electrostatic contribution of positively charged residues to the binding energy was only 30%.[47] They suggested that not only electrostatic interaction but also nonionic interaction, such as hydrogen bonding and van der Waals force, mainly contributed to the free energy for GAG−protein binding. We evaluated the contribution of hydrogen bonds in GAG−protein complexes (Figure 6). No significant correlation was found between RMSDs of docking structures, relative to the crystal, and the number of hydrogen bonds. This suggests that hydrogen bonding is not the only factor in GAG−protein binding and that van der Waals force interaction is also important. Since there are various factors to be considered for the binding free energy, it seems difficult to evaluate the affinity of GAGs for target proteins only from the energy in docking simulations. Consequently, a method for predicting the GAG−protein complex based on structures, namely the present approach, is needed.

In order to examine the causes for generation of structures not close to the crystal's, we focused on residues that are frequently located near the singular models but rarely interact with heparins in the crystal structure. The residues of the protein within 4.0 Å from the singular models were counted in each test case (Figure S2 in the Supporting Information). Acidic or hydrophobic residues were closely examined because their interaction with heparins cannot be straightforwardly explained. In the case of (A) bFGF, many singular models have interaction with K129 and G133. K129 is located in the binding site but rarely interacts with heparin in the crystal structure. Because a side chain of K129 extends outside, a heparin may be attracted to the residue. All 15 structures interacting with G133 also interact with Q134 and K135. These residues provide both a positive charge and hydrogen bond acceptors and donors; therefore, this area is a likely binding site for heparin (Figure S3(a) in the Supporting Information). In the case of (B) antithrombin, 14 residues before E42 are missing in the crystal structure for the simplex, 1E04. The N-terminus is therefore open and likely to interact with a sulfuric acid group of heparin. Attention should be given to the missing atoms when an intact crystal structure is employed for a receptor in docking simulations (Figure S3(b) in the Supporting Information). For case (C) for the Ca^{2+}-binding domain of annexin V, D66 is positioned near R61, and the side chains of D66 and R61 interact with each other. When a heparin is attracted to R61, the heparin will also be trapped near D66 (Figure S3(c) in the Supporting Information). Since Ca^{2+} ions counteract the negative charge of OD1 and OD2 of E70, the heparin interacting with S69 can be positioned near E70 (Figure S3(d) in the Supporting Information). For case (D) with no Ca^{2+}-associating domain of annexin V, OD1 and OD2 of D162 interact with the main chains of V201 and S202. That is, the positively charged side chain extends outside the protein. Therefore, a heparin is likely to approach D162 (Figure S3(e) in the Supporting Information). I245 and P246 are located at a loop near the target domain on the outside of the protein. Either pair of R205 and R206 or K284 and K288 interacts with heparin, and these residues are in the proximity of I245 and P246 (Figure S3(f) in the Supporting Information). Since those residues interact with singular models, we conclude that heparin is likely to be attracted to basic or nonpolar hydrophilic amino residues. In particular, a heparin has a very high binding affinity for Lys and Arg.

In the present study, GOLD 3.1 was not as accurate as AutoDock 3.0 in predicting the structure of GAG−protein complexes (Table 1 and Table S1 in the Supporting Information). Many heparin structures generated by GOLD 3.1 are apt to be bound to the inside of the proteins, despite the fact that heparins are incapable of being compactly packed inside the protein because of their strong negative charge. This can be explained from the GOLD scoring function for a geometry search.[48−50] In docking simulations by GOLD 3.1, the van der Waals term seems to have a particularly strong influence on the docking ligand structures. With an increase in the contact area, the van der Waals contribution becomes large. Hence, the stabilization energy for ligand binding is estimated to be smaller when a ligand adheres to the surface of a protein than when a ligand is inside a protein. In addition, it might be disadvantageous for GOLD to estimate the large electrostatic interaction between a ligand and positively charged residues. Therefore, GOLD 3.1 might be inadequate for docking simulations of GAGs such as heparins because the binding site is on the surface of a protein and the binding is highly influenced by charges.

Coulomb force is an explicit factor of the scoring function in a geometry search of AutoDock 3.0. Therefore, even negatively charged GAGs are evaluable as the ligand in docking simulations. In the present study, the binding sites of docking simulations were determined on the basis of crystal structures. If the binding site is unidentified, calculation of the surface electrostatic potential of a protein will be helpful in searching for probable binding sites for GAGs.

Conclusion

By performing (1) a docking simulation with AutoDock 3.0, (2) energy minimization with AMBER8, and (3) cluster analysis, it is possible to model the complex of a heparin and a GAG-binding protein. An adequate structure for the complex is predictable by this approach if the unliganded protein structure is available. The van der Waals force, hydrogen bonding, and Coulomb force are of considerable importance in the GAG−protein binding; therefore, incorporation of all these terms in docking simulations is highly desirable. The scoring function in the geometry search of AutoDock 3.0 contains all three of these terms; hence, AutoDock 3.0 is appropriate for GAG−protein docking simulations, although careful consideration should be given to the point that the strong influence of Lys and Arg of a target protein sometimes leads to generation of inadequate binding structures.

Acknowledgment. This work was supported by the Japan Science and Technology Agency.

Supporting Information Available: Tables S1−S3 and Figures S1−S3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Paul, N.; Rognan, D. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 521.

(2) Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem.* **1997**, *18*, 1175.

(3) Hoffmann, D.; Kramer, B.; Washio, T.; Steinmetzer, T.; Rarey, M.; Lengauer, T. *J. Med. Chem.* **1999**, *42*, 4422.

(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727.

(5) Lindahl, U.; Lidholt, K.; Spillman, D.; Kjellén, L. *Thromb. Res.* **1994**, *75*, 1.

(6) Yayon, A.; Klagsbrun, M.; Esko, J. D.; Leder, P.; Ornitz, D. M. *Cell* **1991**, *64*, 841.

(7) Prestrelski, S.; Fox, G. M.; Arakawa, T. *Arch. Biochem. Biophys.* **1992**, *293*, 314.

(8) Bjork, I.; Lindahl, V. *Mol. Cell. Biochem.* **1982**, *48*, 161.

(9) Goodsell, D. S.; Morris, G. M.; Halliday, R. S.; Huey, R. *J. Comput. Chem.* **1998**, *19*, 1639.

(10) Goodsell, D. S.; Olson, A. J. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 95.

(11) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293.

(12) Holland J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(13) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849.

(14) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. *J. Med. Chem.* **1989**, *32*, 1083.

(15) Wade, R. C.; Clark, K. J.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 140.

(16) Wade, R. C.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 148.

(17) Goodsell, D. S.; Morris, G. M.; Olson, A. J. *J. Mol. Recognit.* **1996**, *9*, 1.

(18) Bitomsky, W.; Wade, R. *J. Am. Chem. Soc.* **1999**, *121*, 3004.

(19) Case, D. A.; Darden, T. A.; et al. . *AMBER*, *version 8*; Department of Pharmaceutical Chemistry, University of California: San Francisco, CA, 2004.

(20) Faham, S.; Hileman, R. E.; Fromm, J. R.; Linhardt, R. J.; Rees, D. C. *Science* **1996**, *271*, 1116.

(21) Ago, H.; Kitagawa, Y.; Fujishima, A.; Matsuura, Y.; Katsube, Y. *J. Biochem.* **1991**, *110*, 360.

(22) Jin, L.; Abrahams, J. P.; Skinner, R.; Petitou, M.; Pike, R. N.; Carrell, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14683.

(23) Skinner, R.; Abrahams, J. P.; Whisstock, J. C.; Lesk, A. M.; Carrell, R. W.; Wardell, M. R. *J. Mol. Biol.* **1997**, *266*, 601.

(24) Capila, I.; Hernaiz, M. J.; Mo, Y. D.; Mealy, T. R.; Campos, B.; Dedman, J. R.; Linhardt, R. J.; Seaton, B. A. *Structure* **2001**, *9*, 57.

(25) Swairjo, M. A.; Concha, N. O.; Kaetzel, M. A.; Dedman, J. R.; Seaton, B. A. *Nat. Struct. Biol.* **1995**, *2*, 968.

(26) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

(27) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, B. *J. Phys. Chem.* **1995**, *99*, 3832.

(28) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541.

(29) Frisch, M. J.; Trucks, G. W.; et al. *Gaussian 03*, *Revision C.02*; Gaussian Inc.: Wallingford, CT, 2004.

(30) Cornell, W. D.; Cieplak, P.; Bayly, C.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620.

(31) Ode, H.; Neya, S.; Hata, M.; Sugiura, W.; Hoshino, T. *J. Am. Chem. Soc.* **2006**, *128*, 7887.

(32) Sato, Y.; Hata, M.; Neya, S.; Hoshino, T. *J. Phys. Chem.* **2006**, *110*, 22804.

(33) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(34) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824.

(35) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics Modell.* **1996**, *14*, 33.

(36) Sato, Y.; Hata, M.; Neya, S.; Hoshino, T. *J. Phys. Chem.* **2006**, *110*, 22804.

(37) Ode, H.; Matsuyama, S.; Hata, M.; Hoshino, T.; Kakizawa, J.; Sugiura, W. *J. Med. Chem.* **2007**, *50*, 1768.

(38) Ode, H.; Matsuyama, S.; Hata, M.; Neya, S.; Kakizawa, J.; Sugiura, W.; Hoshino, T. *J. Mol. Biol.* **2007**, *370*, 598.

(39) Walters, P.; Dolata, M.; Babel, S. *A Molecular Structure Information Interchange Hub*; Department of Chemistry, University of Arizona: Tucson, AZ (accessed Aug 22, 2005).

(40) Shortle, D.; Simons, K. T.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11158.

(41) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153.

(42) Zagrovic, B.; Pande, V. S. *Biophys. J.* **2004**, *87*, 2240.

(43) AutoDock3.0.5_USGuide.pdf. Molecular Graphics Lab. http://www.scripps.edu/mb/olson/doc/autodock (accessed Nov 23, 2005).

(44) Azra, P.; Cindy, G.; Kenneth, A. J.; Xue-Jun, H.; Robert, J. L. *Glycobiology* **1995**, *5*, 83.

(45) Jin, L.; Abrahams, J. P.; Skinner, R.; Petitou, M.; Pike, R. N.; Carrell, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14683.

(46) Basten, J.; Jaurand, G.; Olde-Hanter, B.; Duchaussoy, P.; Petitoub, M.; van Boeckel, C. A. A. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 905.

(47) Thompson, L. D.; Pantoliano, M. W.; Springer, B. A. *Biochemistry* **1994**, *33*, 3831.

(48) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43.

(49) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457.

(50) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609.

CT700029Q