JCTC Journal of Chemical Theory and Computation

Tracing the Entropy along a Reactive Pathway: The Energy As a Generalized Reaction Coordinate

Carine Michel,[†] Alessandro Laio,*[,‡] and Anne Milet*[,†,#]

Département de Chimie Moléculaire, Chimie Théorique, UMR-5250, ICMG FR-2607, CNRS, Université Joseph Fourier Grenoble I, DU BP 53 38041 Grenoble Cedex 09 France, SISSA, ISAS, Via Beirut 2-4, I-34014 Trieste, Italy, and Institut Universitaire de France, France

Received April 11, 2009

Abstract: By using metadynamics at a temperature T_0 we reconstruct the free energy $F_{T_0}(E,s)$ as a function of the potential energy E and of a geometrical variable s. We show here that from $F_{T_0}(E,s)$ one can estimate the free energy also *at a different temperature*. This allows tracing the entropy and characterizing the properties of molecular systems at all temperatures by a single simulation. We validate this approach on the water dimer dissociation.

Metadynamics[1] is a novel technique that can be used for computing the free energy barriers and exploring new reactions pathways in a wide range of contexts, from protein folding[2] to mineral phase transitions[3] and organometallic reactivity.[4] It is based on biasing the dynamics with a history-dependent potential $V_G(s,t)$ defined in the space of a set of collective variables (CV).[5] In the limit of a long metadynamics run, the free energy surface can be reconstructed as a function of the CV s: $F(s) \sim -V_G(s,t)$. We show here that the potential energy E can also be used as a generalized coordinate for studying chemical reactions. E has already been used as a CV in metadynamics for reconstructing the density of states of an Ising model.[6] It has also been used as a helpful auxiliary collective variable to better explore the configuration space in some nucleation studies.[7] During a chemical process, the potential energy E varies as the system explores new intermediate/transition states. Thus, it is a relevant collective variable. We propose to use E in combination with ordinary geometrical CVs to conduct a chemical study. The advantages of such an extended set of CVs will be shown: tracing the entropy along a reactive pathway and extrapolating the thermodynamic quantities at different temperatures.

Using the potential energy E as a CV together with ordinary geometrical variables s in a metadynamics scheme allows reconstructing, at a temperature T_0, the free energy F_{T_0} as a function of E and s simultaneously. Remarkably $F_{T_0}(E,s)$ contains the relevant information for characterizing the thermodynamic properties of the system at *all the temperatures* T, including the probability to observe the reactants/products and the activation entropy. In fact, for a system of potential energy $E(r)$ we have, within the canonical ensemble

$$\begin{aligned} F_{T_0}(s,E) &= -k_B T_0 \ln \int dr \delta(E - E(r))\delta(s - s(r)) \exp\left(-\frac{1}{T_0}E(r)\right) \\ &= E - k_B T_0 \ln \int dr \delta(E - E(r))\delta(s - s(r)) \\ &= E - k_B T_0 \ln (\Omega(E,s)) \end{aligned}$$

where, in the second passage, we used the properties of the Dirac delta. The density of states

$$\Omega(E,s) = \int dr \delta(E - E(r))\delta(s - s(r))$$

appearing in the last passage is a measure of the number of configurations r that exist at an energy $E(r)$ and for which $s(r) = s$. By definition this quantity does not depend on T_0. Thus, the free energy at a different temperature T can be directly computed from $F_{T_0}(E,s)$

$$F_T(E,s) = E - Tk_B \ln \Omega(E,s) = E + \frac{T}{T_0}(F_{T_0}(E,s) - E) \quad (1)$$

The internal energy profile $U_{T_0}(s)$ and the free energy profile $F_{T_0}(s)$ are also directly derived from $F_{T_0}(E,s)$

$$U_{T_0}(s) = \frac{\int dE E \exp(-F_{T_0}(E,s)/k_B T_0)}{\int dE \exp(-F_{T_0}(E,s)/k_B T_0)} \quad (2)$$

$$F_{T_0}(s) = -k_B T_0 \log\left(\int dE \exp(-F_{T_0}(E,s)/k_B T_0)\right) \quad (3)$$

Finally, the entropy $S_{T_0}(s)$ as a function of the reaction coordinate s is given by $S_{T_0}(s) = (U_{T_0}(s) - F_{T_0}(s))/T_0$. Thus, using eqs 1−3 allows computing the U, F, and S profiles at any temperature T by a run performed at a single temperature T_0.

The entropy obtained in this manner could alternatively be derived computing U as the average of E for fixed s and the free energy F as a function of s alone by thermodynamic integration or one-dimensional metadynamics. However, as it is clear from eqs 2 and 3, internal energy, free energy, and

* Corresponding author e-mail: Anne.milet@ujf-grenoble.fr (A.M.), laio@sissa.it (A.L.).

[†] Université Joseph Fourier Grenoble-I.

[‡] Statistical and Biological Physics, SISSA.

[#] Institut Universitaire de France.

10.1021/ct900177h CCC: $40.75
Published on Web 07/31/2009

entropy depend implicitly on the temperature T_0. Thus, if one wants to characterize the behavior of the system at several temperatures T, one would be forced to repeat the calculation at all T. One can also obtain the entropy as a function of a geometrical parameter within the harmonic and the rigid rotator approximations,[8] but those approximations break down in solvated systems and at high temperature where anharmonicity effects become important. Instead, our approach allows computing the entropy profile at a minimal computational cost without those approximations. Furthermore, it provides F, U, and S profiles in a wide range of temperatures by a single calculation.

In practical application, $F_{T_0}(E,s)$ is estimated by metadynamics that, by construction, explores only a finite region of the CV space. Thus, even if eq 3 is in principle exact, the larger the difference between T and T_0 is, the less the extrapolation will be meaningful. Indeed, the CV regions that are relevant at T and T_0 might be different. We here perform the extrapolation if two criteria are satisfied:

1. The free energy surface $F_{T_0}(E,s)$ has to be sufficiently explored. For a given (E,s), the accumulation of Gaussians must be larger than $2k_BT_0$ to extrapolate $F_{T_0}(E,s)$ into $F_T(E,s)$.

2. In order to compute the profiles $F_T(s)$ and $U_T(s)$, one has to integrate $F_T(E,s)$ along E according to eq 2 and eq 3. Thus, for a given value of s, the extrapolated free energy surface $F_T(E,s)$ has to be sufficiently explored as a function of E in the region around the minimum of F, as this region dominates the integrals. Denoting by $E_0(s)$ the value of E minimizing $F_T(E,s)$ at fixed s, we require that all the values of E for which $F_T(E,s) - F_T(E_0(s),s) < 2k_BT$ are explored. This is imposed requiring the history-dependent potential to be larger than $2k_BT_0$ in this region.

This procedure has been validated on a simple system, the water dimer treated at the DFT level. At $T = 0$, the only stable state is the bound dimer, while at high T, a dissociated state becomes gradually stabilized by configuration entropy. In order to retain the physical meaning of the total energy of the system, we perform our calculations using the Born−Oppenheimer *ab initio* molecular dynamics approach, using the CP2K-QuickStep program.[11] QuickStep is an implementation of the Gaussian Plane Waves (GPW) method based on the Kohn−Sham formulation of density functional theory (DFT). It is a hybrid method using a linear combination of Gaussian-type orbitals to describe the Kohn−Sham orbitals, whereas an auxiliary plane waves basis set is employed to expand the electronic charge density. The basis set used is a quadruple-valence set of Gaussian orbitals with a set of three polarization functions added for all atoms in conjunction with the Goedecker-Teter-Hutter pseudopotentials. The auxiliary PW basis set was defined by a cubic box of $12A^3$ and by a density cutoff of 500 Ry for the larger grid. The use of the BLYP functional for this system has been validated through comparison with experimental results[9] and high-level calculations[10] The four hydrogen atoms have been replaced by deuterium in order to increase the dynamics step to 0.5 fs. Velocity rescaling has been used to enforce a constant temperature T_0.

The free energy has been reconstructed at $T_0 = 100$ K by metadynamics[11] as a function of the potential energy E and a geometrical collective variable, namely the coordination number between the two oxygen atoms, $s = \mathrm{cn(O,O)} = [1-(r/r_0)^3]/[1-(r/r_0)^6]$ where r is the O−O distance and $r_0 = 3.5$ Å. The shape of the added Gaussians is defined by the height (0.063 kcal/mol ≈ 0.3 k_BT_0), the width along s (0.03), and the width along E (0.188 kcal/mol). The Gaussians are added every 20 fs.

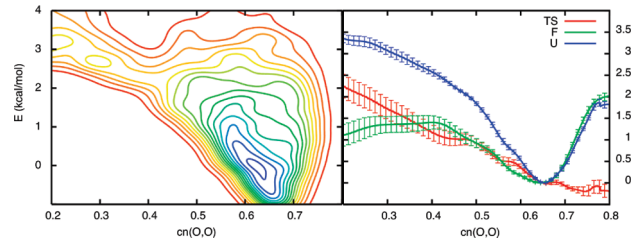


Figure 1. Left panel: Free energy in function of the potential energy E and $s = \mathrm{cn(O,O)}$ at $T_0 = 100$ K reconstructed from a metadynamics run. Right panel: Free energy F, internal energy U, and temperature × entropy T_0S profiles along the reaction coordinate $s = \mathrm{cn(O,O)}$, resulting from the mean value of 5 independent metadynamics runs at $T_0 = 100$ K. Error bars correspond to the standard deviation. High values of cn(O,O) correspond to the bound water dimer; low values of cn(O,O) correspond to the dissociated dimer.

The metadynamics runs have been stopped after 1500 Gaussians have been added: after this time, a diffusive behavior in CV space and several recrossings of the transition state region are observed. The free energy surface $F_{T_0}(E,s)$ is then reconstructed on a 500 × 500 grid in the $(0.2 < \mathrm{cn(O,O)} < 0.8) \times (0\ \mathrm{kcal/mol} < E < 10\ \mathrm{kcal/mol})$ region. The cn(O, O) limits derive from the size of the box, which restrains the accessible range for the lower region, and on repulsion between nuclei, which forbids the system visiting the upper region. All the energies in the following are expressed as a difference with respect to the global total energy minimum. The two-dimensional free energy $F_{T_0 = 100K}(E,s)$ is shown in Figure 1. Then, following the general procedure detailed above, we extracted from $F_{T_0 = 100K}(E,s)$ the internal energy $U_{T_0 = 100K}(s)$, the free energy $F_{T_0 = 100K}(s)$, and the entropy $S_{T_0 = 100K}(s)$ as a function of the geometrical variable s. The result is shown in Figure 1, right panel. The error bars are estimated repeating the procedure in 5 independent but equivalent simulations.[12] All the free energy profiles fall in a range of 0.5 kcal/mol around the mean. Similarly, all the entropy profiles fall in a range of 5 cal/mol/K × 100 K = 0.5 kcal/mol around the mean. From now, the discussion will be based on the more accurate mean profiles.

The free energy profile $F_{T_0 = 100K}(s)$ and the entropy profile $S_{T_0 = 100K}(s)$ obtained at $T_0 = 100$ K are plotted together with the internal energy profile $U_{T_0 = 100K}(s)$ in Figure 1. Both F and U present a minimum at $s = 0.64$ corresponding to the bound water dimer, with an O−O distance $r = 2.89$ Å. For smaller s, along the dissociation path, those two profiles start becoming different: the free energy profile $F(s)$ presents a maximum around $s = 0.40$, whereas the mean energy profile $U(s)$ increases monotonically. It is well-known that the origin of those different behaviors lies in the entropic effect: the higher the temperature is, the more the dissociated dimer is stabilized by configurational entropy. Indeed, the entropy monotonically increases during the dissociation (see Figure 1): from $s = 0.64$ to $s = 0.2$, the dissociation of the water dimer leads to a net entropy increase of $T_0 \times S = 2.2$ kcal/mol at $T_0 = 100$ K.[13] Thus, the entropic contribution (2.2 kcal/mol)

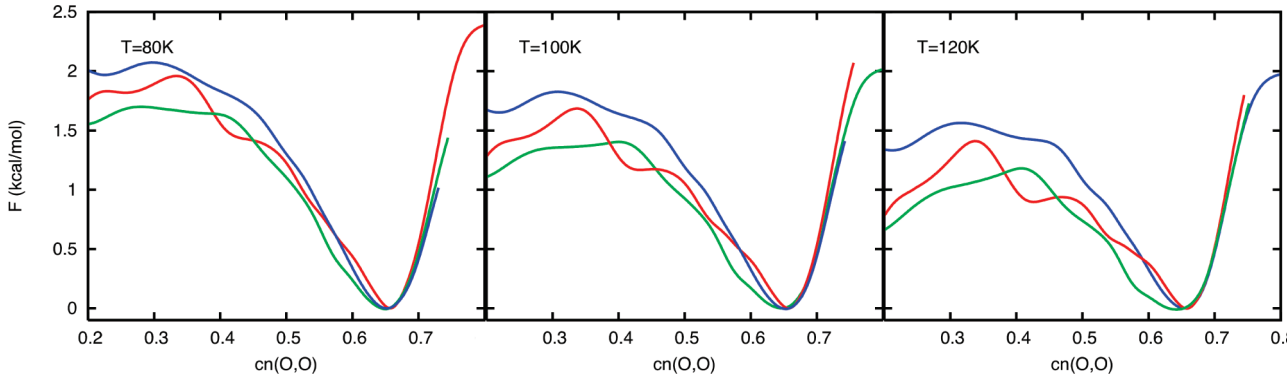


Figure 2. Free energy profiles extrapolated at various temperatures T: 80 K (left panel), 100 K (middle panel), 120 K (right panel). In red, the free energy profile extrapolated from metadynamics runs performed at $T_0 = 80$ K, in green, the free energy profile extrapolated from metadynamics runs performed at $T_0 = 100$ K, in blue, the free energy profile extrapolated from metadynamics runs performed at $T_0 = 120$ K.

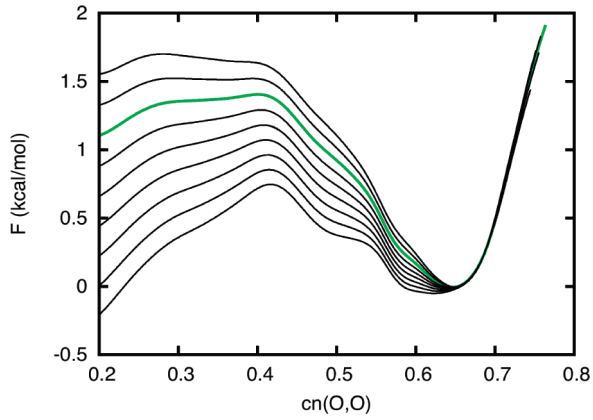


Figure 3. Free energy profiles extrapolated at various temperatures T (from 80 K to 160 K, every 10 K) from the metadynamics runs at $T_0 = 100$ K. In green, the free energy profile at 100 K. Notice the lower profile ($T = 160$ K): the dissociated dimer is more stable than the associated one.

compared with the net energy increase (3.4 kcal/mol) is far from being negligible.

Any extrapolation of a given property at another temperature T based on the entropy profile $S_{T_0}(s)$ has to be taken with caution: this procedure cannot be valid for large changes in temperature, as $S_{T_0}(s)$ depends on T_0. On the other hand, computing the free energy F as a simultaneous function of a geometrical parameter s and the potential energy E allows extrapolating the results over a much broader temperature range. We have performed the same calculations at lower ($T_0 = 80$ K) and higher ($T_0 = 120$ K) temperature. In Figure 2, the free energy profile F_{T_0} at temperature T_0 obtained from a metadynamics at this same temperature T_0 is plotted together with the free energy profiles F_T extrapolated from $F'_{T_0}(E,s)$ obtained at another temperature $T'_0 = T_0 \pm \Delta T$ ($\Delta T = 20$ K and/or 40 K). Those free energy profiles perfectly illustrate the stabilization the dissociated dimer gains at higher temperature. Indeed, this behavior is observed for both the original and the extrapolated profiles. At the beginning of the dissociation process (cn(O,O) > 0.40), the agreement between simulated and extrapolated profiles is particularly good. Once the dimer is dissociated (cn(O,O) < 0.40), the difference between simulated and extrapolated profile remains lower than 0.7 kcal/mol. Following this idea, we have extrapolated the free energy at much higher temperature from the runs at $T_0 = 100$ K: the dissociated dimer is found to be more stable for $T > 150$ K (see Figure 3).

In conclusion, we have here shown that by using the potential energy E as a collective variable in a metadynamics framework, it is possible to address two issues that are considered very challenging in computational chemistry: tracing the entropy along a reactive pathway and extrapolating thermodynamic quantities at different temperatures. If used in combination with other geometrical collective variables s, the potential energy E allows reconstructing the free energy surface $F_{T_0}(E,s)$. From this quantity one can directly obtain the entropy along the reactive pathway. Moreover, using simple thermodynamics identities, from $F_{T_0}(E,s)$ one can estimate the free energy at a different temperature T. This allows extracting from a single simulation the relevant thermodynamic quantities in a wide range of temperatures.

Acknowledgment. We thank the CECIC for providing computer facilities. C.M. thanks P. Fleurat-Lessard for fruitful discussions.

References

(1) Laio, A.; Parrinello, P. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562.

(2) (b) Bussi, G.; Gervasio, F.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435. (c) Bonomi, M.; Gervasio, F.; Tiana, G.; Provasi, D.; Broglia, R. A.; Parrinello, M. *Biophys. J.* **2007**, *93*, 2813.

(3) (a) Martonak, R.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 075503. (b) Organov, A. R.; Martonak, R.; Laio, A.; Raiteri, P.; Parrinello, M. *Nature* **2005**, *438*, 1142.

(4) (a) Stirling, A.; Iannuzzi, M.; Parrinello, M.; Molnar, F.; Bernhart, V.; Luinstra, G. A. *Organometallics* **2005**, *24*, 2533. (b) Michel, C.; Laio, A.; Mohamed, F.; Krack, M.; Parrinello, M.; Milet, A. *Organometallics* **2007**, *26*, 1241.

(5) Any differentiable function of the ionic coordinates can be used as a collective variable, such as distances, angles, etc.

(6) Micheletti, C.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2004**, *92*, 170601.

(7) (a) Donadio, D.; Raiteri, P.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 5421. (b) Trudu, F.; Donadio, D.; Parrinello, M. *Phys. Rev. Lett.* **2006**, *97*, 105701. (c) Quigley, D.; Rodger, P. M. *J. Chem. Phys.* **2008**, *128*, 221101.

(8) McQuarrie, D. A. *Statistical Thermodynamics*; Harper and Row: New York, 1976.

(9) Odutola, J. A.; Dyke, T. R. *J. Chem. Phys.* **1980**, *72*, 5062.

(10) Klopper, W.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227. Xu, X.; Goddard, W. A., III *J. Phys. Chem. A* **2004**, *108*, 2305.

(11) In order to keep the physical meaning of the total energy of the system, we perform metadynamics based on a Born-Oppenheimer ab initio dynamics using the CP2K program. (VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, *J. Comput. Phys. Commun.* **2005**, *167*, 103. The time step is 0.5 fs. The energy has been computed at the DFT level using the BLYP functional. The basis set used was a quadrupole-valence set of Gaussian orbitals with a set of three polarization functions added for all atoms in conjunction with the Goedecker-Teter-Hutter pseudopotentials. The auxiliary PW basis set was defined by a cubic box of 12 $Å^3$ and by a density cutoff of 500 Ry for the larger grid.

(12) For each simulation, the initial velocities have been randomly chosen according to the Boltzmann distribution at T_0.

(13) Within the usual approximations framework, a calculation at the BLYP/6-311+g(3df,3pd) level gives an entropy increase of 21 cal/mol/K at 100 K.

CT900177H

JCTC Journal of Chemical Theory and Computation

A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1

Fabio Pietrucci* and Alessandro Laio

International School for Advanced Studies (SISSA-ISAS), via Beirut 2-4, I-34014 Trieste, Italy

Received April 24, 2009

Abstract: We introduce a new class of collective variables which allow forming efficiently beta-sheet structures in all-atom explicit-solvent simulations of proteins. By this approach we are able to systematically fold a 16-residue beta hairpin using metadynamics on a single replica. Application to the 56-residue SH3 and GB1 proteins show that, starting from extended states, in ∼100 ns tens of structures containing more than 30% beta-sheet are obtained, including parts of the native fold. Using these variables may allow folding moderate size proteins with an accurate explicit solvent description. Moreover, it may allow investigating the presence of misfolded states that are relevant for diseases (e.g., prion and Alzheimer) and studying beta-aggregation (amyloid diseases).

1. Introduction

All-atom explicit solvent simulations are still not competitive with bioinformatics or knowledge-based approaches for protein structure predictions, at least in terms of cost-effectiveness. In order to improve the predictivity of accurate simulations, it will be necessary to address two issues, both of the utmost importance: First, the accuracy of the force fields that, as is well-known, is still far from optimal. Second, the sampling efficiency: even if the "perfect" force field was available, in order to predict the native fold of a protein one should explore several structures and compare their free energies accurately. In this work we address the problem of performing an efficient sampling of protein conformations.

The structure of proteins typically contains a large amount of secondary structure, in the form of alpha helices and beta sheets. The formation of an α helix is a rather simple process, which mainly requires the local alignment of backbone dihedrals in a segment of the protein chain and typically happens on a time scale of about 100 ns.[1] Instead, beta-structures are more complicated, as shown by characteristic formation times at least 1 order of magnitude longer.[1−3] This difference is mainly related to the fact that building of beta-structures requires the proper dihedral arrangement in two distant segments of the protein chain and the simultaneous formation of specific interstrand H-bonds. As a result, simulating by accurate molecular dynamics (MD) with explicit solvent the folding process of proteins containing beta-structure is challenging: already studying a 16-residues beta-hairpin requires significant computational resources.[4−6] This is an important limitation, as beta-structures are present in many proteins and, moreover, are the key structural element in fibrils.[7]

In order to enhance the probability of observing beta-structures one can use an enhanced sampling scheme such as umbrella sampling, thermodynamic integration, or metadynamics.[8] These approaches require choosing an appropriate collective variable (CV) which describes the progress of the conformational transition. For instance, such a variable could be defined using the beta secondary structure definition of DSSP[9] or STRIDE,[10] which is primarily based on the H-bonds pattern. Unfortunately, it is well-known that a precise indicator of a structural property is not necessarily a good *reaction coordinate* for simulating a transition process.[11] Indeed, we tested CVs based on the peculiar H-bond arrangement of beta-structures, that, for example, in the antiparallel β sheet are formed between pairs of residues $(i,i+h)$, $(i+2,i+h-2)$, etc. We observed that using this class of CVs with metadynamics[12] allows the formation of beta-structure but is affected by technical problems when implemented in MD, since it drives the system not only toward well-formed beta strands but also toward unphysical structures with unlikely conformations. This can be understood considering that the formation of even a single beta bridge is a rather complex process that requires establishing selected hydrogen bonds after specific dihedral transitions and after the correct alignment of the two chain segments forming the bridge. A CV taking into account only one of these three aspects may not be effective to bias the formation of beta-sheet structures.

To overcome these limitations, we here introduce a CV for beta-structure which is defined in a different manner, by counting how many pairs of 3-residue segments adopt the correct beta conformation in a given protein structure. The correct conformation is taken simply as the average beta conformation of experimental protein structures. The CV is not tailored on the specific fold of a single protein, but it is meant as a general-purpose CV which can describe beta-structure in all proteins.

* Corresponding author e-mail: fabio.pietrucci@gmail.com.

10.1021/ct900202f CCC: $40.75
Published on Web 08/04/2009

First, we benchmarked the new CV by simulating the C-terminal beta hairpin of protein GB1 in explicit solvent, using a single-replica metadynamics[12,13] simulation. The correct folded state is systematically obtained, together with several misfolded states, at a cheap computational cost. Next, we investigated the conformational space of two larger (56 amino acids) proteins, SH3 and GB1, whose native folds include a large amount of beta secondary structure. These two proteins have been investigated in several experimental and theoretical studies of protein folding.[14−18] However, a compelling simulation of their folding mechanism with accurate explicit-solvent force fields is still lacking. We attempted to fold both SH3 and GB1 by bias-exchange metadynamics[19,20] simulations in explicit solvent, employing the new CVs. Starting from extended states, within ∼100 ns of total simulated time tens of different conformations with a large content of extended beta structure are obtained. Among these, several configurations are obtained which contain the main structural elements of the native state. Bias-exchange metadynamics simulations performed on the same systems but using other variables explore the conformational space ∼10 times less efficiently.

2. The Beta Collective Variable

In order to define a CV for exploring protein beta-sheet structures by metadynamics or other enhanced sampling techniques, we first defined from the PDB database the shape of an ideal building block for beta sheets, i.e. a small beta subunit composed of a few amino acids, which by replication gives rise to the extended beta-strucures. To this aim, we considered the representative proteins of the 20 architecture entries in the "mainly beta" class of the CATH database[21] (PDB codes 1bds, 1gvk, 1h8p, 1i5p, 1itv, 1k7i, 1m3y, 1n7v, 1nh2, 1qre, 1rg8, 1tl2, 1w6s, 1ylh, 2bbk, 2dpf, 2hnu, 2nwf, 3sil, 4bcl). For each protein, we extracted the residues belonging to beta secondary structure (according to the STRIDE[10] definition), and, among them, we extracted all pairs of segments of 3 residues connected by hydrogen bonds. We computed the RMSD of the positions of backbone N, C_α, C, O, and C_β atoms in the 3 + 3 blocks. Antiparallel beta blocks are similar within a RMSD of only 0.048 ± 0.017 nm, parallel beta blocks within 0.066 ± 0.028 nm. Therefore the beta-structures observed in proteins, despite their impressive variety, are composed of 3 + 3 blocks which are remarkably similar. This allows the definition of the ideal (i.e., average) "beta block". We defined the ideal antiparallel and parallel beta blocks by taking the central structure of each pool. In the case of the parallel beta-structure, two equivalent blocks exist, corresponding to a symmetry operation obtained by rotating of 180° both the 3-residue segments around their backbone axis.

Using this definition of beta blocks, we implemented a CV which counts how many 3 + 3 residues units are similar to the "beta block". This CV is defined as a differentiable function of the atomic coordinates in the following manner

$$S = \sum_{\alpha} n[\mathrm{RMSD}(\{\mathbf{R}_i\}_{i\in\Omega_\alpha}, \{\mathbf{R}^0\})] \qquad (1)$$

$$n(\mathrm{RMSD}) = \frac{1 - (\mathrm{RMSD}/0.1)^8}{1 - (\mathrm{RMSD}/0.1)^{12}} \qquad (2)$$

where n is a function switching smoothly between 0 and 1, the RMSD is measured in nm, and $\{\mathbf{R}_i\}_{i\in\Omega_\alpha}$ are the atomic coordinates of a set Ω_α of six residues of the protein, while $\{\mathbf{R}^0\}$ are the corresponding atomic positions of the ideal beta block. In the case of antiparallel beta, all sets Ω_α of residues of the form $(i,i+1,i+2; i+h+2,i+h+1,i+h)$ are summed over in eq 1. For parallel beta, sets $(i,i+1,i+2;i+h,i+h+1,i+h+2)$ are instead considered. For each residue, only backbone N, C_α, C, O, and C_β atoms are included in the RMSD calculation (in Gly residues the C_β is missing and the corresponding hydrogen is used instead).

The same procedure has been applied to define the ideal α helix block formed by six consecutive residues, in order to define a CV measuring the amount of alpha secondary structure. In this case the sum in eq 1 runs over all possible sets Ω_α of six consecutive protein residues $(i, i+1,i+2,i+3,i+4,i+5)$, and $\{\mathbf{R}^0\}$ are the atomic positions of the ideal alpha block. In summary, the CVs $S_{\mathrm{anti}\beta}$, $S_{\mathrm{para}\beta}$, and S_α are approximately proportional to the number of beta/alpha blocks of six residues which are present in a protein 3D structure (Figure 1).

3. Metadynamics Simulation of Beta-Hairpin Folding

We tested on several proteins the ability of the new CVs to generate secondary structure, starting from unfolded conformations and performing metadynamics simulations[12] in explicit solvent. First, we benchmarked our approach on the 16-residues C-terminal beta hairpin of protein GB1 (PDB code 1pgb, Figure 2-A). We used a version of the GROMACS 3.3.1 package[22] modified by us, employing the AMBER03[23] and TIP3P[24] force fields for protein and water, respectively. The protein was solvated by 3373 water molecules in a orthorhombic box of 105.8 nm^3, neutralized by three Na^+ ions. The particle-mesh Ewald method[25,26] was used for long-range electrostatics with a short-range cutoff of 0.8 nm. A cutoff of 0.8 nm was used for the Lennard-Jones interactions. All bond lengths were constrained to their equilibrium length with the LINCS[27] algorithm. The time step for the MD simulation was 2.0 fs. NPT simulations at 340 K and 1 atm were performed by coupling the system to a Nose-Hoover thermostat[28,29] and a Berendsen barostat,[30] both with relaxation time of 1 ps. After 1 ns of

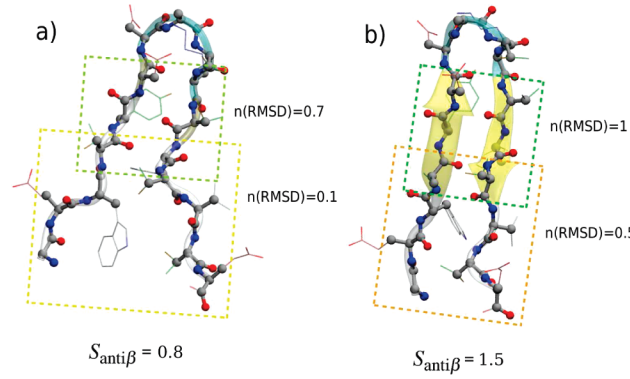


Figure 1. Values assumed by the CV $S_{\mathrm{anti}\beta}$ in the (a) partially formed and (b) almost-completely folded beta hairpin. The dashed rectangles outline 3 + 3 residue beta blocks.

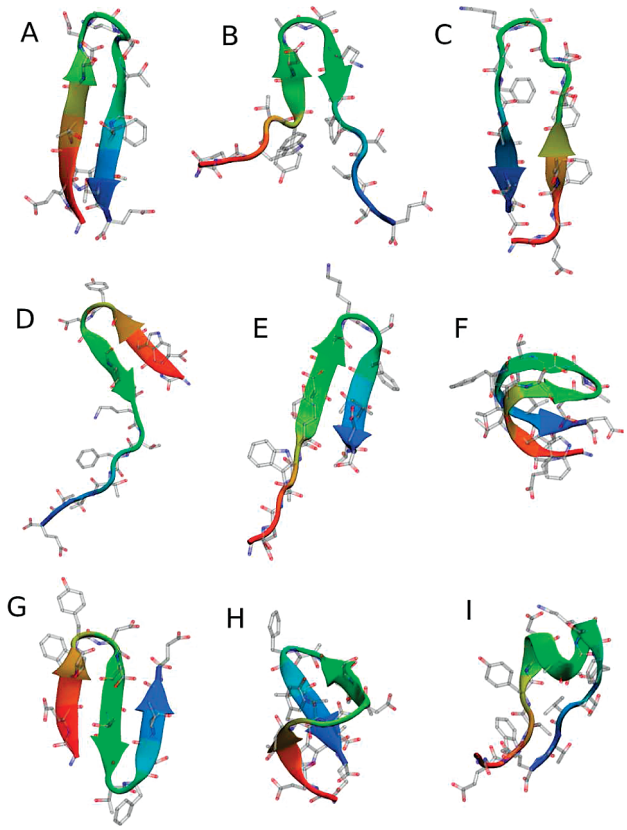


Figure 2. Folded state (A) and several misfolded conformations (B−I) of the 16-residues C-terminal beta-hairpin of GB1, obtained by a single metadynamics simulation of 100 ns. The collective variables $S_{anti\beta}$ and C_α radius of gyration have been biased.

equilibration, the barostat was removed, and the simulations were continued in the NVT ensemble.

Four independent metadynamics simulations of length 100 ns were performed, starting from extended states without secondary content. The CVs antiparallel beta ($S_{anti\beta}$) and radius of gyration of C_α (R_{gyr}) were biased by 2-dimensional Gaussians of height 2 kJ/mol, width 0.1 and 0.05 nm, respectively, adding a Gaussian every 5 ps.

In each simulation, the experimental folded state is observed within 50 ns (backbone RMSD < 0.2 nm, Figure 2-A). Furthermore, several different misfolded states with alpha or beta content are also found, some of which are reported in Figure 2 (panels B−I). This calculation shows that a suitable choice of the reaction coordinate allows folding the beta hairpin and finding misfolded states even using a single replica.

4. Bias-Exchange Metadynamics Simulation of SH3 Folding

Using the CV introduced in this work, we also investigated the conformational space of the larger (56 amino acids) protein src-SH3. The native fold consists of a terminal beta-hairpin packed orthogonally on top of a three-stranded antiparallel beta-sheet, plus a small 3_{10} helix (PDB codes 1srl, Figure 3-A). The protein was solvated by 3641 water molecules in a cubic box of 127.2 nm^3, neutralized by three Na^+ ions. The MD parameters are the same as for the beta hairpin (see above).

We performed a bias-exchange metadynamics[19] simulation at 340 K, employing four replicas and starting from an extended

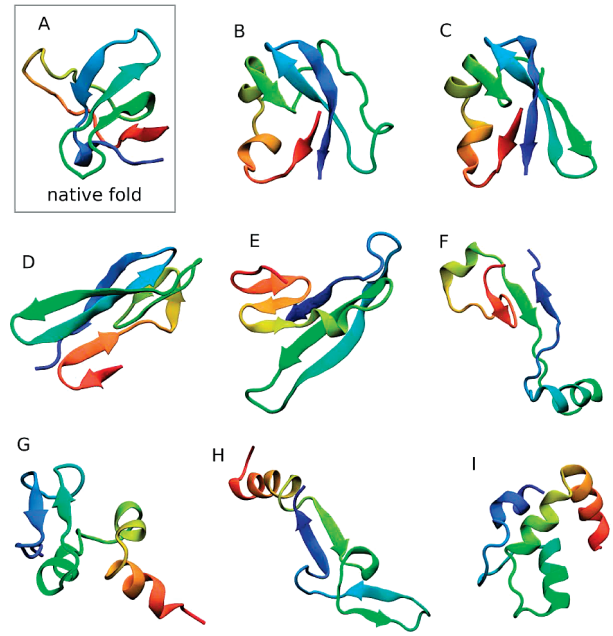


Figure 3. Experimental folded state of SH3 (A) and selected conformations obtained using the CVs introduced in this work in a bias-exchange metadynamics simulation (B−I).

Table 1. Average Number of SH3 Clusters Explored per 100 ns of Total Simulation Time by Bias-Exchange Metadynamics, Using the New CVs or the Old CVs[a]

	old CVs	new CVs
>20% sec. str.	97	203
>30% sec. str.	42	104
>10% beta	4	40
>20% beta	1.7	15
>30% beta	0.1	4

[a] See the text for a definition. The total simulation time is 940 ns for the old CVs and 240 ns for the new CVs. The structures have been clustered using the TM-align algorithm[31] with a threshold of 0.5.

state with no secondary structure. The following CVs have been biased, each one on a different replica: S_α, $S_{anti\beta}$, $S_{para\beta}$, and the radius of gyration of hydrophobic side chain carbons (R_{gyr}). These CVs are not tailored on the specific native state of SH3, but they represent general-purpose reaction coordinates for protein folding. One-dimensional hills of height 2 kJ/mol were added every 5 ps, and exchanges of the bias potentials were attempted every 50 ps. The trajectories were clustered with the TM-align algorithm[31] (threshold 0.5) in order to extract the significantly different structures.

Within 60 ns per replica (total simulation time 240 ns), ~200 different structures with more than 30% content of secondary structure are obtained (see Table 1), a selection of which is reported in Figure 3. In particular, 9 different structures are obtained which include more than 30% extended beta. Among these, structures B and C in Figure 3 clearly contain a substantial part of the native multiple-β sheet of SH3 (Figure 3-A), although with some topological differences.

As a comparison, another bias-exchange metadynamics simulation was performed on the same system, this time without employing the new CVs introduced in this work but biasing the CVs introduced in ref 19: number of backbone H-bonds in the first half of the protein, in the second half, and between the

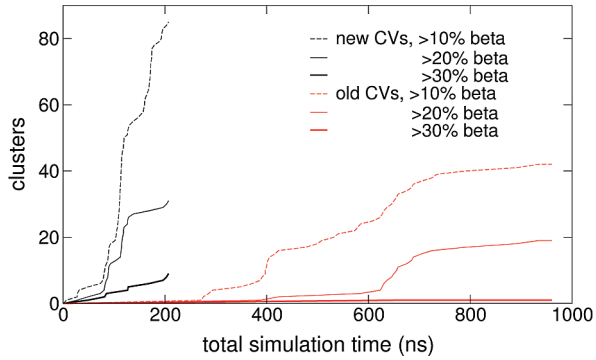


Figure 4. Number of SH3 clusters explored by enhanced sampling molecular dynamics as a function of total simulation time, using the new CVs or the old CVs (see text for definition). Clusterization has been performed using the TM-align algorithm[31] with a threshold of 0.5. The clusters contain an amount of beta-structure between 10% and 30%, following the DSSP definition.[9]

two halves; helicity of the backbone Φ_α (as defined in ref 19) in each fourth of the protein chain; number of contacts among aromatic side chain carbons, or among hydrophobic side chain carbons, defined as $N = \Sigma_{ij}[1 - (R_{ij}/R_0)^8]/[1 - (R_{ij}/R_0)^{14}]$ ($R_0 =$ 0.3 nm, i,j run over the appropriate carbon atoms). Each replica was biased by two-dimensional hills of height 0.5 kJ/mol added every 2 ps; exchanges of the bias potentials were attempted every 2 ps. These variables are referred to as "old CVs" in Table 1. Sixteen replicas have been used, and a simulation of 60 ns per replica was performed starting from extended states. By using these CVs a much smaller number of structures containing beta-sheets is explored per unit simulation-time, compared to the new CVs (Table 1 and Figure 4).

5. Bias-Exchange Metadynamics Simulation of GB1 Folding

As a second application, we studied the 56-residues protein GB1. The native fold consists of a four-strands β sheet formed by two terminal beta-hairpins connected in parallel and packed on top of an α helix (PDB code 1pgb, Figure 5-A). The protein was solvated by 3780 water molecules in a cubic box of 125.0 nm^3, neutralized by four Na^+ ions. The MD parameters are the same as above.

A bias-exchange metadynamics[19] simulation was performed on GB1 at 300 K, using 8 replicas and biasing each of the following CVs on a different replica: $S_{anti\beta}$, $S_{para\beta}$, the helicity of the backbone Φ_α (as defined in ref 19) in each third of the protein chain, and the number of contacts $N = \Sigma_{ij}[1 - (R_{ij}/R_0)^8]/[1 - (R_{ij}/R_0)^{14}]$ ($R_0 = 0.7$ nm) with the summation extended to C_α pairs belonging to first and second, second and third, or first and third segments of the protein chain. One-dimensional hills of height 2.5 kJ/mol were added every 5 ps, and exchanges of the bias potentials were attempted every 50 ps. The trajectories were clustered with the TM-align algorithm[31] as described above.

Starting from extended states, in 80 ns per replica 53 different structures with more than 30% beta content are explored. A selection of the structures is reported in Figure 5, panels B—Q. Remarkably, structures B, D, E, and P contain one or both of the native terminal beta hairpins, whereas structures B, C, D, E, and L contain the native central helix (compare Figure 5-A). Thus, also in this case, using the newly introduced variable allows exploring a very large number of structures with significant secondary content.

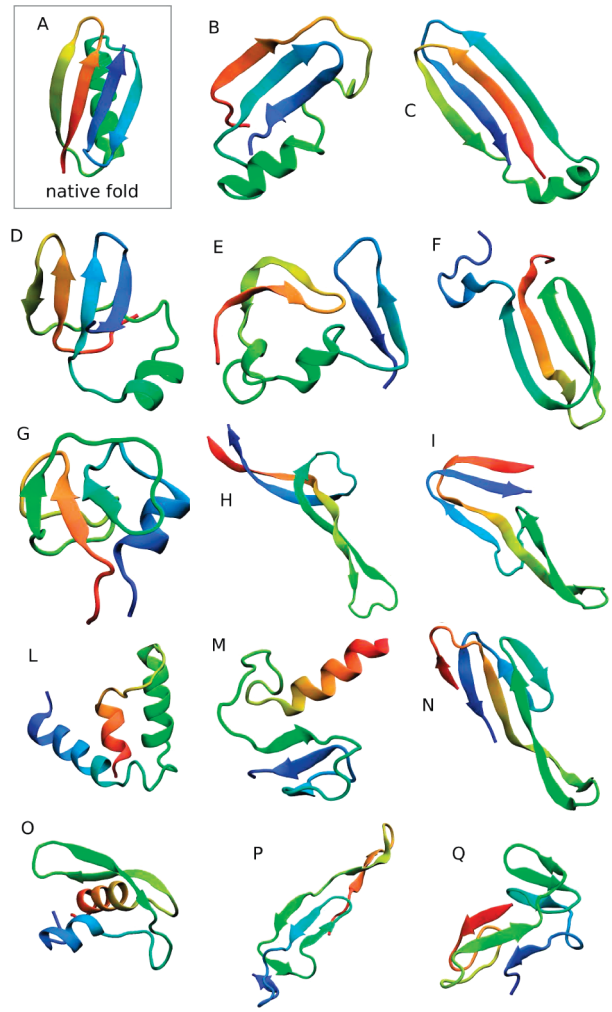


Figure 5. Experimental folded state of GB1 (A) and selected conformations obtained using the CVs introduced in this work in a bias-exchange metadynamics simulation (B—Q).

6. Conclusions

We introduced a new class of collective variables (CVs) specifically designed for observing the formation of beta sheets and alpha helices. The CVs are not tailored specifically on a given protein fold, but they are aimed at describing the content of secondary structure in all proteins. Using these CVs together with an enhanced sampling technique such as umbrella sampling, thermodynamic integration, or (bias-exchange) metadynamics allows exploring quickly a large number of complex alpha- and beta-structures starting from unfolded states and employing accurate explicit-solvent force fields. In particular, it is possible to systematically fold the C-terminal beta-hairpin of protein GB1 employing a single-replica simulation. Application of the new CVs to the 56-residue proteins SH3 and GB1 in explicit solvent shows that bias-exchange metadynamics simulations allow to observe in ~100 ns the formation of tens of different structures with large alpha- and beta-content. Some of these structures contain parts of the elements of the native

fold. For both proteins, however, the exact native fold is not reached within the relatively short span of the simulations.

One should remark that the experimental folding times of GB1 and SH3 are of the order of milliseconds and seconds, respectively, indicating that the atomic rearrangements that have to take place in order to explore the folded state are rather complex. For a comparison, the folding times of GB1 and SH3 are 3 to 6 orders of magnitude larger than the one of advillin, the largest system that has been so far reversibly and reproducibly folded with an all atom force field describing the solvent explicitly. The quality of the new variables introduced here is demonstrated only comparing the number of nontrivial structures that are found in a given simulation time. More extended simulations should be employed to find the experimental folded state and also to estimate the relative free energy of different conformations, e.g. by means of a weighted hystogram analysis of the bias-exchange metadynamics trajectories, as detailed in ref 32.

Still, our results allow for being optimistic about the possibility to completely fold proteins of less than 100 amino acids using simulation times of the order of microseconds. Furthermore, the new variables may help investigating the presence of protein misfolded states and the phenomenon of beta-aggregation, which are relevant to understand the mechanism of several diseases.

Acknowledgment. We acknowledge Pilar Cossio, Fabrizio Marinelli, and Xevi Biarnés for useful discussions. We also acknowledge the grant MIUR PRIN-2006025255 and CINECA for providing computational resources.

References

(1) Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327–359.

(2) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.

(3) Jager, M.; Deechongkit, S.; Koepf, E. K.; Nguyen, H.; Gao, J.; Powers, E. T.; Gruebele, M.; Kelly, J. W. *Biopolymers* **2008**, *90*, 751–758.

(4) Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50–61.

(5) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

(6) Yoda, T.; Sugita, Y.; Okamoto, Y. *Proteins* **2007**, *66*, 846–859.

(7) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.

(8) Dellago, C.; Bolhuis, P. G. *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Springer: Berlin/Heidelberg, 2008; Vol. 221, pp 1−67.

(9) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(10) Frishman, D.; Argos, P. *Proteins* **1995**, *23*, 566–579.

(11) Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.

(12) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(13) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(14) Shea, J. E.; Brooks, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499–535.

(15) Hubner, I. A.; Edmonds, K. A.; Shakhnovich, E. I. *J. Mol. Biol.* **2005**, *349*, 424–434.

(16) He, Y.; Chen, Y.; Alexander, P.; Bryan, P. N.; Orban, J. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14412–14417.

(17) Narzi, D.; Daidone, I.; Amadei, A.; Di Nola, A. *J. Chem. Theory Comput.* **2008**, *4*, 1940–1948.

(18) Hori, N.; Chikenji, G.; Berry, R. S.; Takada, S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 73–78.

(19) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.

(20) Piana, S.; Laio, A.; Marinelli, F.; Troys, M. V.; Bourry, D.; Ampe, C.; Martins, J. C. *J. Mol. Biol.* **2008**, *375*, 460–470.

(21) Cuff, A. L.; Sillitoe, I.; Lewis, T.; Redfern, O. C.; Garratt, R.; Thornton, J.; Orengo, C. A. *Nucleic Acids Res.* **2009**, *37*, D310–D314.

(22) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(23) 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*, 1999–2012.

(24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(25) Darden, T. A.; York, D. *J. Chem. Phys.* **1993**, *98*, 10089.

(26) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(27) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, G. E. M. J. *J. Comput. Chem.* **1997**, *18*, 1463.

(28) Nose, S. *Mol. Phys.* **1984**, *52*, 255.

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

(30) Berendsen, H. J. C.; Postma, J. P. M.; Gusteren, W. F. V.; Nola, A. D.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(31) Zhang, Y.; Skolnick, J. *Nucleic Acids Res.* **2005**, *33*, 2302–2309.

(32) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. *PLoS Comput. Biol.* 2009. in press.

CT900202F

JCTC Journal of Chemical Theory and Computation

Peptide Partitioning and Folding into Lipid Bilayers

Jakob P. Ulmschneider,*,† Jacques P. F. Doux,‡ J. Antoinette Killian,‡ Jeremy C. Smith,§ and Martin B. Ulmschneider*,‡

IWR, University of Heidelberg, Germany, Department of Chemistry, University of Utrecht, Utrecht, The Netherlands, and Oak Ridge National Laboratory, Oak Ridge, Tennessee

Received May 20, 2009

Abstract: The folding and partitioning of WALP peptides into lipid bilayers is characterized using atomic detail molecular dynamics simulations on microsecond time scales. Elevated temperatures are used to increase sampling, and their suitability is validated via circular dichroism experiments. A new united atom parametrization of lipids is employed, adjusted for consistency with the OPLS all-atom force field. In all simulations secondary structure forms rapidly, culminating in the formation of the native trans-membrane helix, which is demonstrated to have the lowest free energy. Partitioning simulations show that peptide insertion into the bilayer is preceded by interfacial folding. These results are in excellent agreement with partitioning theory. In contrast, previous simulations observed unfolded insertion pathways and incorrectly report stable extended configurations inside the membrane. This highlights the importance of accurately tuning and experimentally verifying force field parameters against microsecond time scale phenomena.

How helical peptides fold and integrate into lipid bilayer membranes remains one of the most intriguing processes in biophysics. Unfortunately, these folding pathways cannot currently be directly spatially and temporally resolved in experiments. In principle, computer simulations can provide the required information, as they can now reach >μs time scales. However, simulation accuracy requires careful calibration and verification of force field parameters. We have recently reported a new set of OPLS-UA lipid parameters tuned against a large set of experimental data.[1] Here we present the application of these parameters in ab initio simulations of peptide bilayer folding and insertion simulations. The results demonstrate the importance of accurate parameters in obtaining the correct partitioning pathway.

Partitioning theory,[2] and previous simulations using implicit membrane models,[3,4] strongly suggest a folded insertion pathway for hydrophobic peptides and a stable trans-membrane helix as the native state: the high cost of desolvating exposed peptide bonds (estimated at ~4 kcal/mol/bond)[5] dictates that the transfer of solvated peptides into a hydrocarbon phase should follow a two-stage pathway, where helical segments fold at the phase boundary prior to insertion (see Figure 1A).[6,7] A recent microsecond molecular dynamics simulation[8] directly examined the folding and partitioning of synthetic WALP peptides[9] into explicit lipid bilayer membranes. However, this study reported both an unfolded insertion pathway as well as a nonhelical native state in the bilayer. The disagreement of this result with theory may arise from either errors in the GROMOS96(43a2)/Berger (G96/Berger) parameters, or it might be a consequence of the elevated temperature of 80 °C used in the simulations to enhance sampling.

Since no experimental data are available for WALP at elevated temperatures we have measured the peptide secondary structure as a function of temperature using circular dichroism spectroscopy. These experiments show that the helicity of WALP in DMPC lipid vesicles decreases by less than 4% when the temperature is increased from 25° to 90 °C (see Figure 2, for experimental details see the Supporting Information). Equivalent folding simulations can therefore also be performed at elevated temperatures, as the native state is highly thermostable. This has the advantage of greatly speeding up conformational sampling. The destabilization of the native state helix in ref 8 therefore strongly points toward a problem with either the protein force field or the 'Berger' lipid parameters.[10] To address this problem, we have recently developed a new set of united-atom lipid parameters for use in combination with the OPLS all-atom (OPLS-AA) force field for the peptide.[1] For this, the 'Berger' lipid parameters were modified to permit a nonbonded scale factor of 0.5 for 1−4 interactions, which is the standard for OPLS-AA.[11] In addition, the hydrocarbon Lennard-Jones parameters of the lipid tails were changed, as the strength of the interactions in the original set was too weak. These new lipid parameters were used to study the folding and partitioning of WALP, resulting in markedly different behavior to the previous simulations (Methods are given in the Supporting Information).

* Corresponding author e-mail: jakob@ulmschneider.com (J.P.U.), martin@ulmschneider.com(M.B.U.).

† University of Heidelberg.

‡ University of Utrecht.

§ Oak Ridge National Laboratory.

10.1021/ct900256k CCC: $40.75
Published on Web 07/23/2009

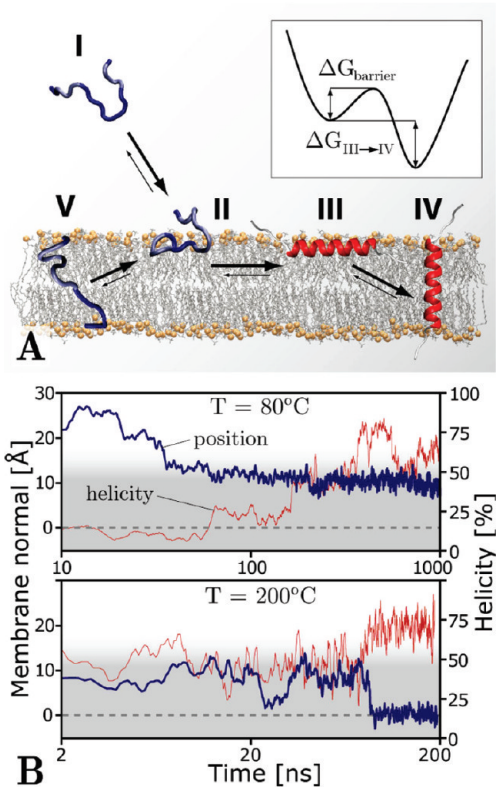


Figure 1. **A:** Schematic of peptide partitioning into lipid bilayer membranes. Hydrophobic peptides are generally unfolded in solution (I). Membrane insertion occurs after interfacial adsorption (II). Two partitioning pathways have been proposed: Unfolded (I→II→V→IV) and folded insertion (I→II→III→IV). Energetic arguments strongly favor a folded partitioning pathway. **B:** Adsorption, folding and partitioning of WALP16 in a DPPC bilayer. The helicity (red) and position along the membrane normal (blue) show peptide absorption and interfacial folding at 80 °C (**upper panel**), followed by folded insertion after ~80 ns at an elevated temperature of 200 °C (**lower panel**).

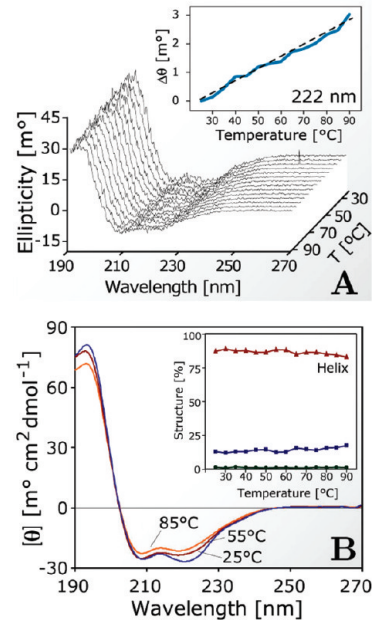


Figure 2. Circular dichroism measurements of the secondary structure of WALP16 in DMPC vesicles over a temperature range of 25−90 °C (for details see the Supporting Information). **A:** The spectra show three distinctive extrema at 208, 222 nm, and 193 nm, which are characteristic of alpha-helices embedded in the lipid bilayer membrane of vesicles. The 222 nm line (inset) shows linear behavior, demonstrating that no melting takes place over the full temperature range. **B:** The position of the peaks in the smoothed spectra show no significant lateral shifts. Secondary structure analysis (inset) shows that the peptide remains helical over the entire temperature range with a loss of helicity of less than 4%. Similar results were obtained in DPPC (manuscript in preparation).

In the simulations two types of bilayer-forming lipid (DPPC, DMPC) were used, and the peptide lengths (WALP16, WALP23) were varied. In addition to the force field changes reported, folding was also studied at lower temperatures (50 °C), where sampling still turned out to be rapid enough to capture folding events within the 1 μs time frame of the simulations.

Two sets of starting configurations were generated, with extended WALP peptides inserted into bilayers in a membrane spanning configuration, and outside the membrane in the solvent. The folding results of the different systems starting with the extended peptide embedded in the membrane (WALP16 in DMPC at 50 °C, WALP16 in DPPC at 80 °C, and WALP23 in DPPC at 80 °C) are summarized in Figure 3. All peptides rapidly form secondary structure, culminating in stable membrane spanning helices within 1 μs.

The helices correspond to the experimentally observed native state. The initial phase of all simulations is characterized by a rapid collapse of the peptide from its extended conformation. Within 50−100 ns, the water-exposed parts of the peptides are buried in the bilayer, and the enthalpy drops by ~10 kcal/mol per residue in all simulations. The collapse is concomitant with a buildup of helical turns (see Figure 4B). The significantly slower sampling rate of WALP16 in DMPC at 50 °C results in the peptide remaining unfolded but fully inserted from 170 to 260 ns, thus enabling energetic separation of the burial from the folding process. The result is an enthalpic stabilization of the native helix over the unfolded inserted conformations of ~2.6 ± 1.8 kcal/mol per residue. This is of the same order of magnitude as the ~4 kcal/mol estimated previously[5,12] and somewhat smaller than the 5−10 kcal/mol reported by Nymeyer et al.[13] Thus, the helix is enthalpically strongly favored over competing structures, in agreement with experimental evidence and in contrast to the results obtained with different force fields.[8]

To assess the thermodynamic stability of the peptides, we obtained the free energy as a function of helicity by plotting the logarithm of a population histogram over the final 500 ns (see Figure 4A, details in the Supporting Information). For the OPLS-AA protein in combination with the newly parametrized lipids unfolded insertion is highly unfavorable, even at high temperatures of 80 °C, and no unfolding is observed on the microsecond time scale of the simulations, indicating the transmembrane helix is the global free energy minimum. These results are independent of the lipid used, with indistinguishable folding patterns for WALP16 embedded into DMPC and DPPC bilayers. Increasing the hydrophobic core length of WALP from 16 to 23 residues showed no marked effect on the folding process, and except for a slight increase in the tilt angle of the native state from 18° ± 5° to 30 ± 9° no difference is

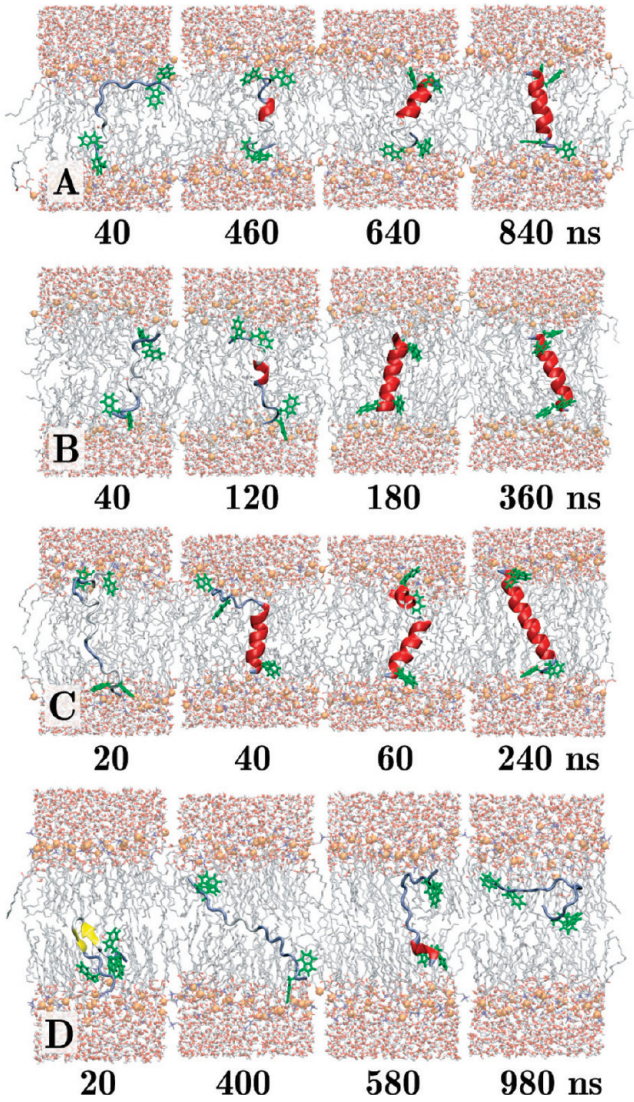


Figure 3. Intramembrane folding of WALP in explicit lipid bilayer membranes with the OPLS-AA force field and newly derived lipid parameters: **A:** WALP16 in DMPC at 50 °C. **B:** WALP16 in DPPC at 80 °C. **C:** WALP23 in DPPC at 80 °C. **D:** WALP16 in DPPC at 50 °C using the G96/Berger force field and the SPC water model. A large number of unfolded and beta-structures are sampled.

discernible. The tilt angle increase is consistent with experimental observations[14] and with results from implicit membrane models.[4,15,16] In addition, no major dependence of the equilibrium properties on the temperature is visible. The folding kinetics of WALP16 at 50 °C is markedly slower with 600 ns required to reach >50% helicity, compared to 140 ns at 80 °C. The folding pathway, however, remains similar.

The present results contrast sharply with previous WALP16 simulations in DMPC and DPPC bilayers.[8] In ref 8 unfolded conformations, stably inserted into the bilayer, dominate at 80 °C, with the peptide oscillating between deeply inserted completely extended and misfolded conformations. The transient formation of a membrane spanning helix after ~1.9 μs did not lead to further energetic stabilization, and the helix remained stable for only ~200 ns before unfolding again.[8] Furthermore, a control simulation starting from an inserted helix was found to unfold after ~300 ns, indicating that it is not stable at 80 °C.

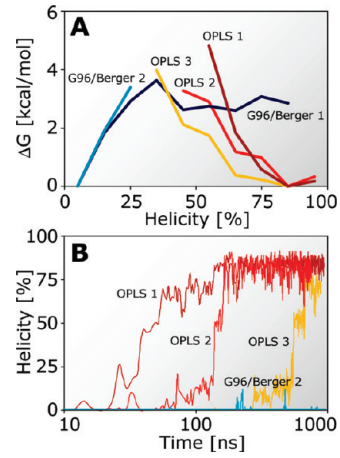


Figure 4. **A:** Free energies as a function of peptide helicity. All OPLS simulations correctly identify the native helical state as the free energy minimum. The G96/Berger systems show unfolded free energy minima. **B:** Helicity against simulation time for the OPLS systems (OPLS 1 = WALP16 in DPPC at 80 °C; OPLS 2 = WALP23 in DPPC at 80 °C; OPLS 3 = WALP16 in DMPC at 50 °C; G96/Berger 1 = WALP16 in DPPC at 80 °C; G96/Berger 2 = WALP16 in DPPC at 50 °C). For the G96/Berger simulations helices form only sporadically and unfold again rapidly.

The comparatively long time scales required to unfold the peptide suggests the state is kinetically trapped, with a thermally accessible barrier to unfolding/refolding. Simulations at 50 °C give similar results, with the peptide remaining inserted in an unfolded configuration (Figure 3D). However, the lack of folding events in this simulation, and the nonreversibility of the folding in all the OPLS simulations indicate that despite the microsecond scale the free energy profiles in Figure 4 should be considered only as rough estimates.

To investigate whether some of the observed differences can be explained by a helical bias in the underlying protein force field, control simulations (500 ns each) of WALP16 starting from an α-helical conformation were performed in water boxes (~2000 water molecules) at 27 and 80 °C, with both the G96 and OPLS-AA force field. As expected for a completely hydrophobic peptide, WALP16 quickly unfolds in all simulations, exposing the polar backbone to the water. The average helicity is ~10% over the final 100 ns in all four simulations. Thus, the difference in stability seen in the partitioning simulations is not due to an intrinsic preference of the OPLS-AA parameters favoring the helical state.

Spontaneous folding and insertion of WALP16 was also studied starting from unstructured conformations placed in bulk solvent. Hydrophobic peptides are thought to follow the 'folded insertion' pathway[6] illustrated in Figure 1A, where the unstructured peptide (I) is first adsorbed to the interface (II), which catalyzes folding (III), finally allowing the peptide to insert (IV). Hydrogen bond formation greatly decreases the cost of partitioning peptide bonds, favoring folded structures in the membrane. This model is supported by our simulations: WALP16 at 80 °C is adsorbed to the surface after ~25 ns, where it forms an interfacial helix after ~400 ns (see Figure 5A, the corresponding helicity and z-position is given in Figure 1B). The helix remains stable, sandwiched between the hydrophobic core and the polar lipid head groups. However, insertion is not

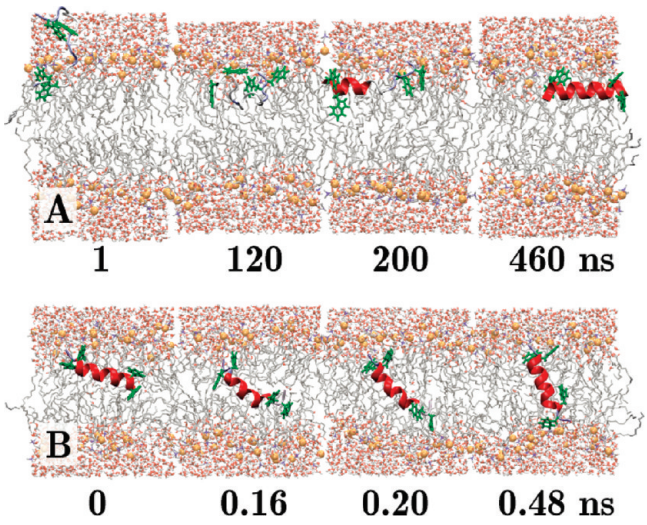


Figure 5. **A:** Adsorption and interfacial folding of WALP16 at 80 °C. **B:** Insertion at 200 °C after 80 ns. The insertion event is rapid taking less than 0.5 ns to complete.

observed within the microsecond time frame of the simulation, indicating the presence of a kinetic barrier to partitioning, due to the need to translocate two bulky TRP residues across the hydrophobic core of the bilayer from one interface to the other, their preferred position along the membrane normal. To probe the thermal surmountability of the barrier, the simulation temperature was increased to 200 °C. Figure 1B shows that WALP remains predominantly helical at this temperature, and the barrier is crossed after ~80 ns. At 80 °C, the transmembrane state of WALP16 is strongly stabilized compared to the interfacial helix by $\Delta H_{\mathrm{III}\rightarrow\mathrm{IV}} = -12 \pm 4.8$ kcal/mol or -0.8 ± 0.3 kcal/mol per residue. From the thermal accessibility of the kinetic barrier at 200 °C we estimate its height to be roughly ~20 kcal/mol.

The present results are very different from the interfacial folding simulations in refs 8 and 13, which observed neither an interfacial helical state nor a barrier to insertion, with WALP favoring the unfolded inserted (V) state. In contrast, the present OPLS simulations show that WALP16 follows the pathway illustrated in Figure 1A. The interfacial insertion barrier is probably not a general feature for all hydrophobic peptides and could vanish if bulky headgroup anchoring side chains are replaced by e.g., alanines or leucines. In this case the surface adsorbed folded state will only be populated during a short transition time, and the whole process might show concomitant folding and insertion. However, unfolded insertion will be unfavorable irrespective of the particular hydrophobic sequence.

The partitioning behavior illustrated in Figure 5 matches previous results on WALP peptides obtained with generalized Born implicit membrane models. These implicit membrane simulations also predicted an interfacial folding path to the final transmembrane helix and at a fraction of the computational cost.[3,4,17] However, implicit models are generally limited by a poor representation of structural features such as the complex lipid headgroup environment and entropic effects due to lipid tail order. Our results demonstrate that ab initio peptide partitioning studies can now also be performed in fully explicit lipid bilayers. This greatly increases both the accuracy and scope of membrane protein simulations. Explicit models also allow for studies of peptide induced bilayer deformations via hydrophobic mismatch, pore formation events, and membrane fusion or lysis. In addition, the lipid composition of the bilayer can be varied easily. Explicit treatment also directly accounts for structural water molecules, which often form part of an intricate system of hydrogen bonds that interconnect helices and protein subunits. Reorganization of these highly dynamic hydrogen bonding networks is one of the key drivers of conformational changes and associated function.

However, great care must be taken to ensure that the protein and lipid force fields are well balanced. In the future, this can be best achieved by comparing experimental peptide partitioning results with direct microsecond time scale folding simulations and thus obtaining improved force field parameters that reproduce partitioning data of whole peptides. The resulting models open the possibility to provide accurate atomic detail insights into complex biophysical membrane processes, such as antimicrobial peptide-induced membrane lysis or spontaneous assembly of membrane proteins from fragments.

Acknowledgment. This research was supported by the Human Frontier Science Program (M.B.U.) and BIOMS (J.P.U.). J.C.S. was supported by a DOE Laboratory Directed Research and Development award.

Supporting Information Available: Methods. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ulmschneider, J. P.; Ulmschneider, M. B. *J. Chem. Theory Comput.* 2009. in press (available online).

(2) White, S. H.; Wimley, W. C. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 319.

(3) Im, W.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6771.

(4) Ulmschneider, M. B.; Ulmschneider, J. P. *Mol. Membr. Biol.* **2008**, *25* (3), 245.

(5) White, S. H. *Adv. Protein Chem.* **2006**, *72*, 157.

(6) Jacobs, R. E.; White, S. H. *Biochemistry* **1989**, *28* (8), 3421.

(7) Popot, J. L.; Engelman, D. M. *Biochemistry* **1990**, *29* (17), 4031.

(8) Ulmschneider, M. B.; Ulmschneider, J. P. *J. Chem. Theory. Comput.* **2008**, *4* (11), 1807.

(9) Killian, J. A. *FEBS Lett.* **2003**, *555* (1), 134.

(10) Berger, O.; Edholm, O.; Jahnig, F. *Biophys. J.* **1997**, *72*, 2002.

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

(12) BenTal, N.; Sitkoff, D.; Topol, I. A.; Yang, A. S.; Burt, S. K.; Honig, B. *J. Phys. Chem. B* **1997**, *101* (3), 450.

(13) Nymeyer, H.; Woolf, T. B.; Garcia, A. E. *Proteins* **2005**, *59* (4), 783.

(14) Ozdirekcan, S.; Etchebest, C.; Killian, J. A.; Fuchs, P. F. *J. Am. Chem. Soc.* **2007**, *129* (49), 15174.

(15) Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. *Proteins* **2007**, *69*, 297.

(16) Sengupta, D.; Meinhold, L.; Langosch, D.; Ullmann, G. M.; Smith, J. C. *Proteins* **2005**, *58* (4), 913.

(17) Ulmschneider, J. P.; Ulmschneider, M. B. *Proteins* **2009**, *75* (3), 586.

CT900256K

JCTC Journal of Chemical Theory and Computation

Improved Hydrogen Bonding at the NDDO-Type Semiempirical Quantum Mechanical/Molecular Mechanical Interface

Qiantao Wang and Richard A. Bryce*

School of Pharmacy and Pharmaceutical Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, U.K.

Received May 25, 2009

Abstract: A semiempirical quantum mechanical (QM)/molecular mechanical (MM) potential with reformulated QM core-MM charge interactions is introduced, specifically to more accurately model hydrogen bonding at the QM/MM interface. Application of this potential using the PM3 Hamiltonian shows improved prediction of geometry and interaction energy for hydrogen bonded small molecule complexes typical of biomolecular interactions, without significantly impacting the modeling of other interaction types. Using this potential, more quantitative prediction of interaction energies is also found at a protein−ligand interface.

1. Introduction

Biological structure and function are dictated in major part by the influence of electrostatic interactions, in particular hydrogen bonds. These interactions typically dominate in substrate recognition, solvent rearrangements on binding and in enzyme reaction mechanisms. An ability to faithfully describe hydrogen bonding is fundamental to accurate biomolecular modeling methods. Ideally, it is desirable to model such complex biological systems at a quantum chemical level. A computationally efficient way to achieve this is *via* quantum mechanical/molecular mechanical (QM/MM) methods, where a quantum region is coupled to a solvent and biomolecular environment.[1] Since their first introduction,[2,3] QM/MM approaches have evolved in their power and popularity alongside improvements in the underlying QM and MM approaches.

QM/MM methods have also been employed to calculate thermodynamic properties, such as free energy reaction profiles, for comparison with experiment.[1] Here, however, in order to account for the large number of conformational states required in statistical mechanical calculations, the use of *ab initio* and density functional methods remains highly expensive. Correspondingly, there has been considerable interest in the development of fast and accurate semiempirical QM methods. Recent approaches include OMx,[4,5] PM3-D,[6] and PDDG/MNDO and PDDG/PM3[7] methods. Some approaches[8−11] seek to address the functional form of the NDDO[12] core−core interaction, contributing to the substantial improvements observed in the chemical accuracy of these methods.

In parallel with developments in QM and MM methodology, considerable effort has been invested into refining the description of the QM/MM interface, which can be modeled via mechanical or electronic embedding.[13] For the latter, within an *ab initio* QM/MM framework, the MM partial point charges enter the Hamiltonian of the QM region, forming electronic and nuclear interactions, while nonelectrostatic QM-MM interactions are modeled using a van der Waals potential. However, the atoms of a NDDO-based semiempirical QM region comprise valence electrons and an atomic core; this QM core subsumes the core electrons and nucleus of the atom into an *s* orbital term scaled by an effective nuclear charge, Z_a. Following the early work of Field et al.,[3] most current semiempirical QM/MM methods analogously treat MM point charges as scaled *s* orbital cores. The resulting QM/MM core−core interaction is of the same form as the QM/QM core−core expression that required correction as described above.

In this work, we therefore revisit the model of the semiempirical QM/MM interface, specifically exploring the effect of reformulating core−core interactions to more accurately model hydrogen bonding interactions in a biological context. In order to evaluate the ability of the method to model biologically important noncovalent interactions, we use the benchmark S22 data set,[14] a representative set of 22 bimolecular complexes with interaction energies determined at the level of CCSD(T) in conjunction with extrapolation methods to estimate the complete basis set (CBS) limit. We also apply our modified NDDO-based semiempirical QM/MM method to modeling noncovalent interactions at a protein−ligand interface.

2. Methods

Here we use the PM3[15] Hamiltonian in conjunction with the AMBER force field.[16] In the commonly adopted approach of Field et al.,[3] the energetic contribution arising from interaction of the core of a given QM atom *a* with MM atom *m* is given by

* Corresponding author phone: (0)161-275-8345; fax: (0)161-275-2481; e-mail: R.A.Bryce@manchester.ac.uk.

10.1021/ct9002674 CCC: $40.75
Published on Web 07/31/2009

$$E_{QM/MM}^{core} = Z_a q_m (s_a s_a, s_m s_m)(1 + e^{-\alpha_a R_{am}} + e^{-\alpha_m R_{am}}) \quad (1)$$

where Z_a is the effective charge of QM core a, q_m is the partial charge on MM atom m, s_a is an s orbital on the QM atom, s_m is a notional s orbital on the MM atom, and R_{am} is the QM-MM interatomic separation. α_m and α_a are parameters, as are $\rho_{a,0}$ and $\rho_{m,0}$ upon which the two-center two-electron integrals in eq 1 rely. After Field et al., α_m and $\rho_{m,0}$ are typically taken as 5.0 $Å^{-1}$ and 0.0 au, respectively.[3] The scaling of the integrals used in modeling core−core repulsions was originally introduced in the context of a purely semiempirical NDDO model and was intended to reflect increased screening by core electrons as two nuclei approach each other. Clearly, the physical premise alters in a QM/MM context, as now the expression contributes to the interaction of a MM partial point charge with a QM core, alongside a Lennard-Jones potential between QM and MM atoms. Our approach is to introduce a straightforward, general modification to this interaction term. Alteration of the final right-hand term of eq 1 leads to the following QM/MM core−core energy contribution by a given QM/MM atom pair

$$E_{QM/MM}^{core} = Z_a q_m (s_a s_a, s_m s_m)\left[1 + \frac{|q_m|}{q_m} \cdot (-e^{-f_1^a \cdot R_{am}} + e^{-f_2^a \cdot R_{am}})\right] \quad (2)$$

where $\rho_{m,0}$ is taken as 0.0 au as before, and f_1^a and f_2^a are exponential scale factors which depend on QM atom a. Optimal values were heuristically determined for f_1^a and f_2^a (Table 1) based on structures and energetics of complexes **1**−**3** of the S22 data set (Table 2).

A consequence of this new expression eq 2, relative to eq 1, is that the interaction of a QM core with a negatively charged (typically heavy) MM atom is enhanced; conversely, this QM/MM interaction is reduced for interaction of a QM core with a positively charged (typically light) MM atom. Use of eq 2 should therefore increase the strength of hydrogen bonding predicted by the semiempirical QM/MM method. All calculations were performed using a modified version of the Amber 9 software package.[16] For the MM molecules of S22 complexes, nonelectrostatic parameters for QM and MM atoms were obtained from the general AMBER force field (GAFF) for organic molecules.[17] No significant benefit was obtained by variation of van der Waals parameters for QM atoms. Atom-centered partial point charges for these molecules were derived using the AM1-BCC method.[18] For HIV-1 protease, parameters were adopted from the Cornell et al. force field for proteins.[19]

3. Results and Discussion

The S22 set[14] contains a range of binary complexes representative of biomolecular interactions (Table 2). Complexes **1**−**7** are bound together principally by hydrogen bonding; complexes **8**−**15** model dispersion-dominant interactions; and **16**−**22** are complexes with both types of interaction present. Starting from the *ab initio* geometries from Hobza et al. (at the MP2/cc-pVTZ level or higher[14]), interaction energies and optimal structures of the complexes were calculated at the PM3/MM level of theory. This led to a mean unsigned error in geometry across the 22 molecules of 0.22 Å (Table 3). The associated mean unsigned error (MUE) in energy relative to CCSD(T) energies at the complete basis set limit was 2.1 kcal/mol (Table 2). What is striking is the poor agreement for the hydrogen bonding complexes **1**−**7**, with mean unsigned errors in geometry and

Table 1. f_1^a and f_2^a Parameters for QM Atom *a* ($Å^{-1}$)

	f_1^a	f_2^a
H	2.2	2.7
H*	3.4	3.6
C	3.4	3.9
N	2.9	3.4
O	3.6	3.6

* When both QM and MM atoms are H.

Table 2. Interaction Energies of S22 Complexes (kcal/mol)[a] Calculated via the PM3/MM* Model (Using Eq *2*), Compared with PM3/MM and Reference Values[14]

		QM region	ref 14	PM3/MM	PM3/MM*
	Hydrogen Bonded Complexes (7)				
1	ammonia dimer (C_{2h})	*ammonia*	−3.17	−2.17	−3.18
2	water dimer (C_s)	*donor*	−5.02	−3.51	−5.25
		acceptor	−5.02	−3.92	−5.22
3	formic acid dimer (C_{2h})	*formic acid*	−18.61	−10.14	−14.31
4	formamide dimer (C_{2h})	*formamide*	−15.96	−8.51	−10.86
5	uracil dimer (C_{2h})	*uracil*	−20.65	−13.56	−16.61
6	2-pyridoxine/ 2-aminopyridine (C_1)	*2-py*	−16.71	−9.61	−13.23
		2-am	−16.71	−9.82	−13.60
7	adenine/thymine WC (C_1)	*adenine*	−16.37	−10.11	−14.36
		thymine	−16.37	−9.18	−12.59
MUE				**5.41**	**2.63**
MSE				**5.41**	**2.54**
	Complexes with Predominant Dispersion Contribution (8)				
8	methane dimer (D_{3d})	*methane*	−0.53	−0.49	−0.51
9	ethene dimer (D_{2d})	*ethene*	−1.51	−0.79	−0.88
10	benzene/methane (C_3)	*benzene*	−1.50	−0.97	−0.99
		methane	−1.50	−1.06	−1.15
11	benzene dimer (C_{2h})	*benzene*	−2.73	−2.56	−2.67
12	pyrazine dimer (C_s)	*pyrazine*	−4.42	−5.33	−5.64
13	uracil dimer (C_2)	*uracil*	−10.12	−9.31	−10.21
14	indole/benzene (C_1)	*indole*	−5.22	−5.57	−6.01
		benzene	−5.22	−3.99	−5.10
15	adenine/thymine stack (C_1)	*adenine*	−12.23	−12.18	−13.46
		thymine	−12.23	−12.87	−14.36
MUE				**0.54**	**0.65**
MSE				**0.19**	**−0.34**
	Mixed Complexes (7)				
16	ethene/ethine (C_{2v})	*ethene*	−1.53	−0.57	−0.62
		ethane	−1.53	−0.67	−0.73
17	benzene/water (C_s)	*benzene*	−3.28	−2.18	−3.57
		water	−3.28	−3.45	−3.79
18	benzene/ammonia (C_s)	*benzene*	−2.35	−1.59	−2.59
		ammonia	−2.35	−2.87	−3.12
19	benzene/HCN (C_s)	*benzene*	−4.46	−2.51	−2.69
		HCN	−4.46	−3.17	−3.30
20	benzene dimer (C_{2v})	*vertical*	−2.74	−1.98	−2.10
		horizontal	−2.74	−1.78	−1.90
21	indole/benzene T-shape (C_1)	*indole*	−5.73	−5.55	−6.00
		benzene	−5.73	−4.10	−5.27
22	phenol dimer (C_1)	*donor*	−7.05	−5.15	−6.17
		acceptor	−7.05	−5.29	−6.74
MUE				**1.06**	**0.70**
MSE				**0.96**	**0.41**
	Overall				
MUE				**2.14**	**1.24**
MSE				**1.99**	**0.78**
r^2				**0.86**	**0.94**

[a] MUE (mean unsigned error); MSE (mean signed error); correlation with reference values, r^2.

Table 3. Interaction Distances of S22 Complexes (in Å)[a] Calculated via the PM3/MM* Model (Using Eq *2*), Compared with PM3/MM and Reference Values[14]

		QM region	ref 14	PM3/MM	PM3/MM*
	Hydrogen Bonded Complexes (7)				
1	ammonia dimer (C_{2h})	*ammonia*	2.504	2.221	2.112
			2.504	3.530	2.179
2	water dimer (C_s)	*donor*	1.952	2.008	1.791
		acceptor	1.952	1.973	1.770
3	formic acid dimer (C_{2h})	*formic acid*	1.670	1.851	1.637
			1.670	1.891	1.739
4	formamide dimer (C_{2h})	*formamide*	1.841	1.987	1.874
			1.841	2.105	1.886
5	uracil dimer (C_{2h})	*uracil*	1.775	1.882	1.792
			1.775	1.941	1.822
6	2-pyridoxine/ 2-aminopyridine (C_1)	*2-py*	1.859	2.055	1.954
			1.874	1.978	1.856
		2-am	1.859	2.031	1.882
			1.874	2.037	1.869
7	adenine/thymine WC (C_1)	*adenine*	1.819	1.978	1.833
			1.929	2.026	1.886
		thymine	1.819	2.014	1.891
			1.929	1.989	1.841
MUE				**0.201**	**0.092**
MSE				**0.169**	**−0.046**
	Complexes with Predominant Dispersion Contribution (8)				
8	methane dimer (D_{3d})	*methane*	3.718	3.639	3.621
9	ethene dimer (D_{2d})	*ethene*	3.718	3.899	3.842
10	benzene/methane (C_3)	*benzene*	3.716	3.907	3.895
		methane	3.716	3.874	3.838
11	benzene dimer (C_{2h})	*benzene*	3.765	3.823	3.806
12	pyrazine dimer (C_s)	*pyrazine*	3.479	3.444	3.410
13	uracil dimer (C_2)	*uracil*	3.166	3.440	3.356
14	indole/benzene (C_1)	*indole*	3.498	4.500	4.535
		benzene	3.498	4.117	4.577
15	adenine/thymine stack (C_1)	*adenine*	3.172	3.403	3.360
		thymine	3.172	3.344	3.275
MUE				**0.273**	**0.294**
MSE				**0.252**	**0.263**
	Mixed Complexes (7)				
16	ethene/ethine (C_{2v})	*ethene*	2.752	3.082	3.021
		ethane	2.752	2.899	2.859
17	benzene/water (C_s)	*benzene*	3.435	3.309	3.010
		water	3.435	3.063	3.012
18	benzene/ammonia (C_s)	*benzene*	3.592	3.463	3.166
		ammonia	3.592	3.170	3.121
19	benzene/HCN (C_s)	*benzene*	3.387	3.497	3.472
		HCN	3.387	3.592	3.575
20	benzene dimer (C_{2v})	*vertical*	3.513	3.728	3.700
		horizontal	3.513	3.769	3.728
21	indole/benzene T-shape (C_1)	*indole*	3.238	3.062	3.020
		benzene	3.238	3.228	3.050
22	phenol dimer (C_1)	*donor*	1.937	2.093	1.906
			4.921	4.705	4.904
		acceptor	1.937	2.024	1.828
			4.921	5.138	5.090
MUE				**0.198**	**0.220**
MSE				**0.017**	**−0.068**
	Overall				
MUE				**0.218**	**0.187**
MSE				**0.135**	**0.022**

[a] MUE (mean unsigned error); MSE (mean signed error). The definition of interaction distances follows ref 6.

binding energy of 0.20 Å and 5.4 kcal/mol, respectively. The corresponding PM3/MM hydrogen bond distances are systematically too long, with a mean signed error (MSE) in intermolecular distance of 0.17 Å (Table 3). Single-point PM3/MM energy calculations at the high level *ab initio* QM geometries do not significantly improve the systematically underestimated interaction energies (Supporting Information).

It is instructive to consider the case of water dimer. The binding energy of water dimer at the CCSD(T)/CBS level is −5.0 kcal/mol (Table 2). At the QM/MM level, where PM3 water is the proton donor, the binding energy is considerably underestimated, at −3.5 kcal/mol (Table 2); we note the same level of underbinding (−3.5 kcal/mol) is exhibited by the entirely PM3 water dimer.[20] Reversing the QM/MM model, such that PM3 water is the acceptor, the interaction is stronger by 0.4 kcal/mol. Applying the modified potential which uses eq 2, denoted here as PM3/MM*, interaction energies of −5.2 kcal/mol are obtained, regardless of whether QM water is donor or acceptor (Table 1). This is in good agreement with the CCSD(T)/CBS value. The improved hydrogen bond energy arises from enhanced $H_{QM}\cdots O_{MM}$ interactions which, due to choice of f_1^a and f_2^a (Table 1), are pronounced relative to $O_{QM}\cdots O_{MM}$ interactions. As observed above, the modified potential also addresses to some extent the variation in predicted hydrogen bond strength, depending on which water is treated as QM. Larger exponentials for f_1^a and f_2^a of 3.6 and 3.6 $Å^{-1}$ respectively when the acceptor oxygen is treated as QM (*i.e.* a $O_{QM}\cdots H_{MM}$ hydrogen bond) ensure a smaller contribution relative to the converse situation of a QM hydrogen donor, where f_1^a and f_2^a are 2.2 and 2.7 $Å^{-1}$, respectively. Thus, this approach reduces the imbalance in the QM/MM hydrogen bond in water dimer, compensating for underpolarization of the QM oxygen. In terms of geometry of the water dimer complex, where QM water is the proton donor, the O−H···O angle improves from 197.3° at PM3/MM to 187.6° at PM3/MM*. This is closer to the angle observed in the CCSD(T)/cc-pVQZ geometry of 172.8°. For QM water as the acceptor, a smaller improvement is seen in the O−H···O angle, from 188.5° to 181.7°. We note that the improved binding energy for water dimer does correspond to O···H distances of 1.79 and 1.77 Å for QM donor and acceptor, respectively (Table 3); these are about 0.15 Å shorter than the CCSD(T)/cc-pVQZ values and 0.20 Å shorter than PM3 values. However, by comparison, very short PM3/MM hydrogen bond distances of 1.62 and 1.66 Å are required to obtain interaction energies of −5.2 and −5.1 kcal/mol for the QM water as hydrogen bond donor and acceptor respectively (achieved through optimization of van der Waals' parameters of the QM atoms).

For ammonia dimer, there exists a sensitive balance of H···H and N···H noncovalent interactions which determine the optimal C_{2h} structure, a tilted geometry with N−H···N angles of 121.9° and H···N distances of 2.50 Å at the CCSD(T)/cc-pVQZ level (Figure 1). The symmetry of $(NH_3)_2$ is lost by optimization *via* PM3/MM, such that the N−H···N angles are 65.4° and 175.3° and the H···N distances are 2.22 and 3.53 Å (Table 3, Figure 1). Through use of the PM3/MM* potential, the tilted C_{2h} symmetry of the $(NH_3)_2$ complex is approximately recovered with H···N distances of 2.11 and 2.18 Å (Table 3, Figure 1), despite the inherently asymmetric treatment of the waters via QM and MM models. In this improved orientation, the binding energy of ammonia dimer correspondingly improves from −2.2 kcal/mol at PM3/MM to the high level *ab initio* value of −3.2 kcal/mol at the PM3/MM* level (Table 2).

The remaining hydrogen bonding complexes **3**−**7** observe geometries in closer agreement with high level *ab initio* values. For example, formamide dimer has distances of 1.87 and 1.89 Å at the PM3/MM* level (Table 3), in better agreement with

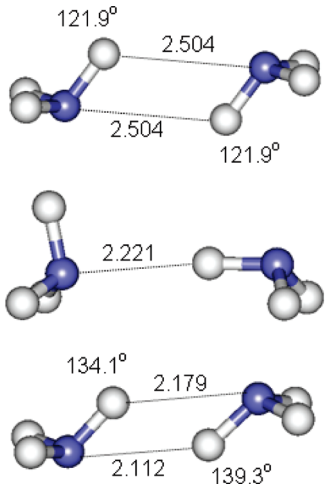


Figure 1. Minimum energy structure of ammonia dimer at the (a) CCSD(T)/aug-cc-pVQZ, (b) PM3/MM, and (c) PM3/MM* levels of theory. Distances in Å.

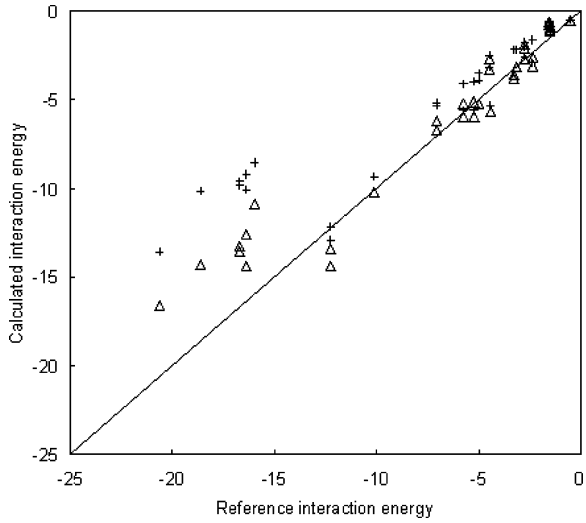


Figure 2. Interaction energies (kcal/mol) of S22 set calculated by CCSD(T)/complete basis set (reference), PM3/MM (+) and PM3/MM* (Δ) levels of theory.

the two CCSD(T)/aug-cc-pVTZ distances of 1.84 Å than PM3/MM values of 1.99 and 2.11 Å. Indeed, for hydrogen bonded systems **1**−**7**, the MUE in geometry improves from 0.20 Å to 0.09 Å, and the MSE reduces from 0.17 Å to −0.05 Å, showing a small underestimation (Table 3). This compares with a more modest overall improvement in geometry, from a MUE of 0.22 Å for the 22 molecules at PM3/MM to 0.19 Å at PM3/MM* (Table 3). Alongside improved geometries, the overall MUE in interaction energy for the S22 set reduces from 2.1 kcal/mol at PM3/MM to 1.2 kcal/mol at PM3/MM* (Table 2); this is illustrated by an improved correlation r^2 with *ab initio* QM binding energies, from 0.86 at PM3/MM to 0.94 at PM3/MM* (Figure 2). For the hydrogen bonded complexes **1**−**7**, the improvement is particularly apparent, with a reduction in MUE from 5.4 kcal/mol at the PM3/MM level to 2.6 kcal/mol *via* PM3/MM* (Table 2).

As expected, less improvement in modeling structures and energetics is observed for dispersion-dominant and mixed interaction sets. However, the errors in energy at the PM3/MM level are already rather smaller relative to hydrogen bonded

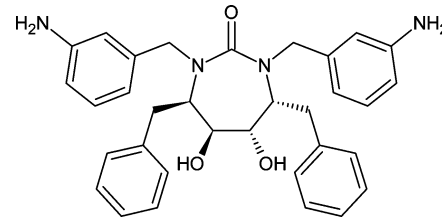


Figure 3. HIV-1 protease inhibitor, mozenavir.

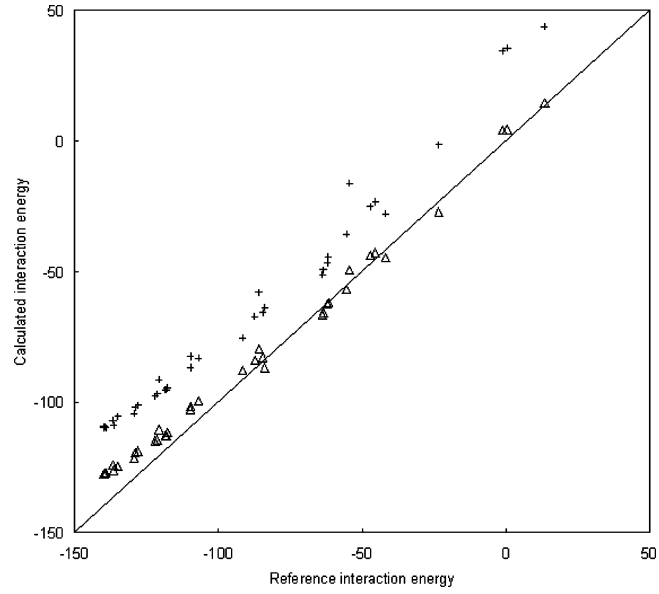


Figure 4. Comparison of PM3/MM (+) and PM3/MM* (Δ) models with HF/6-31G* (reference) QM/MM model in predicting total electrostatic interaction energies (kcal/mol) of native and decoy ligand poses with HIV-1 protease.

Table 4. Mean Unsigned Error (MUE), Mean Signed Error (MSE), Range of Error (Δ*E*), and Correlation r^2 for Electrostatic Interaction Energies of Native and Decoy Ligand/HIV-1 Protease Complexes Computed at the PM3/MM and PM3/MM* Levels of Theory, Relative to Reference *Ab Initio* QM/MM Calculations

method	MUE	MSE	Δ*E*	r^2
PM3/MM	25.20	25.20	25.39	0.85
PM3/MM*	6.29	5.53	15.71	0.96

complexes **1**−**7**. For complexes of mixed interaction type **16**−**22**, a modest decrease in mean unsigned error of binding energy from 1.1 kcal/mol for PM3/MM to 0.7 kcal/mol for PM3/MM* is observed (Table 2). However, no improvement in performance is found for dispersion dominant complexes **8**−**15**, where a small increase in MUE by 0.1 kcal/mol is observed. The mean unsigned errors in geometry for mixed and dispersion-dominant complexes are marginally poorer using the modified QM/MM potential, by 0.02 Å in both cases (Table 3). The subtle dispersive interactions of these complexes are treated in the QM/MM models rather simplistically *via* the QM/MM van der Waals' potential. The largest error in geometry of any bimolecular complex in the S22 set, for both PM3/MM and PM3/MM* models, is found for the indole/benzene stacked dimer (complex **14**). Here, the stacked conformation of the MP2/aug-cc-pVTZ geometry is distorted into a proto-T-shape conformation at the PM3/MM and PM3/MM* levels of theory (complex

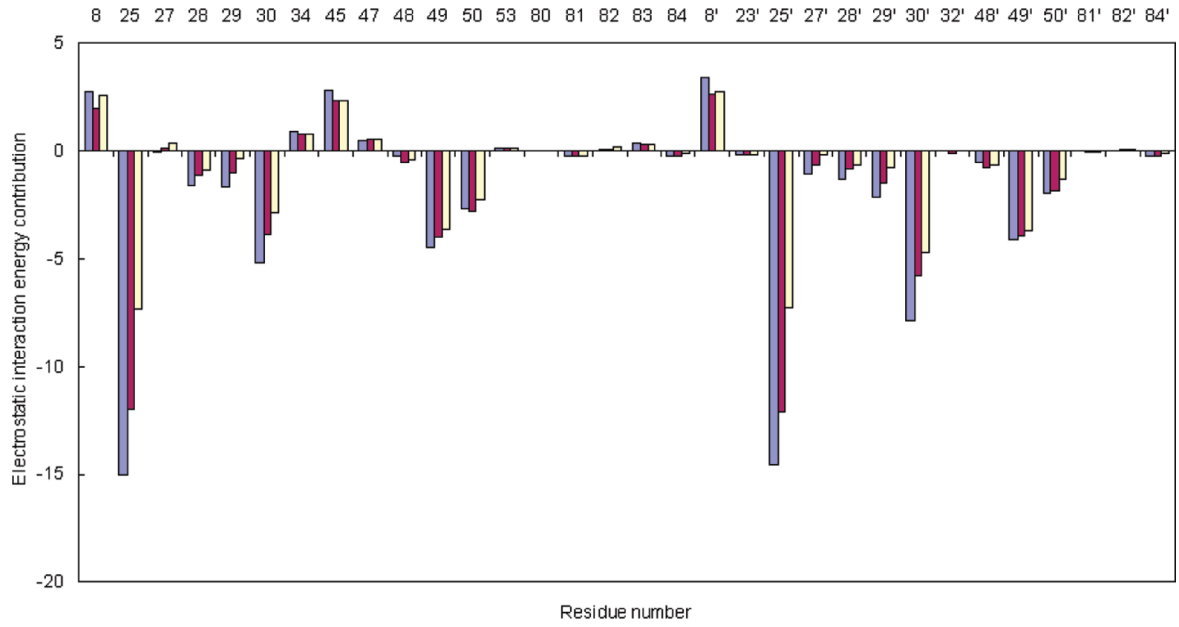


Figure 5. Electrostatic interaction energy contribution (kcal/mol) of selected residues to binding of mozenavir in its native pose, using PM3/MM (white), PM3/MM* (dark gray), and HF/6-31G*:MM (light gray).

21), leading to intercentroid distances of around 1 Å larger than the *ab initio* QM values (Table 3 - note that interaction distances for complexes **14** and **21** are defined differently). This reflects for this complex an imbalance of polar and dispersive interactions in both the semiempirical QM/MM models.

Having adopted a reformulated QM core-MM charge term and applied the resulting PM3/MM* model to the S22 set, we now consider the case of a protein−ligand interaction, specifically that of HIV-1 protease with mozenavir (Figure 3).

The native and 39 non-native structures of the complex have been studied previously by us, using a QM/MM approach, with the QM region (the ligand) described both at the HF/6-31G* and PM3 levels of theory.[21] In accord with available experimental evidence, the catalytic aspartyl dyad was modeled in its neutral form. The ligand conformations adopted in the non-native structures span a root-mean-square deviation in atom positions from 0.15 to 6.86 Å from the native pose. It was found that PM3/MM calculations underestimated the affinity of protein−ligand interactions relative to the *ab initio* QM/MM model quite significantly (Figure 4), with a MSE and MUE of 25.2 kcal/mol relative to HF for the 40 structures.[21]

Application of the modified PM3/MM* potential to the set of mozenavir/HIV-1 protease structures reduces the MSE in interaction energy to 5.5 kcal/mol and MUE to 6.3 kcal/mol; the range in error is also 10 kcal/mol smaller (Table 4). The correlation, r^2, between calculated semiempirical and *ab initio* QM/MM electrostatic interaction energies improves from 0.85 to 0.96.

To obtain insight into contributions of individual amino acid residues to binding mozenavir in its native pose, MM charges of individual key amino acids were deleted and the electrostatic interaction energy difference obtained (Figure 5). Generally more quantitative results are found for PM3/MM* relative to the PM3/MM model. For example, at the *ab initio* QM/MM level, the catalytic residues Asp25 and Asp25′ are predicted to contribute −15.0 and −14.6 kcal/mol, respectively, to the binding of mozenavir (Figure 5). Whereas the PM3/MM model finds contributions of −7.4 and −7.3 kcal/mol for Asp25 and Asp25′, the PM3/MM* method obtains closer agreement with the HF model, with respective estimates of −12.0 and −12.1 kcal/mol. Similar improvement is found for other amino acid contributions, with an overall reduction in MUE from 1.0 to 0.5 kcal/mol. We note that PM3/MM* calculations are in reasonable agreement with HF/6-31G*/MM interaction energies for protein−ligand complexes as well as much higher level CCSD(T)/CBS calculations for the S22 small molecule complexes. Although the reasons for this are not entirely clear, agreement may in part stem from the consistency of the AMBER point charge model with the HF/6-31G* level of theory and from a fortuitous balance struck by the HF/6-31G* model.

4. Conclusions

In this Letter, we have introduced a semiempirical QM/MM potential with reformulated QM core-MM charge interactions. Application of the QM/MM potential shows improved prediction of geometry and interaction energy for hydrogen bonded small molecule complexes typical of biomolecular interactions. This is without significant deterioration in the modeling of dispersive interactions. Further refinement in the approach can be envisaged - for example, here we introduced a specific f_1^a and f_2^a set for H_{QM}-H_{MM} interactions (Table 1). This essentially bond-specific approach could be generalized to other QM-MM atom type pairs, although, in doing so, one introduces more parameters and may reduce the generality of the approach. The modification to the core−core expression we have proposed here in eq 2 is readily applicable to existing semiempirical QM/MM codes. While we have explored here improved handling of hydrogen bonding through this QM/MM core−core expression, we note that an interesting alternative approach could be to recapture the original intent of the core−core expression as used in purely QM approaches, to model only short-range repulsion between atoms (dealt with in current QM/MM formulations principally via the r^{-12} component of the QM/MM Lennard-Jones

potential). Combined with other developments, such as improved QM-QM dispersion (PM3-D, OMx),[5,6] the capability to model metals (*e.g.* PM6[22]), and focused parametrizations (*e.g.* PM3CARB-1 for modeling carbohydrates[23]), semiempirical QM/MM methods remain a powerful cost-effective approach to solving a range of important biochemical and biophysical problems.

Acknowledgment. We thank Jonathan McNamara for useful discussions.

Supporting Information Available: Single point interaction energies of S22 set calculated by PM3 QM/MM models at high level *ab initio* geometries. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.

(2) 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.

(3) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11*, 700.

(4) Kolb, M.; Thiel, W. Beyond the MNDO Model - Methodical Considerations and Numerical Results. *J. Comput. Chem.* **1993**, *14*, 775.

(5) Weber, W.; Thiel, W. Orthogonalization Corrections for Semiempirical Methods. *Theor. Chem. Acc.* **2000**, *103*, 495.

(6) McNamara, J. P.; Hillier, I. H. Semi-Empirical Molecular Orbital Methods Including Dispersion Corrections for the Accurate Prediction of the Full Range of Intermolecular Interactions in Biomolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.

(7) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: Improved Semiempirical Methods. *J. Comput. Chem.* **2002**, *23*, 1601.

(8) Voityuk, A. A.; Rosch, N. AM1/d Parameters for Molybdenum. *J. Phys. Chem. A* **2000**, *104*, 4089.

(9) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. AM1*Parameters for Phosphorus, Sulfur and Chlorine. *J. Mol. Model.* **2003**, *9*, 408.

(10) Sharma, R.; McNamara, J. P.; Raju, R. K.; Vincent, M. A.; Hillier, I. H.; Morgado, C. A. The Interaction of Carbohydrates and Amino Acids With Aromatic Systems Studied by Density Functional and Semi-Empirical Molecular Orbital Calculations With Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2767.

(11) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods IV:Extension of MNDO, AM1, and PM3 to More Main Group Elements. *J. Mol. Model.* **2004**, *10*, 155.

(12) Pople, J. A.; Santry, D. P.; Segal, G. A. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys.* **1965**, *43*, S129.

(13) Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. Combining Quantum Mechanics Methods With Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006**, *2*, 815.

(14) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(15) Stewart, J. J. P. Optimisation of Parameters for Semi-Empirical Methods. 2. Applications. *J. Comput. Chem.* **1989**, *10*, 221.

(16) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668.

(17) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157.

(18) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623.

(19) 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 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(20) Bernal-Uruchurtu, M. I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-Lopez, M. F. Improving Description of Hydrogen Bonds at the Semiempirical Level: Water-Water Interactions As Test Case. *J. Comput. Chem.* **2000**, *21*, 572.

(21) Fong, P.; McNamara, J. P.; Hillier, I. H.; Bryce, R. A. Assessment of QM/MM Scoring Functions for Molecular Docking to HIV-1 Protease. *J. Chem. Inf. Model.* **2009**, *49*, 913.

(22) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173.

(23) McNamara, J. P.; Muslim, A. M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, *394*, 429.

CT9002674

JCTC Journal of Chemical Theory and Computation

Toward a Practical Method for Adaptive QM/MM Simulations

Rosa E. Bulo,*,† Bernd Ensing,‡ Jetze Sikkema,† and Lucas Visscher†

Department of Theoretical Chemistry, Vrije Universiteit Amsterdam, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands and Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Received March 27, 2009

Abstract: We present an accurate adaptive multiscale molecular dynamics method that will enable the detailed study of large molecular systems that mimic experiment. The method treats the reactive regions at the quantum mechanical level and the inactive environment regions at lower levels of accuracy, while at the same time molecules are allowed to flow across the border between active and environment regions. Among many other things, this scheme affords accurate investigation of chemical reactions in solution. A scheme like this ideally fulfills the key criteria applicable to all molecular dynamics simulations: energy conservation and computational efficiency. Approaches that fulfill both criteria can, however, result in complicated potential energy surfaces, creating rapid energy changes when the border between regions is crossed. With the difference-based adaptive solvation potential, a simple approach is introduced that meets the above requirements and reduces fast fluctuations in the potential to a minimum. In cases where none of the current adaptive QM/MM potentials are able to properly describe the system under investigation, we use a continuous force scheme instead, which, while no longer energy conserving, still retains a related conserved quantity along the trajectory. We show that this scheme does not introduce a significant temperature drift on time scales feasible for QM/MM simulations.

I. Introduction

Computational studies play an integral role in many fields of chemistry today. Theoretical investigations are no longer restricted to small organic molecules in the gas phase, but large and complex systems, like solutions, proteins, or solids, are modeled. These complex systems require methods that thoroughly sample configuration space, such as molecular dynamics (MD).[1] For the study of chemical reactions occurring within such large systems, it is advisible to treat changes in the electronic structure explicitly using quantum mechanical models (QM). In this manner the bond breaking and forming processes at the reactive sites can be correctly represented. Since the changes in the electronic structure of a system are usually local, one can describe regions that lie far away from the active site in a simplified manner. A popular approximation is the treatment of those regions in a classical manner using empirical force fields (QM/MM).[2] Treating two regions with different levels of quantum chemical methods (QM/QM), like mixed basis set methods[3] or frozen density embedding (FDE),[4,5] is an alternative option.

A broad range of multiscale methods have been developed, and they have been widely applied to systems with localized active sites. However, problems may occur when one applies these general methods to simulate the time evolution of a multiscale solute−solvent system at the longer time scales that have become accessible in recent years. For example, in a QM/MM simulation of a solution containing both QM and MM solvent molecules, the diffusive nature of the solvent causes QM molecules to move out of the active

* Corresponding author. E-mail: bulo@few.vu.nl.
† Vrije Universiteit Amsterdam.
‡ University of Amsterdam.

10.1021/ct900148e CCC: $40.75
Published on Web 07/30/2009

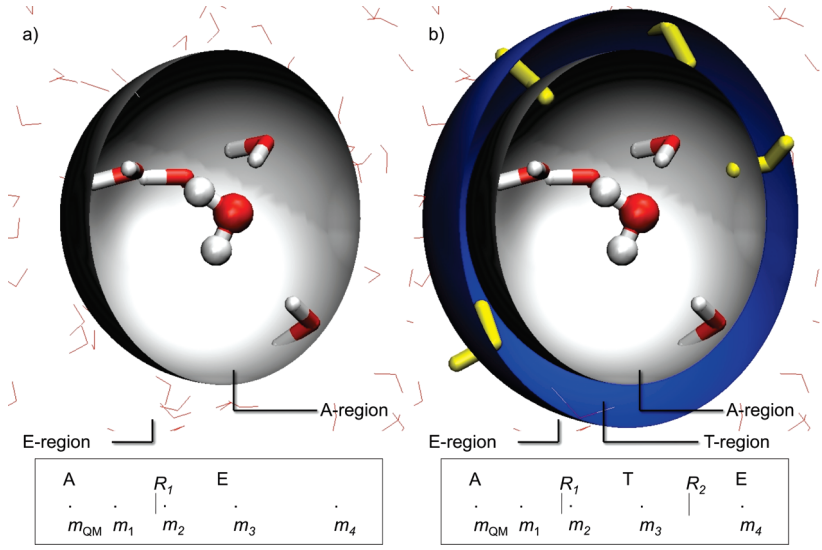


Figure 1. Partitioning of a water box into (a) an A-region (thick water molecules) and E-region (thin lines representing the H_2O) and (b) an A-region, T-region (yellow water molecules) and E-region centered around one water molecule (QM-center: m_{QM}). At the bottom, a schematic 1-dimensional representation of the same partitionings is depicted.

region and MM molecules to move in. This results in a system that treats solvent molecules far away from the active site at a high level of theory, whereas in the sensitive region near the active site, solvent molecules are treated at an inadequate level of theory. To overcome the drawbacks of a rigid partitioning of the system, new methods should be developed that allow the description of diffusive molecules to change 'on the fly' when they cross the border between an active (A) and an environment (E) region.

The simplest version of such a scheme simply chooses an A- and E-region based on a radius around a central active subsystem (Figure 1a). When a solvent molecule crosses this border, its description changes from QM to MM or vice versa. Problematic with this simple scheme is the strong dependence of the absolute potential on the number of atoms that are treated quantum mechanically. Altering the number of QM treated solvent molecules will result in a large and abrupt change in the total energy of the system. Furthermore, because the location of the minima on the QM and MM potential energy surfaces may differ significantly, the system may find itself in a high-energy region after the repartitioning. This effect can lead to sudden large jumps in the forces experienced by the atoms. Indeed, it has been shown that such simulations, even in the NVT ensemble, exhibit extreme accelerations of the atoms.[6] The underlying problem of the simple repartitioning method is the lack of a continuous and smooth potential that defines the forces on the atoms at each time step. Writing the forces as derivatives of such a potential guarantees that Newton's laws apply to the simulations, ensuring reliable statistical data. In practice such simulations are hard to tune because it is difficult to distinguish errors in the setup (such as large time steps) from the inherent flaws of the method.

In order to overcome these problems with the above repartitioning approach, several multiscale methods have been developed which all share a common basis. The system under investigation is no longer sharply divided into an A- and E-region, but contains a transition region (T-region), with inner and outer boundaries R_1 and R_2, (Figure 1b) defined by the distance from the active center. The molecules crossing the T-region gradually change character from QM to MM, allowing the forces to vary in a continuous manner. Among the adaptive QM/MM methods are the 'hot spot' method,[7] ONION-XS,[6] and learn on the fly.[8] These methods combine forces obtained from two QM/MM partitionings, one with and one without the T-region treated at the QM level. They are very efficient, greatly diminish the spurious jumps in the forces, and have resulted in successful applications on a variety of systems.[9] However, for most applications, these methods are not able to fully eliminate discontinuities in the forces. In the absence of a thermostat, this results in strong temperature drifts, and also when a thermostat is present significant deviations from the desired equilibrium situation may occur. Schemes that do entirely eliminate all force discontinuities have been developed for additive pair potential-based (i.e., non-QM) adaptive multiscale setups,[10] in the field of combined atomistic (MM)/coarse grained (CG) simulations.[11]

The problems with the QM/MM methods are rooted in the fact that the molecules in the T-region all acquire a fractional degree of 'QM character' because, unlike combined pair potentials, a QM/MM energy combination can only be defined with integer numbers of molecules in the quantum chemical and classical parts of the system. One effective solution to this problem is to determine at each time step the QM/MM potential energies for all possible partitionings with a different subset of T-region molecules included in the QM calculation. The total potential is then defined as a weighted average of these individual potentials. Recently Heyden et al. introduced such a QM/MM approach, which produced completely continuous forces and was able to conserve the total energy of the system.[12] This approach, however, still has some drawbacks when applied to situations where the QM and the MM energy surfaces exhibit large differences, as will be demonstrated later in this paper.

In this work we present two new developments toward accurate and efficient adaptive QM/MM simulations. In Section II, a new adaptive potential, the difference-based adaptive solvation potential is introduced, which enables energy conserving simulations. Section III compares this potential to the sorted adaptive partitioning potential by Heyden et al. and addresses the advantages of the current approach upon application to chemistry in solution. In Section IV, the common drawback of these two energy conserving methods, the introduction of spurious forces in the T-region, is discussed. A force-based method is introduced as an alternative which, in combination with a bookkeeping algorithm,[10] still retains many of the advantages of a fully Hamiltonian approach. Section V, the Results Section, presents a set of 'proof of principle' simulations and compares results obtained with the two methods discussed in this paper. Finally, Section VI contains our conclusions.

II. Difference-Based Adaptive Solvation

In our adaptive solvation scheme, the targeted fractional 'MM character' of each adaptive molecule (m_i) is captured by a function λ_i that measures the progress of the molecule in the T-region. This function gradually increases from 0 to 1 as a molecule crosses the T-region from the inner border R_1 to the outer border R_2. The progress function applied in this paper, based on the distance r_i of the center of the molecule to the center of the A-region, is presented in eq 1:

$$\lambda_i(r_i) = \begin{cases} 0 & \text{if } r_1 < R \\ \dfrac{(r_i - R_1)^2(3R_2 - R_1 - 2r_i)}{(R_2 - R_1)^3} & \text{if } R_1 \le r_1 \le R_2 \\ 1 & \text{if } r_1 > R_2 \end{cases} \quad (1)$$

The total adaptive potential V^{ad} is defined as a weighted average of QM/MM partitioning energies V_a:

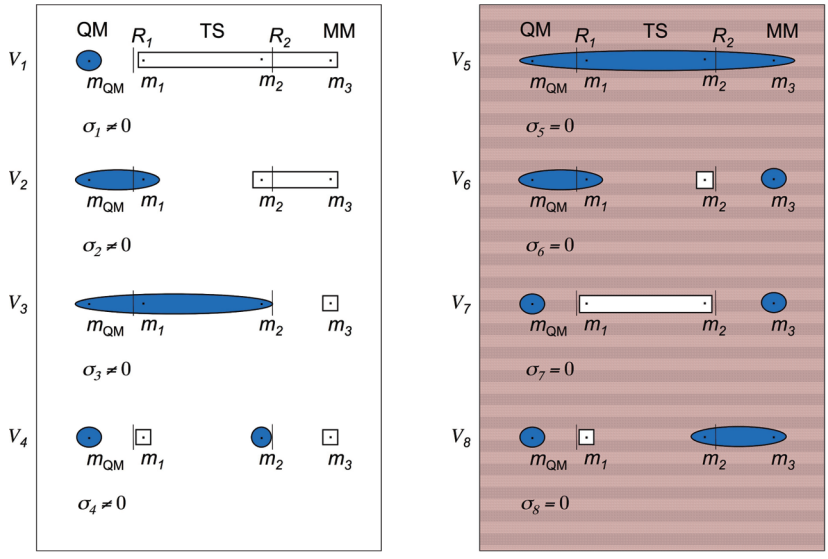
$$V^{ad}(\mathbf{r}) = \sum_a^{2^N} \sigma_a(\mathbf{r}) V_a(\mathbf{r}) \quad (2)$$

in which both the weight σ_a and the potential energy are a function of the coordinates ($\mathbf{r}$) of the N adaptive molecules in the system. The dependence of σ_a on the coordinates will be expressed in terms of the λ_i functions defined above for each of the adaptive molecules. In order to give a detailed description of the method, a few additional definitions are needed. Each QM/MM energy term V_a corresponds to a partitioning of the system into a group of QM molecules (G_a^{QM}) and a group of MM molecules (G_a^{MM}). The sets of λ values for each group will be referred to as $\{\lambda\}_a^{QM}$ and $\{\lambda\}_a^{MM}$.

We can now consider the exact dependence of σ_a on the coordinates. If a total of M molecules is present in the T-region at a given time, then 2^M terms correspond to partitionings that have their QM/MM division line completely inside the T-region. It is those terms that should contribute a non-zero value to the total energy expression of eq 2 (see also Scheme 1). Such a number of energy terms is, however, too large even for the study of small solute molecules. Take for example, the simple test system of one QM water molecule in an MM water liquid. A reasonable approximation would be to define a shell of 4 Å radius as the A-region and surround this by a T-region with a width of 1 Å (Figure 1). At room temperature and pressure, one would then readily find 10 or more water molecules in the T-region, requiring an evaluation of 2^{10} (1 024) relevant QM/MM energy expressions at each time step. At $M = 15$, the number of distinct energy evaluations needed increases to 32 768. Considering that standard MD simulations require several hundred thousand time steps, it is clear that computations such as these will not be feasible.

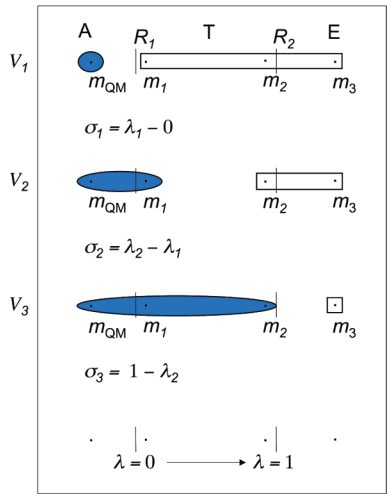
The cause of the unworkable exponential scaling of the intuitive concept described above lies in the inclusion of

Scheme 1. Schematic Representation of the $2^N = 8$ Energy Terms V_a (eq 2) for a System with $N = 3$ Adaptive Molecules[a]



[a] The QM parts of a computation are depicted as blue ellipses, and the MM parts are depicted as white squares. The molecule m_{QM} represents the QM-center and is always computed with a QM method. The terms on the right (red) should have a weight of 0 and will not play a role in the description of the system. Since there are $M = 2$ subsystems in the T-region, $2^M = 4$ terms remain.

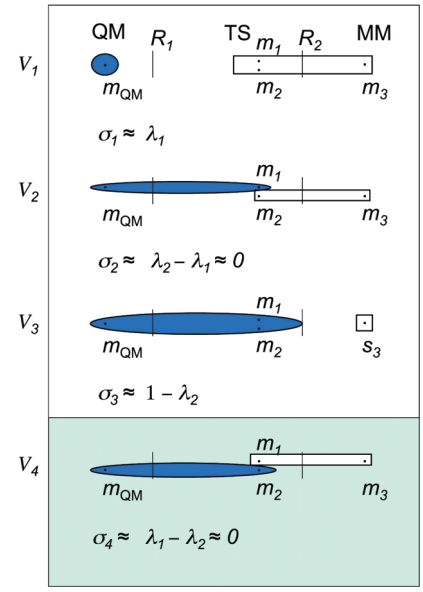
Scheme 2. Schematic Representation of the Contributing Energy Terms V_a for the Difference-Based Adaptive Solvation Method in a System with $N = 3$ Adaptive Molecules



undesirable partitionings that include "almost MM" molecules in the QM description, while "almost QM" molecules are described MM. Ideal σ_a functions should give such energy terms a weight of 0 and focus on the contributions of more "reasonable" partitionings.

In this paper we present a novel potential, which orders the energy terms similarly to the potential by Heyden et al.[12] (Section III) but assigns their weights in a different manner. Both approaches drastically reduce the number of non-zero terms in eq 2 while adhering to a normalization of the weight functions $\sigma_a(\mathbf{r})$. We achieve these goals by first imposing the constraint that energy terms V_a obtain a weight $\sigma_a(\mathbf{r})$ of 0 if there exists a λ in the set $\{\lambda\}_a^{QM}$ that is larger than the smallest λ in the set $\{\lambda\}_a^{MM}$. This amounts to an ordering of the subsystems in the T-region, such that the only contributing partitionings are the following: the term with only the A-region molecules in G^{QM} (V_1 in Scheme 1), the A-region molecules plus the closest T-region molecule in G^{QM} (V_2 in Scheme 1), the A-region molecules plus the two closest T-region molecules in G^{QM} (V_3 in Scheme 1), etc., up to the term in which the entire A- and T-regions are included in G^{QM} (Scheme 2). The exponential scaling of the number of non-zero energy terms is thereby reduced to a linear scaling that is feasible even for systems with large solute molecules. The sorting of energy terms implies that special care needs to be taken in the construction of the weight functions to guarantee the continuity of the potential at instances when two molecules are at comparable distances from the QM center. Suppose we have molecules m_i and m_j with λ_i slightly smaller than λ_j changing their position relative to the QM-center such that λ_i becomes larger than λ_j. At the crossing point of the two λ values, all energy terms that contain m_i in G^{QM} and m_j in G^{MM} cease to contribute, while all previously neglected terms that include m_j in G^{QM} and m_i in G^{MM} become relevant. This could introduce a sudden change that reinvokes the problems encountered in simulations performed without a transition region. We, therefore, need an additional condition: close to degeneracy of the λ values of two molecules, the weights (and the first derivative of the weight

Scheme 3. Schematic Representation of the Contributing Energy Terms V_a for the Difference-Based Adaptive Solvation Method in a System with $N = 3$ Adaptive Molecules[a]



[a] The term V_4 (green) only has a non-zero weight because in the depicted snapshot m_1 and m_2 are at similar distances from the QM-center.

with respect to the molecular coordinates) of the energy terms that assign these molecules to different groups should be zero.

These considerations have led us to define the weight functions σ_a in eq 2 in terms of differences between two individual λ values. These two λ values are the minimum and maximum value for G_a^{MM} and G_a^{QM}, respectively:

$$\sigma_a = \begin{cases} 0 & \text{if } \max(\{\lambda\}_a^{QM}) > \min(\{\lambda\}_a^{MM}) \\ \min(\{\lambda\}_a^{MM}) - \max(\{\lambda\}_a^{QM}) & \text{if } \max(\{\lambda\}_a^{QM}) \leq \min(\{\lambda\}_a^{MM}) \end{cases} \quad (3)$$

Or, completely equivalently, as

$$\sigma_a = \max(\min(\{\lambda\}_a^{MM}) - \max(\{\lambda\}_a^{QM}), 0) \quad (4)$$

For a system with three adaptive molecules, the non-zero σ values take the simple form that is depicted in Scheme 2.

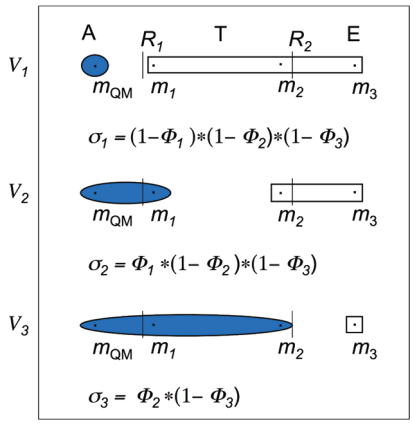
With this simple definition, the weight functions σ_a obey nearly all of the requirements for a smooth and normalized potential.

1) When m_i leaves the T-region at R_1 ($\lambda_i \rightarrow 0$), the sum of weights for the energy terms that include that molecule in G^{QM} equals 1. All weights for the energy terms that contain m_i in G^{MM} become 0. Furthermore, all first derivatives of these weight functions vanish: $\partial\sigma(\mathbf{r})/\partial r_i = 0$.

2) When m_i leaves the T-region at R_2 ($\lambda_i \rightarrow 1$), the sum of weights for the QM/MM terms that contain that molecule in G^{MM} equals 1. All the terms that include the molecule in G^{QM} equal 0. Again, the first derivatives vanish.

3) If m_i exchanges its relative position to the center with m_j, then the energy term V_a with weight $\sigma_a = \lambda_j - \lambda_i$ become

Scheme 4. Schematic Representation of the $M + 1$ Energy Terms V_a with Non-Zero Weights Present in the Total Expression for the Sorted Adaptive Partitioning Potential for a System with M Molecules in the T-Region $(1 - \Phi_3) = 1$



0, and a new energy term V_b with weight $\sigma_b = \lambda_i - \lambda_j$ starts to contribute. Both of these terms equal 0 when $\lambda_i = \lambda_j$, so that the reordering does not cause a discontinuity in the potential. A discontinuity does show up in the derivative of the potential (Figure 2) as a result of the discrete nature of the min()/max() operation.

Because the remaining small discontinuity could affect the dynamics, we investigated a substitution of the discrete min()/max() functions of eq 4 by smooth approximate min()/max() functions. A possible choice for the max() function is the logarithm of a sum of large exponential functions:[13]

$$\max(x_1, ..., x_N) = \frac{1}{\kappa} \ln\left(\sum_{i=1}^{N} e^{\kappa x_i}\right) \tag{5}$$

which can be combined with a similar choice for the min() function (utilizing the fact that $\{\lambda\}_j$ contains only values between 0 and 1):

$$\min(x_1, ..., x_N) = 1 - \frac{1}{\kappa} \ln\left(\sum_{i=1}^{N} e^{\kappa(1-x_i)}\right) \tag{6}$$

The free parameter κ is to be chosen large enough to make small terms in the sum of exponents negligible. These approximate functions smoothen the discontinuity visible in Figure 1 and only give a significant deviation from the true minimum or maximum when two values in the set are similar.

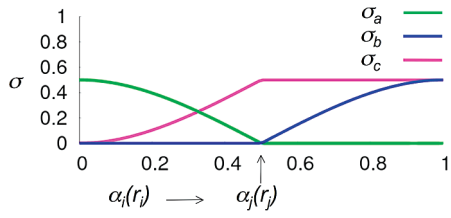


Figure 2. Behavior of the σ functions for a system with only two adaptive subsystems, when m_i moves through a T-region with molecule m_j fixed half-way in the T-region: $\alpha_j(r_j) = 0.5$ with $\alpha_j(r_j) = (r_j - R_1) / (R_2 - R_1)$. σ_a, σ_b, and σ_c correspond to σ_2, σ_4, and σ_1, respectively, in Scheme 3.

A consequence of the introduction of the smooth minimum and maximum functions is a slight increase of the number of partitionings with non-zero weights at instances when two adaptive molecules are at the same distance from the QM center. In this work, we compute energies V_a for which the weight $\sigma_a(\mathbf{r})$ (or its derivatives) exceeds a certain threshold. At most time steps this results in the computation of the minimum of $M + 1$ QM/MM energies (M molecules in the T-region).

While there are to our knowledge currently no faster methods available for doing adaptive QM/MM with fully continuous energies and forces, simulations are still approximately M times slower than a fixed partioning QM/MM simulation. In the target QM/MM simulations, the value M should, thus, be kept small. This is possible because most of the systems of practical interest involve a large active (QM) molecule that contains only a relatively small region where interaction with the surrounding solvent needs to be treated quantum mechanically. Examples are the free coordination space around a transition metal contained in a larger scaffold or the small hydrophilic region in a large hydrophobic (bio)molecule. In these cases, the A-regions may be defined as consisting of one or two atoms on the edge of the active molecule, so that the surrounding T-region will be small enough to keep M manageable. In such cases, the active molecule will have close contacts with MM molecules in the E-region, but this does not affect the scaling. Other ways to keep the costs down will be the development of intelligent restart schemes in which the QM parts of different QM/MM partionings will share their electronic structure information.

III. Sorted Adaptive Partitioning

The difference-based adaptive solvation (DAS) potential is related to the sorted adaptive partitioning potential of Heyden and co-workers.[12] In this sorting approach the same three requirements, as introduced in the previous section, are considered, and for M molecules in the T-region, the total energy is analogously defined as a weighted sum over $M + 1$ terms. The method differs from our setup primarily in the definition of the weights σ_a. Heyden et al. introduce a recursion relation that defines the adaptive potential as a normalized sum of the highest level QM/MM term (V_a, all molecules QM) and another adaptive potential, which excludes the molecule with the least QM character (m_i) from QM description (eq 7). The latter adaptive potential obviously contains terms for all the partitionings of the remaining pool of T-region molecules but can also be written recursively as a weighted average of only two terms, and so on. In contrast to the DAS setup, the expression in eq 7 leads to the definition of the σ from eq 2 as a *product* of the recursive weights Φ_a:

$$V^{\mathrm{ad}}(\mathbf{r}) = V^{\mathrm{ad}(M)}(\mathbf{r}) = \Phi_a(\mathbf{r})V_a(\mathbf{r}) + (1 - \Phi_a(\mathbf{r}))V^{\mathrm{ad}(M-1)}(\mathbf{r}) \tag{7}$$

$$\sigma_a(\mathbf{r}) = \begin{cases} 0 & \text{if } \max(\{\lambda\}_a^{\mathrm{QM}}) > \min(\{\lambda\}_a^{\mathrm{MM}}) \\ \Phi_a(\mathbf{r}) \prod_b (1 - \Phi_b(\mathrm{r})) & \\ \quad b(\max(\{\lambda_b^{\mathrm{QM}}\}) > \max(\{\lambda_a^{\mathrm{QM}}\})) & \text{if } \max(\{\lambda\}_a^{\mathrm{QM}}) \leq \min(\{\lambda\}_a^{\mathrm{MM}}) \end{cases} \quad (8)$$

The auxiliary functions Φ_a can be viewed as progress functions of m_i if this is the solvent molecule in G_a^{QM} with the largest λ value. With $\lambda_i = \max(\{\lambda\}_a^{\mathrm{QM}})$:

$$\Phi_a = \begin{cases} 0 & \text{if } \lambda_i = 1 \\ 1 & \text{if } \lambda_i = 0 \\ (1 + \chi_i)^{-3} & \text{if } 0 < \lambda_i < 1 \end{cases} \quad (9)$$

$$\chi_i = \sum_{j=1}^{i-1} \frac{\lambda_j}{\lambda_i - \lambda_j} + \frac{\lambda_i}{1 - \lambda_i} + \sum_{j=i+1}^{N} \frac{1 - \lambda_j}{\lambda_j - \lambda_i} \lambda_i \quad (10)$$

We note that the smoothing function λ is chosen as a fifth-order spline in the work by Heyden.

Like in DAS, described above, the sorted adaptive partitioning (SAP) potential employs constraints to ensure that "QM character" vanishes if a molecule m_j enters the E-region (eq 9, first line) and the "MM character" vanishes if the molecule enters the A-region. The definition of the functions χ_j (eq 10) ensures the correct behavior at R_1 and R_2. It also ensures vanishing of the contribution of partitionings, assigning molecules at similar distances from the QM-center to different groups.

With this choice of weights, like for the DAS potential, the potential energy V^{ad} is a continuous function of all coordinates. The forces on all atoms are the derivative of this one potential, and the total energy is conserved. As before, the consequence of a solvent molecule m_i passing m_j, while crossing the T-region toward the E-region, is that the weight σ_a of the energy term V_a that treats m_i QM and m_j MM vanishes. Unlike with the DAS potential, it is possible for these weights to have large values just before this situation occurs, leading to a strong variation of the weights in a relatively small region of coordinate space. The effect is especially pronounced in σ_b for the partitioning that treats both molecules MM (Figure 3). The resulting large derivative of the weight functions leads to strong forces on the atoms in the crossing region. A consequence is that commonly chosen time step values in molecular dynamics simulations may no longer be sufficient to conserve the total energy. In fact, as Heyden et al. have shown in an example simulation, the total energy may indeed not be conserved in practice.

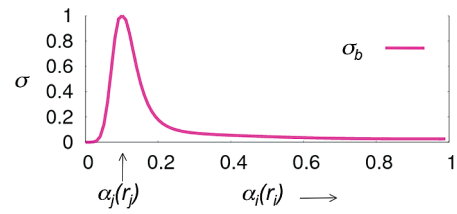


Figure 3. Behavior of a steep σ function for a system with only two adaptive molecules when m_i moves through a T-region with only m_j in it located at $\alpha_j(r_j) = 0.1$ with $\alpha_j(r_j) = (r_j - R_1) / (R_2 - R_1)$. The corresponding partitioning treats both m_i and m_j classically.

IV. Forces in the T-region

In the above, we already mentioned the possibilities for undesirable influences of the chosen weight factors on the computed forces. We did not yet address any problems that can occur when the energy surfaces of the different partitionings differ largely. In order to do so we first consider in more detail the calculation of the forces in simulations with an adaptive potential:

$$V^{\mathrm{ad}}(\mathbf{r}) = \sum_a^{2^N} \sigma_a(\mathbf{r}) V_a(\mathbf{r}) \quad (2)$$

$$F_i = -\frac{\partial V^{\mathrm{ad}}}{\partial r_i} = -\sum_a^{2^N} \sigma_a(\mathbf{r}) \frac{\partial V_a(\mathbf{r})}{\partial r_i} + \frac{\partial \sigma_a(\mathbf{r})}{\partial r_i} V_a(\mathbf{r}) \quad (11)$$

The first term in the force of eq 11 contains simply the weighted force from each contributing partitioning. The second term depends on the derivative of the weight function $\sigma_a(\boldsymbol{r})$ multiplied by the value of the partitioning energy. This introduces a dependence on the relative energies of the different partitionings that is particular to adaptive methods (in simulations with one fixed partitioning force depends only on the slope of the one energy surface). As a consequence, the system may minimize its energy by evolving toward a geometry that maximizes the weight of the partitioning with the lowest absolute energy. This undesirable side effect of adaptive partitioning may be prevented to some extent by aligning the contributing potential energy surfaces as closely as possible. Such a procedure is, however, far from automatic and likely to depend on the particular geometry chosen to define the shifts.

To avoid such errors, it is of interest to consider a deviation from the Hamiltonian approach, in a scheme that interpolates the forces instead of the potential, by using only the first term in eq 11. In contrast to earlier force based methods,[8,9] the weighting functions from the DAS potential guarantee fully continuous forces. This means that many advantages of Hamiltonian approaches can be retained. Integration of the forces over the simulation trajectory should yield a quantity that is conserved throughout the simulation. A so-called bookkeeping term, as used in multiscale atomistic/coarse grain simulations,[10] keeps track of the gain or loss of potential energy throughout the simulation. The conserved quantity is the total energy (using V^{ad} from eq. 2) corrected by this bookkeeping term, and it can be used as a tool to tune other simulation parameters, such as time step and cutoff values.

In the following, we will derive the form of the bookkeeping term mentioned above. In other words, we need to define the corrected potential energy along the path that is associated with the interpolated forces in the first term of eq 11. We can formally add a correction term W^{bk} to eq 2 (eq

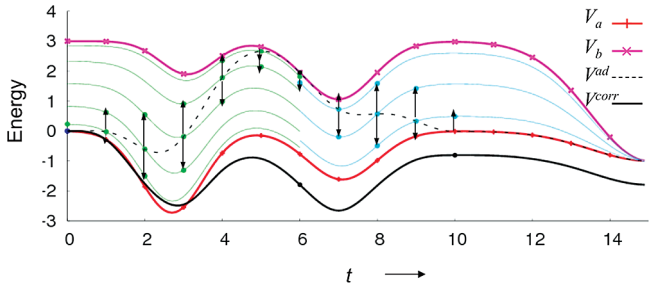
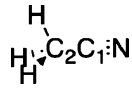


Figure 4. Potential energies along a trajectory. The trajectory follows a loop, with $\mathbf{r}(0) = \mathbf{r}(10)$. V^{ad} is depicted as the thin dashed line. The thick solid line is the potential corrected with the bookkeeping energy contribution.

Table 1. Original and Adjusted Atomic Point Charges in Acetonitrile Force Field



	q original	q modified
N	−0.532	−0.932
C_1	0.481	0.881
C_2	−0.479	−0.479
$(H)_3$	0.177	0.177

12) with a partial derivative that exactly cancels the second term in the forces in eq 11.

$$V^{corr}(t) = V^{ad}(r) + W^{bk}(t) \tag{12}$$

$$F_i^{bk}(\mathbf{r}(t)) = -\frac{\partial W^{bk}(t)}{\partial r_i(t)} = \sum_a^{2^N} \frac{\partial \sigma_a(\mathbf{r})}{\partial r_i} V_a(\mathbf{r}) \tag{13}$$

As an example, we will consider the averaging over two potential energies V_a and V_b, as depicted in Figure 4 as a function of a trajectory with geometry $\mathbf{r}(t)$. We constructed this trajectory such that it forms a loop bringing the system back to its starting coordinates at $t = 10$: ($\mathbf{r}(10) = \mathbf{r}(0)$). The two energy curves have been aligned at an arbitrary point ($t = 15$). The thin dashed line in the figure represents the potential energy V^{ad} as defined in eq 2. Initially the system is in a configuration such that most of the weight is given to V_a, but after a while V_b starts to dominate before the system returns to its initial configuration with V_a again dominant. It is clear from the figure that the shape of the weighted energy curve V^{ad} is mainly determined by the difference in energies of the individual potential energy curves, bearing little resemblance to the shapes of the curves that we aim to average over. With the force-based approach, one utilizes the information gathered during the simulation to define a

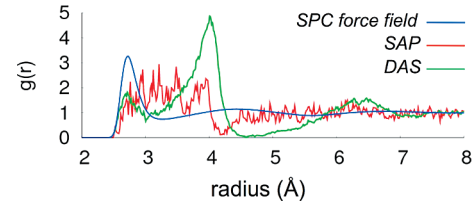


Figure 6. H_2O-OH_2 radial distribution functions around the 'QM'-center for the various methods.

good approximation for the W^{bk} function introduced above. The corrected potential energy curve is depicted in Figure 4 as the black line, and it clearly reflects the shape of the constituent potential energy curves V_a and V_b.

After n time steps along the trajectory, the bookkeeping term W^{bk} can be expressed as the path integral over the force vector from eq 13, as defined in eq 14.

$$\begin{aligned} -W^{bk}(\tau) &= \int_{t=0}^{\tau} F^{bk}(t)\left(\frac{d\mathbf{r}}{dt}\right)dt \\ &= \int_{t=0}^{\tau} \sum_a V_a(t)\left(\vec{\nabla}\sigma_a(t)\left(\frac{d\mathbf{r}}{dt}\right)\right)dt \\ &= \sum_{i=0}^{n} \sum_a V_a(t_i)\left(\frac{\sigma_a(t_{i+1}) - \sigma_a(t_{i-1})}{2\Delta t}\right)\Delta t \\ &= \sum_{i=0}^{n} \sum_a V_a(t_i)\Delta\sigma_a(t_i) \end{aligned} \tag{14}$$

In its discrete form, the bookkeeping term clearly reflects the sum of the accumulated changes in energy at each step along the path. In this manner it corrects for the difference in energy between V_a and V_b, while retaining the different slopes of each of these curves. The value of the correction term is only defined along the path taken by the system during the simulation, and there is no guarantee that the correction term integrates to zero upon revisiting a previous geometry. This feature is also illustrated in Figure 4: at $t = 10$, the geometry is the same as in $t = 0$, but the corrected potential V^{corr} is not. This is due to the fact that the system returns from its largest displacement at $t = 6$ via a route that differs from the initial trajectory. If the system had simply traveled backward via the exact same route, then the corrected energies would have been identical. This reflects how the corrected potential energy V^{corr} depends on the path that is followed and not only on the spatial coordinates of the system.[14] The absence of a single-valued continuous potential in N-dimensional space may result in a small drift of the temperature. However, if this drift is sufficiently small, the simulations will result in a better representation of the system than any other adaptive method.

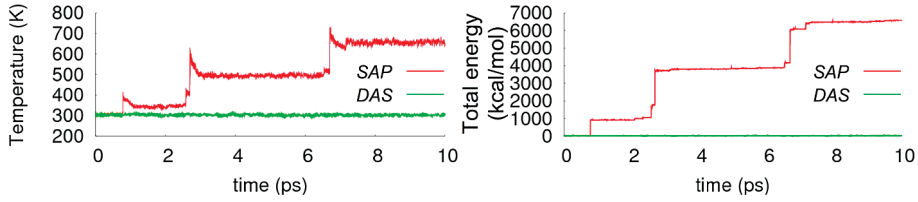


Figure 5. Temperature and total energy during simulation with SAP and DAS potential.

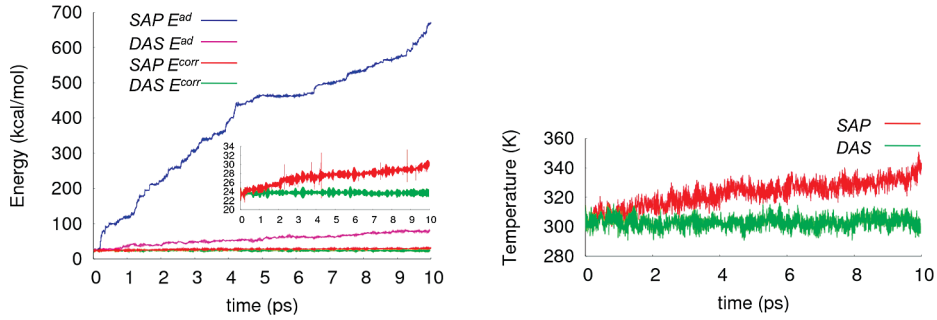


Figure 7. Uncorrected (E^{ad}) and corrected total energies (E^{corr}) obtained with the DAS and SAP interpolated forces and the bookkeeping correction.

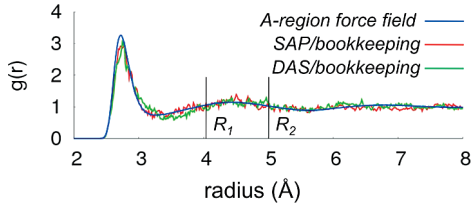


Figure 8. H_2O-OH_2 radial distribution functions obtained with the DAS and SAP interpolated forces.

Table 2. Drift in the Corrected Total Energy for DAS and SAP Force Simulations with Varying Time Steps

	Energy drift in kcal/mol·ps	
ts (fs)	DAS	SAP
0.5	0.017	0.648
0.2	0.003	0.030
0.1	0.022	0.023

V. Results

As an illustrative case study we present NVE simulations for two simple systems, namely water in water and acetonitrile in water. Both systems, containing approximately 3000 water molecules, and around seven water molecules in the active region, can be viewed as representative for target adaptive QM/MM simulations. We used two different classical force fields to represent the two levels of description to provide a quick model for the target QM/MM setup. This also provides us with the possibility to assess the quality of the simulation by comparing the results with the full 'QM' alternative, which would be unfeasible if we had chosen a true QM/MM setup. The NAMD classical molecular dynamics package[15] was used for the energy and force evaluation and the time evolution of the system. For the latter, the program uses a leapfrog algorithm. The composite energies and forces, as described in the previous section, were obtained with a python wrapper script (PyMD) developed for this purpose.

For the MM/MM study of water in water, we started with a box, 30 Å in diameter, containing water that was equilibrated to room temperature and pressure of 400 ps with a time step of 0.5 fs and a flexible TIP3P(Fs) force field.[16] We then evolved the system for 10 ps with a time step of 0.5 fs with both the DAS and SAP potentials in the microcanonical ensemble (NVE). The water in the environment region is described with the same flexible TIP3P(Fs) force field as before, while the water in the active region is

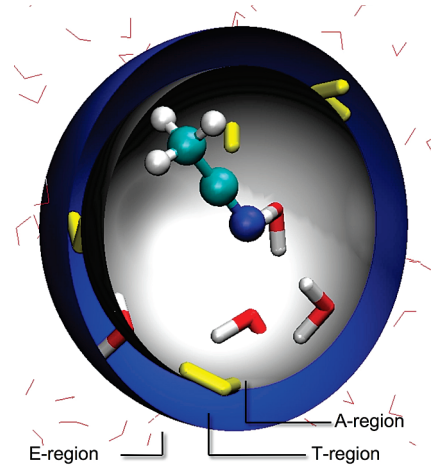


Figure 9. Division of the acetonitrile in water system for adaptive QM/MM.

described with the flexible force field SPC(Fw).[17] The A-region is a sphere with a radius of 4 Å around the oxygen atom of a central water molecule, and the T-region is a shell of 1 Å around the A-region. The 'QM' energy was shifted by the TIP3P(Fs)−SPC(Fw) energy difference for the gas-phase TIP3P(Fs)-optimized solvent molecules, augmented by an additional two-body interaction energy difference, obtained again with the TIP3P(Fs)-optimized coordinates. The aim of the study of acetonitrile in water is to combine an accurate force field in the A-region with a force field in the E-region that is known to fail at a short distance from the solute. As the accurate force field, we chose once again a SPC(Fw) description of water with the AMBER six-center force field for acetonitrile.[18] As the second force field, we chose the TIP3P(Fs) water, in combination with the same AMBER force field for acetonitrile, but this time with a different choice for the atomic charges. We increased the dipole moment of acetonitrile to make the interaction with the solvent artificially bad at short distances. The new partial charges of the acetonitrile atoms are presented in Table 1. The acetonitrile molecule was placed in a water box with a diameter of 30 Å and equilibrated to room temperature and pressure of 400 ps with the flexible TIP3P(Fs) force field. The system was then evolved for 10 ps with a time step of 0.5 fs, using the DAS interpolated forces, with an A-region of 4 Å around the nitrogen atom and a T-region of 1 Å wide.

Water in Water. In order to study the effect of the spurious forces, as discussed in Section IV, we first evolved the system using the full forces from the potential energy

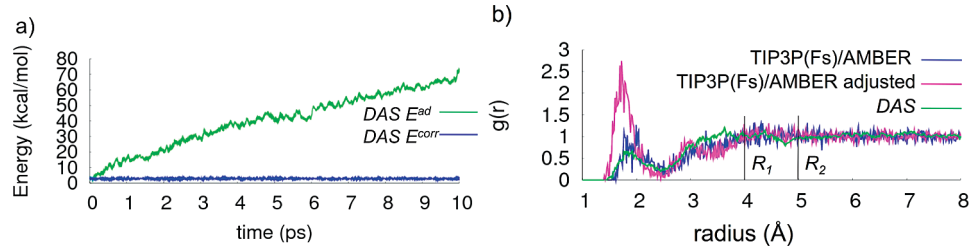


Figure 10. (a) Uncorrected and corrected total energy obtained with the DAS forces. (b) N−H_1 radial distribution functions for a working force field (blue line), a force field that fails at short distances (pink line), and for one obtained with an adaptive mixture of the two in the DAS formulation (green line).

V^{ad}. The adaptive MM/MM simulations with the SAP potential resulted in severe heating of the system in the NVE ensemble. The temperature changes occurred in separate events, corresponding to situations where several water molecules had the same very small λ values. As depicted in Figure 2, this is exactly when very large spurious forces on the atoms can be expected. Simulations with the DAS potential under the exact same circumstances conserved both the temperature and the total energy throughout the run. Figure 5 depicts the behavior of the temperature and total energy throughout the trajectory with both methods.

As expected, the radial distribution functions $g(r)$, obtained from the simulation with the adaptive methods, are not the desired mixture of $g(r)$ from the two force fields used, despite the energy shift that was applied to obtain a reasonable alignment of the potential energy surfaces. We extracted the $g(r)$ from the SAP simulation from the first 1 ps of the simulation where the system still has a temperature of 300 K. Figure 6 shows us how dramatic the error that is introduced for the $g(r)$ between the central water oxygen and all surrounding oxygen atoms. For both methods, the solvent molecules are pushed toward the A-region to maximize the contribution of the partition that describes most molecules with the energetically favorable QM potential. The effect is slightly more pronounced for the DAS result than for that of the SAP result.

We, thus, performed additional simulations using the force-based approach with the weight functions σ_a from the SAP and the DAS methods, respectively, and the bookkeeping correction to the energy, leading to the total energies depicted in Figure 8. The uncorrected total energy obtained with SAP forces displays a strong drift. This is due to the fact that the SAP interpolation is more rapid than that of the DAS approach. The corrected total energy E^{corr} is conserved well only with the DAS forces. Another way of stating this would be to say that, with the SAP forces, the time steps taken in the simulation are too large to give a good approximation of the integral in eq 14. In addition, the DAS approach fully conserves the temperature all throughout the simulation (Figure 7).

Figure 8 shows how the $g(r)$ for both methods are now in perfect agreement with the expected density distribution of water from the combined force fields. With M the number of solvent molecules in the T-region, on average 1.5 times $M + 1$ partitions needs to be computed with the DAS approach, which is an acceptable overhead compared to the number (M + 1) computed in a SAP simulation.

In order to estimate the time step required for a satisfactory performance of the SAP force interpolation, additional short simulations (1 ps) with smaller time steps were performed (Table 2). In all simulations, the bookkeeping correction was applied to the depicted energy. When a time step of 0.2 fs is used, the drift in the bookkeeping quantity with the SAP forces is similarly small to that obtained with a DAS force simulation at a much larger time step.

Acetonitrile in Water. We performed a MM/MM simulation of acetonitrile in water with the DAS forces (Figure 9), applying the bookkeeping correction. Figure 10 depicts the uncorrected and the corrected total energy during the simulation. The figure also shows how the $g(r)$ for (correctly described) acetonitrile in water is reproduced by the adaptive simulation. As expected, the adjusted force field overestimates the number of N−H hydrogen bonds at short distances. This deviation is entirely corrected for in the adaptive formulation.

VI. Conclusions

A new adaptive potential was presented that allows QM/MM molecular dynamics simulations that change the description of solvent molecules on the fly. The scheme defines a potential energy that is smooth enough to allow for a conventionally large time step during the simulation. An interpolated force scheme with fully continuous forces is also introduced, providing correct radial distribution functions and conserving a total quantity along the trajectory. In particular, for large systems of practical interest, these schemes now allow efficient and accurate investigation of processes that involve the dynamical influence of weakly bound solvent molecules.

Acknowledgment. The authors thank The Netherlands Organization for Scientific Research (NWO) for financial support.

References

(1) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation: from Algorithms to Applications*; Academic Press: San Diego, 2002; pp 63−105.

(2) (a) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249. (b) Thole, B. T.; Van Duijnen, P. Th.;. *Theor. Chim. Acta* **1980**, *55*, 307–318. (c) Field, M. J.; Bash, P. A. *J. Comput. Chem.* **1990**, *11*, 700–733. (d) Gao, J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VHC: New York, 1995; Vol 7, pp 119−185; (e) Sherwood, P. In *Modern*

Methods and Algorithms of Quantum Computing; Grotendorst, J. Ed.; John von Neumann Institute for Computing: Jülich, Germany, 2000; pp 257−277; Carloni, P.; (f) Rothlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455. (g) Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2008**, *4*, 2067–2084. (h) Magistrato, A.; DeGrado, W. F.; Laio, A.; Rothlisberger, U.; VandeVondele, J.; Klein, M. L. *J. Phys. Chem. B* **2003**, *107*, 4182–4188.

(3) Moon, S.; Case, D. A. *J. Comput. Chem.* **2006**, *27* (7), 825–836.

(4) (a) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050. (b) Wesolowski, T. A. In *Computational Chemistry: Reviews of Current Trends*; Leszcynski, J., Ed.; World Scientific: Singapore, 2006; Vol*10*, pp 1−82.

(5) (a) Jacob, C. R.; Neugebauer, J.; Jensen, L.; Visscher, L. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2349−2359; (b) Neugebauer, J.; Louwerse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115; (c) Neugebauer, J.; Jacob, C. R.; Wesolowski, T. A.; Baerends, E. J. *J. Phys. Chem. A* **2005**, *109*, 7805–7814. (d) Neugebauer, J. *J. Chem. Phys.* **2007**, *126*, 134116.

(6) Kerdcharoen, T.; Morokuma, K. *Chem. Phys. Lett.* **2002**, *355*, 257–262.

(7) (a) Kerdcharoen, T.; Liedl, K. R.; Rode, B. M. *Chem. Phys.* **1996**, *211*, 313–323. (b) Hofer, T. S.; Pribil, A. B.; Randolf, B. R.; Rode, B. M. *J. Am. Chem. Soc.* **2005**, *127*, 14231–14238. (c) Schwenk, C. F.; Loeffler, H. H.; Rode, B. M. *J. Am. Chem. Soc.* **2003**, *125*, 1618–1624.

(8) Csányi, G.; Albaret, T.; Payne, M. C.; De Vita, A. *Phys. Rev. Lett.* **2004**, *93* (17), 175503.

(9) (a) Rode, B. M.; Schwenk, C. F.; Tongraar, A. *J. Mol. Liq.* **2004**, *110*, 105–122. (b) Rode, B. M.; Hofer, T. S. *Pure Appl. Chem.* **2006**, *78*, 525–539. (c) Rode, B. M.; Hofer, T. S.; Randolf, B. R.; Schwenk, C. F.; Xenides, D.; Vchirawongkwin, V. *Theor. Chim. Acc.* **2006**, *115*, 77–85.

(10) (a) Ensing, B.; Nielsen, S. O.; Moore, P. B.; Klein, M. L.; Parrinello, M. *J. Chem. Theory Comput.* **2007**, *3*, 1100–1105. (b) Praprotnik, M.; Delle Site, L.; Kremer, K. *J. Chem. Phys.* **2005**, *123*, 224106. (c) Praprotnik, M.; Delle Site, L.; Kremer, K. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2006**, *73*, 066701.

(11) (a) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. *Phys. Rev. Lett.* **2005**, *95*, 218102. (b) Neri, M.; Anselmi, C.; Carnevale, V.; Vargiu, A. V.; Carloni, P. *J. Phys.: Condens. Matter* **2006**, *18*, S347.

(12) Heyden, A.; Lin, H.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 2231–2241.

(13) (a) Donadio, D.; Raiteri, P.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109* (12), 5421–5424. (b) Bulo, R. E.; Donadio, D.; Laio, A.; Molnar, F.; Rieger, J.; Parrinello, M. *Macromolecules* **2007**, *40* (9), 3437–3442.

(14) Delle Site, L. *Phys. Rev. E* **2007**, *76*, 047701.

(15) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.

(16) Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 9361–9381.

(17) Wu, Y.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*, 024503.

(18) Grabuledam, X.; Jaime, C.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (10), 901–908.

CT900148E

JCTC Journal of Chemical Theory and Computation

Heavy Halogen Atom Effect on ^{13}C NMR Chemical Shifts in Monohalo Derivatives of Cyclohexane and Pyran. Experimental and Theoretical Study

Alvaro Cunha Neto,[†] Lucas C. Ducati,[†] Roberto Rittner,*[,†] Cláudio F. Tormena,[†] Rubén H. Contreras,[‡] and Gernot Frenking[§]

Chemistry Institute, State University of Campinas, Caixa Postal 6154, 13084-971 Campinas, SP, Brazil, Department of Physics, FCEyN, University of Buenos Aires and CONICET, Buenos Aires, Argentina, and Philipps-Universität Marburg, Hans-Meerwein-Strasse, D-35032 Marburg, Germany

Received December 1, 2008

Abstract: As a first step, a qualitative analysis of the spin−orbit operator was performed to predict the kind of organic compounds, where it could be expected that the SO/FC (spin−orbit/Fermi contact) and SO/SD (spin−orbit/spin dipolar) yield unusually small contributions to the "heavy atom effect" on ^{13}C SCSs (substituent chemical shifts). This analysis led to the conclusion that compounds presenting strong hyperconjugative interactions involving the σ^*_{C-X} orbital (X = halogen) are good examples where such effects can be expected to take place. On the basis of such results, the following set of model compounds was chosen: 2-*eq*-halocyclohexane (**2-*eq***), 2-*ax*-halocyclohexane (**2-*ax***), and 2-*ax*-halopyran (**3**), to measure ^{13}C SCSs. Such experimental values, as well as those of methane and halomethanes taken from the literature, were compared to calculated values at a nonrelativistic approach using B3LYP, and at a relativistic approach with BP86 using scalar ZORA, spin−orbit ZORA, scalar PAULI, and spin−orbit PAULI. Results from relativistic calculations are in agreement with the trends predicted by the qualitative model discussed in this work.

I. Introduction

The heavy halogen atom effect on carbon chemical shifts has now been known for about three decades.[1−7] In the past decade, many articles appeared[8−20] where calculations of this effect are reported using different levels of approximation. The interesting work by Kaupp et al.[21] can be distinguished from the others, because it provides important insights on how the cross term due to SO and FC interactions (SO/FC) is transmitted through the molecule. It was concluded that its propagation is closely analogous to the well-established mechanisms for the transmission of the FC term in indirect spin−spin coupling constants (SSCCs). Although the results reported by Kaupp et al.[21] substantiate this interpretation, it is suggested here that when intending to generalize such results, the role played by the SO operator in defining the SO/FC interaction is somewhat overlooked.

It is easy to obtain a qualitative pictorial representation of some factors affecting the contribution of the SO operator to the heavy atom effect on the ^{13}C substituent chemical shifts (SCS), for C_1 bonded to a halogen atom, σ_{C1-X} (X = Cl, Br, I). Thus, in the next section, a qualitative description of interactions affecting the SO part of the SO/FC and SO/SD terms is given. Such analysis provides an intuitive base to detect the kind of compounds where the "the heavy atom effect" on the $^{13}C_1$ chemical shift could be notably influenced by intramolecular interactions affecting the SO part of both contributions.

On the basis of the discussion presented in section II, it was possible to predict some features, which must be exhibited by some compounds, to show the effects of the heavy halogen atom on $^{13}C_1$ SCSs important enough to be

* Corresponding author e-mail: rittner@iqm.unicamp.br.
[†] State University of Campinas.
[‡] University of Buenos Aires and CONICET.
[§] Philipps-Universität Marburg.

10.1021/ct800520w CCC: $40.75
Published on Web 07/30/2009

amenable to measurement. According to these ideas, the following compounds were chosen for the experimental determination of their $^{13}C_1$ SCSs: *eq- and ax*-halo-cyclohexanes and 2-halo-tetrahydropyran; and the corresponding experimental $^{13}C_1$−SCSs values for the halo-methanes were taken from literature. The experimental results were compared to scalar ZORA, spin−orbit ZORA (SO-ZORA), scalar PAULI, and spin−orbit PAULI (SO-PAULI)[22−24] calculations to verify if the qualitative trends predicted by the approach presented in section II were well-found. For completeness sake, $^{13}C_1$ SCSs GIAO-DFT calculations (Gauge-including atomic orbitals[25] within the framework of nonrelativistic density functional theory) were also performed and compared to those obtained with the ZORA and PAULI methods, which incorporate the relativistic effects into the GIAO-DFT calculation of NMR shielding tensors.

II. Qualitative Analysis of Factors Affecting the SO Part of the SO/FC and SO/SD Terms

Electrons and nuclei are described separately within the Born−Oppenheimer approximation, and an interesting example of its validity is found in NMR spectroscopy. In fact, NMR spectra are obtained inducing transitions between nuclear spin states, which can be described by the Hamiltonian given in eq 1.

$$\hat{H} = -\frac{1}{2\pi}\sum_N \gamma_N \bar{\mathbf{I}}_N \cdot (\overleftrightarrow{1} - \overleftrightarrow{\sigma}_N) \cdot \bar{\mathbf{B}} + \sum_{N\neq M} \bar{\mathbf{I}}_N \cdot (\overleftrightarrow{J}_{NM} + \overleftrightarrow{D}_{NM}) \cdot \bar{\mathbf{I}}_N \quad (1)$$

In the first term of this Hamiltonian, the nuclear magnetic shielding tensor is bilinear both in magnetic moment as in the spectrometer static magnetic field. Therefore, the study of nonrelativistic nuclear magnetic shielding tensors from the electronic molecular part perturbation theory yields

$$\sigma^N_{\alpha\beta} = \left.\frac{\partial^2 E(\mu_N, \mathbf{B})}{\partial \mu_{N,\alpha} \partial B_\beta}\right|_{\mu_N=0;\mathbf{B}=0} \quad (2)$$

where $E(\mu_N,\mathbf{B})$ is the perturbed molecular energy and involves the magnetic field dependent momentum: $\pi = \mathbf{p} + \mathbf{A}$, where $\mathbf{A}$ includes the vector potential of the spectrometer static magnetic field, as well as the sum of those corresponding to the nuclear magnetic moments; N is the nucleus under consideration, whose magnetic moment is μ_N; and $\mathbf{B}$ is the spectrometer static magnetic field. Equation 1 gives the two different contributions, its diamagnetic and paramagnetic parts, $\sigma^N_{\alpha\beta} = \sigma^{N,p}_{\alpha\beta} + \sigma^{N,d}_{\alpha\beta}$.

To take into account the effect of a heavy atom, relativistic corrections are thought to be obtained from approximate solutions of the nonrelativistic Schrödinger equation;[26] that is, the Hamiltonian is taken as

$$H = H_{N-e} + H^{NR}_{KE} + H^{Darwin-1} + H^{SO} \quad (3)$$

where H_{N-e} is the nonrelativistic nucleus−electron attraction; H^{NR}_{KE} is the nonrelativistic kinetic energy term; $H^{Darwin-1}$ is the one-electron Darwin term; and $H^{SO} = H^{SO(1)} + H^{SO(2)}$ corresponds to the one- and two-electron contributions to the spin−orbit Hamiltonian, respectively. The spin−orbit contribution to nuclear magnetic shielding tensor arises from the Hamiltonian operator given in eq 4.

$$H^{SO} = \frac{e^2\hbar}{4m_e^2}\frac{\mu_0}{4\pi} g_e \left(\sum_N Z_N \sum_i \frac{\mathbf{s}_i \cdot \mathbf{l}_{iN}}{r^3_{iN}} - \sum_{i\neq j} \frac{(\mathbf{s}_i + 2\mathbf{s}_j)\cdot \mathbf{l}_{ij}}{r^3_{ij}} \right) \quad (4)$$

where $\mathbf{s}_i$ is the ith electron spin operator; Z_N is the charge of the Nth nucleus; g_e is the electron g-factor; r_{iN} is the distance from the i electron to the N nucleus, $\mathbf{r}_i - \mathbf{R}_N$; $\mathbf{l}_{iN} = (\mathbf{r}_i - \mathbf{R}_N) \times [-i\nabla_i + \mathbf{A}_0(\mathbf{r}_i)]$ is the ith electron angular momentum taken from nucleus N and $\mathbf{l}_{ij} = (\mathbf{r}_i - \mathbf{r}_j) \times [-i\nabla_i + \mathbf{A}_0(\mathbf{r}_i)]$ is its angular momentum with respect to the jth electron. $\mathbf{A}_0$ is the vector potential corresponding to the spectrometer static magnetic field.

The first term in square brackets in $\mathbf{l}_{iN} = (\mathbf{r}_i - \mathbf{R}_N) \times [-i\nabla_i + \mathbf{A}_0(\mathbf{r}_i)]$ is magnetic field independent and, within triple perturbation theory, connects the singlet ground state with triplet excited states, allowing interactions with both the Fermi contact, FC, eq 5, and the spin-dipolar, SD, eq 6, operators:

$$H^N_{FC} = \frac{4\pi}{3}\frac{e\hbar}{m_e}\frac{\mu_0}{4\pi} g_e \sum_i \delta(\mathbf{r}_{iN})\mu_N \cdot \mathbf{s}_i \quad (5)$$

$$H^N_{SD} = \frac{e\hbar}{m_e}\frac{\mu_0}{4\pi} g_e \mu_N \cdot \sum_i \frac{3\mathbf{r}_{iN}\mathbf{r}^T_{iN} - \mathbf{1}r^2_{iN}}{r^5_{iN}} \cdot \mathbf{s}_i \quad (6)$$

yielding third-order terms that undergo orbital interactions with the spectrometer static magnetic field $\mathbf{B}_o$ obtaining four different contributions; see eqs 4−6.

However, keeping in mind that, in this work, only a qualitative description of the heavy atom effect on ^{13}C nuclear magnetic shielding is sought, it can be suggested that only two of those four terms, that is, the one-electron contributions, are considered more important than the two-electron contributions. They will be labeled here as SO/FC and SO/SD contributions to the ^{13}C magnetic shielding constant. In the studied compounds (see below), it is expected that the SO/FC term should be more important than the SO/SD contribution.

Therefore, because this qualitative analysis will be applied mainly to the leading term, the present study is restricted to the SO/FC term. It is recalled that the influence of the FC interaction on the SO/FC term was clearly discussed by Kaupp et al.,[19] and no further comment is worth to be given here. This by no means suggests that in this work the FC contribution to the SO/FC term is being undervalued.

The above considerations indicate that the first step must be a close look at the one-electron part of the H^{SO} Hamiltonian of eq 4, as given in eq 7:

$$H^{SO} = \frac{e^2\hbar}{4m_e^2}\frac{\mu_0}{4\pi} g_e \left(\sum_N Z_N \sum_i \frac{\mathbf{s}_i \cdot \mathbf{l}_i}{r^3_{iN}} \right) \quad (7)$$

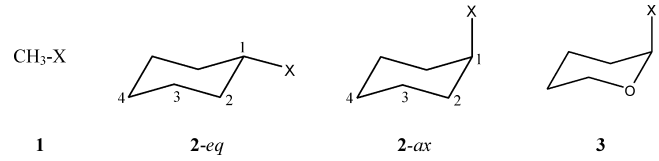
It will be assumed that the triple perturbation theory is applied to both occupied and vacant localized molecular orbitals, LMOs, which were localized through separate procedures.

It is also important to recall that the main aim of this work is to verify if hyperconjugative interactions from bonding or antibonding orbitals involving the carbon atom, whose substituent chemical shift (SCS) value is being analyzed, can affect the "heavy atom" effect. Qualitatively, it can be considered that the valence occupied LMOs behave like the NBO (natural bond orbitals) bonding and lone-pair orbitals, while the valence vacant LMOs behave like the NBO antibonding orbitals, as given by Weinhold et al.'s method.[27,28] To allow an easier way for the qualitative analysis of the SO operator, the gauge origin is taken at the site of the heavy nucleus, and in this way the $\mathbf{l}_i$ operator is, for the present purpose, essentially equal to the rotation operator centered at that nucleus. Therefore, the SO part of the SO/FC cross term must have a significant value whenever the following two conditions hold:

(a) The overlap of a 90° rotated occupied LMO, for example, the lone pairs $LP_2(X)$ or $LP_3(X)$, and the antibonding orbital corresponding to the $\sigma^*_{C_1-X}$ must be significant. A similar contribution originating in $LP_1(X)$, that is, the X lone-pair deepest in energy, is neglected on account of the following two reasons. (i) The rotated $LP_1(X)$ and the $\sigma^*_{C_1-X}$ antibonding orbital overlap to a much lesser extent than the ones involving the rotated $LP_2(X)$ and $LP_3(X)$ LMOs; (ii) the energy gap between $\sigma^*_{C_1-X}$ and $LP_1(X)$ is much larger than the energy gaps between the $\sigma^*_{C_1-X}$ antibonding orbital with either $LP_2(X)$ or $LP_3(X)$.

(b) The relevant energy gaps between the $\sigma^*_{C_1-X}$ antibonding orbital and $LP_2(X)$ and $LP_3(X)$ occupied LMO orbitals are not "very large".

Even though this last assertion cannot be precisely defined, an intuitive description of factors affecting the SO part of the SO/FC term can be sought. Point b indicates that, on the one hand, "heavy atom" effects on ^{13}C chemical shifts are more important for lone-pair bearing atoms like X = iodine than, for instance, X = tetra-coordinated tin atoms because bonding orbitals are much deeper in energy than orbitals representing lone pairs. It is important to recall that in a previous paper,[29] it was observed that in 1-I-bicyclo-[1.1.1]pentane the heavy atom effect on $^{13}C_1$ was estimated as ca. 43.4 ppm, while the analogous value for 1-$Sn(CH_3)_3$-bicyclo[1.1.1]pentane was estimated as ca. −10 ppm. It is important to note that the different sign can be rationalized as originated in the FC part of the SO/FC term [see ref 21] because the Sn magnetogyric ratio is negative for the two most abundant Sn isotopes. On the other hand, any hyperconjugative interaction that increases a relevant energy gap must decrease the corresponding SO part of the SO/FC and the SO/SD cross terms. The simple perturbed molecular orbital theory (PMO)[30] can be used to determine the type of hyperconjugative interactions that affect significantly the relevant energy gaps between the $\sigma^*_{C_1-X}$ antibonding orbitals and the $LP_2(X)$ and $LP_3(X)$ occupied orbitals. They are as follows: (i) any hyperconjugative interaction transferring charge into the $\sigma^*_{C_1-X}$ antibonding orbital, and (ii) hyperconjugative interactions like $LP_2(X) \rightarrow \sigma^*_{C_1-Y}$ and $LP_3(X) \rightarrow \sigma^*_{C_1-Y}$, where Y stands for any atom bonded to C_1 other than X. Interactions of type I push up the σ^*_{C1-X} orbital energy, while those of type ii push down the $LP_2(X)$ and $LP_3(X)$ orbital energies.



1 **2-*eq*** **2-*ax*** **3**

Figure 1. Structures of the studied compounds.

Moreover, it is recalled that the FC term of one-bond SSCCs might be affected by several factors like, for instance, the inductive effect of Y atoms bonded to C_1, the C_1 hybridization, hyperconjugative interactions involving either any bond attached to C_1 or the $\sigma^*_{C_1-X}$ antibonding orbital,[31] etc. This suggests that a competition can take place between the two opposite factors affecting the SO and FC parts of the SO/FC term. Apparently, a case in point is the "heavy atom" effect on C_1 SCSs in 1-X-bicyclo-[1.1.1]pentane (X = F, Cl, Br, I) previously reported.[29,32] Here, in fact, despite the strong hyperconjugative interactions into the $\sigma^*_{C_1-X}$ antibonding orbital, the heavy atom effect on C_1 SCSs is similar to those in halomethanes. It is experimentally known that $^1J_{C1,F}$ in 1-F-bicyclo[1.1.1]-pentane is notably larger, in absolute value, than in other less strained F-alkane derivatives.[33]

III. Selected Compounds To Verify the Qualitatively Predicted Trends in Section II

On the basis of the considerations presented in section II, four representative classes of halo-compounds were chosen to analyze how the "heavy atom" effect on ^{13}C SCSs is affected by changes originating in the SO operator [eq 1], X-methanes (**1**), *eq*- and *ax*-X-cyclohexanes (**2-*eq*** and **2-*ax***), and 2-halo-tetrahydropyrans (**3**), for X = F, Cl, Br, and I (Figure 1).

It is expected that in **1** very small hyperconjugative interactions transferring charge into $\sigma^*_{C_1-X}$ take place. However, important interactions of type $LP(X) \rightarrow \sigma^*_{C-H}$ must be operating, as commented above, widening the relevant energy gaps between the energies of the σ^*_{C-X} antibonding and those occupied $LP_2(X)$ and $LP_3(X)$ lone-pair orbitals.

Comparing **2-*eq*** and **2-*ax*** conformers, it is expected that a hyperconjugative interaction involving the σ^*_{C1-X} antibonding orbital is weaker in the former than in the latter. In compound **3**, the anomeric effect involving the ring oxygen atom defines a strong hyperconjugative interaction into the $\sigma^*_{C_1-X}$ antibonding orbital. Besides, the $\sigma^*_{O-C_1}$ and $\sigma^*_{O-C_5}$ antibonding orbitals in **3** are notably better electron acceptors than $\sigma^*_{C_1-C_1}$ and $\sigma^*_{C_1-C_1}$ antibonding orbitals, in either **2-*eq*** or **2-*ax***. A similar assertion holds for $\sigma^*_{C_{sp}3-H}$ antibonding orbitals in **1** when comparing their electron acceptor ability with **3**. For this reason, it is expected that interactions $LP_2(X) \rightarrow \sigma^*_{C_{sp}3-Y}$ and $LP_3(X) \rightarrow \sigma^*_{C_{sp}3-Y}$ are stronger in **3** than either in **2** or in **1** (Figure 2).

IV. Experimental and Computational Details

Compounds **2** with X = Cl, Br, and I were commercially available, while the fluoro-derivative was synthesized ac-

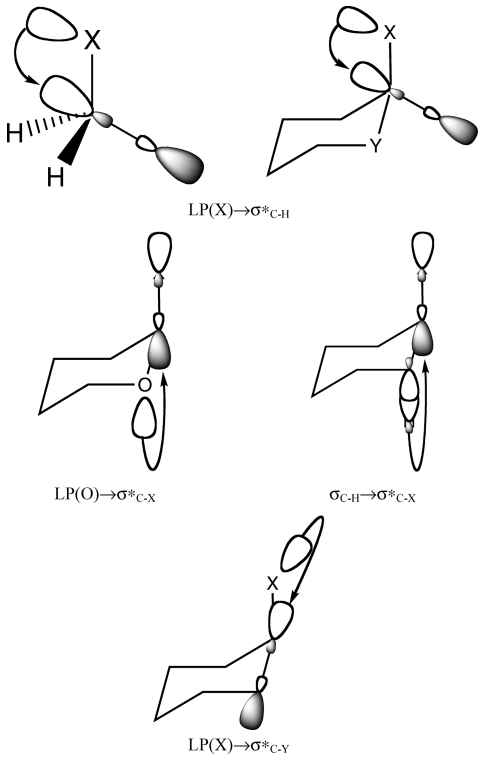


Figure 2. Relevant NBO interactions for methane, cyclohexane, and pyran derivatives, where X = F, Cl, Br, and I, and Y = CH_2 and O.

cording to a literature procedure.[34] Compounds **3** with X = Cl, Br, and I were also synthesized according to literature procedures,[35] and only their *axial* conformers were experimentally observed. Compound **3** with X = F was not studied in this work, because it was not possible to synthesize this compound.

Nonrelativistic calculations, that is, geometry optimization, NBO, and NMR shielding, were carried out at B3LYP level, using cc-pVTZ basis set[36,37] for C, H, O, F, Cl, and Br and Sadlej pVTZ for I.[38] Moreover, the $^1J_{CF}$ SSCCs were also calculated with the B3LYP functional using the EPR-III basis set for C and F, while for H and O the cc-pVTZ basis set was applied using the Gaussian 03 program.[39]

In the relativistic framework, the calculations of ground-state geometries in the relativistic scalar ZORA approach were carried out with the BP86 functional using a triple-ζ doubly polarized Slater-type basis set (TZ2P) with the Amsterdam Density Functional (ADF) package.[40] The corresponding NMR shielding constants were calculated using four different levels of theory: scalar ZORA, SO-ZORA, scalar PAULI, and SO-PAULI.

The ^{13}C substituent chemical shifts (SCS) were obtained as the difference between nuclear shielding constants calculated for each halo-derivative and for the corresponding parent compound and were reported in three different approaches: nonrelativistic at B3LYP level, and the relativistic levels at BP86/TZ2P, which was previously mentioned.

V. Results and Discussion

Experimental and calculated $^{13}C_1$ SCSs for compounds **1**, **2-*eq***, **2-*ax***, and **3** are collected in Table 1, where experimental values for series **1** were taken from the literature,[41] while those for the remaining compounds were measured as part of this work. In Table 1 are also shown the SCS calculated using the following approaches: nonrelativistic B3LYP, scalar ZORA, scalar PAULI, SO-ZORA, and SO-PAULI. For scalar ZORA and scalar PAULI, the NMR shielding was computed by summing the diamagnetic and paramagnetic contributions, and for SO-ZORA and SO-PAULI the SO and FC interaction was taken into account together with diamagnetic and paramagnetic contributions to each SCS, respectively, and it was plotted in Figure 3 for compounds **1**, **2-*eq***, **2-*ax***, and **3**, versus the halogen atom, where it can be observed that the calculated SO contributions are notably smaller for compounds **2-*eq***, **2-*ax***, and **3** than in **1**.

Table 1. Experimental and Theoretical[a] ^{13}C SCS for **1**, **2**, and **3** Halocompounds

		H	F	Cl	Br	I
1[a]	SCS_{exp}	0.0	77.7	27.4	12.3	−18.4
	SCS_{B3LYP}	0.0	81.4	43.5	34.3	18.3
	SCS_{ZORA}[b]	0.0	84.0	40.3	33.2	13.8
	$SCS_{SO\text{-}ZORA}$[c]	0.0	83.4	38.1	20.7	−16.1
	σ_{SO}		0.5	2.2	10.1	29.9
	SCS_{PAULI}[b]	0.0	84.07	40.32	34.18	66.03
	$SCS_{SO\text{-}PAULI}$[d]	0.0	83.56	38.18	22.20	45.33
	σ_{FC}		0.5	2.1	11.9	20.1
	D[e]		57	54	41	34
2-*eq*	SCS_{exp}	0.0	66.7	32.8	25.1	6.1
	SCS_{B3LYP}	0.0	67.5	45.6	43.5	37.6
	SCS_{ZORA}[b]	0.0	69.7	41.8	43.7	35.1
	$SCS_{SO\text{-}ZORA}$[c]	0.0	69.1	39.9	33.7	12.1
	σ_{SO}		0.6	1.9	10.0	22.8
	SCS_{PAULI}[b]	0.0	69.71	41.80	44.40	98.12
	$SCS_{SO\text{-}PAULI}$[d]	0.0	69.14	39.74	33.76	75.75
	σ_{FC}		0.6	2.1	10.6	21.8
	D[e]		95	103	100	91
2-*ax*	SCS_{exp}	0.0	63.5	33.1	28.4	13.8
	SCS_{B3LYP}	0.0	64.7	45.4	44.6	39.8
	SCS_{ZORA}[b]	0.0	66.3	41.5	44.8	37.9
	$SCS_{SO\text{-}ZORA}$[c]	0.0	65.9	40.1	37.6	21.9
	σ_{SO}		0.5	1.5	7.2	15.8
	SCS_{PAULI}[b]	0.0	66.31	41.48	45.34	94.28
	$SCS_{SO\text{-}PAULI}$[d]	0.0	65.75	39.63	36.05	74.85
	σ_{FC}		0.6	1.9	9.2	18.4
	D[e]		101	114	109	108
3	SCS_{exp}	0.0		26.0	25.9	16.7
	SCS_{B3LYP}	0.0	42.2	24.7	34.1	43.6
	SCS_{ZORA}[b]	0.0	42.7	36.6	42.7	42.2
	$SCS_{SO\text{-}ZORA}$[c]	0.0	42.2	35.0	35.2	26.9
	σ_{SO}		0.5	1.6	7.5	15.3
	SCS_{PAULI}[b]	0.0	42.76	36.61	42.86	81.50
	$SCS_{SO\text{-}PAULI}$[d]	0.0	42.15	34.63	33.74	64.11
	σ_{FC}		1.0	2.3	9.4	16.9
	D[e]		154	176	190	212

[a] Experimental values for halomethanes were taken from ref 36. [b] $\sigma = \sigma_{dia} + \sigma_{para}$. [c] $\sigma_{SO\text{-}ZORA} = \sigma_{dia} + \sigma_{para} + \sigma_{so}$. [d] $\sigma_{SO\text{-}PAULI} = \sigma_{dia} + \sigma_{para} + \sigma_{FC}$. [e] D: "Descriptor" of hyperconjugative interactions affecting the SO part of the SO/FC term (see Table 2).

Relevant NBO analyses were carried out for **1**, **2-*eq***, **2-*ax***, and **3** compounds. For the qualitative analysis described in section II, it is important to evaluate hyperconjugative interactions (a) involving the $\sigma^*_{C_1-X}$ antibonding orbital, and (b) involving the $LP_2(X)$ and $LP_3(X)$ lone pair orbitals. In this qualitative study, the former interactions (a) are taken into account globally considering the occupancy of the $\sigma^*_{C_1-X}$ antibonding orbital, while for the latter (b), by considering the sum of the occupancies of both lone-pairs.

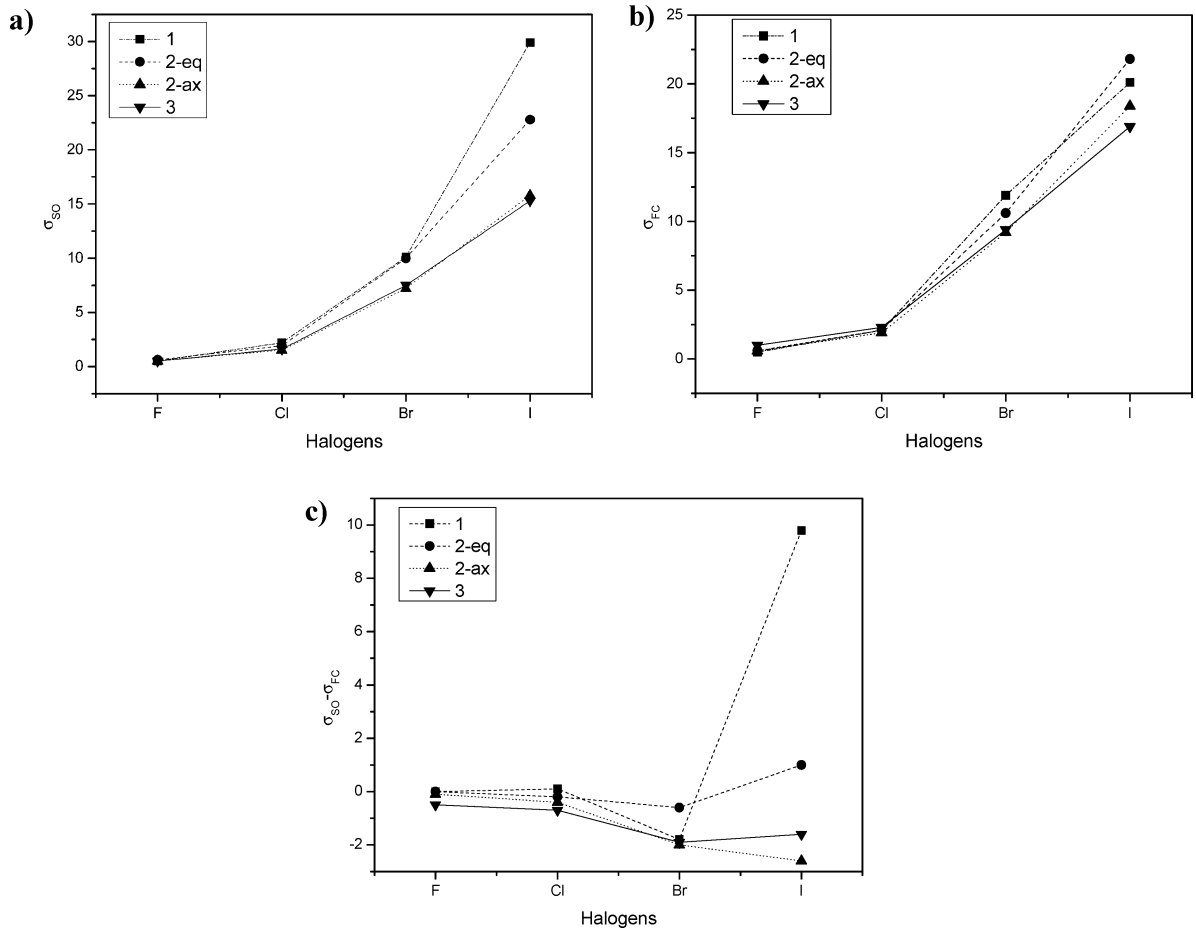


Figure 3. Calculated contribution to ^{13}C SCS for different halo-substituents in methane, **1**, cyclohexane-eq, ***2-eq***, cyclohexane-ax, ***2-ax***, and pyran, **3**: (a) σ_{SO} contribution; (b) σ_{FC} contribution; and (c) its difference, $\sigma_{SO} - \sigma_{FC}$.

Table 2. Relevant NBO Occupancies To Study the Relative Influence of the SO Part of the SO/FC Contribution to the $^{13}C_\alpha$ SCSs in **1**, ***2-eq***, ***2-ax***, and ***3-ax***[a]

	occupancy[b]	F	Cl	Br	I
1	Σnx	−54	−50	−38	−32
	σ^*_{CX}	3	4	3	2
	$\sigma^*_{CX} - \Sigma nx$	+57	+54	+41	+34
2-eq	Σnx	−56	−57	−48	−40
	σ^*_{CX}	39	46	50	51
	$\sigma^*_{CX} - \Sigma nx$	+95	+103	+100	+91
2-ax	Σnx	−56	−55	−45	−39
	σ^*_{CX}	45	59	64	69
	$\sigma^*_{CX} - \Sigma nx$	+101	+114	+109	+108
3	Σnx	−66	−90	−63	−39
	σ^*_{CX}	89	86	127	173
	$\sigma^*_{CX} - \Sigma nx$	+154	+176	+190	+212

[a] Each LP occupancy was obtained by subtracting 2.000 from the calculated occupancy (in units of 10^{-3}), that is, the occupancy of an ideally occupied NBO. σ^*_{C-X} occupancies are also given in units of 10^{-3}. [b] Σnx stands for the sum of the $LP_2(X)$ and $LP_3(X)$ occupancies. $\sigma^*_{CX} - \Sigma nx$ = D, which stands for "descriptor of hyperconjugative interactions affecting the SO part of the SO/FC term" (see Table 1).

In Table 2, such occupancies are given in units of 10^{-3} for the $\sigma^*_{C_1-X}$ antibonding orbital, and for the lone-pair orbitals they are given as the difference between the sum of the calculated occupancies of both lone-pairs and 4.000, which corresponds to the sum of the occupancies of two ideally occupied NBO orbitals. These differences are negative, and they are also expressed in units of 10^{-3}. As commented in section II, both types of interactions tend to decrease the absolute value of the SO/FC term due to the SO influence. Therefore, the sum of the absolute values of both types of occupancies, given in units of 10^{-3}, is taken as a significant, although qualitative, "descriptor" (D), of the influence of hyperconjugative interactions on the SO part of the SO/FC contribution to $^{13}C_1$ SCSs. For this reason, in Table 2, such sums of occupancies (D) are also displayed for the four chosen compounds for X = F, Cl, Br, I. It is recalled that for the same halogen atom, the SO contribution decreases with the increase of the D "descriptor". In all cases, for the same halogen atom, D increases monotonously from compounds **1** to **3**, and, therefore, the qualitative description presented above suggests that the SO decreases along the same series of compounds. In general, this observation is in line with the results displayed in Figure 3, with the exception of compounds **3** where the SO-ZORA calculated SO terms are quite similar to those calculated in ***2-ax***. This observation suggests that two opposite effects are taking place in halogen derivatives of compound **3**; that is, while the SO part of the SO/FC term is notably reduced due to the strong hyperconjugative interactions that take place in this compound, the FC contribution is increased due to the strong inductive effect produced by the ring oxygen atom placed α to C_1. To test if this suggestion is supported by available data, in Table 3 are displayed calculated $^1J_{CF}$ SSCCs (in Hz) at the B3LYP//

Table 3. $^{1}J_{CF}$ SSCCs (in Hz) Calculated at the B3LYP//EPR-III/aug-cc-pVTZ for the Fluorinated Compounds of Series **1**, **2-*eq***, **2-*ax***, and **3**

	1	**2-*eq***	**2-*ax***	**3**
FC	−280.3	−279.0	−273.0	−320.0
SD	23.3	23.4	24.1	24.5
PSO	35.9	29.5	30.6	42.9
DSO	0.4	0.9	0.9	1.1
J_{calc}[a]	−220.7	−225.2	−217.4	−251.5
J_{exp}	−157.5	−169.5	−164.6	

[a] J_{calc} = FC + SD + PSO + DSO.

EPR-III/aug-cc-pVTZ level for the fluorinated compounds of series **1**, **2-*eq***, **2-*ax***, and **3**. In fact, it is observed that the absolute value of the FC term of $^{1}J_{CF}$ SSCC in **3** is notably larger than for the remaining compounds. It is recalled that $^{1}J_{CF}$ SSCCs, in general, are not reproduced accurately within the DFT framework.[42] However, because in this work only a qualitative approach is applied, it is considered that the trend of the calculated FC contribution to $^{1}J_{CF}$ SSCCs is adequate to validate this qualitative analysis on the influence of the SO contribution to the "heavy atom effect" on ^{13}C SCSs.

VI. Concluding Remarks

For the results described in this Article, the spin−orbit operator is analyzed from a qualitative point of view, to estimate how certain types of hyperconjugative interactions would affect the performance of the SO operator to define a notably small "heavy atom effect" on the ^{13}C SCS bonded to a heavy halogen atom. The SO/FC contribution to ^{13}C SCS was calculated within the scalar ZORA, SO-ZORA, scalar PAULI, and SO-PAULI approaches, and the results are compared in Figure 3. It is observed that the SO and FC parts of the SO/FC term are sensitive enough to show observable differences for both *equatorial* and *axial* cyclohexane conformers. It is also important to highlight those strong hyperconjugative effects, which yield an important decrease on the SO part of the SO/FC term, and sometimes can be masked by strong inductive effects increasing the corresponding FC term of spin−spin coupling constants, as in compound **3**.

Acknowledgment. We are grateful to FAPESP (grants 06/03980-2, 05/59649-0, and 06/02783-9) for the financial support of this work and for a fellowship to A.C.N. and scholarship to L.C.D., and to CNPq for financial support and for fellowships (to R.R. and C.F.T.). Financial support from CONICET (PIP 5119/05) and UBATEC (X047) to R.H.C. is gratefully acknowledged.

References

(1) Morishina, I.; Endo, K.; Yonezawa, T. *J. Chem. Phys.* **1973**, *59*, 3356.

(2) Pyykkö, P. *Chem. Phys.* **1977**, *22*, 289.

(3) Cheremisin, A. A.; Schastnev, P. V. *J. Magn. Reson.* **1980**, *40*, 459.

(4) Kidd, R. G. *Annu. Rep. NMR Spectrosc.* **1980**, *10A*, 2.

(5) Pyykkö, P.; Wiessenfeld, L. *Mol. Phys.* **1981**, *43*, 557.

(6) Harris, R. K. *Nuclear Magnetic Resonance Spectroscopy. A Physicochemical View*; Longman: New York, 1986; p 188.

(7) Pyykkö, P.; Görling, A.; Rösch, N. *Mol. Phys.* **1987**, *61*, 195.

(8) Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1996**, *261*, 335.

(9) Kaupp, M.; Malkina, O. L.; Malkin, V. G. *Chem. Phys. Lett.* **1997**, *265*, 55.

(10) Fukuwa, S.; Hada, M.; Fukuda, R.; Tanaka, S.; Nakatsuji, H. *J. Comput. Chem.* **2001**, *22*, 528.

(11) Nakatsuji, H.; Takashima, H.; Hada, M. *Chem. Phys. Lett.* **1995**, *233*, 95.

(12) Wolff, S. K.; Ziegler, T. *J. Chem. Phys.* **1998**, *109*, 895.

(13) Malkina, O. L.; Schimmelpfennig, B.; Kaupp, M.; Hess, B. A.; Chandra, P.; Wahlgren, U.; Malkin, V. G. *Chem. Phys. Lett.* **1998**, *296*, 93.

(14) Vaara, J.; Ruud, K.; Vahtras, O.; Agren, H.; Jokisaari, J. *J. Chem. Phys.* **1998**, *109*, 1212.

(15) Vaara, J.; Ruud, K.; Vahtras, O. *J. Chem. Phys.* **1999**, *111*, 2900.

(16) Kutzelnigg, W. *J. Comput. Chem.* **1999**, *20*, 1199.

(17) (a) Vaara, J.; Malkina, O. L.; Stoll, H.; Malkin, V. G.; Kaupp, M. *J. Chem. Phys.* **2001**, *114*, 61. (b) Vaara, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399.

(18) Melo, J. I.; de Azua, M.C. R.; Giribet, C. G.; Aucar, G. A.; Provasi, P. F. *J. Chem. Phys.* **2004**, *121*, 6798.

(19) Helgaker, T.; Jaszuński, M.; Pecul, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *53*, 249.

(20) Jameson, C. J.; de Dios, A. C. Theoretical aspects of nuclear shielding. In *Specialist Periodical Repors on Nuclear Magnetic Resonance*; Webb, G. A., Ed.; The Royal Society of Chemistry: London, 2007; Vol. 36, p 50; 2008; Vol. 37, p 51.

(21) Kaupp, M.; Malkina, O. L.; Malkin, V. G.; Pyykkö, P. *Chem.-Eur. J.* **1998**, *4*, 118.

(22) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.

(23) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783.

(24) van Lenthe, E.; Ehlers, A.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 8943.

(25) (a) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251. (b) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497.

(26) (a) Fukui, H.; Baba, T.; Inomata, H. *J. Chem. Phys.* **1996**, *105*, 3175. (b) Fukui, H.; Baba, T.; Inomata, H. *J. Chem. Phys.* **1996**, *106*, 2897. (c) Manninen, P.; Lantto, P.; Vaara, J.; Ruud, K. *J. Chem. Phys.* **2003**, *119*, 2623. (d) Manninen, P.; Ruud, K.; Lantto, P.; Vaara, J. *J. Chem. Phys.* **2005**, *122*, 114107. (e) Manninen, P.; Ruud, K.; Lantto, P.; Vaara, J. *J. Chem. Phys.* **2006**, *124*, 149901.

(27) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(28) Carpenter, J. E.; Weinhold, F. *J. Mol. Struct. (THEOCHEM)* **1988**, *169*, 41.

(29) Barone, V.; Peralta, J. E.; Contreras, R. H. *J. Comput. Chem.* **2001**, *22*, 1615.

(30) Dewar, M. J. S.; Dougherty. *The PMO Theory of Organic Chemistry*; Plenum: New York, 1975.

(31) Contreras, R. H.; Esteban, A. L.; Díez, E.; Della, E. W.; Lochert, I. J.; Dos Santos, F. P.; Tormena, C. F. *J. Phys. Chem. A* **2006**, *110*, 4266.

(32) Barone, V.; Contreras, R. H.; Díez, E.; Esteban, A. *Mol. Phys.* **2003**, *101*, 1297.

(33) Della, E. W.; Cotsaris, E.; Hine, P. T. *J. Am. Chem. Soc.* **1981**, *103*, 4131.

(34) Olah, G. A.; Welch, J. T.; Vankar, Y. D.; Nojima, M.; Kerekes, I.; Olah, J. A. *J. Org. Chem.* **1979**, *44*, 3872.

(35) (a) Meltzer, P. C.; Wang, P.; Blundell, P.; Madras, B. K. *J. Med. Chem.* **2003**, *46*, 1538. (b) Zelinski, R.; Yorka, K. *J. Org. Chem.* **1958**, *23*, 640. (c) Keiman, E.; Perez, D.; Sahai, M.; Shvilly, R. *J. Org. Chem.* **1990**, *55*, 2927.

(36) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975.

(37) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.

(38) (a) Sadlej, A. J.; Urban, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *147*, 234. (b) Sadlej, A. J. *Theor. Chim. Acta* **1992**, *79*, 123. (c) Sadlej, A. J. *Theor. Chim. Acta* **1992**, *81*, 45. (d) Sadlej, A. J. *Theor. Chim. Acta* **1992**, *81*, 339.

(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 E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(40) Baerends, E. J.; Autschbach, J.; B'erces, A.; Bickelhaupt, F. M.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; van Faassen, M.; Fischer, L.; Fan, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; McCormack, D. A.; Michalak, A.; Neugebauer, J.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D. Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S.; Woo, T.; Yakovlev, A.; Ziegler, T. *ADF 2006.01, SCM, Theoretical Chemistry*; Vrije Universiteit: Amsterdam, The Netherlands. URL: http://www.scm.com.

(41) Abraham, R. J.; Fisher, J.; Loftus, P. *Introduction to NMR Spectroscopy*; John Wiley & Sons: Chichester, 1988; p 15.

(42) Krivdin, L. B.; Contreras, R. H. *Annu. Rep. NMR Spectrosc.* **2007**, *61*, 133, and references cited therein.

CT800520W

JCTC Journal of Chemical Theory and Computation

All-Electron Scalar Relativistic Basis Sets for the Lanthanides

Dimitrios A. Pantazis and Frank Neese*

Lehrstuhl für Theoretische Chemie, Institut für Physikalische und Theoretische Chemie, Universität Bonn, Wegelerstrasse 12, D-53115 Bonn, Germany, and Max-Planck-Institut für Bioanorganische Chemie, Stiftstrasse 34-36, 45470 Mülheim an der Ruhr, Germany

Received February 22, 2009

Abstract: Segmented all-electron relativistically contracted (SARC) basis sets are constructed for the elements $_{57}$La−$_{71}$Lu and optimized for density functional theory (DFT) applications. The basis sets are intended for use in combination with the DKH2 or ZORA scalar relativistic Hamiltonians for which individually optimized contractions are provided. Significant computational advantages can be realized owing to the loose contraction of the SARC basis sets compared to generally contracted basis sets, while their compact size allows them to replace effective core potentials for routine studies of lanthanide complexes. The new basis sets are evaluated in DFT calculations of the first four ionization energies of the lanthanides. They yield results that accurately reproduce the experimental trends, confirming a balanced treatment of different electronic configurations. The performance of the basis sets is further assessed in molecular systems with a comprehensive study of the lanthanide trihalides. Despite their compact size, the SARC basis sets demonstrate consistent, efficient, and reliable performance and will be especially useful in calculations of molecular properties that require explicit treatment of the core electrons.

Introduction

Despite the traditional "rare earth" misnomer, the elements La−Lu that comprise the 4f block of the periodic table are ubiquitous in nature and are more abundant, in fact, than many transition metals.[1] Owing to their special chemical and physical properties, they are similarly ubiquitous in practical applications and occupy a central place in modern science and technology. Lanthanide complexes find extensive use as catalysts in synthetic chemistry,[2,3] but in general their applications are more often associated with their unique optical and magnetic properties stemming from the partially occupied 4f shell.[4] Thus, they feature prominently in the area of molecular magnetism, especially in the rapidly expanding field of d/f heterometallic chemistry.[5−8] They are of fundamental importance in the technology of lasers, the fabrication of special glasses, and the construction of cathode-ray or plasma displays as well as in the materials for light emitting diodes and optical fibers.[9] In the biomedical field, the luminescence of lanthanides is exploited for labeling purposes in biological assays[10] and nanoparticle bioprobes.[11,12] Lanthanide compounds are also actively researched for therapeutic uses,[13] while the most highlighted application in clinical practice is arguably the dominant use of gadolinium(III) complexes as contrast agents in magnetic resonance imaging.[14−16]

From a quantum chemical perspective, the lanthanides present difficulties in their computational treatment because of the large number of electrons and the necessity to account for significant relativistic effects.[17] In many practical cases, such as in density functional theory (DFT)[18,19] studies of structures and relative energies, both of these issues can be addressed by the use of effective core potentials (ECPs), which are adjusted to reproduce relativistic reference data that go beyond most approximate Hamiltonians.[20] By explicitly treating only the valence electrons, ECPs also serve to decrease the computational requirements while simulta-

* Corresponding author. E-mail: neese@thch.uni-bonn.de.

10.1021/ct900090f CCC: $40.75
Published on Web 08/12/2009

neously incorporating scalar relativistic effects implicitly through parametrization. Note, however, that the definition of "valence electrons" is not obvious for the lanthanides. There are mainly three types of effective core potentials and associated valence basis sets available for lanthanides. These include the averaged relativistic ECPs and the spin−orbit operators of Ross et al. with a [Xe] core,[21] the shape-consistent quasirelativistic ECPs of Cundari and Stevens with a [Kr]d^{10} core,[22] and the energy-consistent quasirelativistic ECPs and basis sets of Dolg,[23−29] which come in a "large-core" (the 4f shell is included in the core) and a "small-core" variety that treats the four, five, and six shell electrons explicitly. In contrast to their heavier 5f congeners, the partially occupied 4f shell of the lanthanides can, for many purposes, be considered as chemically inert and, thus, can be subsumed into the ECP. Even though this eliminates most of the magnetic and electronic subtleties of the lanthanides, it has proven to give acceptable results in studies that focus on the structural features or the estimates of relative energies.[30,31] However, this approach requires a separate potential for each oxidation state or 4f configuration[29] and, in practice, it precludes the modeling of f-centered processes and the treatment of spin−orbit coupling, while it may also mask the potential effects of f shell asphericity on structure.[32]

Thus, the use of large-core ECPs has inherent limitations, and one is faced with the dilemma that reducing the size of the core will allow for a more flexible treatment of the valence region, whereas a larger core allows for a better modeling of the all-important relativistic effects. Deficiencies of ECPs have been pointed out in specific situations. A recent example is the systematic study of the spin-state energies of iron complexes by Swart and co-workers, who showed that the use of ECPs results in spin-state splittings that never converge to the common limit attainable by both Slater- and Gaussian-type all-electron basis sets, regardless of the extension of the valence basis.[33] In a different field of application, Frenking and co-workers have noted that total electron densities derived from ECP calculations are associated with artifacts in the topological analysis.[34] From the field of molecular magnetism, a recent theoretical study of the exchange coupling in Gd^{III}/Cu^{II} systems by Cirera and Ruiz addressed the performance of different relativistic approaches and basis sets, concluding that effective core potentials lead to qualitatively incorrect results; only the combination of the all-electron basis sets with a scalar relativistic Hamiltonian proved capable of producing the correct behavior.[35] All these examples highlight the need for reliable and efficient all-electron alternatives to ECPs, a requirement that is also fundamental when properties of the inner shells are being probed, as in EPR or X-ray absorption experiments.

Thus, either for validating the use of ECPs or circumventing some of their limitations, it is necessary to have all-electron basis sets that allow efficient calculations with the popular scalar relativistic Hamiltonians, such as the zeroth order regular approximation[36−38] (ZORA), the infinite order regular approximation[39] (IORA), and the Douglas−Kroll−Hess[40−44] (DKH) approach. It should be emphasized that nonrelativistic basis sets are not flexible enough in the core region to be used with scalar relativistic Hamiltonians. Besides, even if used completely uncontracted, standard basis sets lack the much higher exponents that are typically required by scalar relativistic Hamiltonians. To the best of our knowledge, existing Slater-type scalar relativistic all-electron basis sets for the lanthanides are limited to the ZORA basis sets available in the Amsterdam density functional code.[45,46] Gaussian-type all-electron relativistic basis sets for the lanthanides include the atomic natural orbital (ANO-RCC) basis sets of Roos et al. for use with the DKH2 Hamiltonian,[47] the DKH3 basis sets of Hirao and co-workers, which are available for both point- and finite-nucleus approximations,[48,49] and the segmented contracted correlating basis sets of Koga and co-workers.[50] These three families of large and high-quality basis sets are suitable for use with correlated ab initio methods and, in combination with such theoretical approaches, are able to deliver highly accurate results for small systems.[47−50] However, their size and construction is not tailored to the more modest requirements of DFT approaches, which show much faster convergence for molecular properties with respect to extension of the basis set.[18] A more serious issue in terms of efficiency for the Roos and Hirao basis sets is their general contraction, since for DFT calculations the generation of two-electron integrals over basis functions dominates the computational effort.

Hence, we feel that it is important to have small standard Gaussian basis sets available that are not generally contracted and are compact enough to be used in day-to-day DFT calculations with the most popular scalar relativistic Hamiltonians. In this paper, we propose such segmented all-electron relativistically contracted (SARC) basis sets, which are constructed for DFT treatments of lanthanide systems in conjunction with the scalar relativistic DKH or ZORA Hamiltonians. The SARC basis sets are sufficiently small in terms of the total number of basis functions so as not to present a grossly inefficient alternative to ECPs for routine DFT studies of large molecules. Exponents of the Gaussian primitives are derived from relatively simple empirical rules, and contraction coefficients are determined separately for both the ZORA and second-order DKH (DKH2) schemes. These two scalar relativistic approximations produce quite distinct shapes for the core orbitals,[51] therefore, the contractions must be adapted to each particular Hamiltonian. The same strategy has been used successfully in the construction of SARC basis sets for third-row transition metals[51] and has been extensively evaluated in calibration studies of first-, second- and third-row transition metal geometries.[52] All the above basis sets are now part of the freely available ORCA program package.[53] The performance of the SARC basis sets is assessed here for both atomic and molecular properties.

Construction of Basis Sets. Restricted open-shell Hartree−Fock (ROHF) calculations were first carried out for each atom in its ground-state configuration in order to obtain the innermost radial expectation values that are subsequently employed in the generation of the new primitive Gaussian functions. These calculations followed Zerner's spin-averaged (SAHF) formalism,[54,55] as implemented in ORCA.[53] Within this approach, the ROHF code averages over all of

Table 1. Radial Expectation Values of Innermost Orbitals (in Bohr) Determined from Spin-Averaged ROHF Calculations

	$\langle r_s \rangle$	$\langle r_p \rangle$	$\langle r_d \rangle$	$\langle r_f \rangle$
La	0.026644	0.097171	0.259694	
Ce	0.026180	0.095340	0.253476	0.983364
Pr	0.025731	0.093569	0.247552	1.008485
Nd	0.025298	0.091867	0.241922	0.960901
Pm	0.024879	0.090226	0.236554	0.920354
Sm	0.024473	0.088641	0.231428	0.884966
Eu	0.024081	0.087111	0.226529	0.853546
Gd	0.023701	0.085636	0.221853	0.783083
Tb	0.023333	0.084202	0.217344	0.813301
Dy	0.022976	0.082819	0.213034	0.793152
Ho	0.022629	0.081480	0.208898	0.773551
Er	0.022293	0.080184	0.204925	0.754670
Tm	0.021967	0.078927	0.201104	0.736583
Yb	0.021651	0.077709	0.197427	0.719308
Lu	0.021343	0.076531	0.193896	0.674575

the states of a given spin for a given configuration. Although Huzinaga's decontracted well-tempered basis sets (WTBS)[56,57] were found to be perfectly adequate for this step, in the case of third-row transition metal atoms,[51] exploratory calculations indicated that a larger basis set would be needed to approach the basis set limit for the lanthanides. Therefore, in this work we have chosen the universal Gaussian basis set (UGBS) of de Castro and Jorge.[58,59] We extended the number of primitive Gaussian functions beyond the range proposed in the original paper in order to create a common basis set that covers consistently all elements with $Z = 57-71$. This led to a (34s24p20d14f) basis set which was used in a completely decontracted form, resulting in 304 functions per atom. We note that this basis set (referred to simply as UGBS in the following) yields total atomic energies that deviate less than 1 mE_h from the nonrelativistic numerical Hartree−Fock values of Koga et al.[60] and, thus, can serve as a suitable reference point for benchmarking. All atoms were considered in their respective ground states; these generally arise from $4f^n6s^2$ configurations with the exception of La, Ce, Gd, and Lu, which have an electron in the 5d shell. In detail, the atomic configurations and ground states are: La ($5d^16s^2$, ^{2}D), Ce ($4f^15d^16s^2$, ^{1}G), Pr ($4f^36s^2$, ^{4}I), Nd ($4f^46s^2$, ^{5}I), Pm ($4f^56s^2$, ^{6}H), Sm ($4f^66s^2$, ^{7}F), Eu ($4f^76s^2$, ^{8}S), Gd ($4f^75d^16s^2$, ^{9}D), Tb ($4f^96s^2$, ^{6}H), Dy ($4f^{10}6s^2$, ^{5}I), Ho ($4f^{11}6s^2$, ^{4}I), Er ($4f^{12}6s^2$, ^{3}H), Tm ($4f^{13}6s^2$, ^{2}F), Yb ($4f^{14}6s^2$, ^{1}S), and Lu ($4f^{14}5d^16s^2$, ^{2}D). The resulting radial expectation values $\langle r_l \rangle$ of the 1s, 2p, 3d, and 4f orbitals are listed in Table 1.

Following the same procedure as previously detailed,[51] the exponents α_l of the tightest s, p, d, and f functions were determined from the radial expectation values as $\alpha_l = 2k_l f_l^2/\pi\langle r_l \rangle^2$, where $f_l = 1$, 4/3, 8/5, and 64/35 for $l =$ s, p, d, and f. Scaling factors k_l of 1 000, 100, 33, and 10 were used for s, p, d, and f functions, respectively. The exponents of these tightest basis functions per atom are listed in Table 2. Notice the regular and smooth variation of the exponents along the series, which implies that the basis sets could be used reliably for comparisons between systems containing different lanthanide atoms. Sets of primitive Gaussians were then constructed for each angular momentum as series of the form $\alpha_i x^{-i}$ ($i = 1, 2, ...$), $x = 2.25$, 2.50, 2.75, and 3.00 for $l =$ s, p, d, and f. Exponent cutoff values were set to 0.02 for s, p, and d functions and 0.2 for f functions, but more diffuse functions can be easily generated from the same formula. We emphasize that the above values for the empirical extrapolation factors x were found to serve well our primary goal of a sufficiently small number of functions that still maintain high performance in terms of atomic and molecular properties, while also ensuring fast convergence for DFT calculations. Of course, a choice of larger values for x can lead to more densely spaced primitives, if this need arises in special applications.

Table 2. Maximum Exponents Per Angular Momentum (in Bohr^{-2}) Used for the Construction of the SARC Basis Sets

	α_s	α_p	α_d	α_f
La	896770.416684	11986.275399	797.462366	
Ce	928839.847608	12451.087955	837.067226	22.012804
Pr	961538.755282	12926.877106	877.609164	20.929801
Nd	994735.795984	13410.301051	918.931840	23.054021
Pm	1028523.650946	13902.541106	961.110752	25.130098
Sm	1062932.540176	14404.172125	1004.158396	27.180085
Eu	1097819.871374	14914.599562	1048.060643	29.217975
Gd	1133304.941480	15432.804306	1092.706223	34.712700
Tb	1169335.032050	15962.936347	1138.514896	32.181136
Dy	1205955.486947	16500.520028	1185.048662	33.836944
Ho	1243224.036239	17047.298098	1232.439088	35.573458
Er	1280982.187936	17602.816491	1280.690360	37.375741
Tm	1319284.994791	18167.970032	1329.819235	39.233820
Yb	1358076.401925	18741.957799	1379.815234	41.140939
Lu	1397555.920017	19323.367785	1430.527887	46.778199

The resulting uncontracted basis sets have the form (23s16p12d6f) comprising 173 primitives in total (131 for La since no f functions are used). In the last step, the innermost six s, five p, four d, and four f primitives were contracted to create the final basis sets with a [18s12p9d3f] pattern and 120 functions. To put this number into perspective, the uncontracted (26s23p17d13f) ANO-RCC basis sets[47] would yield 271 functions, while the (27s23p15d10f) Hirao basis sets[48] correspond to 241 functions. On the other hand, the valence [10s8p5d4f] basis set of Dolg's small-core ECP basis[27] already contains 87 contracted basis functions. Contraction coefficients were obtained by scalar relativistic calculations of the atomic ground states and are specifically optimized for each scalar relativistic Hamiltonian (DKH2 or ZORA). These SARC basis sets are of valence triple-ζ (TZV) quality, and in molecular calculations employing the DKH2 or ZORA Hamiltonians, they should normally be combined with their relativistically contracted counterparts for the other atoms, either the respective SARC basis sets for the third-row transition metals or the appropriate relativistically contracted variants of the Karlsruhe basis sets[51] for other elements that are freely available in the ORCA basis set library. Additional basis functions of higher angular momentum to be used as polarization or correlation functions were generated by scaling the most diffuse existing exponents by 1.25. One additional function has been used for TZVP and three functions for TZVPP-quality basis sets. Complete listings of the basis sets in ready-to-use format are provided in the Supporting Information.

Assessment and Applications. *Evaluation of Basis Set Construction.* Compared to the large uncontracted UGBS, two formal approximations enter in the construction of the new SARC basis sets: the significant reduction in the total

Table 3. Estimated Incompleteness Errors (E_h) from Comparison of the UGBS and SARC Basis Sets[a]

	UGBS (304 functions)	SARC (120 functions)	ΔE
La	–8486.519526	–8485.444981	1.074545
Ce	–8853.276583	–8851.937575	1.339008
Pr	–9229.684172	–9228.320549	1.363623
Nd	–9615.948651	–9614.439880	1.508771
Pm	–10012.152291	–10010.488169	1.664122
Sm	–10418.434516	–10416.603845	1.830671
Eu	–10834.936141	–10832.926564	2.009578
Gd	–11261.687283	–11259.548746	2.138537
Tb	–11698.581889	–11696.151822	2.430067
Dy	–12146.284756	–12143.621665	2.663092
Ho	–12604.772058	–12601.859229	2.912829
Er	–13074.195242	–13071.014408	3.180834
Tm	–13554.708332	–13551.239799	3.468533
Yb	–14046.468030	–14042.690360	3.777670
Lu	–14549.587004	–14545.604380	3.982623

[a] Spin-averaged ROHF calculations with the DKH2 Hamiltonian.

number of primitives and the contraction of the innermost functions in a specific pattern adapted to the scalar relativistic Hamiltonians. Thus, the magnitude of the errors introduced by the above approximations, namely the incompleteness and contraction error, can serve as a preliminary indicator for the internal consistency and construction quality of the SARC basis sets.

In order to determine these errors, we performed SAHF calculations with the DKH2 Hamiltonian for the atomic ground states. In Table 3 we compare the total electronic energies obtained with the [18s12p9d3f] SARC basis sets to those obtained with the (34s24p20d14f) UGBS, the latter being considered a good approximation to the basis set limit. The energy difference is a good estimate of the incompleteness error, which rises monotonically from 1.07 E_h for lanthanum up to 3.98 E_h for lutetium. The effect of contraction is, of course, folded into the incompleteness error, as determined above. Therefore, to obtain an estimate of the contraction error, additional calculations were performed using the SARC basis sets in fully uncontracted form (23s16p12d6f). The difference in the total energies obtained by the contracted and uncontracted SARC versions is found to increase smoothly from a minimum of 33 mE_h for lanthanum to a maximum of 74 mE_h for lutetium, clearly a minor effect compared to the magnitude of the incompleteness error.

The small values and the limited span of both the incompleteness and the contraction errors are remarkable. To appreciate this point we have to refer to the SARC basis sets for the third-row transition metals,[51] which have already been shown to be of excellent quality in practical applications.[52] In that case, the basis sets displayed incompleteness errors ranging from a little over 4.22 up to 7.75 E_h, all values beyond the corresponding error range for the lanthanides. Furthermore, the maximum contraction error in the present case is lower than the lowest contraction error obtained for a third-row transition metal (hafnium, at 79 mE_h). Although incompleteness and contraction errors are not expected to significantly influence molecular properties other than total energies, these fundamental metrics already demonstrate the quality of the proposed basis sets in terms of both size and contraction pattern.

Table 4. Bond Lengths r (pm) and Dissociation Energies D_e (eV) of Lanthanide Diatomics Computed with the DKH2 Hamiltonian, without and with BSSE Counterpoise Corrections (CPC), Compared with Experiment

	PBE0/SARC		PBE0/SARC+CPC		expt[a]	
	r	D_e	r	D_e	r	D_e
LaH ($^1\Sigma$)	202.1	2.57	202.2	2.57	203.2	–
LaF ($^1\Sigma$)	203.0	6.43	203.0	6.40	202.7	6.23
LaO ($^2\Sigma$)	184.2	7.82	184.3	7.80	182.6	8.29
LuH ($^1\Sigma$)	190.3	3.12	190.4	3.12	191.2	3.47
LuF ($^1\Sigma$)	191.3	7.22	191.4	7.20	191.7	5.93[b]
LuO ($^2\Sigma$)	178.1	6.84	178.1	6.81	179.0	7.04

[a] Refs 63 and 64. [b] Estimated.

Nevertheless, it is possible that the existing incompleteness error might give rise to basis set superposition errors in molecular calculations if the incompleteness arises from an insufficient description of the valence orbitals. To clarify this point, we have performed DKH2 calculations with the PBE0 functional on a series of lanthanide hydrides, fluorides, and oxides using the two elements at the ends of the series (La and Lu). The def2-TZVP basis sets were used for H, F, and O. The bond length and dissociation energy of each species were corrected for BSSE errors using the counterpoise correction method of Boys and Bernardi.[61] The results (Table 4) indicate that superposition errors are either minimal or nonexistent, the corrections being at the most 0.1 pm for the bond lengths and 0.03 eV for the dissociation energies. Therefore, the valence orbitals that contribute to the bonding appear to be covered quite well with the SARC basis sets, and thus, the incompleteness error is not associated with the valence space. Note that these conclusions may only hold for DFT, where basis set completeness for the valence space is relatively easy to achieve. Correlated ab initio methods typically require larger and more extensively polarized basis sets than the basis sets proposed here.[62]

In terms of basis set performance for atomic systems, another useful descriptor is the energy of the outermost valence and semicore orbitals (6s, 5d, and 4f) compared to the essentially converged UGBS results. Using the data obtained from the SA-ROHF DKH2 calculations described in the previous section, we observe a very close agreement between the SARC and the UGBS results. Specifically, the energies of the 6s orbitals are practically identical, with maximum deviations of 0.01 eV. For the four elements that have an occupied 5d orbital in their ground-state configuration (La, Ce, Gd, and Lu), the 5d energies also agree within 0.03 eV. Greater discrepancies are observed for the energies of the 4f orbitals, which are stabilized with the SARC basis sets by 0.30 eV on average, but this difference is still too small to cause any concern about the performance of the basis sets in routine molecular applications. These observations are in line with the comments made above regarding the coverage of the valence space and the minimal BSSE errors, and they also confirm that the incompleteness errors reported in Table 3 are not associated with the valence orbitals.

The origin of the incompleteness errors is discovered upon moving close to the nucleus, where the difference in size of the two basis sets leads to more pronounced deviations in

Table 5. Orbital Energies (E_h) and Radial Expectation Values for Lutetium, Obtained from SAHF-DKH2 Calculations with the UGBS and SARC Basis Sets

	UGBS		SARC			
	E	<*r*>	*E*	<*r*>	Δ*E*	Δ<*r*>
1s	–2334.889	0.019	–2334.342	0.019	0.547	0.000
2s	–402.623	0.081	–402.713	0.081	–0.090	0.000
2p	–355.307	0.072	–355.434	0.072	–0.128	0.000
3s	–93.381	0.212	–93.513	0.212	–0.132	0.000
3p	–78.819	0.211	–78.949	0.211	–0.129	0.000
3d	–60.730	0.192	–60.854	0.193	–0.124	0.000
4s	–19.664	0.479	–19.729	0.479	–0.065	0.000
4p	–14.769	0.507	–14.820	0.507	–0.051	0.000
4d	–8.151	0.549	–8.186	0.550	–0.035	0.000
5s	–2.697	1.194	–2.705	1.192	–0.008	–0.002
5p	–1.439	1.408	–1.442	1.406	–0.004	–0.002
4f	–0.822	0.701	–0.816	0.705	0.006	0.004
6s	–0.222	3.910	–0.222	3.905	0.000	–0.005
5d	–0.188	2.739	–0.188	2.742	0.000	0.003

absolute energies, with the most part of the discrepancy attributed to the chemically unimportant 1s orbital. It is possible to reduce the discrepancy by using higher s exponents, but this creates imbalances in other shells and introduces numerical instabilities that can be typically avoided with a finite-nucleus model.[49] Regardless, this energy difference is chemically irrelevant. For the DFT applications the SARC basis sets are aimed at, we cannot think of any situation where convergence to a "basis set limit" scalar relativistic energy would be sought. For such purposes, other methods and basis sets mentioned in the introduction might present better options. The most relevant criterion here is the radial expectation value because this better reflects the scalar relativistic effects that the basis sets aim to capture. Importantly, for all core and semicore orbitals there is coincidence in the $\langle r \rangle$ values. Table 5 presents detailed orbital energies and radial expectation values for the "worst-case" element, lutetium. Based on the properties of atomic orbitals, we conclude that the SARC basis sets are overall well-balanced for their size, yielding orbital features that follow those predicted by the significantly larger UGBS basis set both close to the core and in the valence space.

Ionization Energies. The first ionization energies (IE_1) of the lanthanides are known with high accuracy (within 0.02 eV), and this makes them ideal for benchmarking the new basis sets for atomic systems in a more rigorous way. We evaluate ionization energies using the B3LYP functional as a representative DFT method, since it is one of the typical hybrid functionals to be used in the expected application setting of the SARC basis sets. Across all lanthanides, and similar to the 5d transition series, the first ionization energy is associated with the removal of an electron from the doubly occupied 6s orbital except for lutetium, which loses the single 5d electron to attain a closed-shell configuration. In the case of cerium, a change of configuration is observed with an increase of the 5d occupation number. As expected from the semicore character of the 4f orbitals, their occupation remains intact. Thus, the specific configurations and corresponding ground states of the cations are: La^+ ($5d^2$, ^3F), Ce^+ ($4f^15d^2$, ^4H), Pr^+ ($4f^36s^1$, ^5I), Nd^+ ($4f^46s^1$, ^6I), Pm^+ ($4f^56s^1$, ^7H), Sm^+ ($4f^66s^1$, ^8F), Eu^+ ($4f^76s^1$, ^9S), Gd^+ ($4f^75d^16s^1$, ^{10}D), Tb^+ ($4f^96s^1$, ^7H), Dy^+ ($4f^{10}6s^1$, ^6I), Ho^+ ($4f^{11}6s^1$, ^5I), Er^+ ($4f^{12}6s^1$, ^4H), Tm^+ ($4f^{13}6s^1$, ^3F), Yb^+ ($4f^{14}6s^1$, ^2S), and Lu^+ ($4f^{14}6s^2$, ^1S). Note that the ground state of the Ce atom ($4f^15d^16s^2$, ^1G) is a non-Hund (or "unnatural parity") singlet state, as discussed in depth by Morgan and Kutzelnigg,[65] and as such, it is inaccessible within the current DFT framework. Thus, the triplet state of the neutral Ce atom was used instead for our B3LYP calculation of ionization energies. As anticipated, our results confirm that this pragmatic choice is the logical one for the DFT approach and has no adverse effect on computed quantities.

Table 6. First Ionization Energies (eV) Computed with the B3LYP Functional and the Uncontracted UGBS Basis Set, with and without Relativistic Corrections, Compared with Experimental Values

		nonrelativistic		DKH2		ZORA	
	expt[a]	IE_1	Δ*E*	IE_1	Δ*E*	IE_1	Δ*E*
La	5.58	4.80	–0.78	5.56	–0.02	5.57	–0.01
Ce	5.54	5.02	–0.52	5.46	–0.08	5.47	–0.07
Pr	5.47	5.16	–0.31	5.39	–0.08	5.40	–0.07
Nd	5.52	5.21	–0.31	5.45	–0.07	5.46	–0.06
Pm	5.58	5.25	–0.33	5.51	–0.07	5.52	–0.06
Sm	5.64	4.34	–1.30	5.57	–0.07	5.57	–0.07
Eu	5.67	5.33	–0.34	5.62	–0.05	5.63	–0.04
Gd	6.15	7.06	0.91	6.07	–0.08	6.08	–0.07
Tb	5.86	5.48	–0.38	5.81	–0.05	5.82	–0.04
Dy	5.94	5.55	–0.39	5.92	–0.02	5.93	–0.01
Ho	6.02	5.61	–0.41	6.01	–0.01	6.02	0.00
Er	6.11	5.68	–0.43	6.11	0.00	6.12	0.01
Tm	6.18	5.74	–0.44	6.19	0.01	6.21	0.03
Yb	6.25	5.80	–0.45	6.28	0.03	6.29	0.04
Lu	5.43	6.56	1.13	5.39	–0.04	5.39	–0.04
MAD			0.56		0.05		0.04
rms			0.64		0.05		0.05

[a] Reference 66.

Before we assess the SARC basis sets, it is instructive to see how large is the importance of scalar relativistic effects for this property using the practically complete UGBS basis set. According to the B3LYP results presented in Table 6, nonrelativistic calculations generally underestimate ionization energies with a root-mean-squared (rms) error of 0.64 eV. Interestingly, this error is significantly smaller than that obtained from similar nonrelativistic calculations for the third-row transition metal atoms (1.39 eV). This rms value, however, masks the fact that pronounced nonsystematic failures, such as the qualitatively different ionization process for Gd ($f^8d^1s^1$ to $f^7d^2s^0$ compared with the relativistic $f^7d^1s^2$ to $f^7d^1s^1$), lead to a wide error spread of more than 2.4 eV. Inclusion of scalar relativistic effects with either the DKH2 or the ZORA Hamiltonians produces an evidently more uniform behavior and reduces the rms error down to only 0.05 eV.

In view of the large difference in size between the UGBS and the SARC basis sets (304 and 120 basis functions, respectively), we anticipated that moving to the more compact SARC basis sets might adversely affect the accuracy of the calculated values. However, the results summarized in Table 7 show that the reduction in size, which is accompanied by significant gains in terms of computational cost, does not compromise in any way the accuracy of the

Table 7. First Ionization Energies (eV) Computed with the B3LYP Functional and the SARC Basis Sets, Compared with Experimental Values

		DKH2		ZORA	
	expt[a]	IE_1	ΔE	IE_1	ΔE
La	5.58	5.58	0.00	5.59	0.01
Ce	5.54	5.54	0.00	5.54	0.00
Pr	5.47	5.41	–0.06	5.42	–0.05
Nd	5.52	5.47	–0.05	5.47	–0.05
Pm	5.58	5.52	–0.06	5.53	–0.05
Sm	5.64	5.58	–0.06	5.59	–0.05
Eu	5.67	5.63	–0.04	5.64	–0.03
Gd	6.15	6.08	–0.07	6.09	–0.06
Tb	5.86	5.80	–0.06	5.81	–0.05
Dy	5.94	5.93	–0.01	5.94	0.00
Ho	6.02	6.02	0.00	6.03	0.01
Er	6.11	6.13	0.02	6.13	0.02
Tm	6.18	6.20	0.02	6.22	0.04
Yb	6.25	6.29	0.04	6.30	0.05
Lu	5.43	5.39	–0.04	5.39	–0.04
MAD			0.04		0.03
rms			0.04		0.04

[a] Reference 66.

Table 8. Second, Third, and Fourth Ionization Energies (eV) Computed with the B3LYP Functional and the SARC Basis Sets using the DKH2 Hamiltonian, Compared with Experimental Values

	IE_2		IE_3		IE_4	
	calc	expt[a]	calc	expt[a]	calc	expt[a]
La	11.23	11.06	19.15	19.17	50.63	49.94
Ce	10.72	10.85	20.45	20.20	37.14	36.76
Pr	10.80	10.55	21.77	21.62	38.94	38.98
Nd	10.97	10.73	22.48	22.08	40.82	40.41
Pm	11.14	10.90	22.91	22.28	41.77	41.15
Sm	11.30	11.07	24.13	23.42	42.42	41.35
Eu	11.45	11.24	25.18	24.92	43.93	42.60
Gd	12.36	12.09	20.45	20.62	44.76	44.05
Tb	11.77	11.52	22.22	21.91	40.95	39.79
Dy	11.86	11.67	23.30	22.80	42.30	41.47
Ho	11.99	11.80	23.49	22.84	43.60	42.60
Er	12.12	11.93	23.48	22.74	43.93	42.65
Tm	12.25	12.05	24.50	23.68	44.04	42.69
Yb	12.38	12.19	25.41	25.03	45.26	43.74
Lu	14.03	13.89	21.27	21.07	45.74	45.19
MAD	0.21		0.41		0.86	

[a] Reference 1.

predicted ionization energies. The rms error is even marginally reduced by 0.01 eV, testifying to the well-balanced construction of the basis sets. It should be pointed out that the present all-electron B3LYP/SARC results compare very favorably with previously published high-level calculations of the first ionization energies of lanthanides. For example, multireference averaged coupled-pair functional (ACPF) calculations employing relativistic energy-consistent small-core pseudopotentials[24] with extensive valence basis sets of up to 114 functions achieved mean absolute deviations (MAD) in the range of 0.22−0.24 eV,[26,27] and more recent CASPT2 calculations with the ANO-RCC basis sets achieve an accuracy around 0.1 eV.[47]

For a more comprehensive evaluation of the SARC basis sets, we present the second, third, and fourth ionization energies of the lanthanide atoms in Table 8, computed with the B3LYP and the DKH2 Hamiltonian. Compared with the results obtained for IE_1, larger deviations are observed as successive electrons are removed, with a tendency toward overestimation of ionization energies. Nevertheless, the computed values are still satisfactory and compare well with the results of Cao and Dolg,[26] exactly matching the ACPF accuracy for the chemically most relevant third ionization potential. This is obviously of importance for most practical molecular applications, since the lanthanides are normally found in the +3 oxidation state. We emphasize that the experimental uncertainties associated with the ionization energies increase rapidly from IE_2 to IE_4, often exceeding 1 eV for IE_4 in the middle of the series. This, combined with the fact that the present DFT approach might not treat differential correlation effects on an equal footing, complicates the assessment of individual cases for IE_2 to IE_4. In contrast to the actinides, the effects of spin−orbit coupling in the lanthanides are not usually considered of crucial importance for the calculation of ionization energies because they are smaller in magnitude than the differential correlation effects arising from the different electronic configurations.[26,27,47,67] Note, however, that Liu and Dolg have shown that spin−orbit effects may become as large as 0.5 eV toward the end of the series when the f occupancy is changed.[68] Although inclusion of spin−orbit corrections would be necessary for high-accuracy work, especially when higher ionization energies are calculated, for the methods we use here (B3LYP), it is not expected that inclusion of spin−orbit corrections would lead to systematic improvement of the computed values.[26,27,47,67]

Regardless of the origin of specific deviations in absolute numerical values, the diagram of cumulative ionization energies in Figure 1 demonstrates that all experimentally observed trends across the series are faithfully reproduced by the calculations. Ionization energies generally increase from the lighter to the heavier elements, with specific irregularities related to filled and half-filled f subshell effects. The most characteristic deviations from the trend are the low IE_3 values for gadolinium and lutetium, which attain f^7 and f^{14} configurations at the +3 oxidation state, whereas europium and ytterbium display high IE_3 values owing to the loss of an electron from the corresponding f^7 and f^{14} configurations of their +2 state. Removal of a fourth electron is very costly for lanthanum because of its filled p subshell (xenon configuration) in the +3 oxidation state. On the other hand, low IE_4 values create two deep minima in the cumulative diagram for cerium (p^6 at the +4 state) and terbium (f^7 at the +4 state). This agrees perfectly with chemical facts since cerium, with an IE_4 of approximately 36.8 eV (calculated 37.1 eV), is the only element of the series that has extensive chemistry at the +4 oxidation state. Overall, the computed ionization energies closely follow the experimental patterns and do not reveal any bias for a particular electronic configuration, thus reinforcing our confidence in the ability of the SARC basis sets to cover all chemically relevant oxidation states across the entire 4f series.

Geometries of Trihalide Complexes. The lanthanide trihalides (LnX_3) form a complete and fairly well characterized class of compounds that encompass all elements of the

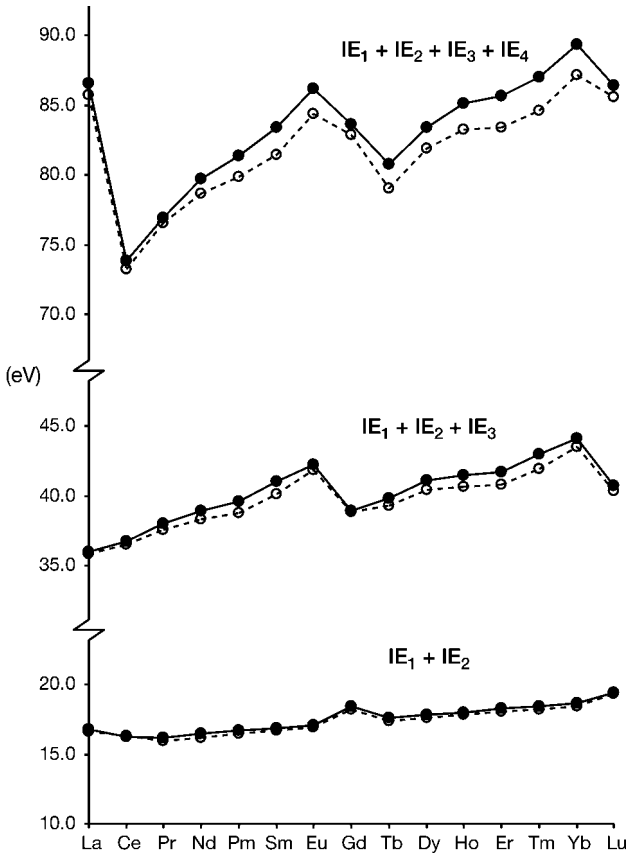


Figure 1. Cumulative ionization energies (eV) of the lanthanide atoms: comparison of experimental (dotted line) and calculated values (solid line). Calculations were performed with the SARC basis sets using the B3LYP functional and the DKH2 Hamiltonian.

4f series;[69] this makes them a suitable reference for testing the performance of the new basis sets with respect to molecular geometries. Several publications presenting theoretical studies of specific members of the family exist, and we refer the reader to the comprehensive review by Kovács and Konings for a detailed overview of the relevant literature.[69] However, to the best of our knowledge, the only study to explicitly address the complete series of lanthanides and halides is that of Cundari et al.,[70] who employed a multiconfigurational (MC) SCF approach with effective core potentials that leave the 4f electrons of the lanthanides in the valence space.[22] Since MCSCF does not attempt to account for dynamic correlation, to make a fair comparison in this paper we will use the same DFT method to compare values obtained with the ECPs and the SARC all-electron basis sets using the DKH2 Hamiltonian. For the ECP calculations, we use the Stuttgart−Bonn small-core (28 electrons in the core) pseudopotentials for the lanthanides[24] combined with the high-quality [10s8p5d4f3g] valence basis sets of Cao and Dolg.[27] We employ the hybrid (25% exact exchange) PBE0 density functional,[71] since this emerged as the top performer for geometries of transition metal complexes in our recent calibration study.[52] In the all-electron calculations, the DKH2 relativistically recontracted TZVP basis sets were used for the halides.[51] No symmetry restrictions were imposed on optimizations. As reference values, we use the recommended equilibrium bond distances proposed by Kovács and Konings in their survey of LnX_3 systems.[69] These values were obtained by a joint analysis of trends in available experimental and theoretical data and are currently regarded as the best available estimates.

Table 9. PBE0/SARC (DKH2) Equilibrium Angles (degrees) of Lanthanide Trifluorides and f-orbital Mulliken Population Analysis

	FLnF angle	f charge	f spin
Ce	113	1.4	1.0
Pr	115	2.4	2.0
Nd	118	3.4	3.0
Pm	120	4.3	4.0
Sm	120	5.3	5.0
Eu	114	6.3	6.0
Gd	117	7.2	6.9
Tb	115	8.2	5.9
Dy	117	9.2	4.9
Ho	120	10.2	3.9
Er	120	11.2	3.0
Tm	120	12.2	2.0
Yb	120	13.2	1.0
Lu	120	14.2	0.0

All Br, I, and Cl compounds are predicted by our PBE0 optimizations to be essentially planar, in agreement with available experimental data that support planar or quasiplanar equilibrium structures for the heavier halides.[69] By contrast, many fluorides converge to pyramidal geometries. Pyramidalization of the trifluoride molecules is a long-standing controversial issue that presents challenges for both experiment and theory. A critical review of experimental studies draws attention to the difficulties and ambiguities associated with the collection and the analysis of data concerning X-Ln-X angles,[69] while the sensitivity of computed parameters used methodological choices has not allowed researchers to reach definitive answers through quantum chemical approaches, either. It is accepted that the degree of pyramidalization in the trihalides depends on a subtle balance of factors that may include the polarizability of the lanthanide, the electronegativity and size of the halide, and the asphericity of the 4f electron shell, as suggested by Molnár and Hargittai.[32] In the case of the fluorides, Kovács and Konings suggest a uniform trend in the F−Ln−F angles with an increase in the Ln atomic number, based on steric considerations. This is at odds with the present PBE0/SARC-DKH2 results, which display marked discontinuities. LaF_3, CeF_3, PrF_3, and NdF_3 optimize to pyramidal geometries, but PmF_3 and SmF_3 are planar; deviations from planarity are again observed for the four subsequent lanthanides (Eu, Gd, Tb, Dy), whereas all trifluorides beyond dysprosium are planar. In Table 9, the angles of all trifluorides are given in detail along with the population analysis for the lanthanide f orbitals, which show a slightly higher occupancy than that expected from the formal ionic model.

We do not wish to place excessive confidence to the results concerning the angles, since this particular parameter is known to be sensitive to methodological choices. However, it is impossible to miss the extraordinary agreement with the predictions of the 4f asphericity model.[32] Considering the shape of the 4f shell, Molnár and Hargittai divided the Ln^{3+} cations into two groups: those with a spherically

Table 10. Equilibrium Bond Distances (pm) of Lanthanide Trihalides: PBE0/ECP and PBE0/SARC (DKH2) Results Compared with the Kovács−Konings Recommended Values

	recommended values[a]				PBE0/ECP				PBE0/SARC			
	F	Cl	Br	I	F	Cl	Br	I	F	Cl	Br	I
Ce	207	252	268	286	209	254	270	292	209	255	270	290
Pr	206	251	266	285	206	252	268	289	207	253	268	288
Nd	205	250	265	284	206	250	266	288	207	252	267	287
Pm	204	249	264	283	205	250	265	287	206	251	265	285
Sm	203	248	263	282	204	249	264	287	204	249	264	285
Eu	202	247	262	281	204	248	265	294	203	249	263	288
Gd	201	245	260	280	201	247	263	285	203	247	262	282
Tb	200	244	259	279	200	245	261	283	201	246	260	280
Dy	199	243	258	278	199	244	259	280	200	245	259	279
Ho	198	242	257	277	199	243	258	279	200	243	258	278
Er	197	241	255	276	199	242	257	278	199	242	257	277
Tm	196	240	254	275	198	241	256	277	198	241	256	276
Yb	195	238	253	274	197	240	255	277	197	240	255	276
Lu	194	237	252	273	196	239	253	275	196	239	254	274

[a] Ref 69; the estimated uncertainty is ± 2 pm.

Table 11. Atomization Energies (eV) of Lanthanide Trihalides: PBE0/SARC (DKH2) Results Compared with Experimental Values

	PBE0/SARC				expt[a]				ΔE			
	F	Cl	Br	I	F	Cl	Br	I	F	Cl	Br	I
Ce	19.8	15.4	13.5	12.5	20.1	15.6	13.6	11.2	–0.3	–0.1	–0.1	1.3
Pr	19.0	14.9	13.0	11.6	19.1	15.2	13.2	10.6	0.0	–0.3	–0.2	1.0
Nd	18.5	14.4	12.5	11.2	19.0	14.5	12.6	9.9	–0.6	–0.1	–0.1	1.3
Pm	17.8	13.7	11.8	10.6								
Sm	16.7	12.5	10.6	9.7	17.3	13.2	11.1	9.1	–0.6	–0.7	–0.5	0.6
Eu	15.8	11.8	10.5	8.8	17.2	13.1	10.8	8.5	–1.4	–1.3	–0.3	0.3
Gd	19.2	15.2	13.3	12.0	19.2	15.1	13.1	10.8	0.0	0.1	0.2	1.2
Tb	18.9	14.9	12.6	11.7	19.0	15.1	12.9	10.7	–0.1	–0.2	–0.3	1.0
Dy	17.9	13.9	11.9	10.7	17.3	14.2	11.8	9.7	0.6	–0.3	0.1	1.0
Ho	17.7	13.7	11.8	10.5	17.2	14.3	11.8	9.4	0.5	–0.6	0.0	1.1
Er	17.6	13.6	11.8	10.3	17.2	14.3	12.1	9.9	0.4	–0.7	–0.3	0.4
Tm	16.8	12.8	10.8	9.5	17.0	13.7	11.1	9.0	–0.3	–0.9	–0.3	0.5
Yb	15.8	11.6	9.9	8.6	16.0	12.4	10.1	7.8	–0.2	–0.8	–0.1	0.8
Lu	19.3	15.4	13.4	12.1	18.4	15.2	12.8	10.7	0.9	0.2	0.6	1.4

[a] Ref 77.

symmetrical or an axially elongated 4f shell, for which no distortion of the LnF_3 compounds from the planar geometry is anticipated (La^{3+}, Pm^{3+}, Sm^{3+}, Gd^{3+}, Er^{3+}, Tm^{3+}, Yb^{3+}, Lu^{3+}), and those with an axially compressed 4f shell, for which pyramidal LnF_3 geometries are expected (Ce^{3+}, Pr^{3+}, Nd^{3+}, Eu^{3+}, Tb^{3+}, Dy^{3+}, Ho^{3+}). Despite the fact that the coincidence between these predictions and the present results is not absolute, we suggest that the level of agreement provides strong support to the notion that a uniform periodicity in bond angles should *not* be expected for the trifluorides.

Focusing now on the bond lengths of the trihalides, the results of our PBE0/SARC-DKH2 optimizations are summarized in Table 10, where they are compared with the recommended values as well as with PBE0/ECP results. With both basis sets, a steady contraction of the Ln−X bond length is predicted with an increasing atomic number in the lanthanide. Both methods also yield almost identical total contractions from CeX_3 to LuX_3, 13−14 pm for the fluorides and 16−17 pm for the heavier halides. For each member of the four halide series, the SARC bond length approaches the Kovács−Konings reference value as closely as the ECP result, illustrating that the same level of accuracy in structural parameters can be expected from either approach. Remarkably, the maximum deviation in the F, Cl, and Br series never exceeds 2 pm. In the case of the iodides, an overestimation of the bond length by approximately 3 pm is evident up to Eu, but after this point the optimized values practically coincide with the reference. A point of particular importance is that the PBE0/SARC mean absolute deviation remains constant at 2 pm regardless of the nature of the halide. This encouraging result highlights the fact that the new basis sets combine well with the relativistically recontracted all-electron basis sets for main group elements that we presented in our previous contribution,[51] ensuring well-balanced and consistent performance not only across the periods but also down the groups of the periodic table.

In order to explore the dependence of the bond lengths on methodological choices, we have repeated the calculations with the nonhybrid (GGA) version of the functional, PBE.[72−74] On the whole, this approach yields similarly good results without significant deterioration compared to those of Table 10 (see Supporting Information), although mean average deviations rise to 2, 4, 4, and 5 pm for the four halide species, respectively. However, two discontinuities appear in the form of pronounced maxima at the europium and ytterbium compounds of all halides, followed by sharp contractions for the subsequent elements gadolinium and lutetium. These abrupt changes can become smaller but in

no case eliminated by extending the halide basis sets, whereas extending the SARC basis set has minimal effect. Hence, we are led to attribute the origin of these discrepancies to the imbalanced treatment of different electronic configurations by the GGA functional and/or its incorrect description of the covalency of the Ln-X bond.[75,76] As shown in Table 10, no such disruptions in the uniform contraction trend appear with the hybrid version of the functional, and only a small hump is noticeable for EuI_3.

Finally, the atomization energies of all species were computed at the same level of theory. The values are compared with experimental atomization energies obtained from Myers[77] in Table 11 and show reasonably good agreement with experiment. The largest deviations are observed for the iodides, where atomization energies are typically overestimated by 1 eV. In contrast, the mean absolute deviations for the F, Cl, and Br series are 0.4, 0.5, and 0.2 eV, respectively. In conclusion, it is clear from the results that a protocol based on the SARC basis sets in combination with the PBE0 functional performs accurately and consistently for the prediction of molecular properties across the lanthanide series.

Summary

Scalar all-electron relativistic (SARC) basis sets have been constructed for the accurate and the affordable treatment of lanthanide systems in conjunction with scalar relativistic Hamiltonians (DKH2 or ZORA). The SARC basis sets are small and compact, so they present a very efficient alternative to effective core potentials in routine DFT studies of chemically relevant systems. Their contraction pattern guarantees their computational efficiency compared to generally contracted relativistic basis sets. Extensive evaluation of the basis sets for the first four atomic ionization potentials of the lanthanides demonstrates that they provide a balanced description of different electronic configurations, not only reproducing the experimental trends but also achieving quantitative accuracy in most cases. Thus, they can be used with confidence for the prediction of energetic properties and the unbiased description of processes involving changes in oxidation state and associated changes in 4f and 5d occupation numbers. Moreover, the excellent results obtained with the SARC basis sets and the PBE0 density functional in a detailed study of the lanthanide trihalides confirm that the applicability of the basis sets can be safely extended to molecular systems. The new basis sets are particularly well suited for the calculation of molecular properties that require or benefit from the explicit treatment of the core electrons. These include not only the study of electron paramagnetic resonance, Mössbauer and X-ray absorption spectra, but also the derivation of electron densities that will be subsequently subjected to topological analysis and the study of the magnetic properties in mixed d/f heterometallic complexes and clusters.

Acknowledgment. We gratefully acknowledge financial support from the DFG priority program 1137 "Molecular Magnetism" and from the Max Planck Society.

Supporting Information Available: Full listings of the SARC basis sets. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cotton, S. *Lanthanide and Actinide Chemistry*. 2nd ed.; John Wiley & Sons: Chichester, U.K., 2006; p 280.

(2) *Lanthanides: Chemistry and Use in Organic Synthesis. Topics in Organometallic Chemistry*, Kobayashi, S., Ed.; Springer-Verlag: Berlin, Germany, 1999; Vol. 2, p 300.

(3) Imamoto, T. *Lanthanides in Organic Synthesis*. Academic Press: London, 1994; p 160.

(4) Bünzli, J.-C. G.; Piguet, C. *Chem. Soc. Rev.* **2005**, *34*, 1048–1077.

(5) Benelli, C.; Gatteschi, D. *Chem. Rev.* **2002**, *102*, 2369–2387.

(6) Bunzli, J. C. G.; Piguet, C. *Chem. Rev.* **2002**, *102*, 1897–1928.

(7) Sakamoto, M.; Manseki, K.; Okawa, H. *Coord. Chem. Rev.* **2001**, *219*, 379–414.

(8) Winpenny, R. E. P. *Chem. Soc. Rev.* **1998**, *27*, 447–452.

(9) Kido, J.; Okamoto, Y. *Chem. Rev.* **2002**, *102*, 2357–2368.

(10) Matsumoto K.; Yuan, J. G. Lanthanide Chelates as Fluorescent Labels for Diagnostics and Biotechnology. In *Metal Ions in Biological Systems*; Sigel, A., Sigel, H., Eds. Marcel Dekker Inc.: New York, 2003; Vol. 40, pp 191−231.

(11) Shen, J.; Sun, L.-D.; Yan, C.-H. *Dalton Trans.* **2008**, 5687–5697.

(12) Bünzli, J.-C. G. *Chem. Lett.* **2009**, *38*, 104–109.

(13) Fricker, S. P. *Chem. Soc. Rev.* **2006**, *35*, 524–533.

(14) Caravan, P.; Ellison, J. J.; McMurry, T. J.; Lauffer, R. B. *Chem. Rev.* **1999**, *99*, 2293–2352.

(15) Bottrill, M.; Kwok, L.; Long, N. J. *Chem. Soc. Rev.* **2006**, *35*, 557–571.

(16) Hermann, P.; Kotek, J.; Kubicek, V.; Lukes, I. *Dalton Trans.* **2008**, 3027–3047.

(17) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563–594.

(18) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*. 2nd ed.; Wiley-VCH: Weinheim, Germany, 2001; p 300.

(19) Parr, R. G.; Yang, W., *Density-Functional Theory of Atoms and Molecules*. Oxford University Press: Oxford, U.K., 1989; p 352.

(20) Dolg, M. Effective core potentials. In *Modern Methods and Algorithms of Quantum Chemistry*, 2 ed.; Grotendorst, J., Ed. John von Neumann Institute for Computing: Jülich, Germany, 2000; Vol. 3, pp 507−540.

(21) Ross, R. B.; Gayen, S.; Ermler, W. C. *J. Chem. Phys.* **1994**, *100*, 8145–8155.

(22) Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993**, *98*, 5555–5565.

(23) Dolg, M.; Stoll, H.; Savin, A.; Preuss, H. *Theor. Chem. Acc.* **1989**, *75*, 173–194.

(24) Dolg, M.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1989**, *90*, 1730–1734.

(25) Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chem. Acc.* **1993**, *85*, 441–450.

(26) Cao, X.; Dolg, M.*J. Chem. Phys.* **2001**, *115*, 7348–7355.

(27) Cao, X.; Dolg, M. *J. Mol. Struct. (THEOCHEM)* **2002**, *581*, 139–147.

(28) Yang, J.; Dolg, M. *Theor. Chem. Acc.* **2005**, *113*, 212–224.

(29) Hülsen, M.; Weigand, A.; Dolg, M. *Theor. Chem. Acc.* **2009**, *122*, 23–29.

(30) Dolg, M.; Stoll, H. Electronic structure calculations for molecules containing lanthanide atoms. In *Handbook of Chemistry and Physics of Rare Earths*, Gschneider, K. A., Eyring, L., Eds.; Elsevier: Amsterdam, The Netherlands, 1996; Vol. 22, pp 607−729.

(31) Frenking, G.; Antes, I.; Böhme, M.; Dapprich, S.; Ehlers, A. W.; Jonas, V.; Neuhaus, A.; Otto, M.; Stegmann, R.; Veldkamp, A.; Vyboishchikov, S. F. *Rev. Comp. Chem.* **1996**, *8*, 63–143.

(32) Molnár, J.; Hargittai, M. *J. Phys. Chem.* **1995**, *99*, 10780–10784.

(33) Güell, M.; Luis, J. M.; Solà, M.; Swart, M. *J. Phys. Chem. A* **2008**, *112*, 6384–6391.

(34) Vyboishchikov, S. F.; Sierraalta, A.; Frenking, G. *J. Comput. Chem.* **1997**, *18*, 416–429.

(35) Cirera, J.; Ruiz, E. *C. R. Chim.* **2008**, *11*, 1227–1234.

(36) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783–9792.

(37) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505–6516.

(38) van Wüllen, C. *J. Chem. Phys.* **1998**, *109*, 392–399.

(39) Dyall, K. G.; van Lenthe, E. *J. Chem. Phys.* **1999**, *111*, 1366–1372.

(40) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89–155.

(41) Hess, B. A. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *32*, 756–763.

(42) Hess, B. A. *Phys. Rev. A: At., Mol., Opt. Phys.* **1986**, *33*, 3742–3748.

(43) Jansen, G.; Hess, B. A. *Phys. Rev. A: At., Mol., Opt. Phys.* **1989**, *39*, 6016–6017.

(44) Wolf, A.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 9215–9226.

(45) *Amsterdam Density Functional (ADF)*, 2007.01; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2007; http://www.scm.com. Accessed September 10, 2008.

(46) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.

(47) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O.; Borin, A. C. *J. Phys. Chem. A* **2008**, *112*, 11431–11435.

(48) Tsuchiya, T.; Abe, M.; Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 4463–4472.

(49) Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2002**, *116*, 8270–8275.

(50) Sekiya, M.; Noro, T.; Miyoshi, E.; Osanai, Y.; Koga, T. *J. Comput. Chem.* **2006**, *27*, 463–470.

(51) Pantazis, D. A.; Chen, X. Y.; Landis, C. R.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 908–919.

(52) Bühl, M.; Reimann, C.; Pantazis, D. A.; Bredow, T.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 1449–1459.

(53) Neese, F. *ORCA - an ab initio, Density Functional and Semiempirical Program Package* , Version 2.6−35 Universität Bonn: Bonn, Germany, 2007.

(54) Stavrev, K. K.; Zerner, M. C. *Int. J. Quantum Chem.* **1997**, *65*, 877–884.

(55) Zerner, M. C. *Int. J. Quantum Chem.* **1989**, *35*, 567–575.

(56) Huzinaga, S.; Kolbukowski, M. *Chem. Phys. Lett.* **1993**, *212*, 260–264.

(57) Huzinaga, S.; Miguel, B. *Chem. Phys. Lett.* **1990**, *175*, 289–291.

(58) Jorge, F. E.; de Castro, E. V. R.; da Silva, A. B. F. *J. Comput. Chem.* **1997**, *18*, 1565–1569.

(59) de Castro, E. V. R.; Jorge, F. E. *J. Chem. Phys.* **1998**, *108*, 5225–5229.

(60) Koga, T.; Watanabe, S.; Thakkar, A. J. *Int. J. Quantum Chem.* **1995**, *54*, 261–263.

(61) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(62) Küchle, W.; Dolg, M.; Stoll, H. *J. Phys. Chem. A* **1997**, *101*, 7128–7133.

(63) Huber, H. P.; Herzberg, G. *Constants of diatomic molecules*. Van Nostrand: New York, 1979; p 565.

(64) Ram, R. S.; Bernath, P. F. *J. Chem. Phys.* **1996**, *104*, 6444–6451.

(65) Morgan, J. D., III.; Kutzelnigg, W. *J. Phys. Chem.* **1993**, *97*, 2425–2434.

(66) *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Ed.; National Institute of Standards and Technology: Gaithersburg, MD, 2005.

(67) Adamo, C.; Maldivi, P. *J. Phys. Chem. A* **1998**, *102*, 6812–6820.

(68) Liu, W.; Dolg, M. *Phys. Rev. A: At., Mol., Opt. Phys.* **1998**, *57*, 1721–1728.

(69) Kovács, A.; Konings, R. J. M. *J. Phys. Chem. Ref. Data* **2004**, *33*, 377–404.

(70) Cundari, T. R.; Sommerer, S. O.; Strohecker, L. A.; Tippett, L. *J. Chem. Phys.* **1995**, *103*, 7058–7063.

(71) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(72) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(73) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(74) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B* **1996**, *54*, 16533–16539.

(75) Dolg, M.; Liu, W.; Kalvoda, S. *Int. J. Quantum Chem.* **2000**, *76*, 359–370.

(76) Ramakrishnan, R.; Matveev, A. V.; Rösch, N. *Chem. Phys. Lett.* **2009**, *468*, 158–161.

(77) Myers, C. E. *Inorg. Chem.* **1975**, *14*, 199–201.

CT900090F

JCTC Journal of Chemical Theory and Computation

An Ab Initio Study of the Structures and Selected Properties of 1,2-Dihydro-1,2-azaborine and Related Molecules

Janet E. Del Bene,*,† Manuel Yáñez,‡ Ibon Alkorta,*,§ and José Elguero§

Department of Chemistry, Youngstown State University, One University Plaza, Youngstown, Ohio 44555, Departamento de Química, C-9, Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain, and Instituto de Química Médica, Consejo Superior de Investigaciones Cientfícas, Juan de la Cierva, 3, E-28006 Madrid, Spain

Received March 18, 2009

Abstract: An ab initio study has been carried out to investigate the effect of replacing $[HC-CH]_n$ linkages in benzene by the isoelectronic $[HN-BH]_n$ linkages for $n = 1$, 2, and 3. Such replacements give rise to azaborine, a set of diazaborines, borazine, and pseudoborazine. These replacements lead to significant rearrangements of electron densities in these molecules due primarily to the introduction of the polar B−N bond. As a result, azaborine and diazaborines exhibit much more localized structures than that of benzene. They are also less aromatic than benzene but have a higher degree of aromaticity than borazine. The bonding patterns can be related to the relative stabilities of the diazaborines. Among these molecules, the most stable isomer contains an N−B−N−B linkage, while the two least stable isomers have either a B−B or a N−N bond. Changes in bonding patterns are also reflected in changes in the N1−B2 coupling constant. When N1 and B2 are bonded to the less electronegative atoms C and B, ^1J(N1−B2) increases relative to borazine, but when either N1 or B2 is bonded to N, ^1J(N1−B2) decreases. Computed NMR chemical shifts and coupling constants are in good agreement with available experimental data.

1. Introduction

There has been a long-standing fascination with molecules in which a HC−CH group is replaced by a HB−NH one, giving an isoelectronic molecule with a B−N bond. A fundamental question is how similar or dissimilar are these two molecules?[1] Obviously, the C−C bond is nonpolar while the B−N bond is very polar, since the B atom is significantly less electronegative than the C atom, whereas the N atom has a higher electronegativity. The introduction of the polar B−N bond can significantly alter the electronic properties of the system. For example, while graphite has a very high conductivity, the isoelectronic boron nitride, which has the graphite structure, is an insulator. Moreover, in a previous paper,[2] we have shown that substituent effects on the bonding properties and vibrational frequencies of iminoboranes are not only different but also opposite at times to substituent effects on the corresponding acetylene derivatives. The dissimilarities between acetylene and iminoborane derivatives are primarily a consequence of the difference in the electronegativities of B and N, which leads to a significant distortion of electron density in the B−N bonding region compared to that of the C−C region.

Of particular interest is the replacement of one HC−CH group of benzene (**1**) by one HN−BH group to form 1,2-dihydro-1,2-azaborine (**2**),[3,4] or the replacement of two HC−CH groups by two HN−BH groups to form the diazaborines (**3**−**7**) and three HC−CH groups to give borazine (**8**) or pseudoborazine (**9**). These molecules are

* Corresponding authors: E-mail: jedelbene@ysu.edu (J.E.D.B.) and ibon@iqm.csic.es (I.A.).

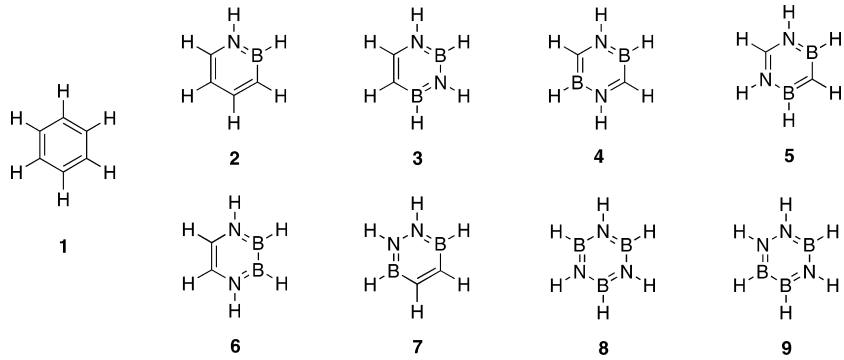
† Youngstown State University.

‡ Universidad Autónoma de Madrid.

§ Consejo Superior de Investigaciones Cientfícas.

10.1021/ct900128v CCC: $40.75
Published on Web 07/31/2009

Chart 1. Benzene and the Azaborines



shown in Chart 1. It is interesting to note that while borazine (**8**) has been known experimentally since 1926,[5] 1,2-dihydro-1,2-azaborine (**2**) was only recently prepared in 2008 by Liu et al.[1] and subsequently investigated by other authors.[6−8] On the basis of their theoretical and experimental data, these authors concluded that 1,2-azaborine has significant aromaticity.[8]

The number of compounds which have substituents on the N and B atoms of rings containing B−N bonds are legion,[9,10] but the parent compounds in which only H atoms are present are not common. For the most part, studies of such molecules have been restricted to theoretical investigations. A MP2/6-31G(d) study of azaborines, including **2**, was reported by Kranz and Clark,[11] who concluded that these compounds present a considerable degree of electron delocalization. The most stable isomer arises when B and N are directly bonded (1,2), with the stability decreasing in the order of 1,2 (**2**) ≫ 1,4 > 1,3. According to Schleyer et al., borazine (**8**) is nonaromatic due to the polar BN bonds and shows little or no evidence of ring currents.[12] In contrast, the harmonic oscillator model of aromaticity (HOMA) and Bird aromaticity indexes calculated by Krygowski and co-workers indicate that the aromaticity of borazine (**8**) is only slightly less than that of benzene.[13] Kar, Elmore, and Scheiner carried out MP2/6-31+G(d,p) studies of B- and N-containing molecules, including several of those illustrated in Chart 1, and concluded that molecules with adjacent B−N bonds are more stable than other isomers.[14] Consistent with this, Dixon and co-workers concluded that **2** was aromatic on the basis of high-level computational data and experimental results.[8] Doerksen and Thakkar calculated the vibrational frequencies and polarizabilities of the compounds shown in Chart 1 as well as azaborinines with nonadjacent B and N atoms.[15] In 2009, Bosdet and Piers published a review in which molecule **2** and N- and/or B-substituted derivatives are described as well as the structures of molecules related to **3** and **5**−**7**.[16] Didehydro derivatives (benzyne derivatives) of some of the molecules in Chart 1 were investigated by Fazen and Burke at various levels of theory up to and including coupled-cluster singles and doubles method, CCSD(T).[17] Ab initio calculations up to and including CCSD(T)/CBS have been used to obtain reliable thermochemical data for borazine and related derivatives.[18] We previously reported an investigation of neutral and anionic BN-containing five-member rings and found that they exhibit a significant degree of electron delocalization.[19] In the present paper, we extend our studies of BN-containing molecules and report a systematic analysis of the structures, energies, bonding, aromaticity, and NMR properties of six-member rings (**2**−**9**) containing one, two, or three HB−NH groups.

2. Theoretical Methods and Computational Details

Molecular geometries have been optimized at second-order Møller−Plesset perturbation theory (MP2)[20−23] with the 6-311+G(d,p) basis set.[24,25] Frequency calculations were performed to confirm that these structures are local minima on their potential surfaces. Improved energies were obtained using the composite G3B3 method implemented in Gaussian-03. All calculations were carried out using the Gaussian-03 package.[26]

The electron densities of these molecules have been analyzed using the atoms in molecules (AIM) theory as implemented in the AIMPAC[27] and AIMAll programs.[28] Calculations of atomic properties were carried out by integration within the atomic basins using the default parameters in these packages, except in those cases where the integrated Laplacian was less than 1×10^{-3} au, and tighter criteria was employed. Previous reports have shown that small errors can occur in the energy and charges of systems if the values of the integrated Laplacian are less than those of the default parameters.[29]

Further analyses of the bonding in these systems were carried out by means of the natural bond orbital (NBO) method.[30] This method offers a picture of the bonding of a given compound as a combination of localized hybrids and lone pairs, which can be obtained by block-diagonalizing the one-particle density matrix. In this analysis, we have allowed for the detection of three center bonds, which are rather common in B-containing compounds. Further insight into the bonding characteristics was obtained by examining interactions between occupied and unoccupied molecular orbitals (MOs) through a second-order perturbation analysis of the Fock matrix and computing Wiberg bond orders. Since the systems under investigation may have significant aromatic character, we have also used the natural resonance theory (NRT)[31] to estimate the contributions of different resonance structures and the corresponding delocalization energy. The delocalization energy is defined as

$$E_{\mathrm{deloc}} = E_{\mathrm{total}} - E_{\mathrm{Lewis}} \qquad (1)$$

where E_{total} is the total energy of the compound, and E_{Lewis} is the energy of a hypothetical Lewis molecule with strictly localized bonds. E_{Lewis} is obtained by removing all off-diagonal elements from the Fock matrix and computing one self-consistent field (SCF) cycle.

The absolute chemical shieldings have been calculated at MP2/6-311+G(d,p) with the GIAO computational method, as implemented in Gaussian-03. The nuclear independent chemical shift (NICS), an aromaticity index based on magnetic criteria, has been evaluated for molecules **1**−**9**. This index is defined as the negative absolute magnetic shielding computed at the center of the ring.[32] The NICS(1) index is calculated 1 Å above the ring center.[33] Rings with highly negative NICS values are aromatic, whereas those with positive values are antiaromatic. The so-called *para*-delocalization index (PDI), proposed by Solà and co-workers, has also been evaluated.[34] This index is defined as the average of Bader's electron delocalization index (DI) of atoms, which are in para positions in six-member rings. This index has been found to correlate strongly with HOMA, the latter being inapplicable for the molecules of interest in this study because of the lack of reference values for B−N bonds.

Spin−spin coupling constants were computed for molecules **2**−**9** using the equation-of-motion coupled-cluster singles and doubles method (EOM-CCSD) in the configuration interaction (CI)-like approximation[35−38] with all electrons correlated. The Ahlrichs qzp basis set[39] was used on ^{13}C and ^{15}N atoms, and the qz2p basis set was used for all ^{1}H atoms. The recently constructed hybrid basis set was used for ^{11}B.[2] This basis set has the same number of contracted functions (6s, 4p, and 1d) as the Ahlrichs qzp basis for C and N and was used previously in studies of B−N, B−H, and B−Li coupling constants.[40−43] In the nonrelativistic approximation, the nuclear spin−spin coupling constant is composed of four terms: the paramagnetic spin−orbit (PSO), diamagnetic spin−orbit (DSO), Fermi-contact (FC), and spin−dipole (SD).[44] All of the terms have been computed for all of the molecules. The EOM-CCSD calculations were carried out using ACES II[45] on the Itanium cluster at the Ohio Supercomputer Center.

3. Results and Discussion

The optimized bond distances for compounds **2**−**9** are given in Figure 1. The MP2/cc-pVTZ results previously reported[8] for **2** are in good agreement with those given here. All molecules in Figure 1 have planar rings except for the molecule **9**. Total energies for molecules **2**−**9**, relative energies, and enthalpies of formation for molecules **2**−**7** and **9** are reported in Table 1. The enthalpies were computed for the isodesmic reactions (eqs 2 and 3) at the G3B3 computational level, using the experimental enthalpy of formation for molecules **1** and **8** (82.8 and −510.03 kJ mol^{-1}, respectively[46]). Due to the errors associated with the heat of formation of molecule **8** (±12 kJ mol^{-1}) and the isodesmic energies at this level (from 4 to 8 kJ mol^{-1}), the calculated heats of formation are estimated to have error bars of ±20 kJ mol^{-1}.

(**2**) $$2C_6H_6 + B_3N_3H_6 \rightarrow 3C_4H_6BN \quad (2)$$

(**3**−**7**) $$C_6H_6 + 2B_3N_3H_6 \rightarrow 3C_2H_6B_2N_2 \quad (3)$$

Substitution of a C−C bond in benzene by a B−N bond leads to a significant rearrangement of electron density due to the polarity of the BN bond. In molecule **2**, the natural net charges on B and N are +0.48 and −0.73e, respectively. With the AIM partitioning, these values are +2.00 and −1.47e, respectively, and are in agreement with previous reports that show very large charges within the AIM methodology for bonds between atoms with very different electronegativities.[47] The Laplacian of the electron density at the BN bcp of **2** is positive, as it is in all molecules **2**−**9**. However, the energy density is negative, indicating that although the bond is very polar, the potential energy density dominates over the kinetic energy density as in typical covalent interactions. Replacement of one or two HC−CH linkages by HB−NH does not destroy the planarity of azaborine or diazaborines. However, pseudoborazine (**9**) is slightly nonplanar. Planar molecule **9** has one imaginary

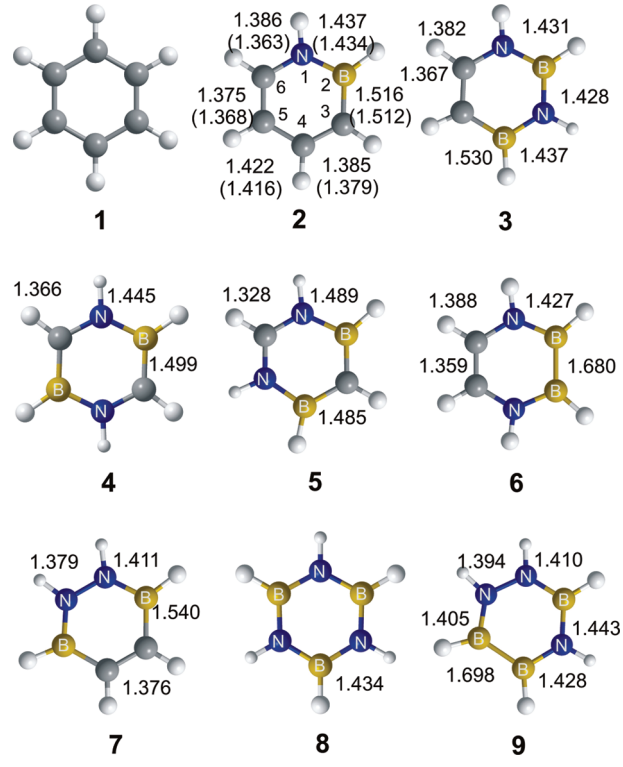


Figure 1. Optimized MP2/6-311+G(d,p) bond distances (Å). The atomic numbering used for these molecules is indicated in **2**. The MP2/cc-pVTZ results reported in ref 8 for compound **2** are given in parentheses.

Table 1. Total MP2/6-311+G(d,p) Energies, Relative Energies, and G3B3 ΔH°_f Values

molecule	symmetry	E_{total} (Hartree)	E_{rel} (kJ mol^{-1})	ΔH°_f (kJ mol^{-1})
2	C_s	−235.009313	–	4.8[a]
3	C_s	−238.484585	0.0	−200.8
4	C_{2h}	−238.424084	158.8	−21.9
5	C_{2v}	−238.428974	146.0	−55.8
6	C_{2v}	−238.412937	188.1	−17.9
7	C_{2v}	−238.399569	223.2	13.4
8	D_{3h}	−242.002367	0.0	–
9	C_1	−241.850656	398.3	−127.5

[a] The predicted ΔH°_f from ref 8 is 12.55 kJ mol^{-1}.

Table 2. MP2/6-311+G(d,p) Aromaticity Indexes

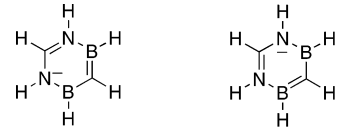
molecule	NICS(0)	NICS(1)	PDI
1	−7.91 (−8.76)[a]	−10.20 (−10.39)[a]	0.059
2	−5.39 (−5.62)[a]	−7.54 (−7.27)[a]	0.049
3	−2.50	−4.29	0.026
4	−4.89	−6.92	0.064
5	−4.10	−6.13	0.037
6	−2.22	−4.38	0.037
7	−2.73	−4.93	0.041
8	−1.50 (−2.02)[a]	−2.51 (−3.01)[a]	0.012
9	−0.51	−2.30	0.023

[a] Values from ref 8.

frequency of −203 cm^{-1} but is only 0.07 kJ mol^{-1} less stable than the equilibrium C_1 structure. As shown previously for five-member rings,[19] one of the most characteristic signatures of a B−N bond is its ability to act as either an electron donor since N is electron rich or an electron acceptor since B is electron poor. This is indeed the case for compound **2**. According to a NRT analysis, although the dominant resonance form is the Kekulé-type structure with a weight of 57%, there are also non-negligible contributions from other resonance forms which involve electron donations from the π_{C3C4} bonding orbital to the π^*_{BN} antibonding orbital and the π_{BN} bonding orbital to the π^*_{C5C6} antibonding orbital. As a result, the BN bond order (1.02) is smaller than that of a typical B═N double bond (1.257 in H_2BNH_2) and close to that of borazine (1.01). In addition, the bond orders of the C3C4 and C5C6 linkages (1.62 and 1.58, respectively) are less than those of typical C═C bonds but greater than that of benzene (1.44). The electron densities at the corresponding bcps are also greater (0.309 and 0.317 au, vs. 0.303 au). A similar conclusion can be derived from the analysis of the delocalization indexes since, as has been shown previously, they are linearly correlated with the bond orders.[48] Hence, the presence of the BN linkage produces a more localized structure than that found in benzene and is reflected in a decrease of about 20% in the delocalization energy. Similar decreases are also found in the absolute values of the NICS and the PDI indexes, as evident from Table 2. Nevertheless, these data indicate that **2** has a significant aromatic character, in agreement with the observations made previously.[1,6−8]

When two HB−NH groups are introduced into the ring, five different isomers (**3**−**7**) can be formed. The most stable isomer (**3**) corresponds to the one in which the two groups are bonded in the order of N−B−N−B, while the least stable isomers have B−B or N−N bonds, as in **6** and **7**, respectively. The NICS and PDI indexes reported in Table 2 clearly show a decrease in aromaticity going from **2** to **3.** The delocalization energy also decreases by 9%, although **3** still exhibits a certain degree of delocalization as indicated by the existence of BNB and NCC three center bonds. Although the NICS and PDI indexes indicate that the aromaticity of **4** and **5** increases relative to **3**, both species are predicted to be much less stable. This may be attributed to the replacement of stronger C−C by weaker C−B bonds. In addition, the existence of a N−B−N−B linkage in **3** also results in the formation of a very strong C═C bond with a bond order of 1.70, which is intermediate between a pure double bond (ethylene bond order of 2.00) and an aromatic C−C bond of benzene (1.44). The slightly greater stability of **5** relative to **4** can be attributed to the strengths of the C−N and C−B bonds in **5** (bond orders of 1.38 and 1.32, respectively) compared to **4** (1.30 and 1.22, respectively). Similar trends are observed for the electron densities at the bcps. In **5**, the electron densities at the C−N and C−B bcps are 0.340 and 0.202 au, respectively, compared to 0.300 and 0.191 au, respectively, for **4.** These bonding differences also reflect the significant contributions (with a weight of 20% each) of the mesomeric forms **5a** and **5b,** shown in Chart 2, which are favored by the presence of a N−C−N linkage. As might be expected, the B−N bonds have very small bond orders (0.846) as a consequence of the N−C−N and B−C−B linkages.

Chart 2. Mesomeric Forms of Molecule 5



5a **5b**

Only in the least stable isomers **6** and **7** are no three center bonds detected in the NBO analyses, although both NICS(1) and PDI indexes suggest a slight increase in aromaticity relative to **3**. In contrast, the delocalization energies of these two molecules are 9% and 14% lower than that of **3**. In any case, it seems clear that aromaticity changes do not correlate with relative stabilities. The lower stabilities of **6** and **7** are mainly due to the presence of a B−B and a N−N bond, respectively. A B−B bond is weaker than a C−B bond, and in addition, both B atoms are bonded to N atoms. This produces a net positive charge (+0.27 e) on both B atoms and a significant electrostatic repulsion between them, which destabilizes **6** relative to **3**. This effect is even stronger for **7**, since in this isomer the negative net charge accumulated on the two bonded N atoms is much larger in absolute value (−0.56 e), and this again results in an increased electrostatic repulsion. As expected, the B−N bond orders (BO) in **6** and **7** (1.13 and 1.14, respectively) are intermediate between H_2BNH_2 (1.257) and borazine (1.02).

When all the C−C bonds of the benzene ring are replaced by B−N bonds, two possible isomers, **8** and **9**, can arise. The structure and aromaticity of borazine (**8**) has been the subject of many previous studies and will not be discussed here. It is worth mentioning, however, that its low aromaticity has been explained in terms of the polarity of the B−N bonds.[12] The very low stability of **9** is clearly due to the presence of one B−B and one N−N bond within the ring, a situation which results in large electrostatic repulsions. No three center bonds are found in **9**, and the calculated delocalization energy is about 19% less than that of borazine. Although its NICS values are reduced, its PDI is almost twice that of borazine. Of the three nonequivalent B−N bonds in **9**, the one exhibiting the most double-bond character is B5−N6 (BO = 1.203), since it is located between a B−B and a N−N bond. The B2−N3 bond has the smallest bond order (0.979) and is part of the N−B−N−B linkage.

The B−N bond lengths vary significantly from 1.405 Å in **9** to 1.489 Å in **5**, as can be seen from Figure 1. The

Chart 3. Dominant Mesomeric Forms of Molecules 6, 7, and 9

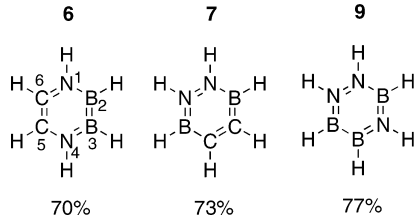
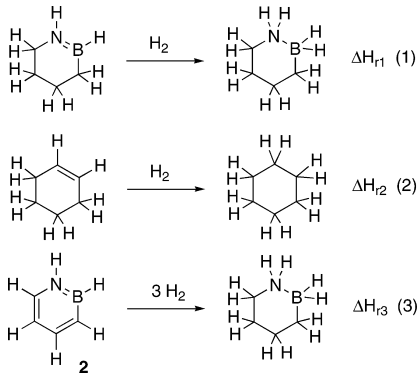


Chart 4. Reaction Enthalpies



shortest B−N bonds have the N atom bonded to another N atom as in **7** and **9**. Conversely, the longest B−N bonds have B and N bonded to C atoms as in **4** and **5**. Insight into these trends can be obtained by recalling that B−B or N−N bonds within the ring significantly favor a Kekulé-type structure, which makes these bonds single bonds and the B−N bonds formally double bonds. Consistently, in **6**, **7**, and **9**, the mesomeric structures shown in Chart 3 dominate with weights of 70−80%. In contrast, the B−N bonds of **5** are long due to the presence of N−C−N and B−C−B linkages, which favor electron delocalization and leave the B−N bond a single bond.

The PDI index is obtained as an average of three individual delocalization values which may vary over a wide range. For the molecules included in this study, the largest delocalization values correspond to those involving a pair of carbon atoms. The maximum value of the DI is that of the *para*-C−C pair in **4**, with a value of 0.123. In contrast, the smallest values are associated with the boron atoms with DI values of B−B pairs ranging between 0.004 and 0.018 in **7** and **3**, respectively.

The resonance energy of benzene is usually calculated as the difference between its hydrogenation enthalpy, i.e., the energy released when benzene is converted to cyclohexane, and the hydrogenation enthalpy of cyclohexatriene, the energy released when hypothetical C_6H_6 with three localized C═C double bonds becomes fully saturated. The hydrogenation enthalpy of the latter is taken as three times the enthalpy change when cyclohexene is converted to cyclohexane.

A similar approach could be used to estimate the resonance energies of the azaborines. Since azaborines contain both C═C and B═N bonds, it is necessary to also evaluate the hydrogenation energy of a B═N bond. This can be done as illustrated by the reactions in Chart 4 for the molecule **2**. ΔH_{r1} is the enthalpy change when the B═N bond becomes

Chart 5. Resonance Energy of Azaborine

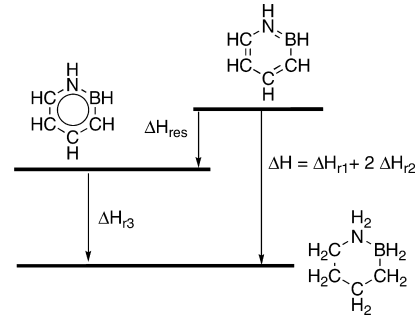


Table 3. G3B3 Resonance Enthalpies (kJ mol^{-1})[a]

compound	resonance enthalpy
2	96
3	57
6	82
7	79
8	6
9	59

[a] The resonance enthalpy for benzene at the same level of theory is 201 kJ mol^{-1}.

hydrogenated. The hydrogenation enthalpy of one C═C double bond is the energy released in reaction 2, ΔH_{r2}, as in benzene. Finally, ΔH_{r3} is the energy released when azaborine is fully hydrogenated. The resonance energy then is simply $\Delta H_{res} = \Delta H_{r1} + 2\ \Delta H_{r2} - \Delta H_{r3}$, as illustrated in Chart 5. By adjusting the weighting coefficients of ΔH_{r1} and ΔH_{r2}, this scheme may also be extended to estimate the resonance energies of compounds **3** and **6**−**9**. However, a similar scheme cannot be applied to compounds **4** and **5** since there are no C−C linkages.

The calculated resonance enthalpies are summarized in Table 3. It is interesting to note that the resonance energy is not negligible for compound **2**, although it is less than half of that calculated for benzene at the same level of theory. This is consistent with a certain aromaticity inherent in the NBO description above and with both the PDI and NICS(1) values. The resonance energies computed for **3, 6**, and **7** clearly indicate a decrease in aromaticity relative to **2** and are consistent with both our previous observations and the calculated PDI and NICS(1) values. The resonance energies also increase slightly in going from **3** to **6**, but decrease slightly going from **6** to **7**, in contrast to the PDI and NICS(1) results.

The resonance energy of borazine suggests that it does have a certain degree of aromaticity, although it is much less than that of compound **2**, in agreement with PDI and NICS(1) descriptions. Also, the resonance energy of borazine is definitely smaller than that of compounds **3, 6**, and **7**, again in agreement with the other aromaticity indexes [PDI and NICS(1)].

Finally, the resonance energy of compound **9** indicates that it is much more aromatic than borazine. This is not consistent with both the dominant weight of the localized resonant forms discussed above or the NICS(1) values, which predict a similar aromaticity for both systems. It should be recognized, however, that the model used to obtain the resonance energies of azaborine, the diazaborines, and borazine has

Table 4. MP2/6-311+G(d,p) Calculated Chemical Shifts (δ, ppm) for Molecules **2**−**9** and Experimental Chemical Shifts for Molecules **1**, **2**, and **8**

atom	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**1** exp.	**2** exp.	**8** exp.	**1**[a]	**2**[a]	**8**[a]
X1	130.04	−203.98	−235.56	−173.07	−205.75	−204.62	−209.84	−250.85	−234.64	128.5		−265.8	135.2	−246.9	−287.6
X2	130.04	26.99	23.57	20.75	29.38	38.56	23.97	27.15	21.24	128.5	31	30.2	135.2	26.9	26.1
X3	130.04	136.21	−225.76	146.19	124.92	38.56	150.19	−250.85	−218.26	128.5	131.6		135.2	139.0	−287.6
X4	130.04	142.08	33.58	−173.07	29.38	−204.62	150.19	27.15	45.58	128.5	144.5	30.2	135.2	151.5	26.1
X5	130.04	116.09	112.89	20.75	−205.75	119.06	23.97	−250.85	30.77	128.5	112.1		135.2	118.8	−287.6
X6	130.04	132.44	145.00	146.19	140.80	119.06	−209.84	27.15	−203.99	128.5	134.7	30.2	135.2	140.2	26.1
H1	7.75	8.42	6.56	9.73	8.08	8.28	7.66	5.58	6.32	7.26	8.44	5.42	7.5	7.8	5.3
H2	7.75	5.49	5.00	5.29	5.51	5.74	4.99	4.94	4.65	7.26	4.90	4.46	7.5	5.4	5.0
H3	7.75	7.61	7.05	8.03	6.99	5.74	8.01	5.58	7.34	7.26	6.92	5.42	7.5	7.4	5.3
H4	7.75	8.04	5.58	9.73	5.51	8.28	8.01	4.94	6.03	7.26	7.70	4.46	7.5	8.0	5.0
H5	7.75	6.83	6.12	5.29	8.08	6.64	4.99	5.58	4.96	7.26	6.43	5.42	7.5	6.6	5.3
H6	7.75	7.52	7.57	8.03	7.75	6.64	7.66	4.94	7.61	7.26	7.40	4.46	7.5	7.3	5.0

[a] B3LYP/Ahlrichs-vtzp values from ref 8.

certain limitations. Although the number of B═N and C═C bonds are accounted for, their bond orders are not those of the reference compounds. Moreover, some of the molecules considered in this study also have B−B or N−N bonds which are not present in the model compounds. These limitations mean that the calculated resonance energies are only approximate.

NMR Chemical Shifts. The calculated absolute chemical shieldings (σ) of molecules **2**−**9** have been transformed to the corresponding chemical shifts (δ) using those obtained for ^{1}H and ^{13}C (in TMS), ^{15}N (NH_3), and ^{11}B (from $BF_3 \cdot OEt_2$ in the gas-phase for which the experimental value is 9.4 ppm).[49] The calculated B3LYP/Ahlrichs-vtzp values from ref 8, those obtained in the present work, and the available experimental values are given in Table 4. Experimental values are available for two of the azaborines, namely **2** and **8**.[8,50] The only value of an experimental ^{15}N chemical shift for the molecules included in this study was measured for borazine (**8**) and was −265.8 ppm.[51] ^{11}B chemical shifts have been obtained for **2** and **8**, and are 31 and 30.2 ppm, respectively.[8,50] Thus, the computed values are consistent with the experimental.

Good correlations are obtained between the experimental and calculated ^{1}H and ^{13}C chemical shifts (for benzene all C and H atoms have been considered), as indicated by eqs 4 and 5 which include TMS ($\delta = 0.00$ ppm).

$$\delta(\text{exp}) = (1.010 \pm 0.006)\delta[\text{MP2/6-311 + G(d,p)}] \qquad n = 34, \quad R^2 = 0.9987 \quad (4)$$

$$\delta(\text{exp}) = (0.945 \pm 0.004)\delta(\text{B3LYP/Ahlrichs-vtzp}) \qquad n = 34, \quad R^2 = 0.9993 \quad (5)$$

Equation 6 shows that the results obtained at these two levels of theory are linearly related.

$$\delta(\text{B3LYP/Ahlrichs-vtzp}) = (4.7 \pm 0.8) + (1.114 \pm 0.008)\delta[\text{MP2/6-311 + G(d,p)}] \qquad n = 37, \quad R^2 = 0.9984 \quad (6)$$

Both theoretical methods yield comparable results, with R^2 slightly better at B3LYP/ Ahlrichs-vtzp, but the slope closer to 1 at MP2/6-311+G(d,p).

NMR Spin−Spin Coupling Constants. Liu and coworkers reported experimental one- and three-bond spin−spin coupling constants as part of their spectroscopic characterization of azaborine (**2**).[8,52] For this molecule, they determined ^1J(B−H) with ^{11}B = 130 Hz and ^1J(N−H) with ^{14}N = 57 Hz. This value for ^1J(N−H) with ^{14}N corresponds to 80 Hz for ^{15}N. In addition, they measured six three-bond H−H couplings, ^3J(H−H). The signs of these coupling constants were not determined. Experimental values of ^1J(B−H) and ^1J(N−H) for borazine (**8**) are 138.4 and 55.1 Hz for ^{11}B and ^{14}N, respectively.[46] The ^{14}N = 55.1 Hz corresponds to ^{15}N = 77.4 Hz. Table 5 reports the computed values of ^1J(N−B), ^1J(N−H), and ^1J(B−H) for molecules **2**−**9**. The computed values of ^1J(B2−H) and ^1J(N1−H) for **2** are 126.4 and −79.6 Hz, and the computed values for **8** are 131.4 and 76.7 Hz, respectively. Thus, the computed values are in good agreement with the experimental values,[8] with the sign of ^1J(N−H)

Table 5. ^1J(N−B), ^1J(N−H), and ^1J(B−H) (Hz) for Azaborine, Diazaborines, Borazine, and Pseudoborazine[a]

molecule no.	^1J(N1−B2)	^1J(B2−N3)	^1J(N3−B4)	^1J(B5−N6)
2	−23.4			
3	−29.6	−27.7	−21.9	
4	−23.4			
5	−16.5			
6	−19.6			
7	−28.9			
8	−26.8			
9[b]	−34.9	−28.0	−18.0	−24.0

molecule no.	^1J(N1−H)	^1J(N3−H)	^1J(N6−H)
2	−79.6 (−80)[c]		
3	−82.4	−75.0	
4	−75.9		
5	−79.6		
6	−78.5		
7	−76.9		
8	−76.7 (−77.4)[d]		
9[b]	−88.6	−73.5	−84.9

molecule no.	^1J(B2−H)	^1J(B4−H)	^1J(B5−H)
2	126.4 (130)[c]		
3	134.5	123.1	
4	129.5		
5	125.4		
6	115.9		
7	112.4		
8	131.4 (138.4)[d]		
9[b]	133.1	114.1	116.4

[a] Coupling constants correspond to ^{1}H, ^{11}B, and ^{15}N isotopes. [b] Values computed for the structure of C_s symmetry. [c] Experimental value from ref 8. [d] Experimental value from ref 50.

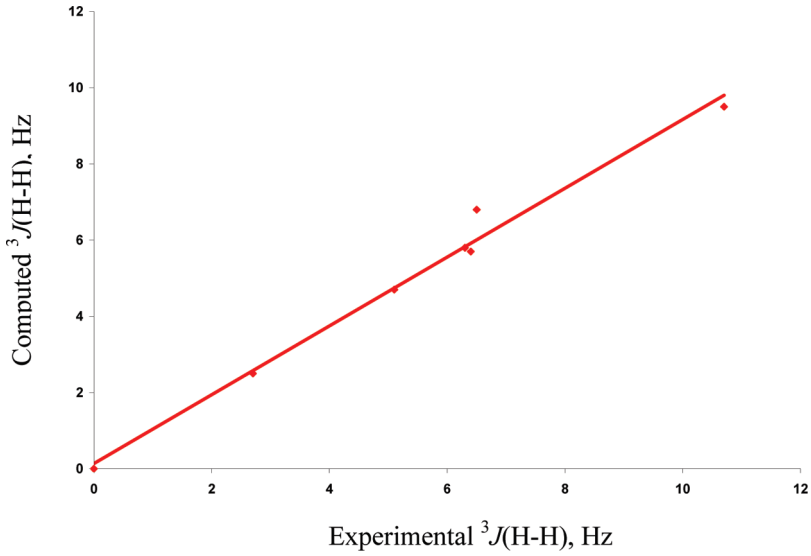


Figure 2. Computed vs experimental values of ^3J(H−H) for azaborine (**2**).

negative and ^1J(B−H) positive. In addition, Figure 2 shows a plot of computed versus experimental values of ^3J(H−H), which are much smaller, varying from 2 to 11 Hz. A reference point at (0,0) has been added. From Figure 2, it can be seen that the computed values are linearly related to the experimental as

$$^3J(\mathrm{H-H})(\mathrm{calc}) = 0.903[^3J(\mathrm{H-H})(\mathrm{exp})] + 0.136 \qquad n = 6, \quad R^2 = 0.986 \quad (7)$$

Thus, for those cases in which comparisons can be made, there is good agreement between the computed and the experimental values of coupling constants for molecule **2**.

Table 5 also lists ^1J(N−B), ^1J(N−H), and ^1J(B−H) for the diazaborines (**3**−**7**), borazine (**8**), and pseudoborazine (**9**). It is advantageous to examine the values of ^1J(N1−B2), ^1J(N1−H), and ^1J(B2−H) for molecules **2**−**7** relative to borazine, since variations in these coupling constants may also be viewed as arising from the substitution of one or two HC−CH groups for one or two HN−BH groups. Relative to ^1J(N1−B2) for borazine, this coupling constant decreases in absolute value for molecules **2** and **4**−**6**, but increases for molecules **3** and **7**. Some insight into these changes can be obtained by examining the bonding patterns in these molecules. In molecules **2** and **4**−**6**, the N1−B2 group is bonded to either two C atoms or one C and one B. In contrast, N1−B2 is bonded to one C and one N in molecules **3** and **7**. Thus, it is apparent that changes in the N1−B2 bond, which occur in these molecules as the atoms bonded to N1 and B2 change, are reflected in the changes in ^1J(N1−B2). When N1 and B2 are bonded to the less electronegative atoms C and B, ^1J(N1−B2) decreases in absolute value, whereas when either N1 or B2 is bonded to a N atom, ^1J(N1−B2) increases. This relationship is also consistent with an increase of ^1J(B2−N3) and a decrease of ^1J(N3−B4) for **3**, and a significant increase in ^1J(N1−B2) for pseudoborazine (**9**) in which both N1 and B2 are bonded to N atoms. Moreover, the increase in ^1J(N1−B2) when either N1 or B2 is bonded to the more electronegative N atom, and the decrease in ^1J(N1−B2) when N1−B2 is bonded to C and B, are consistent with the effects of F and Li substitution in borazine.[40] Fluorine substitution at either N1 or B2 of borazine increases the one-bond N1−B2 coupling constant, whereas Li substitution decreases ^1J(N1−B2). This is also consistent with substituent effects in benzene, in which case ^1J(C−C) increases upon F substitution but decreases with Li substitution. Since values of ^1J(N1−B2) for molecules **2**−**9** are dominated by the Fermi-contact (FC) terms, it is evident that changes in the s electron densities in both the ground and excited states of these molecules must respond to the different bonding patterns in these systems. It should also be noted that the FC term underestimates ^1J(N−B) by about 2 to 3 Hz, the contribution from the PSO term. This situation was also found for borazine and its derivatives.[40]

Relative to its value of −76.7 Hz for borazine, ^1J(N1−H) increases in absolute value in azaborine (**2**) and the diazaborines (**3**−**7**), except for molecule **4**, in which case it decreases but by less than 1 Hz. The largest absolute value of ^1J(N1−H), −82.4 Hz, is found for molecule **3**. The value of ^1J(B2−H) is 131.4 Hz for borazine, and it decreases in the same set of molecules except for molecule **3**, in which case it increases to 134.5 Hz. Moreover, ^1J(N1−B2) also has its greatest absolute value in molecule **3**. What makes molecule **3** unique among these six-member rings? The answer must be related to the retention of the borazine connectivity N1−B2−N3−B4 in molecule **3**, which is also the most stable of the diazaborazine isomers, as noted above. It is also interesting to note that the largest absolute values of both ^1J(N1−B) and ^1J(N1−H) in the entire set of molecules are found for pseudoborazine (**9**), while ^1J(B−H) has its second-largest value in this molecule. Once again, the N1−B2−N3−B4 connectivity is retained in this molecule. No simple relationships could be found between ^1J(N−B), ^1J(N−H), and ^1J(B−H) and the corresponding N−B, N−H, and B−H distances. Moreover, values of

^1J(N−B), ^1J(N−H), and ^1J(B−H) do not appear to correlate with each other.

Conclusions

An ab initio study has been carried out to investigate azaborine, the diazaborines, and borazine. These molecules can be considered as arising when one or more HC−CH linkages in benzene are replaced by HB−NH linkages. Such replacements lead to a significant rearrangement of electron density due primarily to the polarity of the BN bond. Azaborine (**2**) exhibits a much more localized structure than that of benzene, with a dominant contribution (57%) of the Kekulé-type structure. Azaborine and the diazaborines lose aromaticity relative to benzene but are more aromatic than borazine. The bonding characteristics influence the lengths of the B−N bonds in these molecules as well as the relative stabilities of the diazaborines. The most stable isomer has the N−B−N−B linkage, while the least stable have either a B−B or N−N bond.

NMR chemical shifts and coupling constants have been computed and found to be in good agreement with available experimental data. Changes in the N1−B2 bond in these molecules are reflected in the changes in ^1J(N1−B2). When N1 and B2 are bonded to the less electronegative atoms C and B, ^1J(N1−B2) increases in absolute value relative to borazine, but when either N1 or B2 is bonded to N, ^1J(N1−B2) decreases. No simple relationships exist between ^1J(N−B), ^1J(N−H), and ^1J(B−H) and the corresponding N−B, N−H, and B−H distances, respectively. Moreover, values of ^1J(N−B), ^1J(N−H), and ^1J(B−H) do not appear to correlate with each other.

Acknowledgment. This work was supported by the DGI Project Nos. BQU-2003-00894, BQU-2003-06553, BQU-2003-01251, and CTQ2007-61901, Consolider on Molecular Nanoscience CSD2007-00010, and the Project MADRISOLAR, ref S-0505/PPQ/0225 of the Comunidad Autónoma de Madrid. The continuing support of the Ohio Supercomputer Center is gratefully acknowledged.

Supporting Information Available: MP2/6-311+G-(d,p) geometries for molecules **2**−**9**. Also included are the full references for 26 and 45. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Liu, Z.; Marder, T. B. *Angew. Chem., Int. Ed.* **2008**, *47*, 242.

(2) Mó, O.; Yáñez, M.; Martín Pendás, A.; Del Bene, J. E.; Alkorta, I.; Elguero, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3970.

(3) Not to be mistaken with diazaborines, a class of drugs containing a six-membered ring with two nitrogen and one boron atom: Levy, C. W.; Baldock, C.; Wallace, A. J.; Sedelnikova, A. J.; Viner, R. C.; Clough, J. M.; Stultje, A. R.; Slabas, A. R.; Rice, D. W.; Rafferty, J. B. *J. Mol. Biol.* **2001**, *309*, 171.

(4) The anion corresponding to **2** (N^-/B-R) is called 1,2-azaboratabenzene: Pan, J.; Kampf, J. W.; Ashe, A. J. *J. Organomet. Chem.* **2009**, *694*, 1036.

(5) Stock, A.; Pohland, E. *Chem. Ber.* **1926**, *59*, 2215.

(6) Abbey, E. R.; Zakharov, L. N.; Liu, S.-Y. *J. Am. Chem. Soc.* **2008**, *130*, 7250.

(7) Drahl, C. *Chem. Eng. News* **2008**, *86*, 12.

(8) Marwitz, A. J.; Matus, M. H.; Zakharov, L. N.; Dixon, D. A.; Liu, S.-Y. *Angew. Chem., Int. Ed.* **2009**, *48*, 973.

(9) Koch, H.-J.; Roesky, H. W.; Bohra, R.; Noltenmeyer, M.; Schmidt, H.-G. *Angew. Chem., Int. Ed. Engl.* **1992**, *31*, 598.

(10) Ashe, A. J.; Fang, X.; Fang, X.; Kampf, J. W. *Organomet.* **2001**, *20*, 5413.

(11) Kranz, M.; Clark, T. *J. Org. Chem.* **1992**, *57*, 5492.

(12) Schleyer, P. v. R.; Jiao, H.; van Eikema Hommes, J. R.; Malkin, V. G.; Malkina, O. L. *J. Am. Chem. Soc.* **1997**, *119*, 12669.

(13) Madura, I. D.; Krygowski, T. M.; Cyrahski., M. K. *Tetrahedron* **1998**, *54*, 14913.

(14) Kar, T.; Elmore, D. E.; Scheiner, S. *THEOCHEM* **1997**, *392*, 65.

(15) Doerksen, R. J.; Thakkar, A. J. *J. Phys. Chem. A* **1998**, *102*, 4679.

(16) Bosdet, M. J. D.; Piers, W. E. *Can. J. Chem.* **2009**, *87*, 8.

(17) Fazen, P. J.; Burke, L. A. *Inorg. Chem.* **2006**, *45*, 2494.

(18) Matus, M. H.; Anderson, K. D.; Camaioni, D. M.; Autrey, S. T.; Dixon, D. A. *J. Phys. Chem. A* **2007**, *111*, 4411.

(19) Yáñez, M.; Mó, O.; Alkorta, I.; Del Bene, J. E. *J. Chem. Theory Comput.* **2008**, *4*, 1869.

(20) Pople, J. A.; Binkley, J. E.; Seeger, R. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1976**, *10*, 1.

(21) Krishnan, R.; Pople, J. A. *Int. J. Quantum Chem.* **1978**, *14*, 91.

(22) Bartlett, R. J.; Silver, D. M. *J. Chem. Phys.* **1975**, *62*, 3258.

(23) Bartlett, R. J.; Purvis, G. D. *Int. J. Quantum Chem.* **1978**, *14*, 561.

(24) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(25) Spitznagel, G. W.; Clark, T.; Chandrasekhar, J.; Schleyer., P. V. R. *J. Comput. Chem.* **1982**, *3*, 363.

(26) Frisch, M. J., *Gaussian-03*; Gaussian, Inc.: Wallingford, CT, 2004.

(27) Biegler-König, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *3*, 317.

(28) Popelier, P. L. A. with a contribution from R.G.A. Bone (UMIST, Engl, EU), MORPHY98, a topological analysis program; 0.2 ed., 1999.

(29) Alkorta, I.; Picazo, O. *Arkivoc* **2005**, *ix*, 305.

(30) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(31) Glendening, E. D.; Weinhold, F. *J. Comput. Chem.* **1998**, *19*, 593.

(32) Schleyer, P. V. R.; Maerker, C.; Dransfeld, A.; Jiao, H. J.; Hommes, N. J. R. V. *J. Am. Chem. Soc.* **1996**, *118*, 6317.

(33) Schleyer, P. V. R.; Manoharan, M.; Wang, Z. X.; Kiran, B.; Jiao, H. J.; Puchta, R.; Hommes, N. J. R. V. *Org. Lett.* **2001**, *3*, 2465.

(34) Poater, J.; Fradera, X.; Duran, M.; Solà, M. *Chem.—Eur. J.* **2003**, *9*, 400.

(35) Perera, S. A.; Sekino, H.; Bartlett, R. J. *J. Chem. Phys.* **1994**, *101*, 2186.

(36) Perera, S. A.; Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1996**, *104*, 3290.

(37) Perera, S. A.; Bartlett, R. J. *J. Am. Chem. Soc.* **1995**, *117*, 8476.

(38) Perera, S. A.; Bartlett, R. J. *J. Am. Chem. Soc.* **1996**, *118*, 7849.

(39) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(40) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Phys. Chem. A* **2006**, *110*, 9959.

(41) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Phys. Chem. A* **2007**, *111*, 419.

(42) Del Bene, J. E.; Elguero, J.; Alkorta, I.; Yáñez, M.; Mó, O. *J. Chem. Theor. Comp.* **2007**, *3*, 549.

(43) Del Bene, J. E.; Elguero, J. *Magn. Reson. Chem.* **2007**, *45*, 484.

(44) Kirpekar, S.; Jensen, H. J. Aa.; Oddershede, J. *Chem. Phys.* **1994**, *188*, 171.

(45) Stanton, J. F. *ACES II, a program product of the Quantum Theory Project*, University of Florida: Gainesville, FL.

(46) NIST Chemistry Web book http://webbook.nist.gov/chemistry/; June-2005. Accessed 3/15/2009. Chase, M. W., Jr. , *NIST—JANAF Thermochemical Tables, J. Phys. Chem. Ref. Data, Monograph 9*, Fourth ed., **1998**.

(47) Wiberg, K. B.; Rablen, P. R. *J. Comput. Chem.* **1993**, *14*, 1504.

(48) Matta, C. F.; Hernández-Trujillo, J. *J. Phys. Chem. A* **2003**, *107*, 7496. Zhurova, E. A.; Matta, C. F.; Wu, N.; Zhurov, V. V.; Pinkerton, A. A. *J. Am. Chem. Soc.* **2006**, *128*, 8849. Firme, C. L.; Antunes, O. A. C.; Esteves, P. M. *Chem. Phys. Lett.* **2009**, *468*, 129.

(49) Good, C. D.; Ritter, D. M. *J. Am. Chem. Soc.* **1962**, *84*, 1162.

(50) Mellon, E. K.; Coker, B. M.; Dillon, P. B. *Inorg. Chem.* **1972**, *11*, 852.

(51) Duriez, C.; Framery, E.; Toury, B.; Toutois, P.; Miele, P.; Vaultier, M.; Bonnetot, B. *J. Organomet. Chem.* **2002**, *657*, 107.

(52) The Supporting information of ref 8 includes a 2D NMR spectrum (page S8) in which the C3 signal of 1,2-dihydro-1,2-azaborine (**2**) appears as a quadruplet that is not well-resolved. From this signal we have measured a value of $^{1}J(^{11}B2-^{13}C3)$ equal to 74 ± 1 Hz. The calculated EOM-CCSD value is 72.8 Hz.

CT900128V

JCTC Journal of Chemical Theory and Computation

Phenylalanyl-Glycyl-Phenylalanine Tripeptide: A Model System for Aromatic−Aromatic Side Chain Interactions in Proteins

H. Valdes,*,† K. Pluhackova,‡ and P. Hobza*,‡,§

Dpto. Química Física y Analítica, Universidad de Oviedo, C/Julián Clavería, 8, 33006 (Oviedo) Asturias, Spain, Center for Biomolecules and Complex Molecular Systems, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 16610 Prague 6, Czech Republic, and Department of Physical Chemistry, Palacky University, 771 42 Olomouc, Czech Republic

Received April 10, 2009

Abstract: The performance of a wide range of quantum chemical calculations for the ab initio study of realistic model systems of aromatic−aromatic side chain interactions in proteins (in particular those $\pi-\pi$ interactions occurring between adjacent residues along the protein sequence) is here assessed on the phenylalanyl-glycyl-phenylalanine (FGF) tripeptide. Energies and geometries obtained at different levels of theory are compared with CCSD(T)/CBS benchmark energies and RI-MP2/cc-pVTZ benchmark geometries, respectively. Consequently, a protocol of calculation alternative to the very expensive CCSD(T)/CBS is proposed. In addition to this, the preferred orientation of the Phe aromatic side chains is discussed and compared with previous results on the topic.

Introduction

Proteins are linear heteropolymers composed of amino acids. Most proteins fold into unique three-dimensional structures (tertiary structure or native state) in which they perform their biological functions.[1] A folded protein is a complex structure containing very different types of intra- and interresidual interactions. One of the reasons to be interested in the nature of these intramolecular forces is based on the well-known fact that interactions between side chains largely favor the molecule's acquisition of the folded state. Obviously, then a better understanding of the folding process can be gained by studying both the interactions between adjacent residues covalently bound as well as the interactions between noncovalently bound regions. Noncovalent interactions compile a wide variety of weak intermolecular forces (H-bonds, cation$-\pi$ interactions, etc.) among which the $\pi-\pi$ interactions are known to play a relevant role in the stability of proteins. Indeed, around 64% of aromatic side chains in proteins are likely involved in $\pi-\pi$ intramolecular interactions with neighbor aromatic side chains.

Residues buried in the hydrophobic core of any protein are shielded from solvent, and thus, they attain an environment which may be very similar to that in the gas phase. For that reason, noncovalent interactions occurring in the hydrophobic core of any protein may be studied from a quantum chemical point of view. However, such treatment necessarily requires—for obvious reasons—the definition of a much simpler system than a protein; i.e., quantum chemical studies on noncovalent interactions in proteins are typically restricted to some relevant parts of it. In this respect, we have previously focused on the quantum chemical study of the noncovalent interactions between the peptide backbone and the aromatic side chains in di- and tripeptides.[2−6] In the present work, we concentrate on the study of noncovalent interactions between aromatic side chains of adjacent residues along the protein sequence.

Dealing with the quantum chemical computation of aromatic−aromatic interactions in proteins entails mainly two difficulties, the selection of a good prototype model system

* To whom correspondence should be addressed. E-mail: haydee.valdes@marge.uochb.cas.cz.

† Universidad de Oviedo.

‡ Academy of Sciences of the Czech Republic.

§ Palacky University.

10.1021/ct900174f CCC: $40.75
Published on Web 07/28/2009

and, subsequently, the level of theory to be employed for the study. There are two possible ways to select a prototype system of the $\pi-\pi$ interactions of any protein. Having the crystal structure, one would usually concentrate on the minimal possible interacting parts of interest—aromatic side chains in this particular case—and *remove* everything else. Lacking the crystal geometries, one cannot but optimize ab initio the geometries of the model system. This is, however, a quite critical step, since it is questionable how close the minimum found is to the real geometrical disposition of the aromatic side chains in a protein. In other words, the simplest model of the aromatic−aromatic interactions in proteins one could think of is undoubtedly the benzene dimer. Certainly, phenylalanine residues have benzene-like aromatic side chains that could adopt, at least hypothetically, similar geometrical rearrangements to those shown by the benzene dimer. But, undoubtedly, there are some factors affecting the orientation of the aromatic side chains in proteins that are not present in the benzene dimer model, namely, (a) the likely interaction between the aromatic side chains and the peptide backbone; (b) the ϕ and Ψ preferences of allowed regions in protein structures (Ramachandran plot) that determine the local shape assumed by the protein backbone and, thus, the orientation of the side chains, and, finally, (c) peptide bonds linking the residues have a double character that hinders rotation around its axis—guaranteeing that the α carbons are roughly coplanar—and act upon the geometrical disposition of the side chains. Hence, accurately modeling aromatic−aromatic side chain interactions in proteins requires realistic models including all of the above-mentioned factors.

Aromatic−aromatic interactions have been intensively explored up to now by means of rigorous electronic structure computations. Illustrative examples are, for instance, the elaborated works by Urch et al.,[7] Sherrill et al.,[8] and Diederich et al.,[9] where a detailed bibliography on the topic can be found. So, regarding the level of theory to be employed for the study, it is already a very well-known fact that a definitive treatment of the dispersion energy can only be done by using the CCSD(T) method with large basis sets including multiple polarization and diffuse functions;[8] thus, this should be the method to be employed for any study of such characteristics. However, realistic models of aromatic−aromatic interactions in proteins are incompatible with small-size systems, and obviously, such high-level quantum chemical calculations are prohibitive. Then, a level of theory alternative to the very expensive CCSD(T)/large basis set has to be necessarily found.

Plenty of discussion has been done up to now about the performance of the MP2 and the density functional theory (DFT) methods. On one hand, the major disadvantage of the MP2 method is its overestimation of the dispersion energy—relevant in the aromatic−aromatic interactions—when an extended basis set [or even at the complete basis set (CBS) limit] is applied. Reliable energies (and also geometries) are thus frequently obtained when using medium size basis sets. However, this is evidently due to a compensation of errors and it is impossible to rely on this compensation. On the other hand, DFT lacks a proper description of the dispersion interaction. Consequently, many alternatives have appeared quite recently trying to compensate the deficiencies of both methods.[10−16] Hence the question of which level of theory is more suitable for the study of aromatic−aromatic interactions in protein model systems is topical, and for this reason, we present here a theoretical study on the performance of a vast range of levels of theory in comparison with the CCSD(T)/CBS benchmark data for the phenylalanyl-glycyl-phenylalanine (Phe-Gly-Phe, FGF) tripeptide. We have chosen this system since it fulfills the requirements needed for such kind of study, namely, (a) CCSD(T) single point calculations are affordable, though we are in the upper limit of what is nowadays feasible, and (b) it can be considered a realistic model of $\pi-\pi$ interactions occurring in proteins, and more specifically, unlike previous studies where the prototype systems model the aromatic−aromatic interactions occurring between nonconsecutive residues, the system here studied models those $\pi-\pi$ interactions occurring between adjacent residues along the protein sequence.

Additionally, this work deals with another degree of difficulty in comparison with simpler models, e.g., benzene$\cdots$$NH_3$ dimer[17,18] or benzene dimer,[19] since both H-bond and $\pi-\pi$ interactions coexist in the molecule, which shows the need for a method describing correctly and, more important, simultaneously those interactions. Finally, the importance of the basis set superposition error (BSSE) cannot be forgotten.[20−22] When dealing with the calculation of aromatic−aromatic intermolecular interactions, this error is generally corrected by the counterpoise (CP) procedure.[23] However, the case of isolated systems is far more complicated, because neither a well-established method accounting for this error nor a CP-like procedure has yet been developed. Moreover since small peptides are systems of multiconfigurational character, each particular conformation suffers differently from the intramolecular BSSE, which stresses the importance of choosing a level of theory where this error has been, if not erased, at least minimized.

Ultimately, this work aims to gain a better understanding of the $\pi-\pi$ interactions and by extension their role played in proteins from the information obtained on the pure—without any influence of the solvent—aromatic−aromatic interactions provided by calculations in vacuo on isolated small peptides.

Computational Details

As mentioned already in the Introduction, small peptides are systems of multiconformational character, thus showing a very rich conformational landscape. Since only few conformers are typically experimentally detected,[24−26] we have restricted our benchmark study to a set containing 15 energetic and geometrically different conformers. These structures have been selected according to a strategy of calculation previously proven efficient[5] and they constitute the most stable structures in the potential energy surface of FGF. This set contains those conformers observed experimentally and simultaneously; the number of structures included in it is small enough so that high-level calculations can be carried out for all of them. Different levels of theory have then been tested against CCSD(T)/CBS values. A table containing a time scale for the different computational

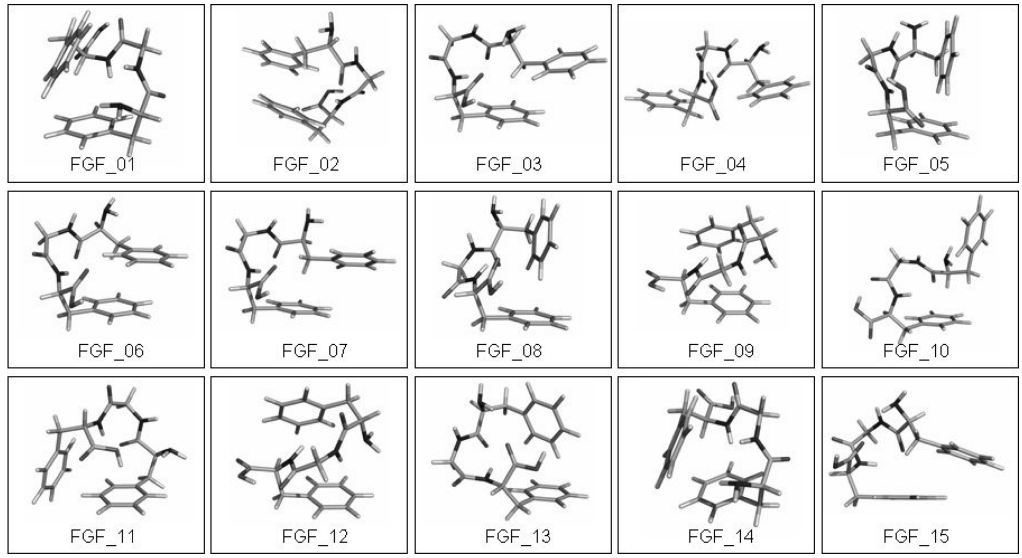


Figure 1. RI-MP2/cc-pVTZ geometries of the 15 most stable structures of FGF tripeptide.

methods employed in the study is included in the Supporting Information (see Table S1).

Empirical Force Field. Single-point energy calculations and geometry optimizations were carried out by means of parm99 empirical force field.[27] B3LYP/cc-pVTZ atomic charges obtained using the restrained electrostatic potential fitting procedure[28] (RESP) have been used for these calculations. Notice that the RESP charges finally used are the average of the RESP charges of six different structures (FGF_02, FGF_03, FGF_04, FGF_05, FGF_10, and FGF_12; see Figure 1).

Tight-Binding Method Extended by an Empirical Dispersion Term (SCC-DF-TB-D). Energy minimizations, single-point energy calculations, and geometry optimizations were obtained by means of the SCC-DF-TB-D[29] method, which includes a term describing the dispersion energy essential for the proper study of peptides containing two aromatic side chains. Additionally, it is a very fast method and particularly interesting for an accurate scanning of the potential energy surface of any small peptide.

Density Functional Theory. Different functionals and basis sets have been tested for geometry optimization, namely, (a) B3LYP[30]/6-31G*,[31] since it is one of the levels of theory commonly used for the study of isolated small peptides; (b) TPSS[32]/6-311++G(3df,3pd)[31] and TPSS-D/6-311++G(3df,3pd), since the latter has been recently proven to perform reasonably well for the study of isolated systems;[3] (c) M06-2X[16]/MIDI!,[33] due to the fact that it belongs to a brand new generation of hybrid meta-generalized-gradient-approximation exchange correlation functionals that include an accurate treatment of medium-range correlation energy that mainly concerns the London dispersion energy; and, finally, (d) M06-L[34]/TZVP[35] since it is much cheaper than M06-2X and consequently, can be combined with larger basis sets without losing computational efficiency.

Additionally, single-point energy calculations on RI-MP2/cc-pVTZ geometries have been also systematically carried out at the following levels of theory: (a) B3LYP/6-311++G(3df,3pd); (b) TPSS/6-311++G(3df,3pd); (c) TPSS-D/6-311++G(3df,3pd); (d) M06-2X/6-311+G(2df,2pd);[31] (e) M06-L/6-311++G(3df,3pd); (f) BH&H[36]/6-311++G-(d,p); and (g) PBE[37]-D/TZVP.

Wave Function Theory (WFT) Calculations. RI-MP2/cc-pVTZ geometries are here considered as the benchmark. CCSD(T) energies where obtained according to the following equation:

$$E_{\mathrm{CBS}}^{\mathrm{CCSD(T)}} = E_{\mathrm{CBS}}^{\mathrm{MP2}} + (E^{\mathrm{CCSD(T)}} - E^{\mathrm{MP2}})\Big|_{\text{6-31G*}} \qquad (1)$$

where MP2/CBS energies were calculated by the extrapolation of the RI-MP2[38,39]/cc-pVTZ[40] and RI-MP2/cc-pVQZ[40] relative energies using the scheme of Helgaker and co-workers.[41] The second term of eq 1 covers the portion of correlation energy beyond the second perturbation order[42] and it has been calculated using a small basis set, as it has been demonstrated that the CCSD(T)-MP2 energy difference depends little on the size of the basis set.[43] Equation 1 has been previously proven as an efficient way to approximate CBS energies for large systems where otherwise such calculations are prohibited.[43] $E_{\mathrm{CBS}}^{\mathrm{MP3}}$ energies were also calculated following eq 1 except for the correlation energy beyond the second order, which is given by the $(E^{\mathrm{MP3}} - E^{\mathrm{MP2}})|_{\text{6-31G*}}$ term. SCS-MP2[14] and SCS(MI)-MP2[15] (with spin-component scaling factors $c_{\mathrm{os}} = 0.40$ and $c_{\mathrm{ss}} = 1.29$) methods have also been tested. For both methods, energies have been extrapolated to the CBS limit using the scheme of Helgaker and co-workers. Additionally, in case of the SCS(MI)-MP2 method, we have also tested another extrapolation scheme suggested by the authors.[15] Since the differences between the results obtained with the different schemes of extrapolation are negligible, for simplicity reasons, we will only show the results obtained using the scheme of Helgaker and co-workers.

Results and Discussion

Figure 1 shows the RI-MP2/cc-pVTZ geometries of the 15 most stable structures in the SCC-DF-TB-D potential energy

Table 1. Root-Mean-Square Deviations (RMSDs) (in Å) between the RI-MP2/cc-pVTZ Benchmark Geometries and Geometries Obtained at Different Levels of Theory[a]

	B3LYP/ 6-31G*	TPSS/ LP_1	TPSS-D/ LP_1	M06-2X/ MIDI!	M06-L/ TZVP	SCC-DF-TB-D	ff99
FGF_01	0.37	0.63	0.14	0.04		0.12	0.21
FGF_02	1.13	1.06	0.82	0.10	0.13	0.08	0.12
FGF_03	0.90	0.88	0.16	0.08	0.15	0.10	0.13
FGF_04	0.30	1.09	0.21	0.14	0.23	0.05	0.12
FGF_05	0.90	0.84	0.29	0.05	0.10	0.05	0.15
FGF_06	2.33	1.09	0.11	0.10	0.14	0.07	0.14
FGF_07	0.78	1.09	0.21	0.08	0.16	0.07	0.32
FGF_08	0.30	0.40	0.12	0.07	0.21	0.09	0.53
FGF_09	0.26	0.51	0.09	0.07	0.06	0.06	0.21
FGF_10	0.62		0.18	0.22	0.39	0.06	0.32
FGF_11	0.45	0.63	0.51	0.14	0.46	0.06	0.13
FGF_12	1.08	0.94	0.31	0.10	0.16	0.07	0.23
FGF_13	0.43	1.66	0.11	0.11	0.13	0.10	0.13
FGF_14	0.41	0.87	0.10	0.10	0.15	0.07	0.16
FGF_15	1.10	0.94	0.17	0.10	0.13	0.08	0.16
average	0.76	0.90	0.24	0.10	0.19	0.08	0.20

[a] Average RMSDs values are also included. LP_1 stands for the 6-311++G(3df,3pd) Pople basis set.

surface of FGF tripeptide. Cartesian coordinates of those structures can be found in the Supporting Information. Additionally, all these coordinates (together with relative energies) can be found in the following web page: www.begdb.com.[44] These geometries have been taken as reference geometries for the assessment of various levels of theory: (a) B3LYP/6-31G*; (b) TPSS/6-311++G(3df,3pd); (c) TPSS-D/6-311++G(3df,3pd); (d) M06-2X/MIDI!; (e) M06-L/TZVP; (f) SCC-DF-TB-D; and (g) geometries obtained using the ff99 force field.

Table 1 collects the root-mean-square deviations (RMSDs) in angstroms obtained for each individual conformer as compared to the benchmark geometries. Average deviations of each particular level of theory are also collected. On the basis of the average deviations shown in Table 1, it is easily inferred that SCC-DF-TB-D and M06-2X/MIDI! show a similar behavior and differ little from the RI-MP2/cc-pVTZ geometries. Indeed, the smaller deviations shown by the SCC-DF-TB-D method suggest that it is highly recommendable for obtaining reliable geometries of aromatic−aromatic protein model systems at an extremely low computational cost. The M06-2X/MIDI! level of theory is equally reliable but certainly more computationally expensive. Additionally, the behavior of the M06-L/TZVP level of theory is slightly worse but still very acceptable and certainly less computational expensive than M06-2X/MIDI!. Also true is that whereas the former two methods rarely show individual conformer deviations larger than 0.2 Å, for the latter, approximately 25% of the individual deviations differ more than 0.2 Å from the benchmark geometries. Notice also that at this level of theory it was not possible to localize (since after 426 steps the convergence criteria was still not satisfied) the FGF_01 structure in the potential energy surface.

Even when on average the ff99 empirical force field performance is comparable to the M06-L/TZVP level of theory, individual geometries obtained using the force field deviate more from the benchmark geometries, as 40% of the conformers show RMSD values higher than 0.2 Å. Consequently, its overall performance is slightly worse than that of the M06-L/TZVP level of theory. Notice, however, that atomic charges of a flexible molecule—like the one here considered—may vary significantly from one conformation to another, obviously affecting the final results. In the present case, the best results were obtained when the RESP charges used were the average of the RESP charges of six structures of different conformations (FGF_02, FGF_03, FGF_04, FGF_05, FGF_10, and FGF_12; see Figure 1).

Geometries obtained at the TPSS-D/6-311++G(3df,3pd) level of theory deviate from the benchmark geometries slightly more than any of the methods previously commented on. However, its overall performance is reasonable, and consequently, geometries obtained at this level should still be reliable. The opposite is true regarding the B3LYP/6-31G* level of theory. Widely used for the study of small peptides, B3LYP/6-31G* gives large RMSD values when compared to the reference geometries, the reason being no other than the poor description of the dispersion interaction given by the functional. A very illustrative example of its different performance in comparison with the TPSS-D functional or the MP2 method can be seen in Figure 2. We cannot forget, though, that the MP2 method suffers from an overestimation of the dispersion energy;[45,46] thus, geometries should be also partially affected by this overestimation, meaning that the comparison between the different methods should be done qualitatively rather than quantitatively. The importance of dispersion in FGF is also supported by the fact that geometries obtained at TPSS/6-311++G(3df,3pd), i.e., removing the dispersion energy from the TPSS functional, gives the largest deviation of all the levels of theory here compared with the RI-MP2/cc-pVTZ values.

Thus, from the analysis of Table 1, one could conclude that any of the levels of theory here discussed—but the B3LYP/6-31G* and TPSS/6-311++G(3df,3pd)—could be in principle used for obtaining reliable molecular structures of aromatic−aromatic protein model systems and its usage should be mostly dependent on the computational resources and time disposal.

Table 2 collects the mean unsigned error (MUE), standard deviation (σ), and maximum unsigned error (Max.) obtained from the comparison between the CCSD(T)/CBS benchmark energies and single-point energies—on the RI-MP2/cc-pVTZ benchmark geometries—obtained at different levels of theory. The statistical values have been calculated as follows. First we have calculated the average energies ($E_{average}$) for each particular method collected in Table 2 as well as the average energy at the CCSD(T)/CBS level of theory. Then, we have obtained the relative energies at all levels of theory subtracting the average energy from the energy of each particular conformer, i.e., $(E_{conformer} - E_{average})^{level\ of\ theory}$ and $(E_{conformer} - E_{average})^{CCSD(T)/CBS}$. Finally, we have calculated the difference $[(E_{conformer} - E_{average})^{CCSD(T)/CBS} - (E_{conformer} - E_{average})^{level\ of\ theory}]$. These last data are those collected in Table 2.

The performance of each method is well-established after the analysis of Table 2. The B3LYP/6-311++G(3df,3pd) level of theory gives the largest errors of all, followed by the empirical force field ff99, and the TPSS/6-311++G-(3df,3pd) level of theory. This is not an unexpected result, though, because dispersion is playing a major role in the

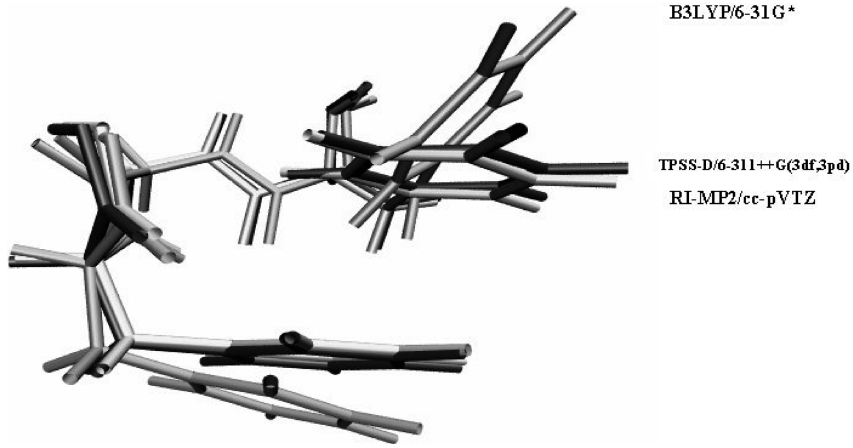


Figure 2. Comparison between the B3LYP/6-31G*, TPSS-D/6-311++G(3df,3pd), and the RI-MP2/cc-pVTZ levels of theory.

Table 2. Mean Unsigned Error (MUE), Standard Deviation (σ), and Maximum Unsigned Error (Max.) Obtained from the Comparison between the CCSD(T)/CBS Benchmark Energies (in kcal/mol) and Single-Point Energies at Different Levels of Theory on RI-MP2/cc-pVTZ Benchmark Geometries

method	MUE	σ	Max.
ff99	1.84	1.55	5.42
SCC-DF-TB-D	0.51	0.39	1.21
B3LYP/6-311++G(3df,3pd)	1.94	1.66	5.80
TPSS/6-311++G(3df,3pd)	1.63	1.72	5.53
TPSS-D/6-311++G(3df,3pd)	0.95	0.51	1.63
PBE-D/TZVP	0.97	0.59	1.91
M06-2X/6-311+G(2df,2p)	0.42	0.38	1.20
M06-L/6-311++G(3df,3pd)	0.39	0.34	0.99
BH&H/6-311++G(d,p)	0.82	0.46	1.51
SCS-MP2/CBS	0.28	0.26	0.86
SCS(MI)-MP2/CBS	0.26	0.28	0.88
MP2/CBS	0.56	0.38	1.37
MP3/CBS	0.40	0.26	0.90

stability of each of the FGF conformers and none of these methods deals properly with dispersion. The next category includes the PBE-D/TZVP, TPSS-D/6-311++G(3df,3pd), and BH&H/6-311++G(d,p) levels of theory with MUEs of approximately 1.0 kcal/mol. It can be seen then that adding dispersion improves the results, which reinforces the former statement about the importance of dispersion in the stability of this system. It also shows that if DFT is to be used for the study of systems of similar characteristics, functionals including dispersion—augmented with dispersion or specifically parametrized to cover dispersion—should be necessarily employed. The SCC-DF-TB-D method, as well as any of the Truhlar's functionals here tested, constitute a significant improvement with respect to any of the levels of theory discussed before. Their statistics are very similar to those given by any of the wave function theory methods. Among the latter, both the SCS(MI)-MP2 and the SCS-MP2 methods show the best performance. This means that the more expensive MP2/CBS and MP3/CBS levels of theory could be avoided, specially, if larger systems were to be calculated.

We have also examined if the global minimum predicted by the different levels of theory here considered is the same. Each individual level of theory except TPSS/6-311++G(3df,3pd), B3LYP/6-311++G(3df,3pd), and f99 predict FGF_01 as the global minimum, in agreement with the benchmark calculations (see Table S2 in the Supporting Information). Notice also that at the BH&H/6-311++G(d,p) level of theory the global minimum predicted is structure FGF_02. However, this structure is only 0.16 kcal/mol more stable than structure FGF_01. Indeed, according to this level of theory, there are four structures within a range of energy of 0.41 kcal/mol and thus it is really difficult to select one as the global minimum. Additionally, in the majority of the levels of theory considered, the 15 structures of the set lie in an interval of energy similar to that of the benchmark calculations (approximately 3.2 kcal/mol). However, this is not the case for the TPSS/6-311++G(3df,3pd), B3LYP/6-311++G(3df,3pd), PBE-D/TZVP, and f99 data. The same holds true for the TPSS-D/6-311++G(3df,3pd) level of theory, where structures lie in a range of 6.00 kcal/mol, suggesting that one should be particularly careful if any selection of the structures is to be done on the basis of the relative energies predicted at this particular level of theory. Also noticeable is that, again at this particular level of theory, structures where an OH···O=C intramolecular H-bond occurs between the carboxylic terminal group and the C=O of the preceding residue (see for instance structure FGF_01) are systematically more stable than structures lacking this particular intramolecular interactions. However, these two families of structures are more interspersed in case of the benchmark calculations. Notice that SCC-DF-TB-D as well as BH&H/6-311++G(d,p) and PBE-D/TZVP show a similar behavior to TPSS-D/6-311++G(3df,3pd) level of theory in this particular respect.

A final comment should be made with respect to our previous paper concerning a similar study, though on a different type of intramolecular interaction within the peptide, i.e., peptide backbone—aromatic side chain interactions.[3] The performance of the methods is consistent for both families of peptides (backbone—aromatic side chain and aromatic—aromatic side chain). Regarding geometries, average RMSD values calculated for both type of systems are almost the same. The only exception is for the B3LYP/6-31G* level of theory, where average RMSD values are larger in the case of the FGF system. Obviously, this is due to the fact that the performance of the B3LYP functional worsens as the aromatic character of the system increases. Regarding

Table 3. Mean Unsigned Error (MUE), Standard Deviation (σ), and Maximum Mean Unsigned Error (Max.) Obtained between the CCSD(T)/CBS Benchmark Energies (in kcal/mol) and Single-Point Energies on the MP2/cc-pVTZ Benchmark Geometries and Geometries Obtained at Each Particular Method

method[a]	MUE	σ	Max.
TPSS/LP_1//RI-MP2/cc-pVTZ	1.63	1.72	5.53
TPSS/LP_1//TPSS/LP_1	1.33	1.00	2.98
TPSS-D/LP_1//RI-MP2/cc-pVTZ	0.95	0.51	1.63
TPSS-D/LP_1//TPSS-D/LP_1	1.12	0.52	1.88
B3LYP/6-31G*//RI-MP2/cc-pVTZ	1.81	1.43	5.35
B3LYP/6-31G*//B3LYP/6-31G*	1.25	1.06	3.19
M06-2X/MIDI!//RI-MP2/cc-pVTZ	0.55	0.44	1.41
M06-2X/MIDI!//M06-2X/MIDI!	0.54	0.38	1.31
M06-L/TZVP//RI-MP2-cc-pVTZ	0.44	0.37	1.28
M06-L/TZVP//M06-L/TZVP	0.44	0.39	1.29
SCC-DF-TB-D//RI-MP2/cc-pVTZ	0.51	0.39	1.21
SCC-DF-TB-D//SCC-DF-TB-D	0.51	0.32	1.05
ff99//RI-MP2/cc-pVTZ	1.84	1.55	5.42
ff99//ff99	1.03	0.70	2.54

[a] LP_1 stands for the 6-311++G(3df,3pd) Pople basis set.

energies, the same trends are observed for both families of peptides. Summarizing, the ff99 force field and standard—not augmented with dispersion or not specifically parametrized to cover dispersion—functionals deviate more from the benchmark energies. Then, DFT improved, i.e., augmented, with dispersion in any possible way performs reasonably well. More specifically, M06-2X and M06-L Truhlar functionals show a better performance than the PBE-D or TPSS-D functionals. Additionally, wave function theory methods show the smallest errors, particularly SCS-MP2 and SCS(MI)-MP2, in comparison with the benchmark data here considered. It should be stressed once again that studying proteins from a quantum chemical point of view necessarily implies restricting, for obvious reasons, the study to those particular areas of interest, e.g. $\pi-\pi$ interactions and peptide backbone−aromatic side chain interactions. However, a global perspective on the topic certainly requires the combination of the conclusions obtained from each individual study. Then, it would be highly unsatisfactory if a certain method worked well, for instance, for peptide backbone−aromatic side chain interactions but failed in the description of $\pi-\pi$ interactions. As we have just seen that the overall performance of the methods is consistent for both $\pi-\pi$ and peptide backbone−aromatic side chain interactions; thus, by choosing the proper method, we can trust that the description of both types of intramolecular interactions is correct.

Table 3 collects the MUE, σ, and Max. obtained from the comparison between single-point energies calculated on the benchmark geometries (RI-MP2/cc-pVTZ) and those geometries included in Table 1 aiming to study the possible influence that the selection of different geometries may have on the final energies. It can be seen that energies calculated using methods covering reasonably well the dispersion energy do not depend much on the geometry chosen for the single-point energy calculations. However, larger differences are found when the TPSS, B3LYP, or ff99 methods are chosen. Interestingly, for these three particular cases, the mean unsigned errors as well as the standard deviations and the maximum error are smaller than those obtained from single-point energies calculated on benchmark geometries, suggesting that a cancellation of error occurs when the former methods are used. Notice that, on one hand, TPSS and B3LYP have the largest geometry deviations compared to RI-MP2/cc-pVTZ geometries, which can have an impact on single-point calculations. On the other hand, ff99 geometries have a low RMSD, but force-field methods are known to be extremely sensitive to changes in the geometry.

Insights into the Preferred Orientation of the Phe Aromatic Side Chains

The FGF tripeptide is a system of multiconformational character where various different types of noncovalent intramolecular interactions are simultaneously present and consequently the quantum chemical method to be used for its study has to deal in a balanced manner with all of these intramolecular forces at once. Thus, examining which method—apart from the very computationally expensive CCSD(T)/CBS—gives the best results at the lowest computational cost is topical and interesting by itself and it constitutes the main scope of this study. However, the study of FGF can also provide interesting biological information. Regarding this point, many different aspects can be analyzed, for instance, the conformational preferences of the peptide backbone or the interaction—via multiple NH$\cdots\pi$ interactions—between the peptide backbone and the aromatic side chains. For the present case, the orientation of the aromatic side chains with respect to each other deserves special attention. Our aim is to shed some light on the *pure*—without the influence of any other factor such as the neighbor residues or hydrophobic effects—intramolecular aromatic−aromatic interactions between the aromatic side chains of residues in the hydrophobic core of globular proteins assuming, as already commented on the introduction, that residues buried in the hydrophobic core attain an environment similar to that in the gas phase. Thus, the conclusions here obtained can be extrapolated up to some point to the behavior of the aromatic side chains of residues within the hydrophobic core of a protein.

Looking at the 15 conformers collected in Figure 1, it is possible to group all these structures into three different categories according to the orientation of the aromatic side chains with respect to each other, namely, (a) *stacked*, those where the aromatic side chains are slipped parallel/parallel to each other, i.e., FGF_02, FGF_06, and FGF_15; (b) *T-shaped*, those where the aromatic side chains are in a T-shaped disposition, i.e., FGF_05, FGF_08, FGF_11, and FGF_13; and (c) *others*, those conformers do not matching any of the criteria mentioned before. According to this geometrical classification, it seems reasonable to conclude then that neither the T-shaped nor the stacked orientations are preferred by the Phe aromatic side chains. Indeed, the number of conformers belonging to both families (three to the first and four to the second) is almost the same. Moreover, according to the data here collected, there is neither a clear preference for the T-shaped family over the stacked or vice versa. The most abundant conformers are those showing a geometrical disposition favoring the maximum number of intramolecular interactions, i.e., those geometries where the

maximum number of H-bonds and NH$\cdots\pi$ and aromatic−aromatic intramolecular interactions are acting together.

We have also calculated the Gibbs relative energies of each particular conformer at $T = 300$ K from TPSS-D/TZVPP ab initio quantum chemical calculations assuming a rigid rotor−harmonic oscillator−ideal gas approximation, since we have already concluded for similar systems that this procedure provides a reliable description of the free energy surface (FES) in a vacuum.[4] According to Gibbs relative energies calculations, these same structures could be ordered as follows (see Table S3 in the Supporting Information): FGF_08 ∼ FGF_07 < FGF_04 < FGF_14 < FGF_06 < FGF_09 < FGF_13 < FGF_12 < FGF_15 < FGF_02 < FGF_03 < FGF_11 < FGF_01 < FGF_05. On the basis of this order, it seems clear that stacked structures are not the most energetically favored. This result may seem in contrast with the work published by Schettino et al.,[47] where it is concluded that, in a hydrophobic environment, such as the protein core, Phe-Phe systems show a slight preference for stacking. However, it should be here explicitly mentioned that these two studies are not straightforwardly comparable, since the prototype systems used by Schettino et al. are simpler than ours. Schettino et al. constructed the models (complexes) for the Phe side chains from the corresponding amino acids by removing the amino and carboxylic groups, and consequently, this work does not take into account the influence of the interactions between the backbone and the aromatic side chains. However, from our study it is clearly inferred that such interactions play a determining role in the final orientation of the aromatic side chains. Indeed, the same conclusion was implicitly obtained by Kollman et al.[19] from a study carried out using benzene and toluene dimers as model systems of aromatic interactions in proteins. Since results obtained with the different model systems were in conflict, Kollman et al. concluded that very simple prototypes hardly model the Phe side chains behavior in proteins. Also interesting is that the structure of the Ac-Phe-Phe-NH_2 system characterized by means of IR/UV double resonance spectroscopy in the gas phase[48] is, as in the case of the FGF peptide, a T-shaped structure.

Summary and Conclusions

Aromatic side chains of proteins often participate in $\pi-\pi$ interactions. Studying $\pi-\pi$ interactions in proteins by means of quantum chemical calculations imposes a restriction on the size of the protein model system to be considered. Too large systems are simply unaffordable from a computational point of view, whereas too small model systems may not be realistic enough and may skip some relevant information, as for instance the geometrical restrictions imposed by the peptide backbone. At the same time, the size of the prototype model influences the level of the quantum chemical calculation. Parallelly, studying model systems of aromatic−aromatic interactions in proteins by quantum chemistry requires the proper treatment of the dispersion energy, which necessarily implies the usage of high-level quantum chemical methods. A satisfactory solution would be then to follow a protocol of calculation that could combine both requirements. In this respect we have shown that geometries optimized at any of the levels of theory here employed—except from B3LYP/6-31G* and TPSS/6-311++G(3df,3pd)—are similar to the RI-MP2/cc-pVTZ geometries (here considered as the benchmark). Particularly, geometries obtained at the SCC-DF-TB-D level of theory are recommended as input geometries for the energy calculations, specially when larger systems are to be calculated.

Energy calculations should never be done using a standard DFT functional that has not been augmented by dispersion or has not been specifically parametrized to cover dispersion energy. These methods also fail in the prediction of the global minimum in the PES. If the size of the system studied is large, then SCC-DF-TB-D or any of the Truhlar functional's here tested should be enough. Otherwise, the final and reliable order of the multiple conformers existing on the potential energy surface of any peptide should be obtained from high-level quantum chemical methods, particularly SCS-MP2 or SCS(MI)-MP2. A necessary condition, which simultaneously deals with the intramolecular basis set superposition error, is the extrapolation of the basis set to the complete basis set limit. Special attention should be paid when selecting the structures according to a specific interval, since TPSS/6-311++G(3df,3pd), B3LYP/6-311++G-(3df,3pd), PBE-D/TZVP, f99 data, and TPSS-D/6-311++G-(3df,3pd) give larger intervals than the remaining methods.

Since the FGF is mainly stabilized by $\pi-\pi$ and H-bond intramolecular interactions, comparing the data obtained at different levels of theory with the benchmark data implicitly tests which method is capable of providing a balanced and accurate treatment of these intramolecular interactions. We have shown that the TPSS-D/6-311++G(3df,3pd), SCC-DF-TB-D, PBE-D/TZVP, and BH&H/6-311++G(d,p), in this order, overstabilize those conformers having an OH$\cdots$O=C intramolecular H-bond.

All the above-mentioned conclusions are in agreement with those obtained after a similar study performed on isolated peptides as model systems of NH$\cdots\pi$ interaction in proteins. This implies that we can combine the results obtained from these two reductionist approaches to obtain a more general overview on the noncovalent interactions occurring in the hydrophobic core of a protein.

FGF is a realistic model system of aromatic−aromatic side chain interactions of adjacent residues along a protein sequence, and consequently, plenty of biological information can be obtained from its study and further used to shed some light on the protein folding process. From the very many structural aspects that can be analyzed, we have focused on the preferred orientation of the aromatic side chains with respect to each other. We have shown that neither the T-shaped nor the stacked orientations are favored. Indeed, for the vast majority of conformers, aromatic side chains adopt a geometrical disposition that favors the maximum number of noncovalent intramolecular interactions.

Benchmark data have been included in the Benchmark Energy & Geometry Database (BEGDB) (http://www.begdb.com/), which aims to provide benchmarks for the testing of many other methods.

Acknowledgment. This work was a part of the research project No. Z40550506 of the Institute of Organic Chemistry

and Biochemistry, Academy of Sciences of the Czech Republic, and it was supported by Grants No. LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. H.V. acknowledges the support of the government of Principado de Asturias under the program Plan de Ciencia, Tecnología e Innovación (PCTI) 2006−2009. We thank K. Berka for his help in evaluating the percentage of aromatic side chains that are likely involved in $\pi-\pi$ intramolecular interactions with neighbour aromatic side chains. A portion of the research described in this paper was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

Supporting Information Available: Cartesian coordinates of all the structures considered in the set, time scale for the different computational methods employed in the study (Table S1), relative energies calculated at different levels of theory (Table S2), and relative Gibbs energies (Table S3). This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Murray, R. F.; Harper, H. W.; Granner, D. K.; Mayes, P. A.; Rodwell, V. W. *In Harper's Illustrated Biochemistry*, 27th ed.; Lange Medical Books/McGraw-Hill: New York, 2006; p 30.

(2) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. *J. Phys. Chem. A* **2007**, *111* (6), 1146.

(3) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747.

(4) Valdes, H.; Spiwok, V.; Rezac, J.; Reha, D.; Abo-Riziq, A. G.; de Vries, M. S.; Hobza, P. *Chem.—Eur. J.* **2008**, *14*, 4886.

(5) Reha, D.; Valdes, H.; Vondrasek, J.; Hobza, P.; Abo-Riziq, A. G.; Crews, B.; de Vries, M. S. *Chem.—Eur. J.* **2005**, *11* (23), 6803.

(6) Valdes, H.; Reha, D.; Hobza, P. *J. Phys. Chem. B* **2006**, *110* (12), 6385.

(7) Hunter, C. A.; Lawson, K. R.; Perkins, J.; Urch, C. *J. Chem. Soc., Perkin Trans.* **2001**, *2*, 651.

(8) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656.

(9) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42* (11), 1210.

(10) Grimme, S. *J. Comput. Chem.* **2004**, *25* (12), 1463.

(11) Grimme, S. *J. Comput. Chem.* **2006**, *27* (15), 1787.

(12) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28* (2), 555.

(13) Distasio, R. A., Jr.; Head-Gordon, M. *Mol. Phys.* **2007**, *1058*, 1073.

(14) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(15) Jung, Y. S.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(16) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(17) Rodham, D. A.; Suzuki, S.; Suenram, R. D.; Lovas, F. J.; Dasgupta, S.; Goddard, W. A.; Blake, G. A. *Nature* **1993**, *362*, 735.

(18) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122* (46), 11450.

(19) Chipot, C.; Jaffe, R.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118* (45), 11217.

(20) Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442* (1−3), 42.

(21) Valdes, H.; Klusak, V.; Pitonak, M.; Exner, O.; Stary, I.; Hobza, P.; Rulisek, L. *J. Comput. Chem.* **2008**, *29*, 861.

(22) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110* (1), 8.

(23) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(24) Bakker, J. M.; Plutzer, C.; Hunig, I.; Haber, T.; Compagnon, I.; von Helden, G.; Meijer, G.; Kleinermanns, K. *Chem. Phys. Chem* **2005**, *6* (1), 120.

(25) Chass, G. A.; Mirasol, R. S.; Setiadi, D. H.; Tang, T. H.; Chin, W.; Mons, M.; Dimicoli, I.; Dognon, J. P.; Viskolcz, B.; Lovas, S.; Penke, B.; Csizmadia, I. G. *J. Phys. Chem. A* **2005**, *109* (24), 5289.

(26) Fricke, H.; Funk, A.; Schrader, T.; Gerhards, M. *Phys. Chem. Chem. Phys.* **2007**, *9* (32), 4592.

(27) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049.

(28) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97* (40), 10269.

(29) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114* (12), 5149.

(30) Becke, A. D. *Phys. Rev. A* **1988**, *38* (6), 3098.

(31) (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650. (b) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P.; von, R. *J. Comput. Chem.* **1983**, *4*, 294. (c) Gill, P. M. W.; Johnson, B. G.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1992**, *197*, 499. (d) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(32) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(33) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93* (5), 281.

(34) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1−18.

(35) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(36) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37* (2), 785.

(37) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78* (7), 1396.

(38) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240* (4), 283.

(39) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97* (1−4), 119.

(40) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96* (9), 6796.

(41) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286* (3−4), 243.

(42) Jurecka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365* (1−2), 89.

(43) Pitonak, M.; Janowski, T.; Neogrady, P.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1761.

(44) Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Rezac, T.; Pitonak, M.; Vondrasek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73* (10), 1261.

(45) Beran, G. J. O.; Head-Gordon, M.; Gwaltney, S. R. *J. Chem. Phys.* **2006**, *124*, 114107.

(46) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100* (48), 18790.

(47) Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V. *J. Am. Chem. Soc.* **2002**, *124*, 6133.

(48) Gloaguen, E.; Valdes, H.; Pagliarulo, F.; Pollet, R.; Tardivel, B.; Hobza, P.; Piuzzi, F.; Mons, M. (Personal communication).

CT900174F

JCTC Journal of Chemical Theory and Computation

Quantitative Assessment of Electrostatic Embedding in Density Functional Theory Calculations of Biomolecular Systems

J.-L. Fattebert,*,† R. J. Law,‡,§ B. Bennion,‡ E. Y. Lau,‡ E. Schwegler,‡ and F. C. Lightstone‡

Center for Applied Scientific Computing, and Physical & Life Sciences, Lawrence Livermore National Laboratory, Livermore, California 94551

Received April 28, 2009

Abstract: We evaluate the accuracy of density functional theory quantum calculations of biomolecular subsystems using a simple electrostatic embedding scheme. Our scheme is based on dividing the system of interest into a primary and secondary subsystem. A finite difference discretization of the Kohn−Sham equations is used for the primary subsystem, while its electrostatic environment is modeled with a simple one-electron potential. Force-field atomic partial charges are used to generate smeared Gaussian charge densities and to model the secondary subsystem. We illustrate the utility of this approach with calculations of truncated dipeptide chains. We analyze quantitatively the accuracy of this approach by calculating atomic forces and comparing results with full QM calculations. The impact of the choice made in terminating dangling bonds at the frontier of the QM region is also investigated.

1. Introduction

Enzymes are the most proficient catalyst known. They are able to accelerate reactions that take as long as 78 million years to 26 ms under ambient conditions corresponding to a rate enhancement of 10^{17}. How enzymes are able to lower the reaction energy barrier is still being debated and a highly active area of study (see for example refs 1−4). Quantum mechanical calculations of the solvated enzyme with reactant would show how the energy barrier is being lowered, but even with modern computers this type of calculation is prohibitive. Most quantum mechanical studies of enzymes use a model system of the active site, typically a few residue side chains and reactant, in the gas phase or with a continuum solvent (see for example refs 5 and 6). Although these calculations can provide useful insights, they do not take into account the heterogeneous nature of an enzyme active site which usually provides a significant contribution in lowering a reaction's activation barrier or the role dynamics can play in catalysis.[7,8]

Proteins are composed of amino acids linked into a linear sequence by amide bonds, also known as peptide bonds, between the amino group of one amino acid and the carboxyl group of the next amino acid. The side chains of the amino acids are diverse and can be polar or nonpolar, acidic or basic, and aromatic or aliphatic. They are capable of carrying a net electrical charge depending on the nature of the surrounding environment. These charges generate long-range Coulomb potentials which affect the electronic structure tens of angstroms away. In modeling an environment made of proteins and water, Coulomb effects will dominate long-range interactions and can influence the active site of a protein.

Because of the very high computational cost of quantum calculations, modeling protein systems at the quantum mechanical level is restricted to rather small systems. While the active site, where a quantum description is required to describe the chemical reaction, is often not very large, a realistic environment surrounding that region is necessary if one hopes to attain useful information out of a small active site calculation.[8] Frequently, the active site is within the

* To whom correspondence should be addressed. Email: fatteber1@llnl.gov.

† Center for Applied Scientific Computing.

‡ Physical & Life Sciences.

§ Current address: Evotec, Oxford, U.K.

10.1021/ct900209y CCC: $40.75
Published on Web 07/31/2009

whole protein such that isolating the active site for a quantum calculations (also called primary subsystem) would involve "cutting" the active site from the remaining part of the protein. It is very likely that covalent bonds will need to be cut. A common method to create the isolated active site is to cut the covalent bonds between the active site and the "environment" (also called the secondary subsystem) and terminate these covalent bonds in the primary subsystem with hydrogens (link atoms) to satisfy valency.

Combined quantum mechanics/molecular mechanics (QM/MM) approaches have become widely used (see for example refs 9 and 10 for recent reviews) for modeling chemical reactions in large protein systems. They allow the simulation of a large effective system using quantum modeling for only a small subsystem, the active site, while the rest of the system is treated using a classical force field approach. Past studies using these methods were limited in the size of the QM region and utilized semiempirical Hamiltonians such as AM1 or PM3. More recently, the self-consistent-charge density functional tight-binding method is gaining popularity for biological QM/MM studies.[11] However, all these methods still face serious technical difficulties in describing the interface between the QM and the MM regions when covalent bonds have to be cut, which is often the case in biochemical applications. These QM/MM methods also require coupling of two completely different models which leads to very complex simulation codes or codes coupling.

In this paper, we evaluate the accuracy of a simpler model, which avoids the whole MM force-field bonded and van der Waals interactions machinery, and is easy to implement in a QM code. We investigate the quantitative effect of electrostatically embedding a QM subsystem into a classical biomolecular mechanics system simulation. The QM computation for the active site is carried out in the presence of a simple one-electron potential describing the electrostatic interaction with the atoms of the secondary subsystem which accounts for the influence of the protein environment. In our approach, the partial point charges associated to the secondary subsystem in a force field model are used to build smeared Gaussian charge distributions which are then used to compute an effective electrostatic potential. Such a model obviously leads to inaccurate forces for QM atoms at the boundary of the primary subsystem where bonded interactions with MM atoms would dominate, and the coordinates of these atoms should be frozen in a molecular dynamics simulation. We focus however on the accuracy of forces deeper inside the QM region, at the active site of interest. A similar study was carried out by Solt et al.[12] in a full QM/MM framework where they attempt to assess in a systematic manner simulation errors at the center of the QM region as a function of the size of that region. They showed that even with a full MM coupling, there are significant errors in the forces affecting atoms near the edge of the QM region.

By reducing the secondary subsystem to a simple electrostatic environment, the interaction between primary and secondary subsystems is greatly simplified. In particular this scheme avoids the introduction of additional parameters necessary for MM calculations such as the van der Waals interaction. While the dominance of the electrostatic effects is widely recognized in modeling a biomolecular environment, a direct accuracy comparison with respect to a fully QM approach is rarely found in the scientific literature. Also, partial atomic charges are typically parametrized and static for MM calculations, and the use of those for the purpose of QM embedding needs to be validated.

To address the difficulty of cutting and terminating the boundary region between the active site and the remaining protein, we evaluate the quality of various schemes to partition a protein system into a primary and secondary subsystem. We adopt the commonly used link-atom scheme,[13] saturating cut dangling bonds at the QM boundary with a hydrogen atom. While there are no definitive rules about how to partition a biomolecular systems into QM and MM parts, some bonds are better candidates than other bonds for cutting and capping with a link atom. Since capping atoms are usually H atoms (or pseudo H atoms) that form a σ bond with the QM system, the most favorable bonds to cut are bonds with the similar characteristics.[13] We will discuss and numerically evaluate the accuracy of cutting covalent bonds at various locations near the amide moiety in the protein backbone.

The purpose of this paper is to validate this simple electrostatic embedding approach by a quantitative evaluation of the perturbation introduced on a quantum mechanical subsystem of interest when part of it is replaced by a one-electron potential. Since our goal is in molecular dynamics simulations of the QM system, we focus on the accuracy of QM atomic forces away from the QM boundary when peptide bonds are cut and the secondary subsystem is modeled as an electrostatic potential. We study small systems such as dipeptides for which we can carry out a full reference QM calculation. We also pay particular attention to the problem of finding the optimal location to cut covalent bonds. This provides an assessment of the accuracy of the scheme we use to construct the one-electron potential describing the electrostatic environment, as well as a quantitative comparison between various link-atom schemes used to cut and terminate covalent bonds.

2. Computational Methods

We use density functional theory (DFT) as the QM model. While DFT has known deficiencies, in particular, to properly describe intermolecular interactions, especially van der Waals forces, it is currently one the most suitable quantum approaches for large scale electronic structure calculations in chemistry and solid-state physics (see for example ref 14).

QM calculations were carried out using a pseudopotential finite difference approach for discretizing the Kohn−Sham equations. Electronic wave functions, potentials, and the electronic density were described by their values at each grid point of a uniform real-space mesh covering the computational domain, and a fourth-order finite difference scheme is used to evaluate the Laplacian in the Hamiltonian operator.[15,16] Exchange and correlation were modeled using the PBE functional.[17] We used norm-conserving pseudopotentials expressed in a Kleinman−Bylander form.[18] We assumed that we have a globally neutral periodic system and use periodic boundary conditions.

Table 1. Radii Used To Generate Gaussian Charge Distributions

species	r_a (Å)
H	0.40
C	0.73
N	0.71
O	0.80

One advantage of such a QM approach is the lack of basis set superposition error often present in LCAO approaches. It allows us in particular to easily assess the accuracy of our model by comparing full QM calculations with subsystems simulations. The DFT real-space finite difference approach is also very adequate for parallelism and can scale on a large number of processors,[15] thus enabling calculations with a relatively large quantum region. For large systems, linear scaling can be achieved using a localized representation of the occupied Kohn−Sham orbitals subspace.[16,19] Note that Takahashi et al.[20] have used such an approach previously for QM/MM calculations of molecules in solution.

For the molecular systems in which we are interested, we extracted a subsystem and treated this subsystem at the quantum mechanical level. This requires properly terminating all the covalent bonds cut during the extraction and also appropriately modeling the atoms outside the QM boundary. Cutting out peptide chains was done at various locations along the chain. A few possible choices were evaluated and are described in the next section.

To model the part of the system not included in the QM simulation, we used a simple electrostatic embedding. We employed the partial charges associated to each atom in a classical force field approach.[21] In several classical force fields, atoms belong to a *charge group* of total charge −1, 0, or 1. To avoid overlap between QM link atoms and MM charges, we excluded charge groups that overlap with link atoms.[22−24] Only charge groups that were fully cut out of the QM region and did not overlap with the termini were included in the list of charges used to generate the electrostatic potential. That usually meant excluding charges from neutral CHNH groups (backbone α carbon+hydrogen and amide nitrogen+hydrogen) immediately connected to a cut peptide bond. One special case was the glycine residue where the "core" charge group also includes an additional H atom (the side chain).

Instead of directly using the point charges associated to atoms outside the QM region, we used smeared charges. For each atom a contributing to the electrostatic potential, we associated a Gaussian charge distribution

$$\rho_a(\mathbf{r}) = Z_a \frac{e^{-(\mathbf{r} - \mathbf{R}_a)^2/r_a^2}}{\pi^{3/2} r_a^3} \tag{1}$$

where Z_a is the classical force field partial charge of the atom, $\mathbf{R}_a$ is its position, and r_a is a radius associated to each atomic species. Values close to the covalent radii can be typically used for that purpose. We used the values listed in Table 1.

Such charge smearing has been used previously, in particular for charges close to the QM boundary in order to soften the interaction between atoms that are separated by only a short distance.[25,26] Indeed, at short distance, the validity of a point charge representation is questionable. In the context of this paper, as well as in a plane wave approach,[27] such a smearing is a convenient way of generating a charge density and a resulting Coulomb potential representable on a uniform mesh.

A total charge density associated with the atoms in the secondary subsystem was computed by summing up all these Gaussian charges

$$\rho_{\text{ext}}(\mathbf{r}) = \sum_a \rho_a(\mathbf{r}) \tag{2}$$

This total charge density was evaluated at every grid point of the real-space mesh used in the quantum calculation. The resulting Coulomb potential was calculated by solving the Poisson problem

$$-\nabla^2 V_{\text{ext}}(\mathbf{r}) = 4\pi\rho_{\text{ext}}(\mathbf{r}) \tag{3}$$

with periodic boundary conditions, by discretizing the Poisson equation on the real-space mesh, the same used for electronic wave functions, by finite differences and solving the resulting linear system by the multigrid method.[28] Note that even if ρ_{ext} may not correspond to a neutral charge, we assume that the sum $\rho_{\text{ext}} + \rho$, where ρ is the charge density of the QM part, is charge neutral. Thus, one can solve eq 3 for the whole system to get the total electrostatic potential without difficulties related to periodic boundary conditions by adding a uniform neutralizing background charge to ρ_{ext} and subtracting it from ρ in the QM calculation. V_{ext} was computed only once at the beginning of the calculation and added to the Kohn−Sham potential in the QM calculation.

Since our goal is in performing molecular dynamics for the QM system, we focus in the next section on force calculations and compare forces computed in truncated systems in reference to full QM calculations. Because we did not include MM atoms, we are obviously missing important forces between atoms linked by covalent bonds at the QM frontier. To address this problem in a molecular dynamics or geometry optimization context, the coordinates of some atoms at the QM frontier should be frozen so that these erroneous forces would not be active.

As is the case when solving the Kohn−Sham equations with a plane wave basis, the electronic wave functions in a real-space finite difference method are in general not restricted to be local in space. As pointed out in Laio et al.,[27] carrying out such a computation in an external Coulomb potential due to MM charges can potentially lead to the so-called electron spill-out effect. While one must be aware of this issue, the calculations reported in this paper using the parameters tabulated in Table 1 showed no measurable evidence of this phenomena.

3. Dipeptide Studies

Ferre et al. showed that it is very difficult to apply the link-atom scheme when cutting amide bonds, such as peptide bonds in a protein.[29] To investigate various cutting procedures, we first quantitatively evaluate the effect of cutting and terminating a dipeptide with no net charge (Ala-Gly).

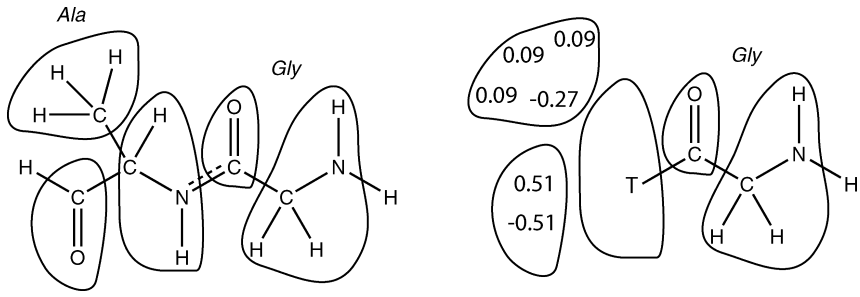


Figure 1. Left: Dipeptide Ala-Gly and the corresponding charge groups (denoted by solid contours). Right: glycine and charges substituted for alanine.

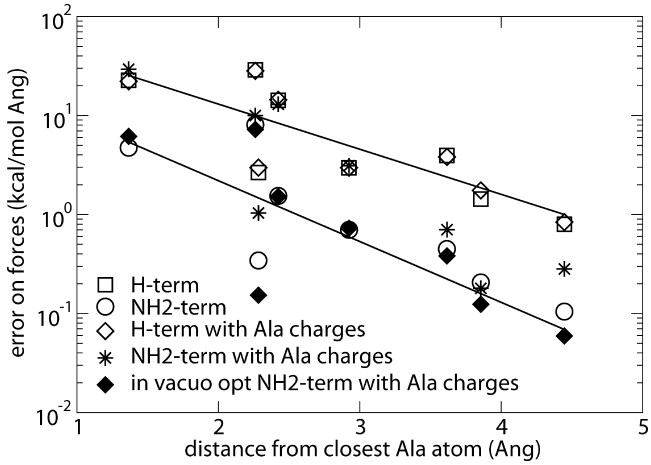


Figure 2. Absolute error on atomic forces in the glycine residue when removing alanine. The lines indicate exponential fits for cases with smallest error (in vacuo opt NH_2-term with Ala Charges) and largest error (H-term with Ala charges). The average errors are 9.7, 2.0, 9.7, 7.2, and 2.1 kcal/mol·Å for the five sets of plotted data. The first data points on the left correspond to the carbon atom of alanine bonded to the termini. The next three data points correspond to the oxygen, the C_α of Gly, and a hydrogen atom bonded to the C_α.

To measure the effects of the Coulomb potential associated to partial charges of atoms cut out of the QM subsystem, we next study dipeptides with net charges (Arg-Asp).

3.1. Alanine-Glycine. We consider the dipeptide H-Ala-Gly-H (Figure 1). In this form (with the H-termini), the compound is stable in gas phase and is charge neutral. We evaluate the effect of cutting the alanine out of the QM calculation. More precisely, we evaluate the atomic forces on the glycine peptide atoms after cutting out the alanine residue while properly terminating the cut peptide bond and compare these forces with those computed for the full QM system. We compare the results using the termini T = H and T = NH_2. Note that a NH_2 termini is equivalent to cutting the peptide chain one bond away from the peptide bond and capping the dangling bond with a H termini. We also look at the quantitative effect of including the Coulomb potential associated to the atoms cut out. The Coulomb potential does not include the charges of the CHNH group removed since some of the atoms would overlap with the termini.

Figure 2 shows the error introduced on the forces when cutting out part of the QM system (the Ala residue). For all the atoms in the glycine part, the error is shown as a function of their distance to the (removed) alanine atoms. Numerical

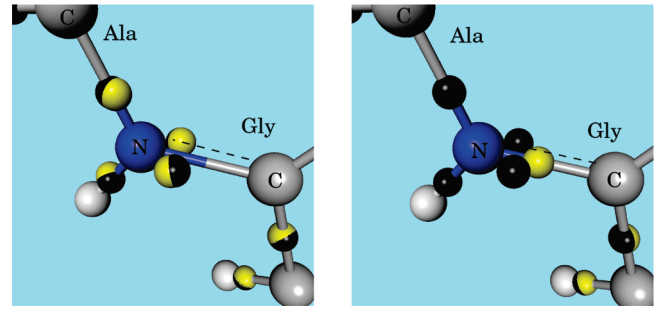


Figure 3. Centers of maximally localized Wannier functions near the peptide bond in Ala-Gly: full QM dipeptide calculation result (black balls) compared with the glycine calculation with NH_2 (left) and H (right) termini (yellow balls). The C−N−C sequence shown correspond to the C−N peptide bond (glycine on right) followed by the C_α of alanine (removed in truncated system).

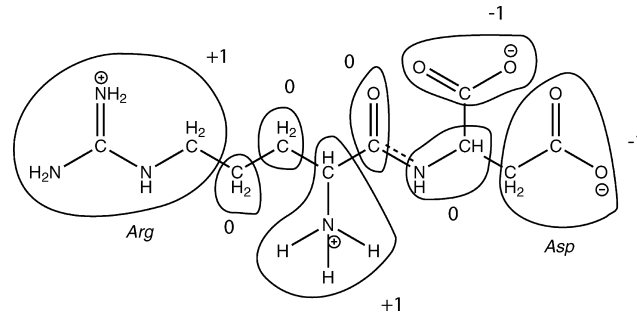


Figure 4. Dipeptide Arg-Asp and constituting charge groups.

results show that the most important aspect of the cut is to properly terminate the cut peptide bond using an NH_2 group instead of a simple atom. In comparison, the effect of including the electrostatic field due to the partial charges of the removed atoms is very small. This is due to the fact that the peptide bond is not as simple as the C−H bond. The peptide bond has two resonance forms which confers partial double character. Analyzing the electronic structure of the NH_2- and H-terminated glycine using maximally localized Wannier functions,[30] one can see the different orbital character from the resulting two termini. The NH_2-terminated glycine has Wannier function centers that overlap those Wannier function centers of the full system (Figure 3, left). The terminated glycine, on the other hand, has a bond-centered Wannier function center compared to the two off-bond Wannier function centers of the full system (Figure 3, right). This confirms the observation by Ferre et al.[29] that a

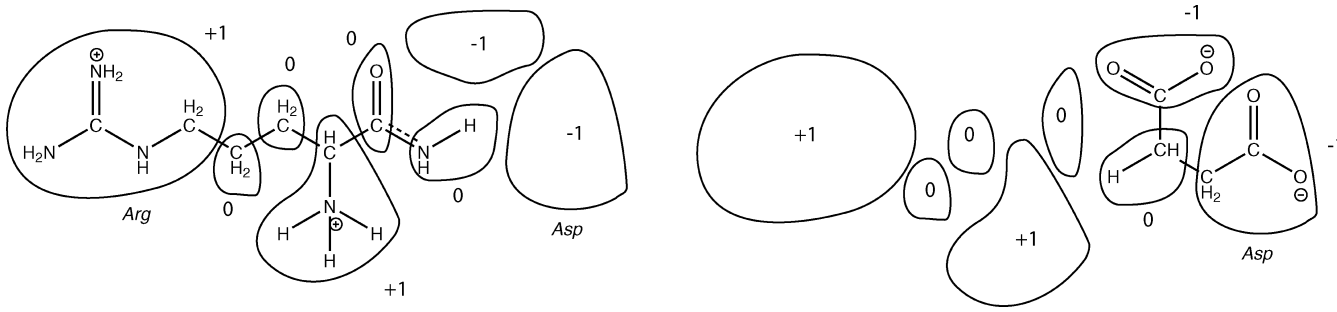


Figure 5. Truncated Arg-Asp dipeptide. Left: Arg peptide after removing Asp and replacing it with an NH_2 terminus. Right: Asp peptide after removing Arg and substituting a H terminus for the dangling NH group.

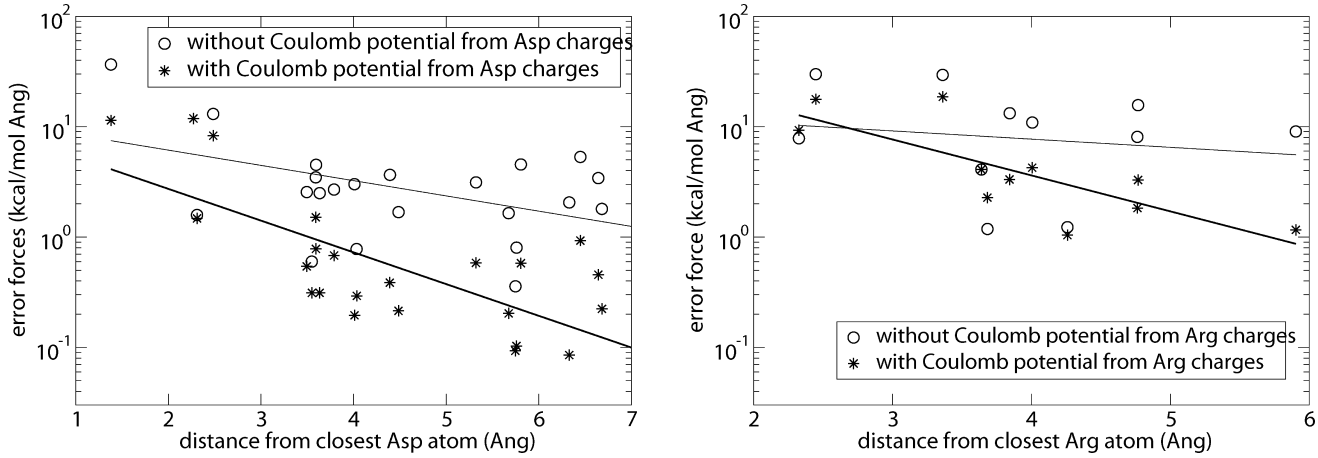


Figure 6. Absolute error on atomic forces on remaining atoms in Arg-Asp when removing Asp (Left) or Arg (right), as a function of distance from the closest atom in the removed peptide. The lines indicate exponential fits for each set of results.

cut peptide bond capped with a simple H terminus is not a suitable approach.

Another interesting result is that optimizing the geometry of the termini in the electrostatic field leads to worse geometry than optimizing in vacuo because the short distance between the termini and the Gaussian charges modeling the environment results in much too strong a coupling. Thus, optimization of termini will be done in vacuo for the test systems presented in the rest of this paper, and bond cuts will be chosen one bond away from the peptide bonds.

3.2. Arginine-Aspartate. We now consider peptides with net electrical charges, both at their termini and in the side chain. We start with the zwitterionic Arg-Asp dipeptide (see Figure 4) solvated in water. The atomic configuration was taken from a snapshot of a classical molecular dynamics simulation. To define the reference system, we first replace the water molecules (solvent) by the Coulomb potential resulting from their partial charges according to the procedure described in Section 2. This potential is sufficiently good to have a stable quantum system made of a dipeptide with net local charges. A DFT calculation is carried out and atomic forces are computed for the whole QM system. We verify that each maximally localized Wannier center for this system closely matches the corresponding centers of the system solvated in water within 0.025 Å accuracy. Next, we cut out one of the two residues and replace it with the Coulomb potential resulting from the partial charges associated to the atoms we have removed according to the procedure described in Section 2.

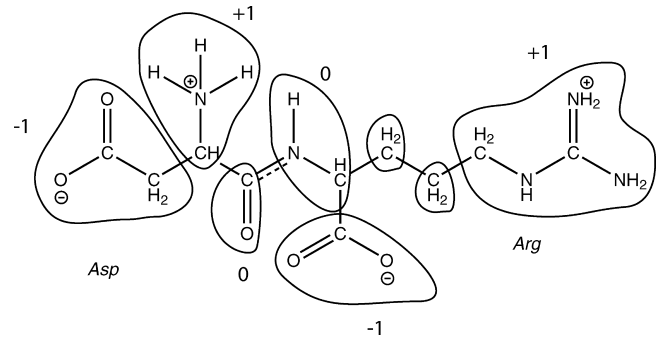


Figure 7. Dipeptide Asp-Arg and constitutive charge groups.

We began by removing Asp from the QM region (see Figure 5). The Arg residue remains intact with a net charge of +2, while the Asp is replaced by an NH_2 terminus and a Gaussian charge distribution. The NH_2 terminus is equivalent to a cut one bond away from the peptide bond capped with a H atom. The geometry of the NH_2 terminus is optimized without the electrostatic field generated by Asp charges, according to the observation made in Section 3.1. Then, the Coulomb potential of the Asp charges minus the CHNH group, which overlaps with the terminus, is turned on (net charge −2). DFT forces were evaluated and compared with the reference calculation for the full dipeptide. Additionally, the DFT forces were compared to the result obtained when ignoring the Coulomb potential due to Asp charges (see Figure 6). The inclusion of the Coulomb potential represents

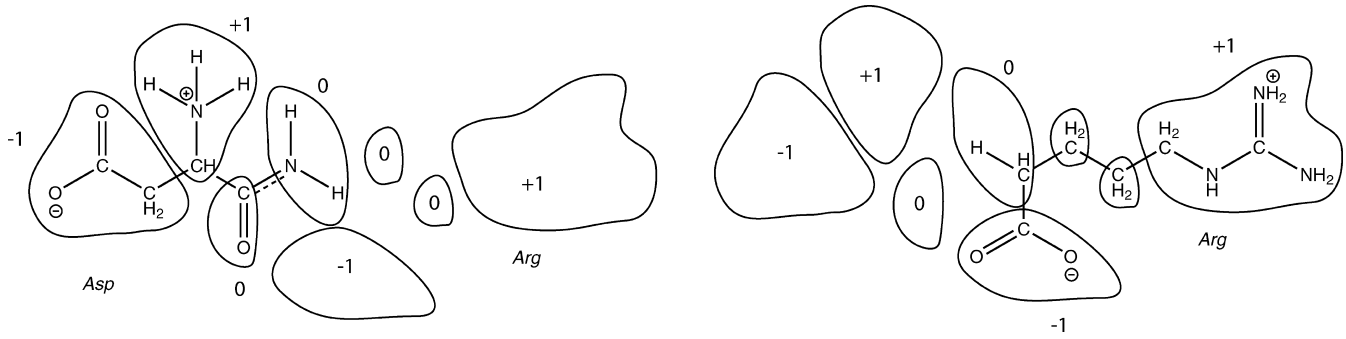


Figure 8. Truncated Asp-Arg dipeptide. Left: NH_2-terminated Asp residue after removing Arg. Right: Terminated Arg peptide after removing Asp.

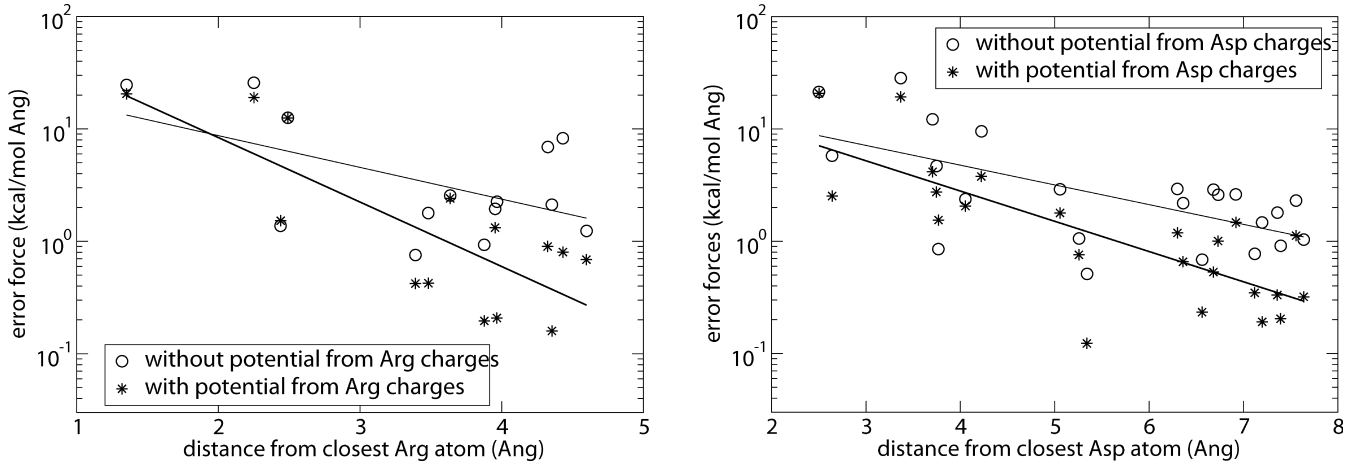


Figure 9. Absolute error on atomic forces on remaining atoms in Asp-Arg when cutting out Arg (Left) or Asp (right), as a function of distance from the closest atom in cut-out peptide. The lines indicate exponential fits for each set of results.

a clear improvement over its deletion. The error is reduced from 5.1 to 1.6 kcal/mol·Å on average, and the trend shows an even larger improvement for atoms further away from the QM boundary. Not surprisingly, the error due to missing bonded forces dominate near the termini, while the Coulomb effect is longer range. We also conclude that the error on forces decays quickly and is within a tolerable accuracy about 3.5 Å away from the removed atoms. For comparison, geometry optimization algorithms for quantum systems typically use tolerance of the order of 5×10^{-4} au. (0.6 kcal/mol·Å).

We also carry out the same study for the case where Arg is removed and Asp is treated at the QM level. The cut is made inside Asp where the NH terminal group is replaced by a single H (see Figure 5). Again, we see a clear improvement when including the Coulomb potential associated to the Arg charges (Figure 6). The error is reduced from 11.9 to 6.1 kcal/mol·Å on average, but the trend shows a larger improvement for atoms further away from the QM boundary. Error is reduced by a factor 10 at a distance 6.5 Å from the closest Asp atom.

3.3. Aspartate-Arginine. We now consider the same peptide chain, but in reverse order, that is Asp on the left with a NH_3^+ terminus and Arg on the right with a COO^- terminus. In this case both residues have no net charge because their respective terminus neutralizes their charge. We have, however, on both sides a large dipole due to two groups with net charges compensating for each other (see Figure 7). The same study as in the previous section is carried out for this system. Cut-out systems are shown in Figure 8. Numerical accuracy of forces is plotted in Figure 9. While no net charge is cut out this time, the importance of including the long-range Coulomb effect of the removed atoms remains. Inclusion of Coulomb charges from removed atoms reduces the error from 6.7 to 4.4 kcal/mol·Å for the first system, and from 4.9 to 2.9 kcal/mol·Å for the second one. Again, we see that the error is dominated by missing bonded forces close to the QM boundary, while taking into account long-range Coulomb effects allows us to significantly reduce the error away from the boundary.

4. Conclusion

In this paper we have presented an electrostatic embedding methodology for reducing QM calculations of protein systems to a small QM subsystem. The approach can be decomposed to five points: (1) cut and properly terminate peptide bonds, (2) optimize the termini atoms for the resulting QM system, (3) make a list of removed atoms, excluding those in charge groups overlapping with QM system, (4) associate each of these atoms to a Gaussian charge distribution, of charge equivalent to partial charges associated in a classical force field, (5) compute Coulomb potential associated to the sum of these charge distributions and use it as an external potential for QM calculation.

We found that the following scheme was appropriate to cut peptide bonds between the QM inside "in" region and the outside "out" region: a peptide bond "in-CO−NH-out" is cut and terminated as "in-CO−NH_2", while the peptide bond "in-NH−CO-out" is terminated as "in-H. This scheme leads to much more accurate results than a simple cut of the peptide bond capped with a H link atom.

Our quantitative results show that the inclusion of the Coulomb potentials resulting from partial charges of atoms cut out of the primary system improves systematically and dramatically the accuracy of the calculated forces inside the QM region compared to an "in vacuo" environment. Specifically, we observe that forces on QM atoms 3.5 Å or further away from the QM boundary are quite accurate when using this procedure.

The present study shows that with a good electrostatic embedding model, QM atoms two to three bonds away from the QM boundary experience forces very close to those they would experience in a full QM calculation. This suggests that a simple electrostatic embedding model such as the one described in this paper is appropriate to simulate the environment for quantum biomolecular subsystems. When mechanical coupling with MM atoms is not included in the calculation, one can freeze the atomic coordinates of a "shell" of quantum atoms at the frontier of the quantum region. The resulting quantum subsystem is thus mechanically constrained at its boundary and embedded in a long-range external electrostatic potential, mimicking the surrounding protein.

As mentioned in Section 2, a generalized gradient approximation (GGA) was used to model DFT exchange and correlation in our quantum calculations (PBE functional[17]). The methodology presented in this paper for dealing with long electrostatic effects and environment modeling is however not limited to this functional, and recent (and future) progress toward better functionals could be incorporated into the quantum modeling to improve the level of theory. In particular, biological systems such as those considered in this paper could benefit from a better treatment of dispersion energy using for instance a van der Waals density functional.[31] Hybrid-GGA such as B3LYP,[32] based on exact-exchange energy terms, are also quite popular in quantum modeling of biomolecular systems. While exact-exchange is not as straightforward to implement in a real-space finite difference context as it is for a Gaussian basis set, recent progress in the field has shown that it is now possible to incorporate and use such functionals in a plane waves code.[33] Even if exact-exchange in this context is computationally quite expensive compared to local density or generalized gradient approximations, the methodology described in ref 33 is directly applicable to a real-space finite difference approach and should be considered as a possible future research direction.

Acknowledgment. We thank the Defense Threat Reduction Agency (BA07TAS072) for financial support. J.-L.F. would like to also acknowledge support by the Office of Science, U.S. Department of Energy, SciDAC Grant DE-FC02-06ER46262 for the development of the MGmol code used for the research presented in this paper. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

(1) Bruice, T. C. *Chem. Rev.* **2006**, *106*, 3119–3139.

(2) Mulholland, A. J. *Biochem. Soc. Trans.* **2008**, *036*, 22–26.

(3) Antoniou, D.; Basner, J.; Nunez, S.; Schwartz, S. D. *Chem. Rev.* **2006**, *106*, 3170–3187.

(4) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev.* **2006**, *106*, 3210–3235.

(5) Siegbahn, P. E. M. *Q. Rev. Biophys.* **2003**, *36*, 91–145.

(6) Himo, F. *Theor. Chem. Acc.* **2006**, *116*, 232–240.

(7) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Huebner, C. G.; Kern, D. *Nature* **2007**, *450*, 838–844.

(8) Mladenovic, M.; Arnone, M.; Fink, R.; Engels, B. *J. Phys. Chem. B* **2009**, *113*, 5072–5082.

(9) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(10) Senn, H.; Thiel, W. QM/MM Methods for Biological Systems. In *Atomistic Approaches in Modern Biology*; Springer: Berlin/Heidelberg, 2007; Vol. 268, pp 173−290.

(11) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316–325.

(12) Solt, I.; Kulhánek, P.; Simon, I.; Winfield, S.; Payne, M. C.; Csányi, G.; Fuxreiter, M. *J. Phys. Chem. B* **2009**, *113*, 5728–5735.

(13) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(14) Hafner, J.; Wolverton, C.; Ceder, G. *MRS Bull.* **2006**, *31*, 659–668.

(15) Briggs, E. L.; Sullivan, D. J.; Bernholc, J. *Phys. Rev. B* **1996**, *54*, 14362–14375.

(16) Fattebert, J.-L.; Bernholc, J. *Phys. Rev. B* **2000**, *62*, 1713–1722.

(17) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(18) Kleinman, L.; Bylander, D. *Phys. Rev. Lett.* **1982**, *48*, 1425–1428.

(19) Fattebert, J.-L.; Gygi, F. *Phys. Rev. B* **2006**, *73*, 115124.

(20) Takahashi, H.; Hori, T.; Hashimoto, H.; Nitta, T. *J. Comput. Chem.* **2001**, *22*, 1252–1261.

(21) MacKerell, A. D., Jr; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; John Wiley & Sons: Chichester, 1998; Vol. 1, pp 271−277.

(22) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189–1200.

(23) Antes, I.; Thiel, W. On the treatment of link atoms in hybrid methods. In *Combined Quantum Mechanical and Molecular Mechanical Methods*; American Chemical Society: Washington, DC, 1998; pp 50−65.

(24) Lyne, P.; Hodoscek, M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 3462–3471.

(25) Logunov, I.; Schulten, K. Quantum chemistry of in situ retinal: Study of the spectral properties and dark adaptation of bacteriorhodopsin. In Proceedings of the Ecole de Physique des Houches; Bicout, D., Field, M. J., Eds.; Springer: Paris, 1995; pp 235−256.

(26) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43–52.

(27) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.

(28) Briggs, W. L.; Henson, V. E.; McCormick, S. F. *A multigrid tutorial*, 2nd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, 2000.

(29) Ferre, N.; Olivucci, M. *Journal of Molecular Structure: THEOCHEM* **2003**, *632*, 71–82.

(30) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12865.

(31) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. B* **2007**, *76*, 125112.

(32) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(33) Wu, X.; Selloni, A.; Car, R. *Phys. Rev. B* **2009**, *79*, 085102.

CT900209Y

JCTC Journal of Chemical Theory and Computation

Theoretical Study of the Structure and Electronic Properties of $Si_3O_n^-$ and $Si_6O_n^-$ ($n = 1-6$) Clusters. Fragmentation and Formation Patterns

William Tiznado,*,† Ofelia B. Oña,*,‡ María C. Caputo,‡ Marta B. Ferraro,‡ and Patricio Fuentealba§

Departamento de Ciencias Químicas, Facultad de Ecología y Recursos Naturales, Universidad Andres Bello, Av. República 275, Santiago-Chile, Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria - Pab. I., Argentina, and Departamento de Física, Universidad de Chile, Las Palmeras 3425, Santiago-Chile

Received May 12, 2009

Abstract: A theoretical study of two series of small clusters, $Si_3O_n^-$ and $Si_6O_n^-$ ($n = 1-6$), has been carried out. The minimum energy structures were produced adding an electron to neutral species followed by relaxation at the B3LYP-6-311G(2d) level. The vertical ionization energies (VIEs) were computed using the electron propagator theory (EPT) in two approximations, Unrestricted Outer Valence Green Functions (UOVGF) and partial third-order approximation (P3). In the series $Si_3O_n^-$ the theoretical VIEs of the minimum energy structures agree well with experimental data. For the second series there are not experimental VIEs, and the theoretical results are predictions. The performance of EPT methodologies in conjunction with all-electron or pseudopotentials (PP) calculations is analyzed. The conjunction of P3 and PP approximation proves to be the most efficient and economical methodology to calculate the VIEs of small anionic silicon oxide clusters. In the series $Si_6O_n^-$ different channels of fragmentation have been calculated. The results suggest that the fragments do not have drastic geometric changes and the anionic fragment corresponds to the atoms where the spin density of the initial large cluster is localized. The Fukui function calculated over selected optimized fragments predicts adequately the interaction between them to form large stable clusters.

I. Introduction

In the past years it has been recognized that complex molecules and atomic clusters (ACs) often possess unique properties, which make them interesting objects of research. The unique properties of ACs are intimately related to its geometric and electronic structure. Therefore, a deep understanding of these properties can be essential for various practical applications including the design and formation of new nanostructures as well as the understanding of fundamental issues, such as functioning of quantum and thermodynamic laws (for a review, see refs1−5). The understanding of the principles of assembly and functioning of complex systems like nanoclusters is an open research field, and there is a large number of experimental and theoretical works that approach these problems from different perspectives.[1−5] Even though experimentally accessible quantities are often highly sensitive to cluster structures, there is not a general experimental method for determining such cluster structures. Therefore, detailed theoretical analysis in conjunction with the experimental results is necessary to determine them. In this context, photoelectron spectroscopy (PES) combined with theoretical calculations is one of the most powerful techniques to assign the geometric and electronic structures of clusters. The photoelectron spectra provide information

* Corresponding author E-mail: wtiznado@unab.cl (W.T.), ofelia@df.uba.ar (O.B.O.).

† Universidad Andres Bello.

‡ Universidad de Buenos Aires.

§ Universidad de Chile.

10.1021/ct900320r CCC: $40.75
Published on Web 07/31/2009

about the vertical ionization energies (VIEs) which are directly related to the electronic structure of the system and consequently with its geometric structure, which is generally the global minimum on the corresponding potential energy surfaces. Once the minimum energy structures have been identified, the VIEs can be theoretically calculated and compared with the experimental available data. In the past, this was performed by molecular orbital calculations (MO) and Koopmans' theorem, where the ionization energy is approximated as the negative of the one-electron MO energy.[6] Both correlated and uncorrelated orbital energies respectively based on density functional and Hartree−Fock methods may produce large errors in determining ionization energies and in some cases even give an erroneous ordering of the final electronic states[7−9]

A quasiparticle approximation in electron propagator theory (EPT) is a good methodology to determine vertical ionization energies which are comparable to experimental data. The most used approximation of EPT, known as the outer valence Green Function method, was developed by Cederbaum and co-workers for closed shell systems (ROVGF)[8,9] and by Ortiz and co-workers for open shell systems (UOVGF).[10−12] An efficient approximation to them, the partial third-order approximation, P3, was also developed by Ortiz et al.[13,14]

Silica is one of the most abundant materials on earth.[15] Its nanoparticles are an interesting object of study because of their importance in technological applications such as microelectronics, optics, glass manufacture, catalysis support, and fiber-optic communications.[16,17] Recently, it has been experimentally proved that silicon oxide clusters have a fundamental role in the growth of silicon nanowires.[18] In this context, the study of small clusters based on silicon and oxygen atoms is an interesting research field which is supported by a large number of experimental and theoretical publications.[19−30]

The present work focuses on applying different theoretical methodologies in the study of two sets of small anionic silicon oxide clusters, $Si_3O_n^-$ and $Si_6O_n^-$ ($n = 1-6$). These clusters have been chosen because they involve two series of oxidized silicon clusters where, as the number of oxygen atoms increases, they evolve from silicon rich clusters to oxygen rich clusters for the first set and to 1:1 (silicon and oxygen relation) clusters for the second set. For the series $Si_3O_n^-$, experimentally well resolved photoelectron spectra[18] will be used to evaluate the capability of EPT based methodologies, UOVGF,[10−12] and partial third-order approximation (P3)[13,14] to yield accurate results. For the series $Si_6O_n^-$ a theoretical study[30] about the global minimum will be also used to compare with our results.

In section III, the results are reported in a sequential way. First, we verified the minimum energy structures as reported in previous studies.[30] Then, we calculated the VIEs of the more stable isomers by means of the EPT methodologies. Theoretical assignment of the experimental photoelectron spectra was only possible for the first set of clusters, whereas in the second set the calculations are only predictive due to the absence of experimental data to compare with. In the second series of clusters ($Si_6O_n^-$) the minimum energy fragmentation channels were evaluated. Finally, we studied the formation of the largest clusters ($Si_6O_n^-$) from the interaction of the stabilized fragments. As predictor of the most probable interacting region we used condensed in atoms Fukui functions.

II. Methodologies

A. Anionic Structures. In order to determine the most stable isomers of $Si_3O_n^-$ and $Si_6O_n^-$ we employed the neutral structures taken from refs 20 and 29. For both sets of clusters we produced the anionic states adding one electron to each stable neutral isomer followed by B3LYP[31]/6-311G(2d)[32,33] relaxation to their corresponding local anionic minima.

B. Electron Propagator. Two approximated electron propagator methods have been applied to the VIEs predictions: the unrestricted outer valence Green's function (UOVGF) approximation[8,9,34] and the partial third-order approximation (P3)[13,14] which has clear computational advantages over UOVGF methods. They are used in combination with all-electron and pseudopotential (PP) calculations.

Whereas all-electron calculations employed the 6-311G-(2d)[32,33] basis sets, the Stuttgart pseudopotential for Si and O was combined with its corresponding basis functions.[35] Because the latter basis set does not contain diffuse and polarization functions, it was augmented with the most diffuse *s* and *p* functions and two contracted *d*-polarization functions of Sadlej's basis set.[36] This augmented basis set has already proven to be accurate enough for the calculation of the dipole polarizabilities of neutral silicon clusters and for VIEs calculations of small anionic silicon clusters.[37]

C. Local Reactivity Descriptor. The local reactivity descriptor used in this work is the Fukui function which was calculated by the finite difference approximation using atomic charges as proposed by Yang and Mortier,[38] $f_k^+ = q_k(m+1) - q_k(m)$ and $f_k^- = q_k(m) - q_k(m-1)$, where f_k^+ and f_k^- are the acceptor (electrophilic) and donor (nucleophilic) Fukui functions condensed at atom k respectively, m is the total electron number of the studied system, and $q_k(m)$, $q_k(m+1)$, and $q_k(m-1)$ are the atomic charges evaluated at the atom k in the geometry of the studied system by adding 0, 1, and −1 electrons, respectively. The atomic charges were obtained from two different methods, natural population analisys (NPA)[39−41] and ChelpG[42] (CHG). The last one was evaluated to include information about the possible electrostatic effects on the interactions.

All the calculations were done using the GAUSSIAN 03 package of programs.[43] The pictures of the structures and spin density isosurfaces were performed with the Molekel 4.3 visualization program.[44]

III. Results and Discussion

A. Search of the Minimum Energy Structures. *$Si_3O_n^-$ Clusters.* It was found that in all cases the previously reported structures[20] for the neutral clusters are the most stable ones. In the case of the $Si_3O_3^-$ cluster the planar C_s structure (see Figure 1-SI) was found as a local minimum isomer which is 0.09 eV above the global minimum structure

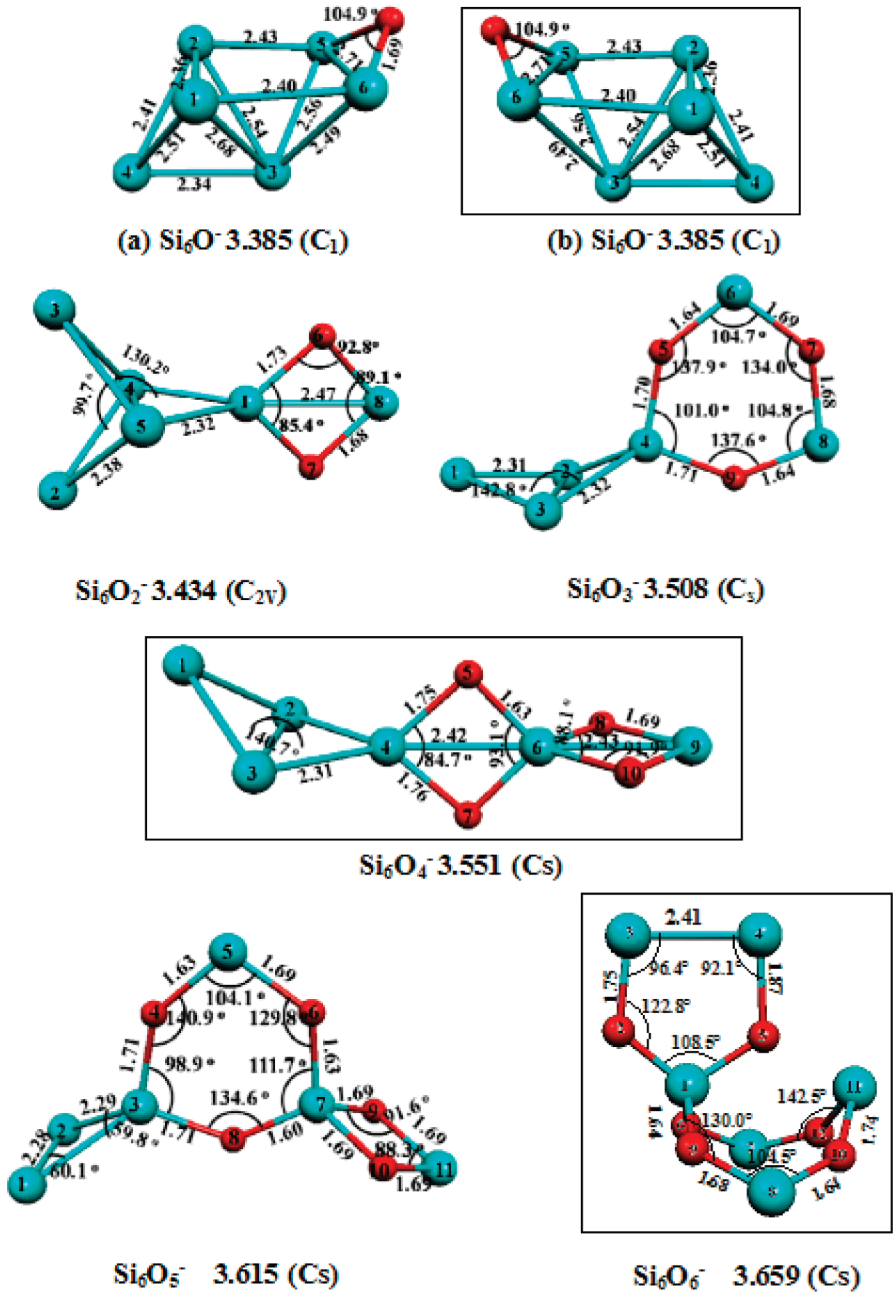


Figure 1. Structures of the ground state for anionic $Si_6O_n^-$ ($n = 1-6$) clusters obtained by adding one electron to these reported in ref 29 and followed by B3LYP/6-311G(2d) local optimization. Absolute binding energies per atom (BE/atom) are in eV, based on a Si atomic energy of −7847.211 eV, an anionic Si atomic energy −7847.193, and O atomic energy of −2038.703 eV ($BE = E(Si_6O_n^-) - 5E(Si) - E(Si^-) - nE(O_2/2)$). The new minimal structures found in this work are enclosed in frames.

at the MP2/cc-pVTZ[45,46] level, including the zero point energy correction in the energy calculations.

$Si_6O_n^-$ Clusters. After exploring different isomers we found some new structures with respect to those previously reported by Zang et al.[30] The clusters and the principal structural data are reported in Figure 1. Two isoenergetic Si_6O^- clusters were found, denoted as (a) and (b) in the figure, which are not superposable mirror structures of the same isomer. Zang et al.[30] only reported the Si_6O^-(a) ground state. Within our scheme the new $Si_6O_4^-$ and $Si_6O_6^-$ structures with C_s symmetry are the global minimum. They are 0.13 and 0.12 eV more stable (at the B3LYP/6-311G(2d) level) than the other isomers, respectively.

Calculation of the VIEs Using UOVGF and P3 Methodologies. Tables 1 and 2 present results of the theoretical calculation of the final state orbital assignments, VIEs, and pole strengths (in parentheses) of the $Si_3O_n^-$ and $Si_6O_n^-$ ($n = 1-6$) clusters in comparison with other theoretical predictions and experimental values. The expectation values of $\langle S^2 \rangle$ (total spin) for the reference Slater determinant are less than 0.8 (see Table 1-S), and the pole strengths are greater than 0.85 in all the studied systems. These results validate the quality of the electron propagator methods used in the present work to predict the VIEs.

$Si_3O_n^-$ Clusters. There is a good agreement between UOVGF and P3 results. In the all-electron results the higher

Table 1. Comparison of Experimental and Pseudopotential Calculations of Vertical Electron Detachment Energies (VEDEs) in eV for the $Si_3O_n^-$ ($n = 1-6$) Series[c]

				all-electron		pseudopotential			
system	initial state	final state	orbital	6-311G(2d) UOVGF(p)[a]	P36-311G(2d) (p)[a]	UOVGF(p)[a]	P3(p)[a]	experiment[b]	$\Delta E(E\nu+1\text{-}E\nu)$
		$^3B_2$	$5a_1$	2.67(0.89)	2.68(0.88)	2.91(0.88)	2.79(0.88)	3.1	
Si_3O^-	$^2B_2$	$^1A_2$	$2b_1$	2.41(0.89)	2.50(0.88)	2.62(0.88)	2.59(0.88)	2.6	
		$^3A_2$	$2b_1$	2.25(0.90)	2.34(0.89)	2.46(0.89)	2.43(0.88)	2.5	
		$^1A_1$	$3b_2$	1.58(0.89)	1.65(0.88)	1.79(0.88)	1.73(0.87)	1.96(0.06)	1.87
		$^3B_1$	$5a_1$	4.90(0.89)	4.73(0.89)	5.15(0.88)	4.83(0.89)		
$Si_3O_2^-$	$^2B_1$	$^1A_1$	$2b_1$	1.99(0.91)	1.94(0.91)	2.23(0.90)	2.07(0.90)	2.2	
		$^1B_1$	$6a_1$	1.97(0.89)	1.81(0.89)	2.20(0.88)	1.94(0.89)	2.1	
		$^3B_1$	$6a_1$	1.74(0.89)	1.60(0.89)	1.98(0.88)	1.73(0.88)	2.03(0.06)	1.83
		$^1A''$	$6a''$	4.15(0.87)	3.92(0.87)	4.40(0.86)	4.02(0.86)	4.2	
$Si_3O_3^-$	$^2A'$	$^3A'$	$9a'$	3.79(0.86)	3.62(0.87)	4.0(0.86)	3.68(0.86)	3.9	
		$^3A''$	$6a''$	3.71(0.85)	3.49(0.85)	3.94(0.84)	3.60(0.85)	3.8	
		$^1A'$	$10a'$	1.23(0.91)	1.17(0.91)	1.46(0.90)	1.28(0.91)	1.50(0.10)	1.54
		$^3B_2$	$7a_1$	6.50(0.91)	6.42(0.91)	6.71(0.91)	6.52(0.91)		
$Si_3O_4^-$	$^2B_2$	$^1B_2$	$8a_1$	5.07(0.87)	4.88(0.87)	5.36(0.86)	5.02(0.86)		
		$^3B_2$	$8a_1$	4.32(0.90)	4.11(0.90)	4.59(0.89)	4.27(0.89)	4.4	
		$^1A_1$	$5b_2$	0.98(0.93)	0.88(0.93)	1.21(0.92)	1.03(0.93)	1.055(0.050)	1.07
		$^1A''$	$7a''$	5.87(0.93)	5.48(0.91)	6.38(0.94)	5.66(0.91)		
$Si_3O_5^-$	$^2A'$	$^3A''$	$7a''$	5.76(0.94)	5.34(0.91)	6.25(0.94)	5.51(0.91)		
		$^3A'$	$14a'$	5.67(0.93)	5.29(0.91)	6.20(0.93)	5.51 (091)		
		$^1A'$	$15a'$	3.20(0.93)	3.27(0.92)	3.54(0.93)	3.45(0.92)	3.1(0.1)	3.11
		$^1A''$	$8a''$	6.36(0.93)	5.99(0.91)	6.85(0.94)	6.17(0.91)		
$Si_3O_6^-$	$^2A'$	$^3A''$	7a''	6.24(0.94)	5.85(0.92)	6.72(0.94)	6.02(0.91)		
		$^3A'$	$16a'$	6.15(0.93)	5.80(0.91)	6.68(0.93)	6.02(0.91)		
		$^1A'$	$17a'$	3.73(0.93)	3.85(0.92)	4.07(0.93)	4.03(0.92)	>3.5	3.68

[a] Pole strengths. [b] Anion photoelectron spectra.[20] [c] Theoretical predictions have been calculated with the OVGF approximation of electron propagator theory using the 6-311G(2d) basis set.

Table 2. Pseudopotential Calculations of Vertical Electron Detachment Energies (VEDEs) in eV for the $Si_6O_n^-$ ($n = 1-6$) Series[b]

system	initial state	final state	orbital	all-electron 6-311G(2d) UOVGF (p)[a]	pseudopotential (p)[a]
Si_6O^-	^2A	^3A	$14a$	3.30(0.89)	3.57(0.87)
		^1A	$15a$	3.14(0.89)	3.37(0.87)
		^3A	$15a$	2.99(0.89)	3.22(0.87)
		^1A	$16a$	2.67(0.89)	2.90(0.88)
$Si_6O_2^-$	$^2A_1$	$^1B_2$	$5b_2$	3.55(0.88)	3.57(0.87)
		$^3A_2$	$2a_2$	3.4290.88)	3.55(0.87)
		$^3B_2$	$5b_2$	3.20(0.88)	3.27(0.87)
		$^1A_1$	$9a_1$	3.09(0.89)	3.11(0.88)
$Si_6O_3^-$	$^2A''$	$^3A''$	$15a'$	3.61(0.88)	3.63(0.87)
		$^1A''$	$16a'$	3.49(0.88)	3.53(0.87)
		$^3A''$	$16a'$	3.31(0.88)	3.35(0.87)
		$^1A'$	$6a''$	2.11(0.88)	2.18(0.87)
$Si_6O_4^-$	$^2A''$	$^3A''$	$15a'$	3.80(0.88)	3.83(0.87)
		$^1A''$	$16a'$	3.70(0.88)	3.79(0.87)
		$^3A''$	$16a'$	3.54(0.89)	3.63(0.87)
		$^1A'$	$9a''$	2.38(0.88)	2.48(0.87)
$Si_6O_5^-$	$^2A''$	$^3A''$	$19a'$	3.17(0.89)	3.17(0.88)
		$^1A''$	$18a'$	3.02(0.88)	3.01(0.88)
		$^3A''$	$18a'$	2.82(0.88)	2.79(0.88)
		$^1A'$	$9a''$	2.71(0.88)	2.76(0.87)
$Si_6O_6^-$	$^2A''$	$^3A''$	$18a'$	4.73(0.89)	4.51(0.89)
		$^1A'$	$12a''$	3.21(0.89)	3.17(0.90)
		$^1A''$	$19a'$	3.15(0.91)	3.11(0.89)
		$^3A''$	$19a'$	2.99(0.88)	2.91(0.88)

[a] Pole strengths. [b] Theoretical predictions have been calculated with the UOVGF approximation of electron propagator theory using the 6-311G(2d) basis set.

differences are of the order of 0.4 eV between the two approximations (UOVGF and P3), and in all cases, with the exception of Si_3O^-, the VIEs-UOVGF are higher than their respective P3 values. The pseudopotential (PP) results in both approximations (UOVGF and P3) are higher than the all-electron ones, and they are closer to the experimental values. We compared our theoretical estimations with the high values of each band in the experimental spectra.[20] The UOVGF/PP results reproduce better the experimental values, but in general P3/PP seems to be an adequate methodology in predictions of VIEs. These results could be extrapolated to the study of larger silicon oxides clusters with the most economical methodology tested in this work. The last column of Table 1 shows the results obtained using finite difference

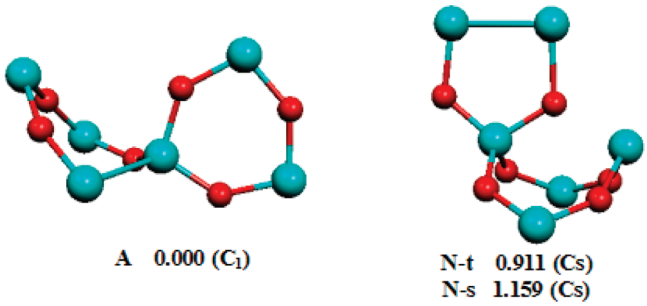


Figure 2. Structures of the most stable neutral Si_6O_6 (A) reported in ref 29 and Si_6O_6 (N) obtained by subtracting one electron to $Si_6O_6^-$ and followed by B3LYP/6-311G(2d) optimization. Singlet spin state is N-s and triplet spin state is N-t. We report the relative energy of each cluster with respect to structure A.

approximations. It means that the vertical ionization energy is calculated as the difference between the anion and the neutral cluster at the geometry of the anion. The values compare very well with the experimental ones and those calculated using EPT (UOVGF and P3) methods. However, in the finite difference approximation one needs to do two calculations and obtains only the first transition.

In Si_3O^-, $Si_3O_3^-$-$Si_3O_6^-$, the first vertical ionization energy calculated by the electron propagator methodologies corresponds to the ejection of the nonpaired electron to yield the corresponding closed shell neutral species. On the contrary, $Si_3O_2^-$ presents a first ionization energy corresponding to the transition toward the triplet state of the neutral cluster. However, the transition to the singlet state is only 0.22 eV higher in energy at the UOVGF/PP level of calculation. This is in agreement with the discussion by Lai-Sheng Wang et al.[20] They used energy differences between the anionic cluster and the singlet and triplet neutral ones (three calculations) for their assignments.

The planar C_s, $Si_3O_3^-$, was found to be closer in energy to the one of minimum energy (Part A of the discussion). In order to evaluate the possible experimental evidence of this cluster, we calculated its VIEs at UOVGF/PP approximation. The calculated VIEs of the most external valence electrons are $\alpha(4a'')$ 1.05 eV, $\beta(12a')$ 4.02 eV, and $\beta(11a')$ 4.63 eV, and the pole strengths values are between 0.85 and 0.91. All the predicted VIEs are in the ranges of the peaks of the Wang's experimental spectra.[20]

$Si_6O_n^-$ Clusters. The VIEs calculated at UOVGF/all-electron and P3/PP levels of theory are reported in Table 2. The results obtained by all-electron and PP calculations show the same systematic trend presented in the $Si_3O_n^-$ clusters, and the P3/PP values are higher than their OVGF/all-electron counterpart.

In the clusters $Si_6O_n^-$ ($n = 1-5$) the first VIEs correspond, from our calculations, to the ejection of the nonpaired electron, giving as a result a closed shell neutral cluster. In the case of the $Si_6O_6^-$ the calculations show that the first ionization gives as a result a neutral cluster in a triplet ground state. In order to verify whether this open shell cluster is the most stable, the geometry was relaxed to the minimum and compared with the neutral structure reported in the literature.[29,30] We optimized the system at the B3LYP/6311 g(2d) level, which is displayed as N-t (N structure-triple spin state) in Figure 2. We display in the same figure the most stable structure for neutral Si_6O_6 (A), which is a singlet. The N-t state is 0.911 eV less stable than isomer A, but 0.248 eV is more stable than N-s, an isomer with the same geometry as N-t but in the singlet state. In Figure 2 all energies are referred as to structure A.

Diffuse functions are not included in the all-electron UOVGF and P3 results, inclusions likely to lead to larger electron binding energies, and, therefore, the pseudopotential calculations which include diffuse functions give closer agreement with the experimental data.

C. Study of the Fragments. *Spin Density Localization.* The spin density isosurfaces show that the spin density is mainly localized over the silicon atoms in the $Si_3O_n^-$ series (Figure 2-SI) and over the silicon rich fragment in the $Si_6O_n^-$ series. In the $Si_3O_5^-$ and $Si_3O_6^-$ clusters the spin density is mainly localized over the terminal silicon atoms and their surrounding oxygen atoms. These Si atoms have a sp^3 like hybridization which gives them an electronegative character. This is in agreement with the high first vertical ionization energy of these clusters. A relation of the spin density localization and the value of the first ionization energy in

Table 3. Fragmentation Energies (FE) and Fragmentation Channels for the Most Stable Anion $Si_6O_n^-$ ($n = 1-6$) Clusters Obtained at the B3LYP/6-311G(2d) Level of Theory

Si_6O_m	$\rightarrow Si_kO_l^-$	$+ Si_{6-k}O_{m-l}$	FE (eV)	Si_6O_{4m}	$\rightarrow Si_kO_l^-$	$+ Si_{6-k}O_{m-l}$	FE (eV)
Si_6O^-	Si_5^-	SiO	1.437	$Si_6O_4^-$	Si_3^-	Si_3O_4	1.989
	Si_4^-	Si_2O	3.369		Si_3O^-	Si_3O_3	2.471
	Si_3^-	Si_3O	3.895		$Si_4O_2^-$	Si_2O_2	3.043
	Si_5	SiO^-	4.067		$Si_3O_3^-$	Si_3O	3.513
	Si_4	Si_2O^-	4.077		$Si_3O_4^-$	Si_3	3.587
$Si_6O_2^-$				$Si_6O_5^-$	Si_2^-	Si_4O_5	2.389
	Si_4^-	Si_2O_2	1.956		$Si_3O_2^-$	Si_3O_3	2.456
	$Si_2O_2^-$	Si_4	3.014		$Si_5O_4^-$	SiO	2.570
	Si_3^-	Si_3O_2	3.383		Si_3O^-	Si_3O_4	2.783
	Si_5^-	SiO_2	3.591		$Si_3O_3^-$	Si_3O_2	3.480
	$Si_3O_2^-$	Si_3	3.843		$Si_4O_4^-$	Si_2O	3.669
$Si_6O_3^-$	Si_3^-	Si_3O_3	1.998	$Si_6O_6^-$	$Si_3O_3^-$	Si_3O_3	2.135
	$Si_5O_2^-$	SiO	2.734		$Si_4O_4^-$	Si_2O_2	2.620
	Si_4^-	Si_2O_3	3.172		$Si_3O_2^-$	Si_3O_4	2.652
	$Si_3O_3^-$	Si_3	3.482		$Si_2O_2^-$	Si_4O_4	3.120
	Si_3O^-	Si_3O_2	4.021		$Si_3O_4^-$	Si_3O_2	3.789

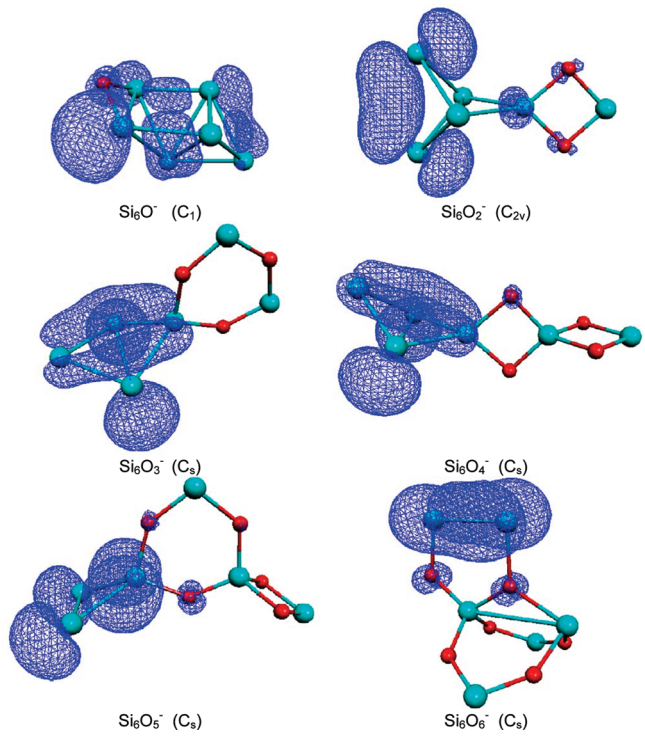


Figure 3. Spin density of the ground state for anionic $Si_6O_n^-$ ($n = 1-6$) clusters obtained at the B3LYP/6-311G(2d) level of theory.

the $Si_6O_n^-$ series is not obvious because the spin density is more delocalized over all the silicon rich fragment.

Fragmentation Energies. We have studied some fragmentation channels for the most stable structures of the $Si_6O_n^-$ ($n = 1-6$) clusters. The energy associated with each fragmentation channel $Si_6O_n^- \rightarrow Si_kO_l^- + Si_{6-k}O_{n-l}$ is defined as $FE = E(Si_kO_l^-) + E(Si_{6-k}O_{n-l}) - E(Si_6O_n^-)$, and the results are shown in Table 3. Basically, we focused our analysis in those fragmentation channels where one anionic and one neutral fragment are produced. We reported only five lower-energy fragmentation channels. It is interesting to remark that the more energetically favored channels are those where the anionic fragment coincides with the spin density localization in the unfragmented cluster (see Figure 3 and Table 3). Then, energetically favored fragmentation products contain the anionic Si_3^-, Si_4^-, Si_5^-, $Si_3O_2^-$, and $Si_3O_3^-$ and the neutral SiO, Si_2O_2, Si_3O_2, Si_3O_3, Si_3O_4, and Si_4O_4 clusters.

For Si_6O^-, the most favorable fragmentation channel is $Si_6O^- \rightarrow Si_5^- + SiO$. This channel presents similarities with fragmentation of Si_nO ($n = 5-10$) clusters where $Si_nO \rightarrow Si_n + SiO$ is the most favorable fragment pathway as discussed by H. Wang et al.[47]

$Si_6O_2^-$ presents $Si_6O_2^- \rightarrow Si_4^- + Si_2O_2$ as the most favorable fragmentation channel. So, we could say that $Si_6O_2^-$ is formed by Si_4^- and Si_2O_2, where the pure silicon cluster is an approximation to the ground state of the anionic Si_4^- cluster. Therefore, the extra electron is not localized on the Si_2O_2 fragment in agreement with the localization of the spin density (Figure 3).

The most favorable fragmentation channel for $Si_6O_3^-$ is $Si_6O_3^- \rightarrow Si_3^- + Si_3O_3$, and its spin density shows localization on the Si_3^- fragment. Therefore, the ground state of $Si_6O_3^-$ can be seen as formed by the fragments Si_3^- and Si_3O_3. The second fragmentation channel, with a FE 0.736 eV higher than the first one, is $Si_6O_3^- \rightarrow Si_5O_2^- + SiO$ where the SiO molecule is neutral again.

The most favorable fragmentation channel for $Si_6O_4^-$ contains Si_3^- and Si_3O_4 as products. $Si_6O_4^- \rightarrow Si_3O^- + Si_3O_3$ is 0.482 eV higher than the minimum FE, 1.989 eV.

The most favorable fragmentation channel for $Si_6O_5^-$ is $Si_6O_5^- \rightarrow Si_2^- + Si_4O_5$, although $Si_6O_5^- \rightarrow Si_3O_2^- + Si_3O_3$ and $Si_6O_5^- \rightarrow Si_5O_4^- + SiO$ are only 0.076 and 0.190 eV higher than the minimum FE, 2.380 eV.

Finally, in the case of $Si_6O_6^-$ the most favorable fragmentation channel is $Si_6O_6^- \rightarrow Si_3O_3^- + Si_3O_3$, but there are two fragment pathways with FE close to the minimum FE, 2.135 eV. $Si_6O_6^- \rightarrow Si_4O_4^- + Si_2O_2$ has a FE higher, by 0.032 eV, with respect to the fragmentation channel $Si_6O_6^- \rightarrow Si_3O_2^- + Si_3O_4$, 2.620 eV.

In most cases SiO, Si_2O_2, and Si_3O_3 appear as a product. It seems that Si_2O_2, Si_3O_3, and SiO preserve their stability within the anionic systems. Note that there is a relation between the most favorable fragment channel and the localization of the spin density. The anionic fragments are silicon-rich clusters. This behavior is similar in the spin density.

In Table 4 the principal vibrational modes for $Si_6O_n^-$ structures are reported. In the first column, the principal vibrational frequencies are shown. We tried to identify these modes with the corresponding vibrational modes of the fragmentation byproducts. Therefore, we present vibration modes of the fragment products measured by Anderson et al.[19] and calculated at the B3LYP/6-311G(2d) level of theory. Note that the vibrational modes of $Si_6O_2^-$, $Si_6O_3^-$, and $Si_6O_5^-$ are close to the vibrational modes of the fragment product of the most favorable fragmentation channel. $Si_6O_4^-$ contains the vibrational mode of Si_2O_2, and the fragmentation channel with this product is higher by 1.053 eV than the minimum

Table 4. Vibration Frequencies for $Si_6O_n^-$ ($n = 1-6$) Clusters and Some of Their Fragmentation Products[e]

Si_6O_m[a]	vibration frequencies principal modes		fragmentation product[b]
Si_6O^-	730.7		
$Si_6O_2^-$	750	766.6[c]	*(Si_2O_2)*
		766.3[d]	
$Si_6O_3^-$	*965.7*	957.6[c]	(Si_3O_3)
		972.6[d]	
$Si_6O_4^-$	*726.0*	766.6[c]	*(Si_2O_2)*
		766.3[d]	
	1017.7	1008.4[c]	*(SiO_2)*
$Si_6O_5^-$	*1044.5*	1260.8[c]	*(SiO)*
		1223.9[d]	
	984.4	957.6[c]	*(Si_3O_3)*
		972.6[d]	
	831.7	801.3[c]	*(Si_2O_2)*
		804.7[d]	
$Si_6O_6^-$	*974.9*	845.8[c]	*($Si_4O_4^-$)*
	996.2	1008.4[c]	*(SiO_2)*

[a] This work. [b] Indicated in $Si_6O_m^- \rightarrow Si_kO_l^- + Si_{6-k}O_{m-l}$. [c] The geometries were optimized at the B3LYP/6-311G(2d) level of theory. [d] Vibration frequencies from the mass spectra reported in ref 19. [e] We report the frequencies for principal modes of the most stable isomer of each cluster.

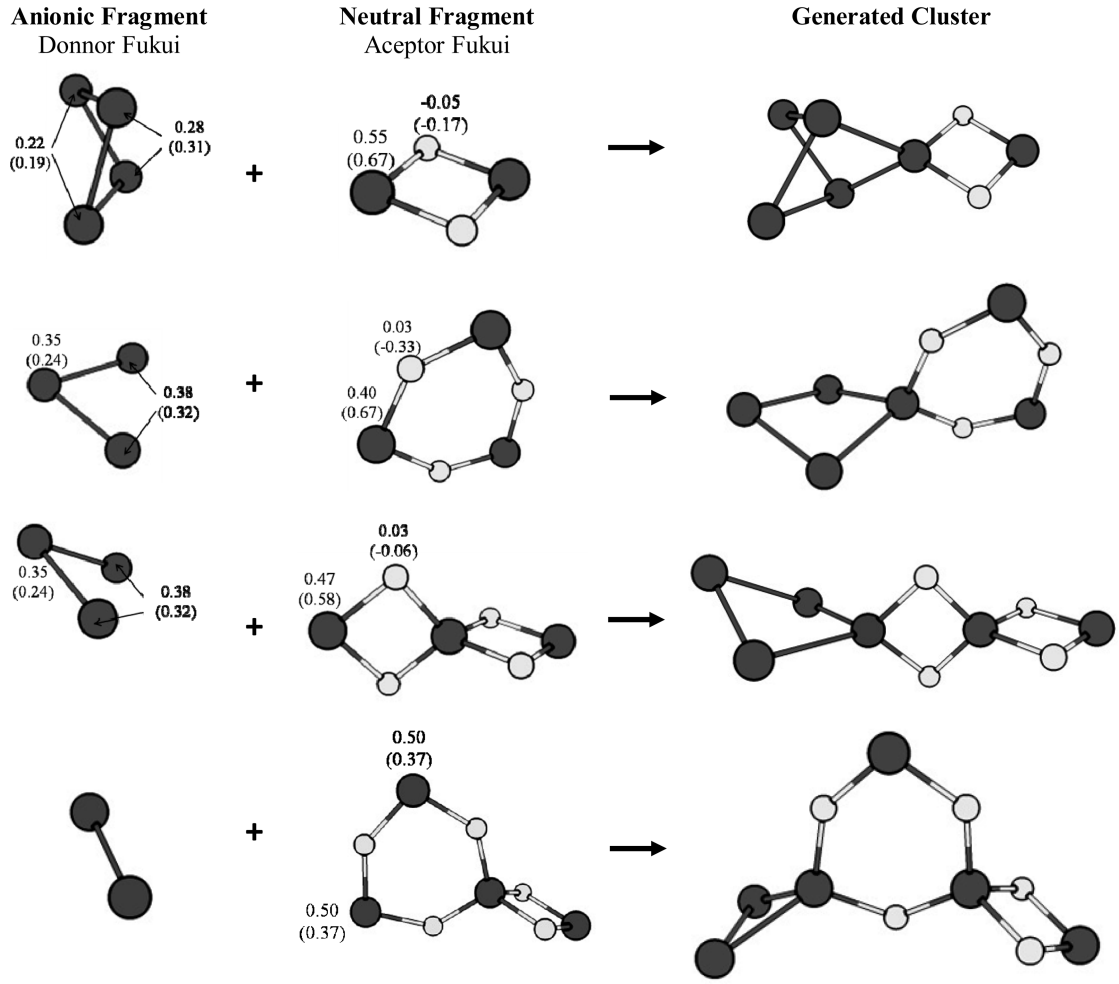


Figure 4. Reactions of formations of large anionic clusters from optimized fragments. Calculated atomic Fukui functions over the fragments, using NPA and CHG charges (in parentheses).

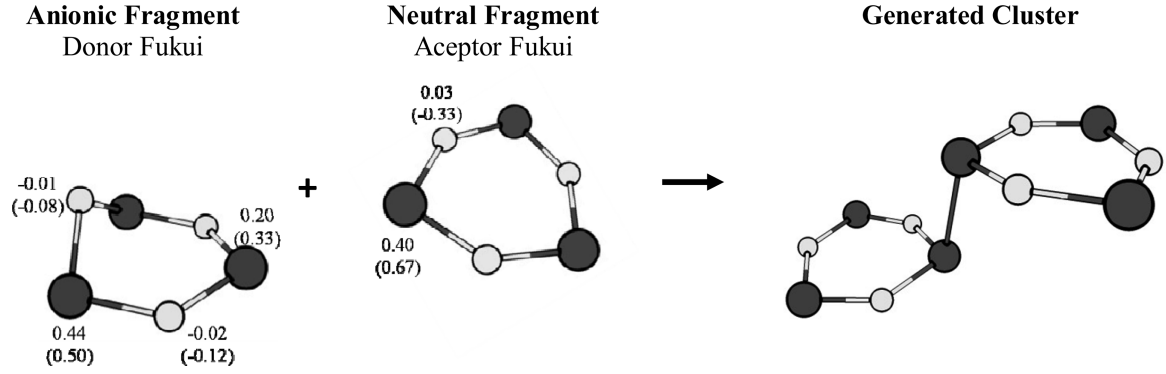


Figure 5. Reactions of formation of $Si_6O_6^-$ D_{2h} clusters from optimized fragments. Calculated atomic Fukui functions over the fragments, using NPA and CHG charges (in parentheses).

FE, 1.989 eV. $Si_6O_6^-$ has one vibrational mode corresponding to that of the $Si_4O_4^-$ product. The fragmentation channel involving this product is only 0.485 eV higher than the minimum FE, 2.134 eV. In most of the cases, the vibrational modes of the fragments present in the most favorable fragmentation channel are similar to the principal modes of the studied anionic systems.

Building Large Clusters from the Small Stabilized Fragments. Previously we have analyzed the fragmentation patterns of $Si_6O_n^-$ series and obtained some characteristic fragments which after energy relaxation are stables structures. In the experimental conditions it is possible that fragmentation and formation are competitive reactions, and thermodynamically the formation reactions in these systems are favored (energy channel analysis). We include an analysis based on the Fukui function with the aim to rationalize the clusters formation from smaller ones, which allows us to identify the most favorable site to interact and therefore to predict the formation of larger systems.

A set of the most characteristic fragments was optimized and calculated the respective donor, f_r^-, and acceptor, f_r^+, condensed Fukui functions, where the donor fragments are

those with negative charge and the acceptor are the neutral ones. The results are shown in the first two columns of Figure 4. Following the selective reactivity criteria the fragments should react in those regions where the respective Fukui values are maxima. The relaxed structures are shown in the last column of Figure 4, and one can see that the structures for $Si_6O_n^-$ ($n = 1-5$) could be obtained by reaction between two small stable fragments.

A formation picture based on Fukui function analysis of the $Si_6O_6^-$ cluster is more complicated. The fragments $Si_3O_3^-$ and Si_3O_3 are the most energetically favored fragments of this cluster (Table 2). The reaction of these two small clusters produce a stable $Si_6O_6^-$ isomer with D_{2h} symmetry (Figure 4) which is 0.52 eV less stable with respect to the minimum energy. The apparent limitation of the Fukui function in predicting the more stable product was discussed by Tiznado et al.[48] in the reaction of hydrogen atoms with silicon clusters. The more stable reactant did not always generate the more stable product.

A complementary analysis of the relaxation pathways which connects different structural isomers with the global minimum structure will be necessary to complement the information obtained from the Fukui function analysis.

IV. Conclusions

A theoretical study of the two series of small clusters $Si_3O_n^-$ and $Si_6O_n^-$ ($n = 1-6$) has been carried out. It has been found that planar $Si_3O_3^-$ D_{3h} cluster is closer in energy to the minimum energy isomer, and its calculated VIEs agree with the experimental ones obtained from the PES experiments. Hence, the experimental presence of this isomer should be not discarded.

Two not previously reported minimum energy isomers were found, $Si_6O_4^-$ and $Si_6O_6^-$. The most stable isomer Si_6O^- is present as two isoenergetic not superimposable mirror images. The P3 methodology in conjunction with the use of effective core pseudopotentials proves to be adequate to calculate the VIEs in small anionic silicon oxides clusters. This was verified comparing to the more complete UOVGF approximation and to the experimental values in $Si_3O_n^-$ ($n = 1-6$) clusters. The propagator methodology establishes that the triplet state of the Si_6O_6 cluster is more stable than the closed shell configuration obtained from the ionization of the most stable anionic cluster. However, the minimum energy structure of the neutral cluster is closed shell. The studied fragmentation channels determine the most stable fragments. In all cases the anionic fragments agree with the localization of the spin density in the large cluster. The spin density is mainly localized on the silicon rich fragment.

The donor and acceptor Fukui functions predict the best interactions between a small anionic silicon cluster and a small neutral silicon oxide cluster to form $Si_6O_n^-$ ($n = 2-5$) global minimum structures, but the $Si_6O_6^-$, obtained from the Fukui function predictions, is not the minimum energy one. A detailed study of the energy pathways which connects different isomers will be necessary to complement the predictions obtained from the Fukui function.

Acknowledgments. Financial support for this work from the University of Buenos Aires, and the Argentinian CONICET, is sincerely acknowledged. W. Tiznado acknowledged Fondecyt's support for its post doctoral fellowship, project No. 3080042, and Dr. P. Jaque for interesting discussions. Part of this work has also been supported by Fondecyt Grants 1080184.

Supporting Information Available: Cartesian coordinates and absolute energies of $Si_3O_3^-$ clusters. Spin density of the ground state for anionic $Si_3O_n^-$ ($n = 1-6$) clusters. Hartree−Fock expectation value of the total spin, $<S^2>$, for $Si_3O_n^-$ and $Si_6O_n^-$ clusters ($n = 1-6$). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Connerade, J. P.; Solov'yov, A. V.; Greiner, W. *Europhysics News* **2002**, *33*, 200.

(2) *NATO Advanced Study Institute, Session LXXIII, Summer School "Atomic Clusters and Nanoparticles"*; Guet, C., Hobza, P., Spiegelman, F., David, F. Eds.; Les Houches, France, July 2−28, 2000, EDP Sciences and Springer Verlag: Berlin, New York, London, Paris, Tokyo, 2001.

(3) de Heer, W. A. *Rev. Mod. Phys.* **1993**, *65*, 611.

(4) *Clusters of Atoms and Molecules, Theory, Experiment and Clusters of Atoms*; Haberland, H., Ed.; Springer Series in Chemical Physics, Springer: Berlin, Heidelberg, New York, 1994; Vol. 52.

(5) *Metal Clusters*; Ekardt, W., Ed.; Wiley: New York, 1999; pp 29−68.

(6) Koopmans, T. *Physica* **1933**, *1*, 104.

(7) Cederbaum, L. S.; Schirmer, J.; Domcke, W.; von Niesse, W. *Int. Quantum Chem.* **1978**, *14*, 593.

(8) Cederbaum, L. S. *J. Phys. B* **1975**, *8*, 290.

(9) von Niessen, W.; Shirmer, J.; Cederbaum, L. S. *Comput. Phys. Rep.* **1984**, *1*, 57.

(10) Zakrzewski, V. G.; Ortiz, J. V. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1994**, *28*, 23.

(11) (a) Zakrzewski, V. G.; Ortiz, J. V. *Int. J. Quantum Chem.* **1995**, *53*, 583. (b) Ortiz, J. V. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1989**, *23*, 321. (c) Lin, J. S.; Ortiz, J. V. *Chem. Phys. Lett.* **1990**, *171*, 197.

(12) (a) Zakrzewski, V. G.; Ortiz, J. V.; Nichols, J. A.; Heryadi, D.; Yeager, D. L.; Golab, J. T. *Int. J. Quantum Chem.* **1996**, *60*, 29. (b) Ortiz, J. V. *Adv. Quantum Chem.* **1999**, *35*, 33.

(13) Ortiz, J. V. *J. Chem. Phys.* **1996**, *104*, 7599.

(14) Ortiz, J. V.; Zakrzewski, V. G.; Dolgounitcheva, O. In *Conceptual Trends in Quantum Chemistry*; Kryachko, E. S., Ed.; Kluwer: Dordrecht, 1997; Vol. 3, p 465.

(15) Holmes, D. L. *Elements of Physical Geology*; Ronald Press: New York, 1969; Chapter *3*.

(16) Morey, G. W. *The Properties of Glass*, 2nd ed.; Reinhold: New York, 1954; Chapter *1*.

(17) Desurvire, E. *Phys. Today* **1994**, *47*, 20.

(18) Wang, N.; Tang, Y. H.; Zhang, Y. F.; Lee, C. S.; Lee, S. T. *Phys. Rev. B* **1998**, *58*, R16024.

(19) Anderson, J. S.; Ogden, J. S. *J. Chem. Phys.* **1969**, *51*, 4189.

(20) Wang, L. S.; Nicholas, J. B.; Dupuis, M.; Wu, H.; Colson, S. D. *Phys. Rev. Lett.* **1997**, *78*, 4450.

(21) Wang, L. S.; Desai, S. R.; Wu, H.; Nicholas, J. B. Z. *Phys. D* **1997**, *40*, 36.

(22) Wang, L. S.; Wu, H.; Desai, S. R.; Fan, J.; Colson, S. D. *J. Phys. Chem.* **1996**, *100*, 8697.

(23) Chelikowsky, J. R. *Phys. Rev. B* **1998**, *57*, 3333.

(24) Snyder, L. C.; Raghavachari, K. J. *J. Chem. Phys.* **1984**, *80*, 5076.

(25) Zhang, R. Q.; Chu, T. S.; Cheung, H. F.; Wang, N.; Lee, S. T. *Phys. Rev. B* **2001**, *64*, 113304.

(26) Song, J.; Choi, M. *Phys. Rev. B* **2002**, *65*, 241302 R.

(27) Nayak, S. K.; Rao, B. K.; Jena, P. *J. Chem. Phys.* **1998**, *109*, 1245.

(28) Chu, T. S.; Zhang, R. Q.; Cheung, H. F. *J. Chem. Phys.* **2001**, *105*, 1705.

(29) Caputo, M. C.; Oña, O. B.; Ferraro, M. B. *J. Chem. Phys.* **2009**, *130*, 134115.

(30) Zang, Q. J.; Su, Z. M.; Lu, W. C.; Wang, C. Z.; Ho, K. M. *J. Phys. Chem. A* **2006**, *110*, 8151.

(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(32) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.

(33) Krishnan, B. R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(34) Ortiz, J. V. In *Computational Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific: Singapore, 1997; Vol. 2, p 1.

(35) Igel-Mann, G.; Stoll, H.; Preuss, H. *Mol. Phys.* **1988**, *65*, 1321.

(36) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1998**, *53*, 1995.

(37) Tiznado, W. A.; Fuentealba, P.; Ortiz, J. V. *J. Chem. Phys.* **2005**, *123*, 144314.

(38) Yang, W.; Mortier, W. J. *J. Am. Chem. Soc.* **1986**, *108*, 5708.

(39) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular Interactions from a Natural Bond Orbital, Donor-Acceptor Viewpoint. *Chem. Rev.* **1988**, *88* (6), 899–926.

(40) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural-Population Analysis. *J. Chem. Phys.* **1985**, *83* (2), 735–746.

(41) Reed, A. E.; Weinhold, F. Natural Bond Orbital Analysis of Near-Hartree-Fock Water Dimer. *J. Chem. Phys.* **1983**, *78* (6), 4066–4073.

(42) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials - The Need for High Sampling Density in Formamide Conformational-Analysis. *J. Comput. Chem.* **1990**, *11* (3), 361–373.

(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, 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 E.01*; Gaussian, Inc.: Wallington, CT, 2004.

(44) Flükiger, P.; Lüthi, H. P.; Portmann, S.; Weber, J. Swiss National Supercomputing Centre CSCS, Manno, Switzerland, 2000.

(45) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(46) Woon, D. E.; Dunning, T. H., Jr. *Chem. Phys.* **1993**, *98*, 1358.

(47) Wang, H.; Sun, J.; Lu, W. C.; Sun, C. C.; Wang, C. Z.; Ho, K. M. *J. Phys. Chem. C* **2008**, *112*, 7097.

(48) Tiznado, W.; Oña, O. B.; Bazterra, V. E.; et al. *J. Chem. Phys.* **2005**, *123*, 214302.

CT900320R

JCTC Journal of Chemical Theory and Computation

Characterization of the Chemical Behavior of the Low Excited States through a Local Chemical Potential

Christophe Morell,*,† Vanessa Labet,† André Grand,† Paul W. Ayers,‡ Frank De Proft,§ Paul Geerlings,§ and Henry Chermette⊥

Commissariat à L'Énergie Atomique Grenoble, 17, Rue des Martyrs, F-38054 Grenoble Cedex 9, France, Department of Chemistry, McMaster University, Hamilton, Ontario, L8S 4M1, Eenheid Algemene Chemie, Vrije Universiteit Brussel, Faculteit Wetenschappen Pleinlaan 2, 1050 Brussels, Belgium, Member of the QCMM Alliance Ghent-Brussels Belgium, and Université de Lyon, Université Lyon 1, Sciences Analytiques CNRS UMR 5180 Chimie Physique Théorique, F-69622 Villeurbanne Cedex, France

Received May 17, 2009

Abstract: Exploiting the locality of the chemical potential of an excited state when it is evaluated using the ground-state density functional theory (DFT), a new descriptor for excited states has been proposed. This index is based on the assumption that the relaxation of the electronic density drives the chemical reactivity of excited states. The sign of the descriptor characterizes the electrophilic or nucleophilic behavior of the atomic regions. A relation between the new descriptor and the dual descriptor is derived and provides a posteriori justification of its use to rationalize the Woodward−Hoffmann rules for photochemical reactions within the conceptual DFT. Finally, the descriptor is successfully applied to some [2 + 2] photocycloadditions, like Paterno−Büchi reactions.

1. Introduction

During the past two decades, conceptual density functional theory (DFT)[1−4] has been a fruitful paradigm for the analysis of chemical processes. Through successive derivatives of the energy, with respect to either the number of electrons or the external potential, different reactivity and selectivity descriptors have been designed to account for the outcome of chemical reactions.[5,6] More importantly, descriptors previously proposed on an intuitive basis found a sharp definition within conceptual DFT. For instance, the electronegativity defined by Mulliken[7] has been identified as the opposite of the first derivative of the energy with respect to the number of electrons.[8] Another example is the proposal to measure the chemical hardness through the second derivative of the energy with respect to the number of electrons.[9] Besides, different older theories, such as the Frontier Molecular Orbital (FMO) initiated by Fukui[10−15] and the Hard and Soft Acids and Bases (HSAB) proposed by Pearson,[16−22] have been unified within conceptual DFT.

However, some chemical processes are still out of the reach of conceptual DFT. Indeed, due to the lack of a suitably formulated excited-state DFT, all the chemical reactions that involve an excited molecule are difficult to rationalize. Even though some reactions have been investigated by transposing,[23−25] the traditional local descriptors for the ground state to the reactivity of the excited state, no formal theory has been designed to support their use. It must be noticed that a preliminary study of time-dependent DFT to design local descriptors has been proposed by Chattaraj and co-workers.[26−28] The purpose of this paper is then to provide some insight into both the reactivity and selectivity of low excited states using the locality of the excited state's chemical potential when evaluated using the energy density functional for ground states.

The paper is organized as follows. In Section 2, using the true density of an excited state as a trial function for the ground-state density, a local descriptor is proposed to characterize the electrophilicity or nucleophilicity of a region

* Corresponding author. E-mail: Christophe.morell@ujf-grenoble.fr.
† Commissariat a L'Energie Atomique Grenoble.
‡ McMaster University.
§ Vrije Universiteit Brussel.
⊥ Université de Lyon.

10.1021/ct900248a CCC: $40.75
Published on Web 08/05/2009

within an excited state. From a comparison between the proposed and dual descriptors, a justification for the use of the negative dual descriptor to rationalize the Woodward−Hoffmann rules appears. After summarizing the computational methods employed, the concepts developed in Section 2 are tested for prototypical excited-state processes, more specifically for the regioselectivity of [2 + 2] photocycloaddition and Paterno−Büchi reactions. The paper ends with a perspective on what has been achieved, and what the outstanding issues are for future work.

2. Chemical Potential of a Hot Excited State

The generation of an excited state can be seen as a two-step process. In the first step, a hot excited state (HES) is created by a vertical electronic transition in which the electronic density changes, while the external potential remains identical. Such processes are governed by the Franck−Condon principle.[29,30] The HES is actually both electronically and vibrationally excited. During the second step, the molecular geometry (external potential) relaxes, adapting itself to the new electronic density. This gives an optimized excited state (OES). In this section, an approximate formula for the chemical potential of the HES is presented. The analysis of this chemical potential formula provides a way to characterize the chemical reactivity of the different atomic sites within a HES.

Traditional DFT is grounded on the first and the second Hohenberg−Kohn theorems.[31] Unfortunately, the practical use of those theorems has only been established for the electronic ground state, and they are not believed to be valid for excited states. However, the situation is not as bad as it seems because different formulations of excited-state DFT have been published.[32−35] Indeed, since the electronic density of the ground state determines both the external potential and the number of electrons, it determines the Hamiltonian and, consequently, all of the eigenstates of the electronic system. One can even define a functional of the electronic density that gives the exact energy for all the stationary electronic excited states.[36] The most popular excited-state DFT is the formulation of Levy and Nagy.[37] In this formulation, the energy is computed as

$$E_k = \int \rho_k(\vec{r})v(\vec{r})\mathrm{d}\vec{r} + F_k[\rho_k(\vec{r}), v(\vec{r})] \qquad (1)$$

in which $\rho_k(\vec{r})$ is the exact electronic density of the excited state.

The bifunctional F_k is defined as

$$F_k[\rho_k(\vec{r}), v(\vec{r})] = \min_{\psi_k \to \rho_k} \langle \psi_k | T + V_{ee} | \psi_k \rangle$$

where it is understood that the k^{th} eigenfunction ψ_k is orthogonal to wave functions of the ground and lower excited states of $v(\vec{r})$. This assumption yields quite an important property: Unlike the Hohenberg−Kohn functional, the bifunctional $F_k[\rho_k(\vec{r}),v(\vec{r})]$ is not universal and depends upon the external potential through the electronic density of the ground state. However, this property of the bifunctional F_k does not affect the definition of local descriptors. Even though F_k is not universal, the electronic density is still given by the first derivative of the energy of that excited state with respect to the external potential at constant number of electron N:

$$\left(\frac{\delta E_k}{\delta v(\vec{r})}\right)_N = \rho_k(\vec{r}) + \left(\frac{\delta F_k[\rho_k(\vec{r}), v(\vec{r})]}{\delta v(\vec{r})}\right)_N \qquad (2)$$

Indeed, using the Hellman−Feynman theorem, Ayers and Levy[38] have shown that the derivative of the bifunctional with respect to the external potential vanishes and, therefore, $(\delta E_k/(\delta v(\vec{r})))_N = \rho_k(\vec{r})$. Consequently, the Fukui functions and the dual descriptor can be generalized as second and third crossed derivatives of the energy. However, as the relaxation of the excited density toward the density of the ground state might be a phenomenon powerful enough to drive the reactivity of the excited state, the ground-state local descriptors could be not appropriate.

On the one hand, the variational principle applied to eq 1 under the constraint that $\int \rho_k(\vec{r})\mathrm{d}\vec{r} = N_0$ yields for all the stationary densities to

$$\left(\frac{\delta E_k}{\delta \rho(\vec{r})}\right)_N = \mu_k = v_k(\vec{r}) + \left(\frac{\delta F_k[\rho_k(\vec{r}), v(\vec{r})]}{\delta \rho(\vec{r})}\right)_N \qquad (3)$$

in which μ_k is a global property of the system that generalizes the chemical potential to the excited state k. If the bifunctional F_k for the excited states were known in a practical form, then the nonlocality of the quantity $v_k(\vec{r}) + (\delta F_k[\rho_k(\vec{r}),v(\vec{r})]/\delta\rho(\vec{r}))_N$ would be an important criterion for finding stationary densities. Indeed, the nonlocality of the chemical potential is a consequence of the stationary character of the exact electronic density of the state k. In other words, μ_k is global only for the true excited-state density.

On the other hand, given the electronic density of a chosen excited state, its use as a trial density in the ground-state DFT must, with rare exception,[8,39] lead to a nonconstant chemical potential:

$$\left(\frac{\delta E}{\delta \rho_k(\vec{r})}\right) = \lambda_k(\vec{r}) = v(\vec{r}) + \frac{\delta F_{\text{HK}}[\rho_k(\vec{r})]}{\delta \rho_k(\vec{r})} \qquad (4)$$

The nonconstant chemical potential captures the tendency of the excited-state electronic density to relax toward the ground-state density. It is this trend we desire to exploit. In particular, we will use the equation:

$$\mathrm{d}E_k = \int \lambda_k(\vec{r})\delta\rho_k(\vec{r})\mathrm{d}\vec{r} \qquad (5)$$

noting that $\mathrm{d}E_k$ does not refer to an excitation energy. Equation 5 is a lower bound to the true change in energy accompanying excitation.

This local chemical potential is related to the true global chemical potential of the ground state as

$$\lambda_k(\vec{r}) = \mu_0 + \mathrm{V}(\vec{r}) \qquad (6)$$

with $V(\vec{r}) = (\delta F_{HK}[\rho_k(\vec{r})]/\delta\rho(\vec{r})) - (\delta F_{HK}[\rho_0(\vec{r})]/\delta\rho(\vec{r}))$, where $\delta F_{\text{HK}}[\rho_k(\vec{r})]/\delta\rho(\vec{r})$ is the electronic density potential for the system of interest and $\delta F_{\text{HK}}[\rho_0(\vec{r})]/\delta\rho(\vec{r})$ is the electronic density potential for which $\rho_0(\vec{r})$ is the N-electron ground-state density. The chemical meaning of the quantity $V(\vec{r})$ is clearly related to electronic relaxation: regions associated with positive values of $V(\vec{r})$ will decrease their electronic

densities and can be considered nucleophilic. Conversely, regions with negative values of $V(\vec{r})$ will increase their electronic densities and can be considered electrophilic.

When a chemical reaction induces the relaxation of the electron density in an excited state, then the sign of $V(\vec{r})$ provides a way to identify the nucleophilic and electrophilic sites within the excited-state molecule. Evaluating $V(\vec{r})$ requires finding a relationship between the excited- and ground-state chemical potentials. The most obvious way to do this is to construct the excited-state density by distorting the ground-state density. Then, using the Taylor series for the universal functional:

$$\left.\frac{\delta F_{\mathrm{HK}}[\rho_k(\vec{r})]}{\delta\rho(\vec{r})}\right|_{v(r)} = \left.\frac{\delta F_{\mathrm{HK}}[\rho_0(\vec{r}) + \Delta_0^k\rho(\vec{r})]}{\delta\rho(\vec{r})}\right|_{v(r)} = \frac{\delta F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})} + \int \frac{\delta^2 F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})\delta\rho(\vec{r}')}\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}' + \frac{1}{2}\iint \frac{\delta^3 F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r}')\delta\rho(\vec{r})\delta\rho(\vec{r}'')}\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}\,\Delta_0^k\rho(\vec{r}'')\mathrm{d}\vec{r}'' + \ldots \tag{7}$$

Here, $\Delta_0^k\rho(\vec{r})$ represents the electronic density difference between the excited state labeled k and the ground state. The local chemical potential of the excited state is

$$\lambda_k(\vec{r}) = v(\vec{r}) + \frac{\delta F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})} + \int \frac{\delta^2 F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})\delta\rho(\vec{r}')}\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}' + \frac{1}{2}\iint \frac{\delta^3 F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r}')\delta\rho(\vec{r})\delta\rho(\vec{r}'')}\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}'\Delta_0^k\rho(\vec{r}'')\mathrm{d}\vec{r}'' \tag{8}$$

The first derivative occurring in eq 8 can be evaluated as

$$\frac{\delta F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})} = \mu_0 - v(\vec{r})$$

The second derivative occurring in eq 8 is the hardness kernel, which is commonly decomposed as

$$\frac{\delta^2 F_{\mathrm{HK}}[\rho_0(\vec{r})]}{\delta\rho(\vec{r})\delta\rho(\vec{r}')} = \eta_0(\vec{r},\vec{r}') = \frac{1}{|\vec{r} - \vec{r}'|} + R(\vec{r},\vec{r}') \tag{9}$$

The first term of the right-hand side of eq 9 is the Coulomb contribution; the second term gathers the kinetic, exchange, and correlation contributions.

The third derivative occurring in eq 8 is the (second-order) hyperhardness kernel:

$$\frac{\delta^3 F[\rho_0(\vec{r})]}{\delta\rho(\vec{r})\delta\rho(\vec{r}')\delta\rho(\vec{r}'')} = \eta_0^{(2)}(\vec{r},\vec{r}',\vec{r}'')$$

Substituting these results into eq 8 gives

$$\lambda_k(\vec{r}) = \mu_0 + \int \frac{\Delta_0^k\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}\mathrm{d}\vec{r} + \int R(\vec{r},\vec{r}')\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}' + \frac{1}{2}\iint \eta_0^{(2)}(\vec{r},\vec{r}',\vec{r}'')\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}'\Delta_0^k\rho(\vec{r}'')\mathrm{d}\vec{r}'' + \ldots \tag{11}$$

Note that if the electronic densities of the ground and excited states are similar, their chemical potentials are also similar. For this reason, one expects the deviation of $\lambda(\vec{r})$ from μ_0 to be larger for higher-energy excited states. Referring back to the interpretation of $V(\vec{r}) = \lambda(\vec{r}) - \mu_0$, this suggests that the electrophilicity/nucleophilicity (ergo, the total reactivity) of excited states increases as the resemblance of the excited-state density to the ground-state density decreases.

While a first- or second-order truncation of the Taylor series might be sufficient for low-lying excited states, it is unlikely to be accurate for highly excited states. The ambition of the present paper is limited to the lower excited states, for which the following second-order truncation is *assumed* to be *qualitatively* accurate:

$$\lambda_k(\vec{r}) = \mu_0 + \int \frac{\Delta_0^k\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}d\vec{r} + \int R(\vec{r},\vec{r}')\Delta_0^k\rho(\vec{r}')\mathrm{d}\vec{r}' \tag{12}$$

If one assumes that the approximation $R(\vec{r},\vec{r}') \approx 0$ suffices for qualitative purposes,[40] then eq 12 becomes

$$\lambda_k(\vec{r}) = \mu_0 + \int \frac{\Delta_0^k\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}\mathrm{d}\vec{r} \tag{13}$$

With:

$$V(\vec{r}) \approx \int \frac{\Delta_0^k\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}\mathrm{d}\vec{r}'$$

In eq 13, the difference between the chemical potential of the excited and ground states is seen to be equal to the difference between the electrostatic potential of the ground and excited states.

Notice that eq 13 is consistent with the fact that when the molecule density shifts toward the ground state from the excited state, the energy decreases

$$\mathrm{d}E_k^0 \approx \int \lambda_k(\vec{r})(-\Delta_0^k\rho(\vec{r}))\mathrm{d}\vec{r}$$

$$\approx -\iint \frac{\Delta_0^k\rho(\vec{r})\Delta_0^k\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}\mathrm{d}\vec{r}\mathrm{d}\vec{r}'$$

A favorable chemical process involving the excited-state will cause its energy to decrease. Therefore, for an "allowed" excited-state reaction:

$$\delta E_k \approx \int \lambda_k(r)\delta\rho_k(\vec{r})\mathrm{d}\vec{r} < 0 \tag{14}$$

Introducing the reduced expression for the local chemical potential one has

$$\delta E_k \approx \int \mu_0\delta\rho_k(\vec{r})\mathrm{d}\vec{r} + \int V(\vec{r})\delta\rho_k(\vec{r})\mathrm{d}\vec{r} \tag{15}$$

The first term of the right-hand side of eq 15 is a global reactivity term; it is zero unless there is electron transfer to/from the system. The second term provides information about regioselectivity. A site in which the excited-state electrostatic potential is lower than the ground-state electrostatic potential attracts electrons. A site in which the

excited-state electrostatic potential is higher than the ground-state electrostatic potential repels electrons. The sign of eq 15, thus, provides a physically consistent characterization of both the electrophilicity and the nucleophilicity of a chemical site.

Trying to rationalize the Grochala, Albrecht, and Hoffmann rule (GAH), Ayers and Parr[41] proposed an approximate relation between the densities of the first excited state $\rho_{es}(\vec{r})$, the ground state $\rho_{gs}(\vec{r})$, and the radical cation $\rho_{-}(\vec{r})$ and anion $\rho_{+}(\vec{r})$:

$$\rho_{gs}(\vec{r}) + \rho_{es}(\vec{r}) - \rho_{+}(\vec{r}) - \rho_{-}(\vec{r}) \approx 0 \qquad (16)$$

Within DFT, the relations between the Fukui functions and the densities of the cation and the anion are

$$\rho_{+}(\vec{r}) \approx \rho_{gs}(\vec{r}) + f^{+}(\vec{r}) \qquad (17)$$

$$\rho_{-}(\vec{r}) \approx \rho_{gs}(\vec{r}) - f^{-}(\vec{r}) \qquad (18)$$

Substitution of eqs 17 and 18 into eq 16 leads to

$$\rho_{es}(\vec{r}) \approx \rho_{gs}(\vec{r}) + f^{+}(\vec{r}) - f^{-}(\vec{r}) \qquad (19)$$

Identifying the dual descriptor[42,43] as the difference between the electrophilic and nucleophilic Fukui functions leads to

$$\Delta_0^1\rho(\vec{r}) \approx \Delta f(\vec{r}) \qquad (20)$$

This result can be understood through a simple molecular orbital picture. Indeed, it corresponds to consider merely the first excited-state density as the one produced when one electron from the highest occupied molecular orbital (HOMO) is promoted to the lowest unoccupied molecular orbital (LUMO). Ayers and Parr further argue that eq 20 partially accommodates relaxation effects. Following the GAH rule, the first excited state combines the characteristics of both the cation and the anion. As expected, $\Delta_0^1\rho_k(\vec{r})$ and $\Delta f(\vec{r})$ are quite similar (see Figure 1).

So for the specific case of the first excited state, one can define the dual potential as

$$V^{[\Delta f]}(\vec{r}) = \int \frac{\Delta f(\vec{r}')}{|\vec{r} - \vec{r}'|} \mathrm{d}\vec{r}' \qquad (21)$$

Then eq 15 reads

$$\delta E_1 \approx \mu_0 \mathrm{d}N + \int V^{[\Delta f]}(\vec{r})\delta\rho_1(\vec{r})\mathrm{d}\vec{r} \qquad (22)$$

in which $\delta\rho_1(\vec{r})$ stands for a variation of the first excited-state's electronic density. The second term of eq 22 can be seen as the contribution due to the variations of the electronic density weighted by the potential created by the dual descriptor. In this context, the best way to stabilize the excited state is to comply with the following criterion at each point $\vec{r}$:

$$\delta\rho_1(\vec{r})V^{[\Delta f]}(\vec{r}) < 0 \qquad (23)$$

This criterion enables us to sort out both the electrophilic and the nucleophilic atomic sites within an excited-state molecule. Indeed, electrophilic sites will have a positive variation of their electronic densities and, according to eq 23, the dual potential should be negative in those sites. Conversely, nucleophilic sites will experience a decreasing of their electronic densities and, according to the same criterion, should display positive values of the dual potential. The meanings of the signs of the dual potential for the first excited state are the opposite of the ones of the dual descriptor for the ground state.

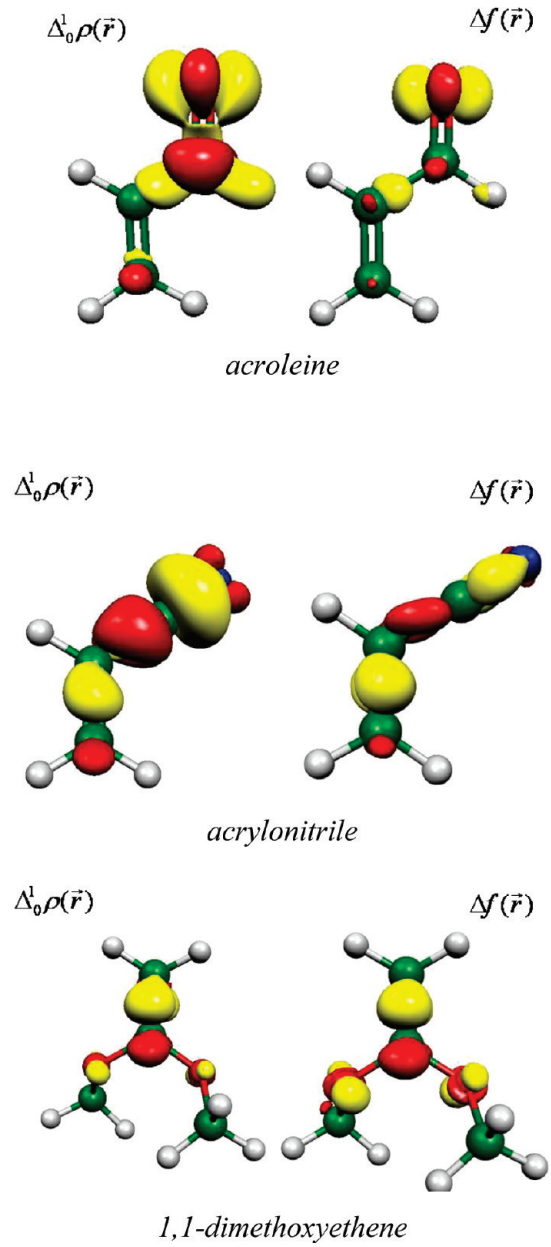


Figure 1. Comparison between $\Delta_0^1\rho(\vec{r})$ and $\Delta f(\vec{r})$.

Figure 2 displays both the dual potential and descriptor for some simple molecules. It can be seen that the general shapes of both functions are quite identical, except for the location of the nodes that separate the positive from negative regions. To explain those similarities, one can consider the dual potential at point $\vec{r}$ as the weighted sum of the dual descriptor in all the other points $\vec{r}'$. The weighting coefficient has two important properties: (i) it is always positive, and (ii) the closer the point r′, the higher its coefficient, and therefore, the more important its contribution to the dual potential. Taking into account those properties, it is no longer a surprise that both the dual potential

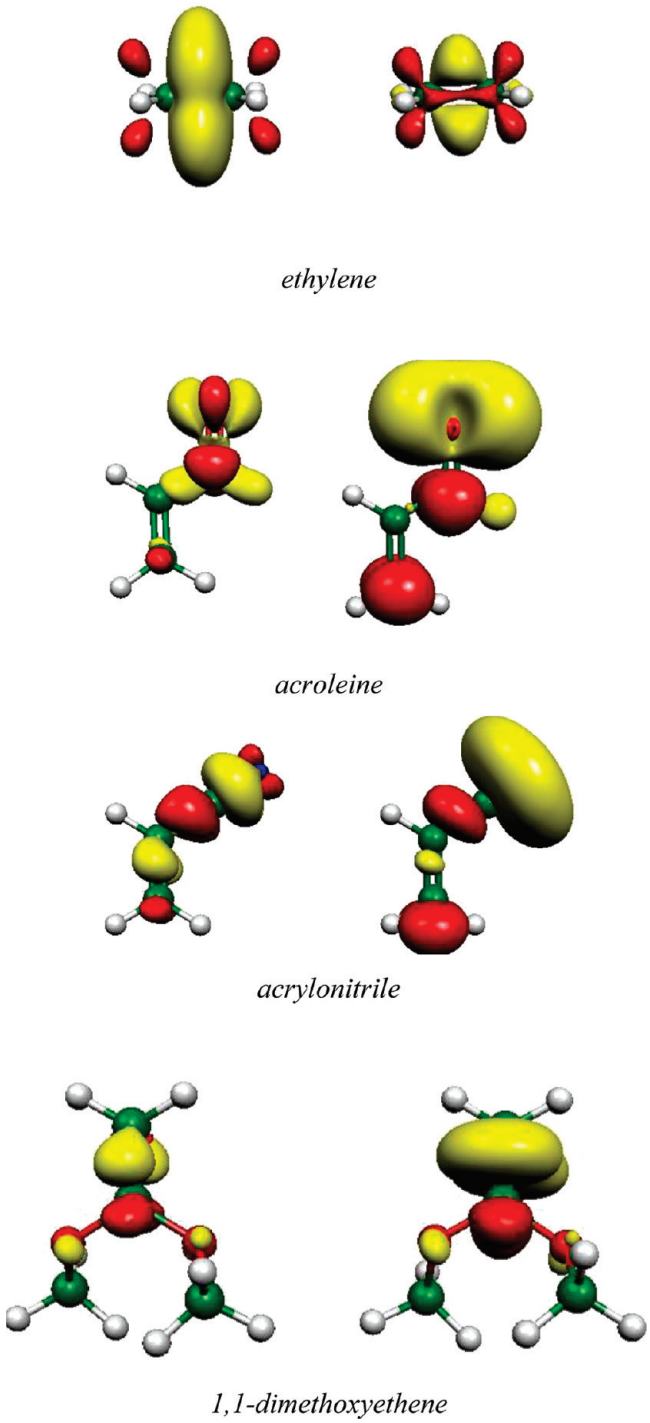


Figure 2. Comparison between $\Delta_0^1\rho(\vec{r})$ and $V^{\Delta_0^1\rho}(\vec{r}) = \int(\Delta_0^1\rho(r'))/(|\vec{r} - \vec{r}'|)\mathrm{d}\vec{r}'$.

descriptor give the same description of the chemical behavior of the atomic sites within a molecule. A more elegant explanation of this similarity can be considered through the Poisson equation directly resulting from eq 21: $\nabla V^{\Delta f}(\vec{r}) = -4\pi\,\Delta f(\vec{r})$. As the dual potential and descriptor have approximately similar shapes, eq 22 is a posteriori theoretical justification of the use of the opposite of the dual descriptor as an index to characterize the behavior of reactive sites of the first excited state.[44]

The reactions studied in the application section always involve one molecule in either the first or second excited state and one molecule in the ground state. According to the concepts developed in this section, the meaning of the sign of the electronic densities difference for the excited state is the opposite of the ones of the dual descriptor for the ground-state species. Therefore, the best interaction between molecules is obtained when regions with the same sign are aligned. Unfavorable interactions are occurring between regions with opposite signs of the density difference potential and dual descriptor. To express the previous consideration from a physical point of view the following criterion is proposed:

Table 1. Favorable and Unfavorable Interactions between a Molecule in Both its Ground and Excited States

ground state $\Delta f(\vec{r})$		excited state $\Delta_0^k\rho(\vec{r}')$		
sign	character	sign	character	interactions
+	electrophilic	+	nucleophilic	favorable
−	nucleophilic	−	electrophilic	favorable
±	el/nu	±	el/nu	unfavorable

$$\delta E_1 \propto -\iint \frac{\Delta_0^k\rho(\vec{r}')\Delta f^{gs}(\vec{r})}{|r - r'|}\mathrm{d}\vec{r}\,\mathrm{d}\vec{r}' \qquad (24)$$

From eq 24 it is easily seen that the energy goes downward when regions of the dual descriptor and electronic density difference of same sign are aligned. Those results are summarized in Table 1. Using this criterion, we will show that the regioselectivity of different photocycloadditions can be predicted.

3. Computational Details

All the studied molecules have been fully optimized at the B3LYP/6-31G** level of theory using the Gaussian 03 software.[45] Both the excited and ground state electronic densities have been obtained through a CIS calculation with the same basis set. Then the 3-dimension cube files of the difference have been performed using the cubman facility program. The dual descriptor for the ground state has been calculated from the density of the radical cation and anion using the formula:

$$\Delta f(\vec{r}) = \rho_{N+1}(\vec{r}) + \rho_{N-1}(\vec{r}) - 2\rho_N(\vec{r})$$

For all the isodensity maps, the positive regions are colored in red, while the negative regions are colored in yellow.

4. Application to Photocycloadditions

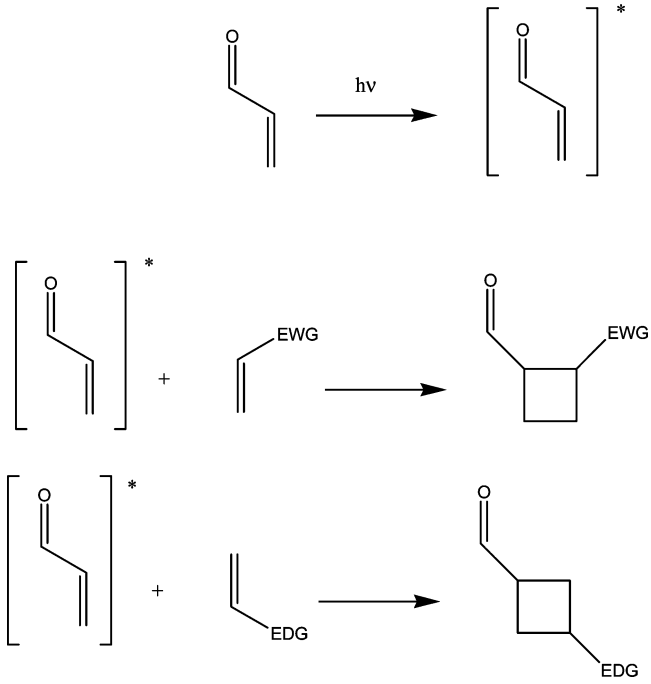
4.1. Regioselectivity of [2 + 2] Photoreactions of Acroleine with Olefins. The [2 + 2] cycloaddition between α, β unsaturated ketones and aldehydes with substituted olefins is one of the most versatile ways to produce cyclobutane derivatives. There are two main factors that control the regioselectivity of this reaction:

1) The steric interaction between the alkene and the α, β unsaturated ketone or aldehyde.

2) The electronic interaction between both reactants.

In the following study, only the reactions known to be controlled by electronic effects are discussed. According to either the electron-donating or electron-withdrawing character of the substituent on the alkene, the head-to-head or the head-

Scheme 1. Schematic Representation of the Regioselectivity of the [2 + 2] Photocycloaddition



to-tail cyclobutane is formed.[46] As can be seen in Scheme 1, when the acroleine reacts with electron-rich alkenes the head-to-tail adduct is preferred. Conversely, the reaction between the acroleine with electron-poor alkenes leads preferentially to the head-to-head cyclobutane. Different conceptual DFT approaches of this reaction[47] have been performed through the softness-matching criterion and reaction force with rather good success.[48] The aim of this subpart is to show how the concept developed in Section 2 can be applied to predict the outcome of the reaction.

Even though specific calculations have been performed to exemplify the usefulness of the electronic density difference, the regioselectivity of those reactions can be elegantly rationalized without any calculation, only using a back-of-the-envelope analysis. Generally, the patterns of the dual descriptor for electron-poor alkenes are asymmetric, with the nucleophilic zones ($\Delta f(\vec{r}) < 0$) centered on the substituted carbon, while the electrophilic character ($\Delta f(\vec{r}) > 0$) is located mainly on the unsubstituted end. As acroleine can be seen as an electron-poor alkene, the position of the positive and negative zones should be located in the same areas for its electronic densities difference $\Delta\rho(\vec{r})$. Since acroleine reacts in its first excited state and following the rules gathered in Table 1, the best interaction between acroleine and electron-poor alkenes is achieved through a head-to-head interaction (see Figure 3). The inverse conclusions can be drawn up for electron-rich alkenes, since their dual descriptors are the opposite of electron-poor alkenes ones. These qualitative conclusions are confirmed by the calculated map of dual descriptors and electronic densities differences.

It is well established that the acroleine reactive state is the one corresponding to a $\pi \rightarrow \pi^*$ transition. The corresponding electronic densities difference $[\Delta_{\pi}^{\pi^*}\rho(\vec{r})]$ has been calculated for both acroleine and cyclohex-2-enone. It is important to notice that this electronic densities difference

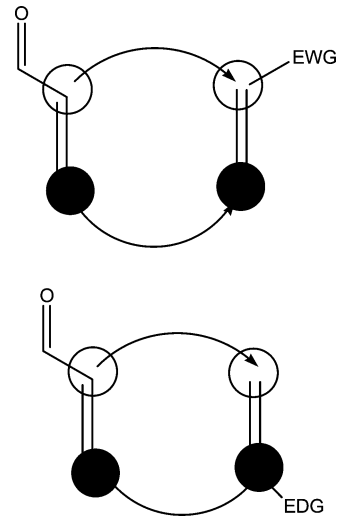


Figure 3. Schematic interaction between the dual descriptor of acroleine and either electron-rich or -poor alkenes. Dark circles correspond to positive value of the dual descriptor, while white circles are attributed to negative values of the dual descriptor.

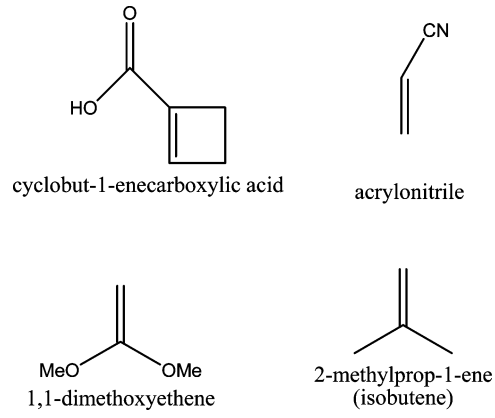


Figure 4. Name and structures of the studied alkenes.

$[\Delta_{\pi}^{\pi^*}\rho(\vec{r})]$ is not equivalent to the dual descriptor for the ground state, since the π orbital is not the HOMO. However, it appears that the positions of positive and negative values on the double bond are similar to the ones obtained for the dual potential. On the other hand, the ground-state dual descriptor calculations have been performed for different electron-rich and -poor alkenes: allene, acrylonitrile, isobutene, and 1,1-dimethoxyethene. The structures of all the compounds are drawn in Figure 4. The positions of positive and negative regions of the dual descriptor within the electron-rich and -poor alkenes follow the qualitative rule given previously. Namely, for electron-poor alkenes, the positive region is located on the unsubstituted end, while the negative area is centered on the substituted carbon. This is due to the asymmetry created on the HOMO and LUMO wave function by the electron-withdrawing group. Conversely, the effect of the electron-donating group (EDG) upon the shape of the dual descriptor is exactly the opposite. The negative regions are located on the unsubstituted carbon, while the positive regions cover mainly the substituted end of the double bond. Following the criterion defined in Section 2, the best interaction between the α, β unsaturated carbonyl compound and the alkene leads to the head-to-tail adduct for the

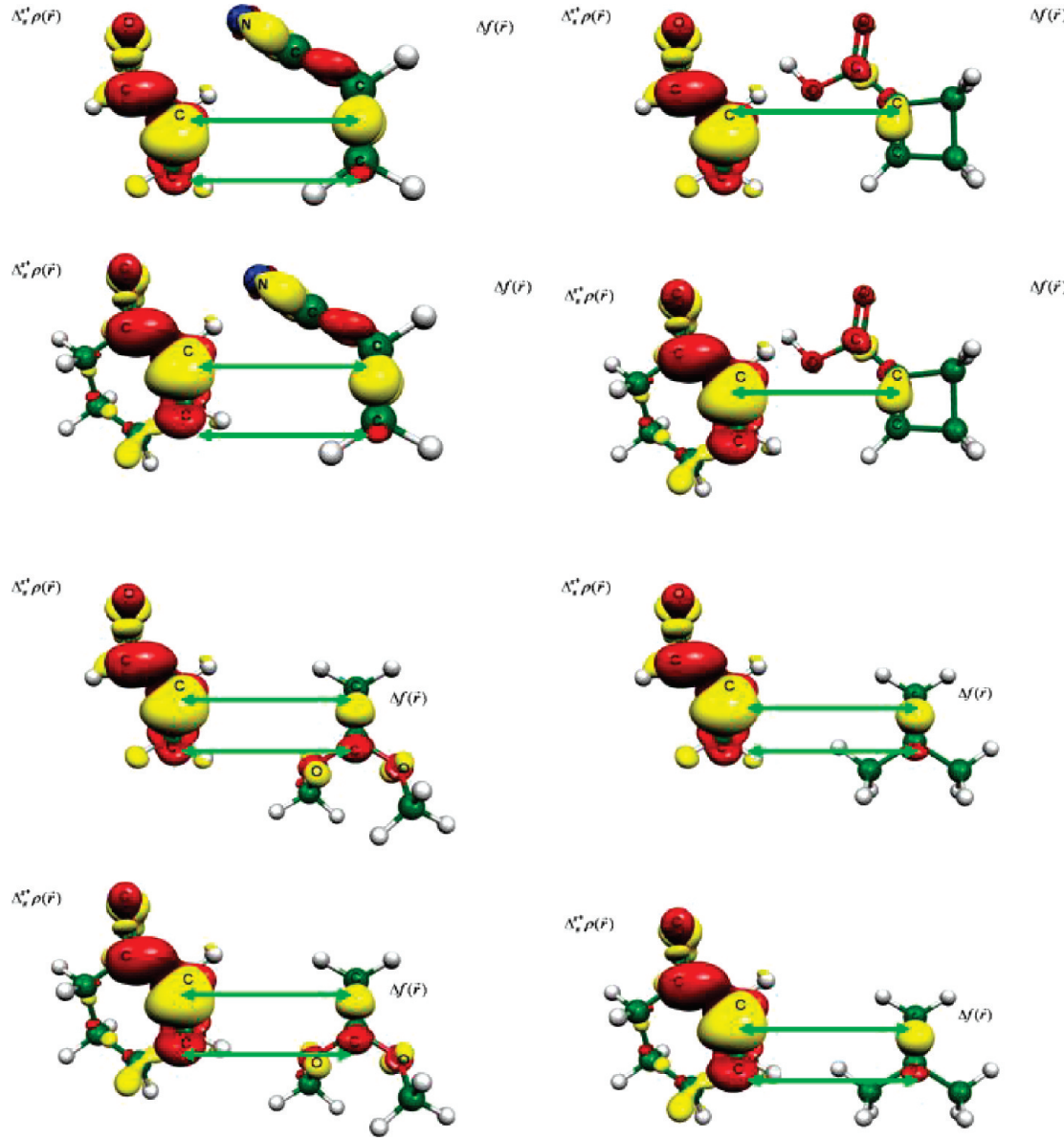


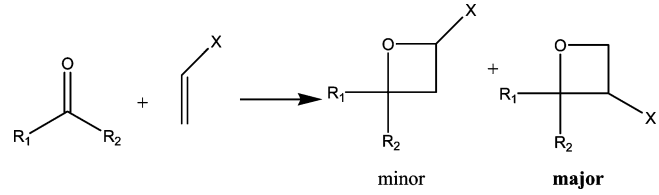
Figure 5. Favorable interactions between acroleine and cyclohex-2-enone with electron-poor (upper pictures) and -rich (lower pictures) alkenes. The isosurfaces are displayed for $\Delta f(\vec{r}) = 0.01$au and $\Delta_{\pi}^{\pi^*}\rho(\vec{r}) = 0.005$ au. The electron-poor alkenes taken are acrylonitryle and cyclobut-1-ene carboxylic acid. The electron-rich alkenes are 2-methylprop-1-ene and 1,1-dimethoxyethene.

electron-rich alkenes and head-to-head for electron-poor alkenes, as can be seen in Figure 5. This is in total agreement with experimental results.

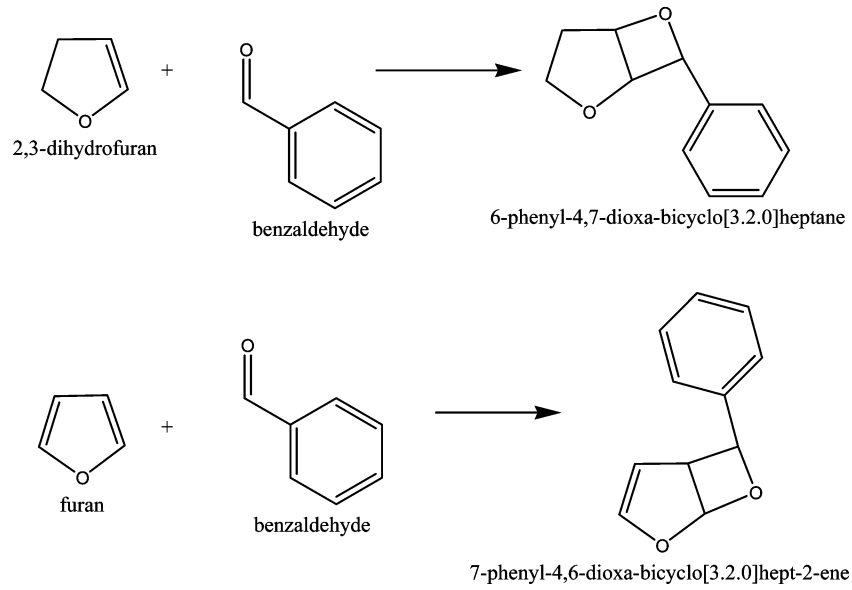
The next section is dedicated to the regio-and stereoselectivity of the Paterno−Büchi reaction.

4.2. Application to the Regio- and the Stereoselectivity of the Paterno−Büchi Reaction. The Paterno−Büchi (PB) is a versatile method to produce oxetane (see Scheme 2), i.e. four-membered oxygen heterocyclic rings with a good control over the regio- and stereoselectivity. This reaction involves a ketone or an aldehyde in its excited state and an olefin in its ground state. Numerous papers have been published on either the preparative or mechanistic aspect of the PB reaction. From all the results gathered, it is now generally admitted that the active excited state arises from the n→π^* transition. In this specific case, and contrary to the [2 + 2] photocycloaddition, the electronic densities difference is roughly equal to the dual descriptor ($\Delta_n^{\pi^*}\rho(\vec{r})$ $\approx \Delta f(\vec{r})$). The PB reaction is particularly rapid when the ketone or aldehyde reacts with an electron-rich alkene and is highly regioselective. In this section, we will study the regioselectivity of the reaction between benzaldehyde on the one hand and furan, or 2,3-dihydrofuran, on the other hand. For those cases, it has been shown experimentally that only one regioisomer is produced[49−51] (see Scheme 3). The PB reaction can, therefore, be called regiospecific. Aiming to check the predictive capability of our method, the dual

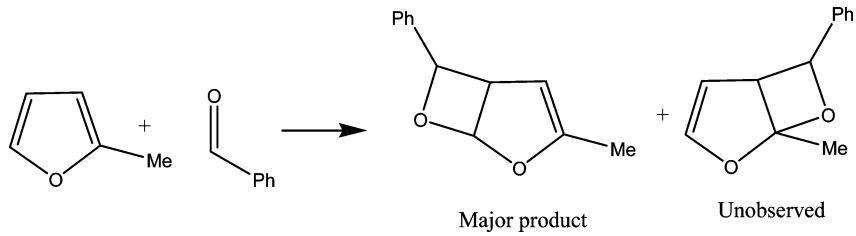
Scheme 2. Schematic Representation of the Paterno-Büchi Reaction



Scheme 3. Schematic Representation of the Regioselectivity of the Paterno-Büchi Reaction



Scheme 4. Schematic Representation of the Reaction between 2-Methylfuran and Benzaldehyde



descriptor for furan and its derivative and the density difference $\Delta_n^{\pi^*}\rho(\vec{r})$ for benzaldehyde have been calculated. The results are displayed in Figure 6. As already said, the best interactions between both molecules are achieved when regions with same sign are aligned. One can easily infer from Figure 6 that, in both cases, the correct regioisomer is predicted. As guessed, the negative region of the $\Delta_n^{\pi^*}\rho(\vec{r})$ of benzaldehyde is located on the oxygen, while the positive region is mainly on the carbonyl carbon. It looks very close to what was expected from the dual descriptor of benzaldehyde. For furan, the negative region is located between Cα and Cβ, namely between the double bond, while the positive region is in between the Cβ, i.e., located on the simple carbon−carbon bond. Therefore, the best interaction between furan and benzaldehyde is obtained when the oxygen of benzaldehyde reacts with the Cα of the furan and when the carbonyl carbon bonds with the Cβ. This must lead to the formation of 7-phenyl-4,6-dioxa-bicyclo[3.2.0] hept-2-ene.

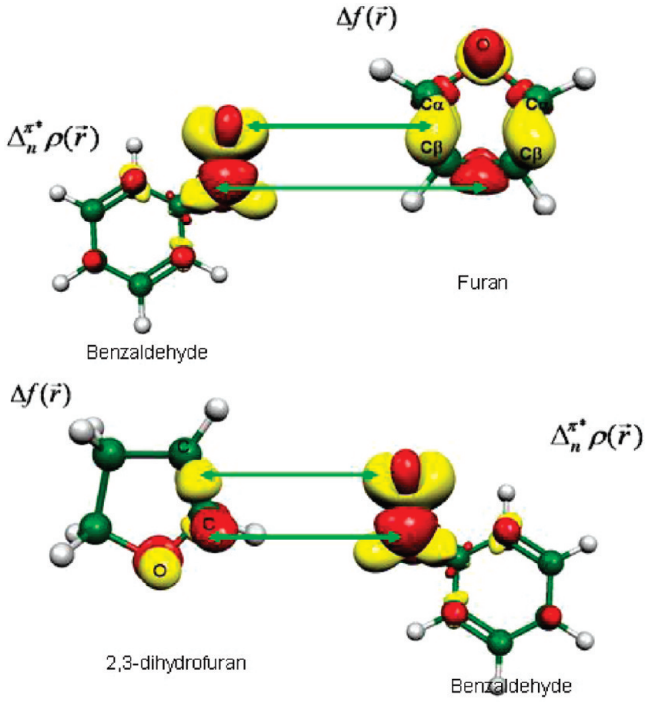


Figure 6. Favorable interaction between an excited benzaldehyde $\Delta_n^{\pi^*}\rho(\vec{r}) = 0.01$au with furan and 2,3-dihydrofuran $\Delta f(\vec{r}) = 0.01$au.

Since the position of the positive and negative region of the dual descriptor of 2,3-dihydrofuran are the opposite of the one obtained on furan, the best interaction is achieved when the oxygen of benzaldehyde reacts with Cβ, while the carbonyl carbon bonds with Cα. Again, the right regioisomer is predicted: 6-phenyl-4,7-dioxa-bicyclo[3.2.0]hept-2-ene.

When furan is methylated in position one, two regioisomers should be produced as shown in Scheme 4. From an analysis based on the frontier molecular orbital theory, D'auria et al.[52] concluded that the reaction is likely to take place in both positions Cα. However, this prediction is not in agreement with experimental results. Indeed, only the unsubstituted Cα carbon is attacked. This result can of course

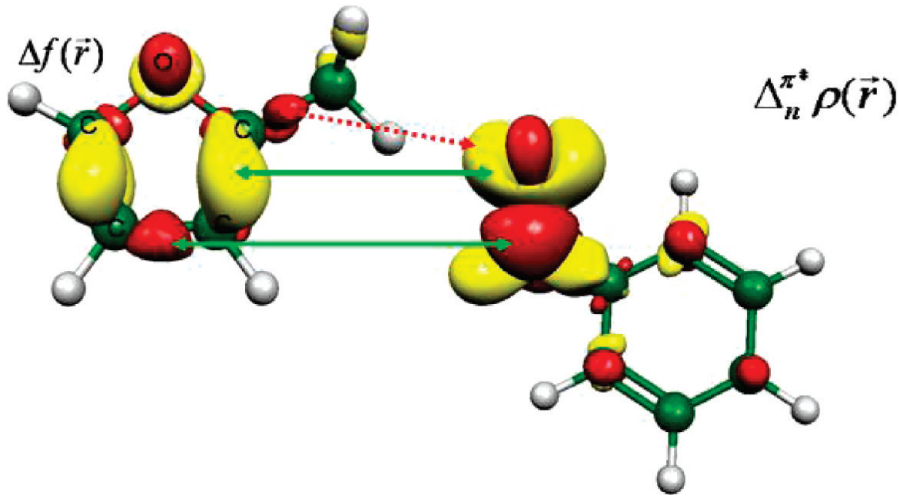
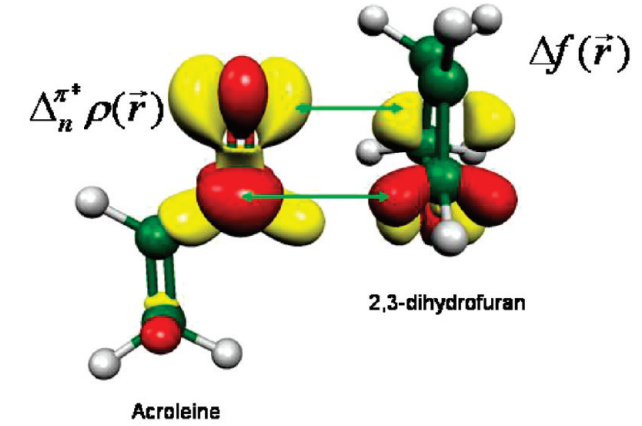


Figure 7. Favorable and unfavorable interaction between an excited benzaldehyde $\Delta_n^{\pi^*}\rho(\vec{r}) = 0.01$au and 2-methylfurane $\Delta f(\vec{r}) = 0.01$au.



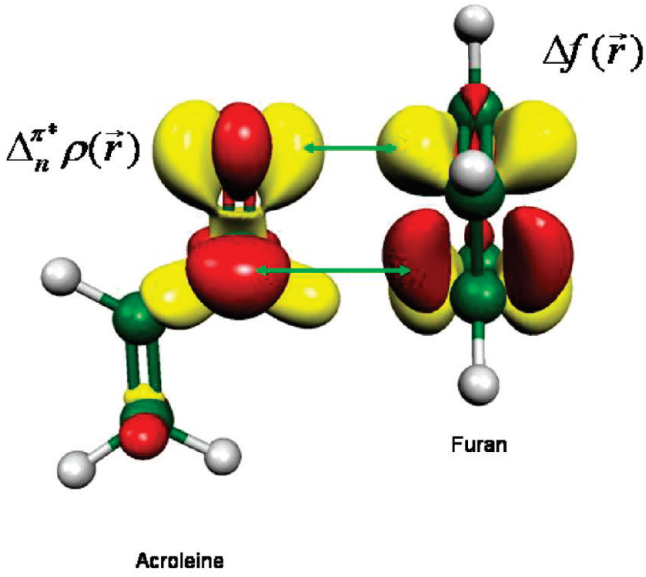


Figure 8. Best approach geometry between an excited acroleine and either furan and 2,3-dihydrofuran.

be explained by the steric hindrance that occurs in the substituted carbon. Still an analysis based on the $\Delta_n^{\pi^*}\rho(\vec{r})$ can provide an additional explanation. As can be seen in Figure 7, an electrophilic zone [$\Delta_n^{\pi^*}\rho(\vec{r}) > 0$, red] is located between the methyl group and the Cα. As the benzaldehyde approaches the substituted Cα carbon of methyl-furan, the favorable interaction between the oxygen of benzaldehyde with the Cα carbon is counterbalanced by the unfavorable interaction between the electrophilic zone near the Cα and the oxygen of the benzaldehyde. This unfavorable interaction is an additional effect that favors the experimentally observed isomer.

The geometry of approach can also be obtained from the interaction study between $\Delta_n^{\pi^*}\rho(\vec{r})$ benzaldehyde and the $\Delta f(\vec{r})$ furan. Indeed, it is well-known that in PB reactions, the molecules approach eachother with a face to edge geometry. As can be seen in Figure 8, this approach seems to be the one that maximizes the favorable interaction between both molecules.

5. Summary and Conclusions

In this paper, using the true excited-state density as a trial function for the ground-state DFT, a new local reactivity descriptor has been proposed. The main assumption is that the relaxation of the electronic density is a powerful enough phenomenon to drive the chemical behavior of an excited molecule. The locality of the chemical potential of an excited state is used to characterize the philicity of an atomic site within an excited molecule. It appears that for the first excited state the local chemical potential and the dual descriptor are related by a Poisson equation and look similar. This provides a posteriori justification of the use of the dual descriptor to rationalize the Woodward−Hoffmann rules for photochemical reactions. Finally, the application of the proposed concept of local chemical potential to [2 + 2] photocycloaddition and to Paterno-Büchi reaction predicts the correct regioselectivity. This paper can be considered as a first step toward a real conceptual DFT of excited states. To achieve this goal in full, however, it is necessary to derive a reliable time-independent functional of the energy that fully describes excited states. The local and global descriptors already defined for the ground state could then be generalized to excited states by differentiation of the excited-state functional.

Acknowledgment. C.M., V.L. and A.G. thank CEA-Grenoble for financial support. P.W.A. thanks the Canada

Research Chair, NSERC, and Sharcnet. C.M. and V.L. thank Professor M. E. Casida for helpful discussions.

References

(1) Geerlings, P.; De Proft, F.; Langenaeaker, W. *Chem. Rev.* **2003**, *103*, 1793.

(2) Gazquez, J. L. *J. Mex. Chem. Soc.* **2008**, *52*, 3.

(3) Liu, S. *Acta. Phys. Chim. Sin.* **2009**, *25*, 1.

(4) Chermette, H. *J. Comput. Chem.* **1999**, *20*, 129.

(5) Morell, C.; Hocquet, A.; Grand, A.; Jamard-Gregoire, B. *J. Mol. Struct.* **2008**, *46*, 849.

(6) Morell, C.; Ayers, P. W.; Grand, A.; Guttierez-Oliva, S.; Toro-Labbé, A. *Phys. Chem. Chem. Phys.* **2008**, *20*, 7239.

(7) Mulliken, R. S. *J. Chem. Phys.* **1934**, *2*, 782.

(8) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801.

(9) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512.

(10) Fukui, K.; Yonezawa, T.; Nagata, C. *Bull. Chem. Soc. Jpn.* **1954**, *27*, 423.

(11) Fukui, K.; Yonezawa, T.; Nagata, C. *J. Chem. Phys.* **1953**, *21*, 174.

(12) Fukui, K.; Yonezawa, T.; Shingu, J. *J. Chem. Phys.* **1952**, *20*, 722.

(13) Parr, R. G.; Yang, W. T. *J. Am. Chem. Soc.* **1984**, *106*, 4049.

(14) Yang, W. T.; Parr, R. G. *J. Chem. Phys.* **1984**, *81*, 2862.

(15) Ayers, P. W.; Levy, M. *Theo. Chem. Acc.* **2000**, *103*, 353.

(16) Pearson, R. G. *Science* **1966**, *151*, 172.

(17) Pearson, R. G. *J. Am. Chem. Soc.* **1963**, *85*, 3533.

(18) Ayers, P. W.; Parr, R. G.; Pearson, R. G. *J. Chem. Phys.* **2006**, *124*, 194107.

(19) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc* **1983**, *105*, 7512.

(20) Chattaraj, P. K.; Lee, H.; Parr, R. G. *J. Am. Chem. Soc.* **1991**, *113*, 1855.

(21) Ayers, P. W. *Faraday Discuss.* **2007**, *135*, 161.

(22) Ayers, P. W. *J. Chem. Phys.* **2005**, *122*, 141102.

(23) De Proft, F.; Fias, S.; Van Alsenoy, C.; Geerlings, P. *J. Phys. Chem. A* **2005**, *109*, 6335.

(24) Sengupta, D.; Chandra, A. K.; Nguyen, M. T. *J. Org. Chem.* **1997**, *62*, 6404.

(25) Mendez, F.; Garcia-Garibay, M. A. *J. Org. Chem.* **1999**, *64*, 7061.

(26) Chattaraj, P. K.; Poddar, A. *J. Phys. Chem. A* **1998**, *102*, 9944.

(27) Chattaraj, P. K.; Poddar, A. *J. Phys. Chem. A* **1999**, *103*, 1274.

(28) Chattaraj, P. K.; Poddar, A. *J. Phys. Chem. A* **1999**, *103*, 8691.

(29) Frank, J. *Faraday Soc.* **1926**, *21*, 536.

(30) Condon, E. *Phys. Rev.* **1926**, *27*, 640.

(31) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(32) Nagy, A. *Int. J. Quantum Chem.* **1998**, *70*, 681.

(33) Chattaraj, P. K.; Ghosh, S. K.; Liu, S.; Parr, R. G. *Int. J. Quantum Chem.* **1996**, *60*, 535.

(34) Gross, E. K. U.; Oliveira, L. N.; Kohn, W. *Phys. Rev. A* **1988**, *37*, 2809.

(35) Gross, E. K. U.; Oliveira, L. N.; Kohn, W. *Phys. Rev. A* **1988**, *37*, 2806.

(36) Ayers, P. W. Ph.D. Thesis dissertation.

(37) Levy, M.; Nagy, A. *Phys. Rev. Lett.* **1999**, *83*, 4361.

(38) Ayers, P. W.; Levy, M. Submitted.

(39) Ayers, P. W. *Theo. Chem. Acc.* **2007**, *118*, 371.

(40) Liu, S.; De Proft, F.; Parr, R. G. *J. Phys. Chem. A* **1997**, *101*, 6991.

(41) Ayers, P. W.; Parr, R. G. *J. Phys. Chem. A* **2000**, *104*, 2211.

(42) Morell, C.; Grand, A.; Toro-Labbé, A. *J. Phys. Chem. A* **2005**, *109*, 205.

(43) Morell, C.; Grand, A.; Toro-Labbé, A. *Chem. Phys. Lett.* **2006**, *425*, 342.

(44) Ayers, P. W.; Morell, C.; De Proft, F.; Geerlings, P. *Eur. J. Chem.* **2007**, *13*, 8240.

(45) *Gaussian 03, Revision B.3*, 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, Inc.: Pittsburgh, PA, 2003.

(46) Broeker, J. L.; Eksterowicz, J. E.; Belk, A. J.; Houk, K. N. *J. Am. Chem. Soc.* **1995**, *117*, 1847.

(47) De Proft, F.; Fias, S.; Van Alsenoy, C.; Geerlings, P. *J. Phys. Chem. A* **2005**, *109*, 6335.

(48) Jaque, P.; Toro-Labbé, A.; Geerlings, P.; De Proft, F. *J. Phys. Chem. A* **2009**, *113*, 332.

(49) Ogata, M.; Watanabe, H.; Kano, H. *Tetrahedron Lett.* **1967**, 533.

(50) Schenck, G. O.; Hartman, W.; Steinmetz, R. *Chem. Ber.* **1963**, *96*, 498.

(51) Gagnaire, D.; Payo-Subiza, E. *Bull. Soc. Chim. Fr.* **1963**, 2623.

(52) D'auria, M.; Emunuele, L.; Racioppi, R. *J. Photochem. Photobiol.* **2005**, *163*, 103.

CT900248A

JCTC Journal of Chemical Theory and Computation

Solvent Dependence of ^{14}N Nuclear Magnetic Resonance Chemical Shielding Constants as a Test of the Accuracy of the Computed Polarization of Solute Electron Densities by the Solvent

Raphael F. Ribeiro, Aleksandr V. Marenich, Christopher J. Cramer,* and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota 55455-0431

Received May 20, 2009

Abstract: Although continuum solvation models have now been shown to provide good quantitative accuracy for calculating free energies of solvation, questions remain about the accuracy of the perturbed solute electron densities and properties computed from them. Here we examine those questions by applying the SM8, SM8AD, SMD, and IEF-PCM continuum solvation models in combination with the M06-L density functional to compute the ^{14}N magnetic resonance nuclear shieldings of CH_3CN, CH_3NO_2, CH_3NCS, and CH_3ONO_2 in multiple solvents, and we analyze the dependence of the chemical shifts on solvent dielectric constant. We examine the dependence of the computed chemical shifts on the definition of the molecular cavity (both united-atom models and models based on superposed individual atomic spheres) and three kinds of treatments of the electrostatics, namely the generalized Born approximation with the Coulomb field approximation, the generalized Born model with asymmetric descreening, and models based on approximate numerical solution schemes for the nonhomogeneous Poisson equation. Our most systematic analyses are based on the computation of relative ^{14}N chemical shifts in a series of solvents, and we compare calculated shielding constants relative to those in CCl_4 for various solvation models and density functionals. While differences in the overall results are found to be reasonably small for different solvation models and functionals, the SM*x* models SM8, and SM8AD, using the same cavity definitions (which for these models means the same atomic radii) as those employed for the calculation of free energies of solvation, exhibit the best agreement with experiment for every functional tested. This suggests that in addition to predicting accurate free energies of solvation, the SM8 and SM8AD generalized Born models also describe the solute polarization in a manner reasonably consistent with experimental ^{14}N nuclear magnetic resonance spectroscopy. Models based on the nonhomogeneous Poisson equation show slightly reduced accuracy. Scaling the intrinsic Coulomb radii to larger values (as has sometimes been suggested in the past) does not uniformly improve the results for any kind of solvent model; furthermore it uniformly degrades the results for generalized Born models. Use of a basis set that increases the outlying charge diminishes the accuracy of continuum models that solve the nonhomogeneous Poisson equation, which we ascribe to the inability of the numerical schemes for approximately solving the nonhomogeneous Poisson equation to fully account for the effects of electronic charge outside the solute cavity.

1. Introduction

Nuclear magnetic resonance (NMR) shielding parameters are very sensitive to the molecular electronic structure,[1−3] and for molecules in solution they can serve as a probe of the solute response to intermolecular interactions. Buckingham et al.[4] proposed a rationalization of solvent effects on chemical shifts by separating these effects into four contributions: (i) bulk magnetic susceptibility of the medium, (ii) anisotropy in the molecular magnetic susceptibility of the solvent, (iii) van der Waals forces, and (iv) polar interactions, which were considered to include hydrogen bonds. The

* Corresponding author e-mails: cramer@umn.edu (C.J.C.); truhlar@unm.edu (D.G.T.).

10.1021/ct900258f CCC: $40.75
Published on Web 07/30/2009

Kamlet-Taft system of solvent properties[5,6] approaches the problem of solvent shifts and shielding parameters in a different way by considering the shielding parameter in cyclohexane as a reference and adding four different contributions for other solvents related to (i) the hydrogen-bond donor strength of the solvent, (ii) the solvent hydrogen-bond acceptor character, (iii) solvent polarizability, and (iv) a correction for superpolarizability of aromatic and highly chlorinated solvents.

Nitrogen shielding has received a great deal of attention in the study of solvent shifts since the lone pair on nitrogen often leads to particularly large shifts in electron density in response to an environment.[7−14] For this reason, several attempts were made to rationalize a broad range of experimental results for this nucleus, and a variety of schemes have been used for the computation of the solvent influence on nitrogen NMR shielding parameters.[7−12,15−26] These methods differ quantitatively in the way that the various contributions mentioned above are taken into account.

Modeling the solvent as a continuum has proven to be useful in predicting free energies of solvation provided that the solvent is characterized not only by the bulk dielectric constant but also by solvent and solute-dependent interfacial surface tensions that account for cavitation, dispersion, and local changes to solvent structure such as hydrogen bonds to the solute, solute disruption of the bulk-solvent hydrogen bonding network, and local changes to the solvent dielectric constant, especially in the first solvent shell.[27−34] However, the partition of free energies of solvation into bulk electrostatics effects and interfacial effects is not unique.[35] Since the former effects are usually[36,37] included by a self-consistent reaction field (SCRF), which affects the solute electronic density and properties, but the latter are usually treated post-SCF, which neglects their effects on the solute wave function or density and hence on solute properties, the success of a model for free energies of solvation does not guarantee its success for properties such as nuclear shieldings in solution. Nevertheless, it is disappointing that approaches based on continuum solvation models tested so far do not consistently explain many of the experimental results for nitrogen chemical shifts in different solvents, even though sometimes the trends are correct. To understand this situation better, the present study evaluates the performance of the newest models of the SM*x* series of solvation models (SM8,[31] SM8AD,[34] and SMD[33]) for solvent shifts and compares the results to those obtained with continuum solvent models[38−44] present in popular quantum chemistry codes[45,46] and with the work of Zhan and Chipman[16] in order to evaluate their absolute ability and relative ability to account for this response property and to ascertain whether any of the parametrization methods or electrostatic treatments that have been employed are systematically better than any of the others.

In section 2 we describe the test sets. Relevant theoretical background is summarized in section 3, while section 4 gives computational details, and section 5 provides the methodologies employed. Sections 6 and 7 contain results and discussion, and section 8 summarizes the main conclusions.

2. Test Sets

In the current article we compare and discuss nuclear shielding changes due to solvent-to-solvent transfer. In principle, theory could also be tested for gas-to-solvent shifts, but there are larger uncertainties associated with experimental measurements of these quantities, and they also pose a substantially greater challenge to the underlying electronic structure theory, which obscures the testing of solvation models. Thus we assess the performance of solvation models only for solvent-to-solvent shifts. Specifically, we restrict ourselves to well-established experimental results for acetonitrile (CH_3CN),[9] nitromethane (CH_3NO_2),[7] methyl nitrate (CH_3ONO_2),[8] and methyl isothiocyanate (CH_3NCS)[8] in a range of solvents. These four solutes are characterized by diverse nitrogen functional groups for which ^{14}N chemical shifts have been measured in multiple solvents with external neat nitromethane as a reference. Corrections for bulk solvent magnetic susceptibility were made[7−9] in the acquisition of the experimental data.

3. Theory

3.A. Solvation Models. Continuum solvation models represent a solvated molecule at an atomic level of detail inside a molecule-sized electrostatic cavity surrounded by a dielectric medium that represents the solvent. In some older work the cavity was a sphere, but here, as in most modern work, the cavity is solute-shaped and either is a superposition of atomic spheres with empirical radii or is a superposition of united atoms (a nonhydrogenic atom and its attached hydrogens), again with sizes defined by empirical parameters. In the electrostatic theory of dielectric media, the medium has associated with it a relative permittivity ε, which is a scalar function of position for isotropic nonhomogeneous media. Following the convention most popular in the chemical literature, ε will be called the dielectric constant. The charge distribution of the solute (charge density) induces polarization in the surrounding dielectric medium, and the self-consistently determined interaction between the solute charge distribution and the electric polarization field of the solvent, when adjusted for the energetic cost of polarizing the solute and the solvent, constitutes what is called the electrostatic contribution to the free energy of solvation. More properly, when one makes the usual assumption that the dielectric constant in regions containing solvent is given by the bulk value and then neglects the difference of the local dielectric constant in the near-solute region from its bulk value, this should be called the bulk-electrostatic contribution. The electric potential due to the polarized dielectric continuum and the polarization of the solute equals the total potential minus the electrostatic potential[47] of the gas-phase solute molecule. The total electric potential satisfies the nonhomogeneous Poisson equation (NPE) for electrostatics

$$\nabla \cdot (\varepsilon \nabla \Phi) = -4\pi\rho_f \qquad (1)$$

where ρ_f is the solute charge density, and the word "nonhomogeneous" refers to the specification of the dielectric constant as different inside the solute cavity (where it is given the value unity) and in the solvent region (where it is given

the bulk value). Therefore, the reaction field can be obtained self-consistently by numerical integration of the NPE coupled to the quantum mechanical electron density of the solute molecule. From the reaction field one calculates the free energy change corresponding to the solvation process. The bulk-electrostatic contribution to the free energy of solvation is given by[48]

$$\Delta G_{\mathrm{EP}} = \left\langle \Psi | H^{(0)} - \frac{e}{2}\phi | \Psi \right\rangle + \frac{e}{2}\sum_k Z_k \phi_k - \langle \Psi^{(0)} | H^{(0)} | \Psi^{(0)} \rangle \tag{2}$$

where e is the atomic unit of charge, ϕ_k is the reaction field evaluated at atom k, Z_k is the atomic number of atom k, $H^{(0)}$ and $\Psi^{(0)}$ are the solute electronic Hamiltonian and electronic wave function, respectively, in the gas phase, and Ψ is the polarized solute electronic wave function in solution. This equation includes the polarization of the solvent by the solute and the distortion of the solute that is induced by this polarization effect. Note that when one uses density functional theory, there is no wave function so the wave function appearing in eq 2 is replaced by the Kohn−Sham determinant of the noninteracting reference system, and the Hamiltonian operator is replaced by the appropriate density functional analog.

NPE solvers that employ the continuous charge density (without approximating it by distributed point charges or multipoles) are called density-based solvation models. Other implicit solvation models solve the NPE using alternative representations of the continuous density, for example, single- or multicenter multipolar expansions. An alternative continuum model, the generalized Born (GB) approximation,[27,49−56] does not start with the NPE but instead employs a starting point based on Coulomb's law and represents the solute as a collection of point charges (a distributed monopole approximation), located at the nuclear positions.

In the present study, we employ six solvation models. Four of them are based on the Integral-Equation-Formalism[39,40,42] of the Polarizable Continuum Model[38] (IEF-PCM) algorithm for solving the NPE using the polarized continuous quantum mechanical charge density of the solute. The first of these, called the SMD[33] model (SM is a general prefix for solvation models developed in our group, and "D" in the name stands for "density"), and the other three IEF-PCM-based models[38,41,43,45] differ in the atomic radii used to define the boundary between solute and solvent. The other two models are Solvation Model 8 (SM8)[31] and Solvation Model 8 with Asymmetric Descreening (SM8AD),[34] both of which utilize the GB approximation for bulk electrostatics based on self-consistently polarized class IV[57] partial atomic charges. In most of the calculations here we use the CM4M charge model,[58] although these models can also be employed with a more general CM4[30] charge model.

We can now revisit a key issue already raised in the Introduction. In all the SMx models employed here (x = D, 8, or 8AD), the observable fixed-concentration solvation free energy is partitioned algorithmically into two components. The first component is the bulk electrostatic contribution resulting from the interaction of a solute with its reaction field, which is the electric field produced by the polarized charge density that the solute induces in the solvent. This component is treated self-consistently. The second component is called the cavity-dispersion-solvent-structure (CDS) term and is the contribution arising from short-range interactions between the solute and solvent molecules in the first solvation shell. This contribution is a sum of terms that are proportional (with geometry-dependent proportionality constants called atomic surface tensions) to the solvent-accessible surface areas of the individual atoms of the solute.[27−34,54,55] In all SMx models, the CDS term is not included in the self-consistent reaction field procedure, and, therefore, it does not alter the resulting electronic wave function and it has no effect on system properties (such as dipole moments or NMR chemical shifts) other than the free energy of solvation. The other widely available parametrizations based on PCM[37,41,43,45] also involve non-bulk-electrostatic terms (usually called cavity-dispersion-repulsion terms) that do not influence the calculation of solute properties (note that "non-bulk-electrostatic" means terms that are not due to bulk electrostatics, not electrostatic terms that are nonbulk in character). Thus, we will consider only the bulk-electrostatic problem hereafter.

A key issue in all implicit solvation models is the boundary between the solute cavity where $\varepsilon < \varepsilon_s$ and the solvent continuum where $\varepsilon = \varepsilon_s$. In the SMD, SM8, and SM8AD models, or in PCM models employing scaled Bondi[59] radii (see below), the boundary between the solute cavity and the solvent dielectric continuum is defined to enclose a superposition of nuclear-centered spheres with radii ρ_{z_k}, which are called intrinsic Coulomb radii. The intrinsic Coulomb radii depend only on the atomic numbers Z_k of the atoms. This boundary forms a so-called solvent-accessible surface (SAS). The IEF-PCM-based calculations use the following approximation of the reaction field at an arbitrary position $\mathbf{r}$ within the SAS

$$\phi(\mathbf{r}) = \sum_m \frac{q_m}{|\mathbf{r} - \mathbf{r}_m|} \tag{3}$$

where $\mathbf{r}_m$ is the position of the center of an element m of surface area on the solute−solvent boundary (such elements are called tessarae), and q_m is an apparent surface charge on element m. In contrast, the GB approximation within the SM8 and SM8AD protocols is equivalent to approximating the reaction field distribution as

$$\phi_k = \sum_{k'} \frac{q_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|} f_{kk'} \tag{4}$$

where $\mathbf{r}_k$ and $\mathbf{r}_{k'}$ are evaluated only at atomic positions, $q_{k'}$ is a partial charge on atom k', and $f_{kk'}$ is a function to be specified. The value of a single term in eq 4 is called a Coulomb integral.

One successful function $f_{kk'}$ for approximating the Coulomb integrals is the dielectric descreening approximation of Still et al.,[53] which yields

$$f_{kk'} = -\left(1 - \frac{1}{\varepsilon_s}\right) \frac{r_{kk'}}{\sqrt{r_{kk'}^2 + \alpha_k \alpha_{k'} \exp(-r_{kk'}^2/d\alpha_k\alpha_{k'})}} \tag{5}$$

where

$$r_{kk'} \equiv |\mathbf{r}_k - \mathbf{r}_{k'}| \tag{6}$$

and where d is a parameter, ε_s is the dielectric constant of the bulk solvent, and α_k is the descreened atomic radius of atom k; α_k represents an appropriately weighted average distance of atom k from the solvent, and it is called a Born radius. The number of elements m in eq 3 is in principle increased to convergence, whereas the number of terms k' in eq 4 is equal to the number of atoms in the solute. A relation connecting eqs 3−6 is given as follows[37]

$$\sum_m q_m = -\left(1 - \frac{1}{\varepsilon_s}\right)\sum_{k'} q_{k'} \tag{7}$$

The SM8 model treats dielectric descreening effects by the Coulomb-field approximation[53] such that a partial effective charge in the solute interacts with the solvent by a charge-induced dipole interaction that varies as r^{-4}, where r is the distance between the partial atomic charge and a volume element of the continuum solvent. The Coulomb-field approximation leads to the following formula for the Born radius[53−55]

$$\alpha_k = \left(\frac{1}{R'} + \int_{\rho_{Z_k}}^{R'} \frac{A_k(r)}{4\pi r^4}\,\mathrm{d}r\right)^{-1} \tag{8}$$

In eq 8, 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, ρ_{Z_k} is the intrinsic Coulomb radius of atom k, and $A_k(r)$ is the exposed area of a sphere of radius r that is centered on atom k. This area depends on the geometry of the solute and the radii of the spheres centered on all the other atoms in the solute.

Grycuk has shown recently[56] that, when the partial atomic charge is asymmetrically situated in the molecule, i.e., not located at the center of the molecule, one can apparently estimate the dielectric descreening more accurately by using a shorter-range function proportional to r^{-6}. Therefore the SM8AD model uses an alternative functional form for the Born radius α_k, which is given by[56]

$$\alpha_k = \left(\frac{1}{R'^3} + \int_{\rho_{Z_k}}^{R'} \frac{3A_k(r)}{4\pi r^6}\mathrm{d}r\right)^{-1/3} \tag{9}$$

The SM8AD model is an extension of the SM8 model that replaces eq 8 by eq 9 to better account for the asymmetric descreening (AD).

SM8, SMD, and SM8AD have atomic radii optimized to free energies of solvation for ions. The four PCM-based models we consider are SMD and three older models whose bulk electrostatic treatment differs from SMD only in the choice of the solute−solvent boundary: (i) PCM-UA0, where UA0 is a united atom topological model;[44] (ii) PCM-UAKS, which also uses united atoms,[41] but in this case optimized for the calculation of free energies of solvation with Kohn−Sham density functional theory;[45] (iii) and PCM-1.2B where 1.2B denotes that the cavities are based on atomic spheres defined by Bondi's atomic radii times 1.2. The Bondi radii[59] are widely used atomic van der Waals radii, and the scale factor used here is 1.2, which, plus or minus ∼0.05, is a widely used value adopted by several groups.[37,60−64] The PCM-UA0 and PCM-1.2B models are effectively defined for all solvents because the only solvent-dependent parameter upon which there is a non-negligible dependence of the electrostatic response is ε. However, the electrostatic response of the UAKS model depends significantly not only on ε but also on α, which is a scale factor for solute radii. For the solvents considered in this article, α is 1.4 for cyclohexane, benzene, carbon tetrachloride, chloroform, dichloromethane, acetone, acetonitrile, and dimethylsulfoxide, 1.3 for diethyl ether, and 1.2 for ethanol, methanol, and water; and it is undefined for hexane, 1,4-dioxane, 2,2,2-trifluoro-ethanol, *N,N*-dimethylformamide, and ethylene glycol. We note that changing α from 1.2 to 1.4 can have a large effect on the chemical shift, often as much as 2−3 ppm.

We will also compare to some calculations of Zhan and Chipman[16] so it is useful to place their methods and nomenclature in the context of what was discussed above. Zhan and Chipman point out that the usual approximate solutions of the NPE in terms of an apparent polarization charge density on the cavity surface are valid only when the solute charge is entirely within the surface, which is never the case for real solutes because all atomic and molecular wave functions have exponentially decreasing tails. For real solutes, one also needs to introduce an apparent polarization charge density in the volume outside the cavity. Zhan and Chipman use SPE to denote methods that neglect this and SVPE to denote methods that include it. Chipman also points out that one can simulate volume polarization by adding additional apparent surface charge density,[25,65,66] and he calls this SS(V)PE.[25] Although a simple surface-polarization NPE solver would correspond to the SPE approach,[25] the IEF-PCM model includes an implicit correction for the effect of solute charge outside the cavity, and it can be formulated so as to be exactly the same as SS(V)PE.[37]

3.B. Nuclear Magnetic Resonance Chemical Shielding. NMR shielding constant tensors are calculated using a variational perturbative approach.[67−69] For a nucleus K they are calculated by

$$\sigma_k = 1 + \left.\frac{\partial^2 G(B,I)}{\partial I_k \partial B}\right|_{B=0,I_k=0} \tag{10}$$

where I_k is the nuclear magnetic moment, B is the magnitude of the applied magnetic field, and G is the free energy of the system. For the gauge invariance of the origin problem,[67] the most popular approach is to work with gauge invariant atomic orbitals (GIAO),[70−72] which includes complex factors in the basis functions and gives rise to gauge-invariant molecular properties. Other common methods to deal with this issue in quantum chemistry codes include Individual Gauge for Localized Orbitals (IGLO)[73] and Continuous Set of Gauge Transformations (CSGT)[74] (which is equivalent to the CTOCD-DZ scheme of Lazzeretti[75,76]).

The chemical shift is defined as minus the difference between the nuclear shielding constant and a reference value.

4. Computational Details

4.A. Software. All the calculations in this study were done using a locally modified version[77,78] of the *Gaussian03*[45] electronic structure suite.

4.B. DFT/M06-L. For our electronic structure model we chose to use density functional theory (DFT) and in particular the meta-GGA local functional M06-L.[79] We based this choice on a recent study[80] indicating that M06-L predicted NMR chemical shielding constants the most accurately in a comparison of several modern functionals for a diverse data set.

4.C. CSGT. The Continuous Set of Gauge Transformations[74] method was used in this work to calculate shielding constants, since the more popular GIAO[70−72] approach is not implemented in *Gaussian03* for meta-GGA or hybrid meta-GGA density functionals. Cheeseman et al.[81] have shown that CSGT and GIAO give very similar results, except that CSGT seems to have a slower convergence with respect to basis set.

4.D. Basis Set and Geometries. It is known that calculations of NMR chemical shielding tensors are very sensitive to basis set size and quality. However, it is also recognized that there is no reliable population analysis based on calculations that employ large basis sets.[82] Therefore, the CM4 and CM4M charge models, used to compute q_k values in the SM8 and SM8AD solvation models, do not have parameters for triple-ζ or larger basis sets. To learn more about basis set requirements, various exploratory studies were carried out. In section 6.A we present the influence of basis sets on the chemical shift of acetonitrile in four different solvents. Section 6.D.3 gives a different perspective by analyzing the effects of using a large basis set including diffuse functions and a smaller basis set with all core functions decontracted, in the computations of relative nitrogen shielding constants for the whole set of experimental data used in this work.

Since there is disagreement in the literature[16,19,24,83,84] concerning the influence of geometry relaxation effects on solvent shifts, we similarly studied acetonitrile in four different solvents with and without solution-phase geometry optimization. These results are in section 6.B.

5. Methods

5.A. Fitting Approach. In the first part of this study, we followed an approach similar to one taken by Zhan and Chipman[16] to uniformly shift the calculated results so as best to compare with the experimental data in solution and in particular with the *change* in chemical shift as a function of solvent. The convention adopted here is the one in which the positive direction of the scale corresponds to increasing magnetic shielding (upfield).

The first step taken was to calculate the nitrogen isotropic chemical shielding for the four selected solutes in four or five arbitrary solvents by SM8, SMD, and SM8AD. Since all of the experimental data used were relative to external neat nitromethane, we uniformly shifted the calculated results by a constant amount in order to correct, in a least-squares sense, for this and systematic errors (e.g., basis set incompleteness, approximate density functional, etc.). For each solvation model and solute, a shifting constant labeled D was obtained. The new values of relative shielding constants ($\sigma_{calc} + D$) were then fit to the following equation

$$\sigma = B + A/\varepsilon \tag{11}$$

where σ is equal to the calculated shielding $\sigma_{calc} + D$, B and A are optimized parameters, and ε is the dielectric constant of the medium. If σ were a function only of the dielectric constant, and if only polarization effects were important, $B + A$ could be regarded as the gas phase value, but since that is not true we cannot guarantee any physical meaning to them.

The parameters A and B were separately fit for each solute and method. An equation describing the dielectric dependence of each solute for each of the SMx models was found, and the results were plotted as a function of $(\varepsilon\text{-}1)/\varepsilon$. The same procedure was applied for each of the models with intrinsic Coulomb radii scaled by 0.8 and 1.2 to gauge the sensitivity of the results to these parameters and because scaling the radii by different factors had previously been determined sometimes to give improved results in prior studies with different solvation models.[16,21,22]

Since the wave function of the solute is affected only by bulk electrostatic effects in the solvation models used here, A and B reflect only polarization contributions to the solvent shift. Other specific contributions such as the high anisotropy of molecular magnetizability in some solvents such as benzene,[4] van der Waals forces, and the deviation of hydrogen bonding from a purely electrostatic effect would be implicitly included in the parameters A and B were the fitting procedure of eq 11 to be applied to experimental data as opposed to the theoretical data derived from the continuum solvation models. Our goals here are to determine the degree to which the three SMx models predict the correct variation of the shielding constants compared to experiment, to examine more closely those cases where they fail, to identify possible corrections in such cases, and to see in general how the predicted results are affected by scaling the intrinsic Coulomb radii.

5.B. Solvent Shifts Relative to CCl_4. Based on the results obtained with the method described in section 5.A, we formulated an extended approach to compute ^{14}N solvent shifts vs carbon tetrachloride as a reference solvent. In particular, for solutes where the nitrogen atom is likely to be involved in hydrogen bonds as an acceptor and the solvent has an Abraham's hydrogen bond acidity[85] $\alpha \geq 0.5$ (note that our α is called $\sum\alpha_2$ by Abraham), we suggest calculations with liquid-phase optimized solute−solvent clusters to improve accuracy. A complete discussion of this approach with results for the SMx models, a comparison with other implicit solvation methods, and the effects of scaling the radii by an increasing factor are presented in section 6.D.

6. Results and Discussion

We use SMD in sections 6.A and 6.B, SMD, SM8, and SM8AD in section 6.C, and all six solvation models in section 6.D. Experimentalists interested only in the solvent

Table 1. Calculated Relative ^{14}N Shielding Constants (in ppm) of Acetonitrile with Gas-Phase Geometry Optimized by M06-L/6-311++G(2df,2p)

solvent shift	6-31G(d)	6-31+G(d,p)	6-311++G(3df,3p)	aug-pcS-3
cyclohexane-water	12.8	14.0	16.2	16.6
cyclohexane-acetone	11.6	12.7	14.7	15.0
cyclohexane-chloroform	7.0	7.6	8.7	9.0
chloroform-water	5.8	6.4	7.4	7.6
chloroform-acetone	4.7	5.1	5.9	6.1
acetone-water	1.2	1.3	1.5	1.5

shifts on NMR shielding constants may skip sections 6.A−6.C and go straight to 6.D.

6.A. Basis Set Dependence of Nuclear Shielding. Since we are concerned with solvent shifts, an exploratory study was done to examine the effect of basis sets on this relative quantity. We selected acetonitrile as the solute, with the gas-phase geometry optimized by M06-L/6-311++G(2df,2p), and we calculated the nitrogen shielding constants using SMD as the solvent model in four different solvents, namely: cyclohexane, chloroform, water, and acetone, with each of the following basis sets: 6-31G(d), 6-31+G(d,p), 6-311++G-(3df,3p),[86−90] and the quadruple-ζ aug-pcS-3,[91−93] which was specially designed for the computation of nuclear magnetic resonance shielding constants by density functional methods.

The results are presented in Table 1; the largest differences between the solvent-to-solvent shifts calculated with aug-pcS-3 and 6-31G(d) are 3.78 and 3.42 ppm, but the ordering of the solvent-to-solvent shifts is consistent between the two basis sets. Since 6-31G(d) has well-established CM4M parameters, which are necessary for the computation of q_k values in eq 4, we decided to use this basis set for all chemical shift calculations, unless stated otherwise.

Irrespective of the convenience of the 6-31G(d) basis set for use with the SM8 and SM8AD models, the choice of this smaller basis set may also be more accurate for the SMD and PCM calculations, in spite of the usual rule that "bigger is better" when it comes to basis sets for electronic structure calculations. In particular we note that Zhan and Chipman[16] compared the performance of an SPE continuum solvation model (which is defined in section 3.A) that employs a surface-charge formalism to represent the solvent reaction to solute charge inside the cavity to that of an SVPE model that also includes polarization due to solute charge in the volume external to the molecular cavity. They observed that the former led to inaccurate chemical shift variations unless unrealistically large cavities were employed, while a "normal" cavity (defined as one that could also be put to routine use for computing free energies of solvation) was able to give realistic results only when volume polarization was included.[16] This problem should be mitigated in the SMD and PCM models examined in this article because they are all based on IEF-PCM, which (as discussed in section 3.A) should be able to simulate volume polarization. Nevertheless this observation suggests that the use of larger basis sets, particularly those that include diffuse functions that permit additional electronic charge density outside a typical solute cavity, may lead to instability in SMD-based and PCM-based chemical shift calculations. In contrast, decontracting the core basis functions centered on the atoms present in the system cannot lead to increased charge penetration, as the portion of the electronic density described by these functions is certainly inside the molecular cavity. Therefore, this approach should give improved chemical shifts. We will assess these points further in section 6.D.3.

Table 2. Relative Shielding Constants (in ppm) of Acetonitrile Calculated Using the Gas-Phase and Liquid-Phase Geometries

solvent shift	liquid geometry	gaseous geometry
cyclohexane-water	13.1	12.8
cyclohexane-acetone	11.9	11.6
cyclohexane-chloroform	7.1	7.0
chloroform-water	5.9	5.8
chloroform-acetone	4.8	4.7
acetone-water	1.2	1.2

6.B. Geometry Relaxation Effects. In order to decide whether or not to optimize the geometry of the solute in the liquid phase, we carried out an exploratory study to look at geometry relaxation effects on the chemical shift. We selected acetonitrile as our test solute because it has the largest range of solvent-to-solvent shifts (23 ppm) of the four solutes in the test set.[7−9] Acetonitrile was optimized in cyclohexane, water, acetone, and chloroform at the SMD/M06-L/6-311++G(2df,2p) level, and the NMR shielding constants were calculated in each solvent at the SMD/M06-L level but with the 6-31G(d) basis set. The trends shown in Table 2 make it clear that optimizing the geometry in the different liquids has little quantitative effect on the solvent-to-solvent shifts. Therefore, all the geometries used here were optimized at the gas phase M06-L/6-311++G(2df,2p) level, unless mentioned otherwise.

6.C. Dependence of the Chemical Shift on Dielectric Constant. *6.C.1. Solvent Shifts of Acetonitrile.* ^{14}N chemical shift data have been reported[9] for acetonitrile in fourteen solvents with neat nitromethane as an external reference. Table 3 and Figure 1 compare computed and experimental results. Each curve describing the dielectric dependence of the CH_3CN ^{14}N chemical shift (using eq 11) was fitted to the computed isotropic shielding constants in CCl_4 (ε = 2.2280), CH_2Cl_2 (ε = 8.93), CH_3CN (ε = 35.688), and $CHCl_3$ (ε = 4.7113).[94] We show here only the plots for standard models. Plots for models having scaled radii are presented in the Supporting Information. Mean unsigned errors between theory and experiment were calculated using the following equation

$$\mathrm{MUE} = \sum_{i=1}^{N} \frac{|\sigma_i - \sigma_{\mathrm{exp},i}|}{N} \qquad (12)$$

Table 3. Fitting Parameters of Eq *11* and MUEs (ppm) for ^{14}N Chemical Shifts of Acetonitrile

	D	*B*	*A*	MUE	MUEX[b]
			$F = 1$[a]		
SM8	86.09	136.97	−20.31	3.4	2.2
SMD	83.35	139.53	−25.92	3.5	2.7
SM8AD	89.10	135.96	−15.26	3.4	2.1
			$F = 0.8$		
SM8	75.02	141.18	−41.29	4.6	4.2
SMD	71.85	142.54	−48.10	6.1	5.3
SM8AD	81.34	138.74	−29.12	5.3	2.6
			$F = 1.2$		
SM8	91.02	135.33	−12.11	3.5	2.1
SMD	89.21	136.22	−16.53	3.3	2.0
SM8AD	92.94	134.75	−9.21	3.7	2.3

[a] *F* is a scaling factor for the intrinsic Coulomb radii in each of the models, where unity corresponds to the standard models. [b] Mean unsigned errors excluding the solvents water and TFE.

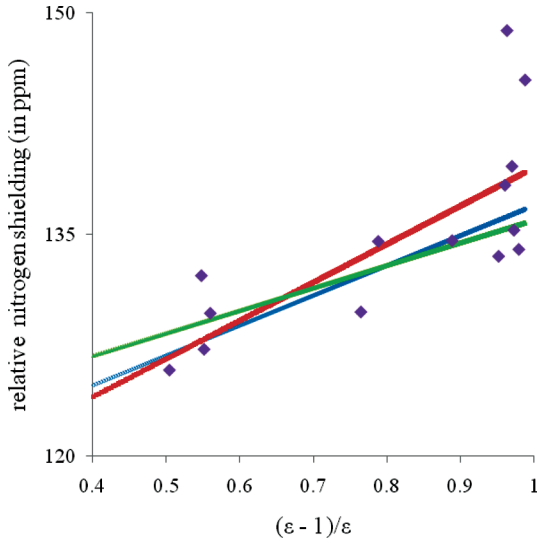


Figure 1. Acetonitrile nitrogen shielding as a function of solvent dielectric constant. Diamonds represent experimental results, the red curve represents eq 11 fitted to SMD results, the blue curve represents eq 11 fitted to SM8 results, and the green curve represents eq 11 fitted to SM8AD results.

where σ_i refers to the value predicted by eq 11 for a solvent having dielectric constant ε, σ_{exp} is the experimental value,[9] and *N* is the number of experimental results for a given molecule.

Table 3 indicates that scaling the standard intrinsic Coulomb radii of the SM8 and SM8AD models either fails to improve the predicted results or improves them by at most a very small amount. For SMD, on the other hand, increasing the radii by a factor of 1.2 does show an improvement in the MUE. The SM*x* models show correct qualitative dielectric dependence of the ^{14}N chemical shielding in solute acetonitrile. There are, however, large discrepancies between the experimental and computed results for water and 2,2,2-trifluoroethanol (TFE). Both of these solvents are strong hydrogen bond donors, and it seems reasonable to assume that hydrogen bonding to the nitrile nitrogen lone pair may be responsible for this disagreement. Thus, the MUEs for acetonitrile were also calculated excluding water and TFE (last column of Table 3), resulting in an error reduction from

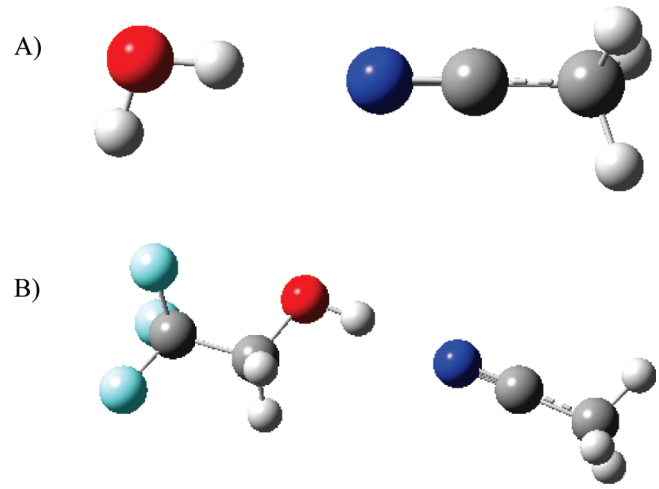


Figure 2. Liquid-phase optimized solute−solvent clusters for A) acetonitrile-water and B) acetonitrile-2,2,2-trifluoroethanol.

Table 4. Acetonitrile ^{14}N Shielding Constants (ppm) in Water and TFE Relative to CCl_4 as Solvent

solvent	SM8	SMD	SM8AD	experiment
		Cluster of Acetonitrile with Solvent		
water	14.1	15.7	13.3	18.2
2,2,2-trifluoroethanol	17.4	18.3	17.5	21.6
		Only Implicit Solvent		
water	8.1	11.1	5.9	18.2
2,2,2-trifluoroethanol	7.5	11.0	5.5	21.6

34% to 39% for the models with standard radii. For water and TFE, a correction in the calculated shifts to account for hydrogen bonding is discussed in the next section.

If we exclude water and TFE, SM8AD delivers the best performance for solvent-to-solvent shifts in the case of acetonitrile using standard radii. SM8 is nearly as accurate, and SMD is the most accurate of all if the cavity radii are scaled by a factor of 1.2.

6.C.2. Cluster Calculations of Hydrogen-Bonding Corrections to the Solvent Shift. The data for CH_3CN suggest that hydrogen bonds play a role in nuclear shielding that is not well described by continuum solvation models. To evaluate this point further, we optimized acetonitrile-water and acetonitrile-TFE clusters in their respective liquid phases at the SM8/M06-L/6-31G(d) level and the shielding constants were calculated for the clusters with each of the SM*x* models using the 6-31G(d) basis set. Figure 2 shows the geometries of the optimized clusters in solution, and Table 4 presents the results of this approach compared to the fully implicit solvent case.

The fully implicit solvent approach significantly underestimates the shieldings and also predicts higher shielding constants for water than for TFE, which fails to agree with the experimental data. The cluster calculations are in significantly improved agreement with experiment and also give the correct order for water vs TFE. SMD is the model that is most quantitatively accurate here, but SM8 and SM8AD also indicate the importance of clustering.

6.C.3. Solvent Shifts of Methyl Isothiocyanate. Measurements of ^{14}N chemical shifts in CH_3NCS have been reported

Table 5. Fitting Parameters[a] and MUEs (ppm) for ^{14}N Shielding Constants of Methyl Isothiocyanate

	D	*B*	*A*	MUE
		$F = 1$[b]		
SM8	105.99	290.37	7.50	1.5
SMD	109.16	288.68	13.68	1.9
SM8AD	106.60	290.01	8.79	1.6
		$F = 0.8$		
SM8	106.99	289.44	10.92	1.7
SMD	113.26	286.77	21.47	2.7
SM8AD	108.34	288.54	14.22	1.9
		$F = 1.2$		
SM8	104.88	290.99	5.20	1.5
SMD	106.60	290.04	8.68	1.6
SM8AD	105.38	290.74	6.13	1.5

[a] Equation 11. [b] F is a scaling factor for the intrinsic Coulomb radii in each of the models, where $F = 1$ corresponds to the standard models.

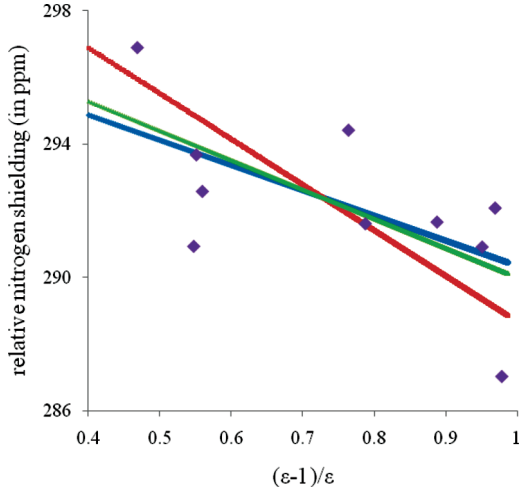


Figure 3. Nitrogen shielding in CH_3NCS as a function of solvent dielectric constant. Diamonds represent experimental results, the red curve represents eq 11 fitted to SMD results, the blue curve represents eq 11 fitted to SM8 results, and the green curve represents eq 11 fitted to SM8AD results.

in ten solvents.[8] The experimental data range over 9.9 ppm. Fits to eq 11 were based on the calculated ^{14}N data in *n*-hexane ($\varepsilon = 1.8819$), 1,4-dioxane ($\varepsilon = 2.2099$), $CHCl_3$ ($\varepsilon = 4.7113$), CH_2Cl_2 ($\varepsilon = 8.93$), and acetone ($\varepsilon = 20.493$).[94] Table 5 and Figure 3 present the results.

Again, scaling the model radii fails to have much effect on the mean accuracy of the SM8 or SM8AD results, but a scaling factor of 1.2 leads to improved accuracy for SMD. The SM*x* models are all consistent with the experimental trend of decreasing chemical shift with increasing dielectric constant. Such behavior contrasts with that of the SPE and SVPE models of Zhan and Chipman,[16] who hypothesized that untreated nonelectrostatic effects were required to capture this trend. The cavities for SPE and SVPE are chosen as surfaces of constant electron density; this is in contrast to the SM*x* models, which use fixed radii for individual atoms. One possible explanation is that in the case of MeNCS, the isodensity surface varies in ways that reverses the inverse relationship between chemical shift and the dielectric function seen here for the constant radius models, and this causes the isodensity models to fail to reproduce the direction of the experimental trend.

Table 6. Fitting Parameters of Eq *11* and MUEs (ppm) for ^{14}N Chemical Shifts of Methyl Nitrate

	D	*B*	*A*	MUE
		$F = 1$[a]		
SM8	89.41	38.90	4.67	0.9
SMD	89.60	38.67	5.63	0.8
SM8AD	89.87	38.73	5.41	0.8
		$F = 0.8$		
SM8	91.17	38.18	7.83	0.6
SMD	92.09	37.64	10.22	0.7
SM8AD	91.86	37.89	9.10	0.6
		$F = 1.2$		
SM8	88.45	39.16	3.49	1.0
SMD	88.45	39.12	3.66	1.0
SM8AD	88.57	39.19	3.38	1.0

[a] F is a scaling factor for the intrinsic Coulomb radii in each of the models; $F = 1$ corresponds to the standard models.

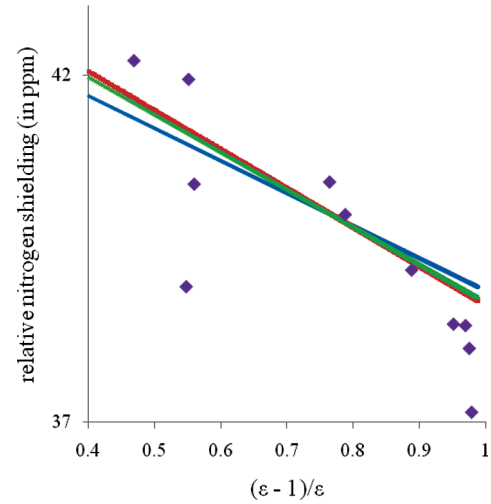


Figure 4. Nitrogen shielding of CH_3ONO_2 as a function of solvent dielectric constant. Diamonds represent experimental results, the red curve represents eq 11 fitted to SMD results, the blue curve represents eq 11 fitted to SM8 results, and the green curve represents eq 11 fitted to SM8AD results.

6.C.4. Solvent Shifts of Methyl Nitrate. For CH_3ONO_2, nitrogen NMR data have been reported in 12 solvents.[8] The experimental data range over only 5.1 ppm, i.e., this solute is the least sensitive to solvent of the four considered here. Fits to eq 11 were based on the calculated ^{14}N data in *n*-hexane ($\varepsilon = 1.8819$), $CHCl_3$ ($\varepsilon = 4.7113$), CH_2Cl_2 ($\varepsilon = 8.93$), and acetone ($\varepsilon = 20.493$).[94] The results found are presented in Table 6 and Figure 4.

The use of radii scaled by a factor of 0.8 leads to improvements in the accuracy of all three SM*x* models, but the sensitivity of the MUEs to radii is rather small, varying by at most about 0.4 ppm over the scaling range from 0.8 to 1.2. We note that the methyl nitrate data set does contain measurements for 1,2-ethanediol and 2,2,2-trifluoroethanol as solvents. However, while these solvents have $\alpha \geq 0.5$, no unusual deviation is observed from the continuum predictions, consistent with the nitrogen atom in the nitrate functional group being a poor hydrogen-bond acceptor.

Table 7. Fitting Parameters and MUEs (ppm) for ^{14}N Chemical Shifts of Nitromethane

	D	*B*	*A*	MUE
		F = 1.0[a]		
SM8	79.78	−0.16	13.19	1.5
SMD	78.29	−0.18	13.27	1.5
SM8AD	80.57	−0.42	14.51	1.6
		F = 0.8		
SM8	86.32	−2.39	25.01	2.3
SMD	84.68	−2.38	24.91	2.3
SM8AD	87.79	−2.92	27.81	2.7
		F = 1.2		
SM8	76.36	0.80	8.02	1.6
SMD	75.52	0.76	8.24	1.6
SM8AD	76.58	0.74	8.34	1.6

[a] *F* is the scaling factor for the intrinsic Coulomb radii; *D*, *B*, and *A* are defined by eq 11.

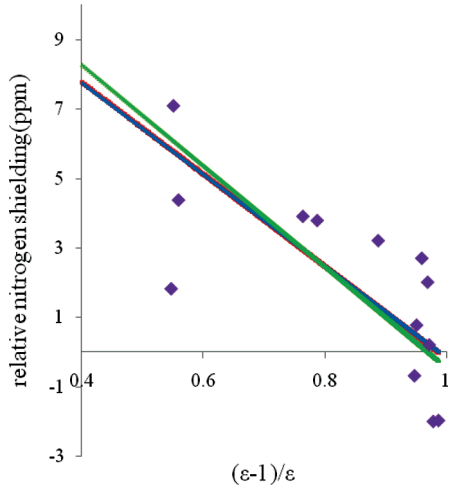


Figure 5. Nitrogen shielding in CH_3NO_2 as a function of solvent dielectric constant. Diamonds represent experimental results, the red curve represents eq 11 fitted to SMD results, the blue curve represents eq 11 fitted to SM8 results, and the green curve represents eq 11 fitted to SM8AD results.

6.C.5. Solvent Shifts of Nitromethane. For CH_3NO_2, NMR experiments have been carried out in thirteen solvents, and the experimental data range over 9.1 ppm.[7] Fits to eq 11 were based on the calculated ^{14}N data in benzene (ε = 2.2706), diethyl ether (ε = 4.24), acetone (ε = 20.493), and acetonitrile (ε = 35.688).[94] Table 7 and Figure 5 present the results.

Scaling the cavity radii does not improve the results for any SM*x* model, although scaling by a factor of 1.2 does not degrade the results much either. Considering Figure 5, the discrepancy between experiment and theory is somewhat larger for nitromethane than for the other three solutes studied here, but the largest error is for 1,4-dioxane as solvent (4.1 ppm). The source of this error is not clear, as no particular structural effects associated with this solvent are obvious.

6.C.6. General Discussion of the Fitting Approach. Table 8 shows the combined MUE and RMSE for each of the SM*x* models tested, including the models with scaled radii but excluding acetonitrile in water and 2,2,2-trifluoroethanol. For the SM8 and SM8AD models, the best results are obtained with the standard model radii. For the SMD model, a quantitative improvement of about 0.2 ppm is obtained after scaling of the radii by a factor of 1.2. This same scaling decreases the accuracy of the SM8 and SM8AD models by small margins. All of the models exhibit significantly poorer performance when a scale factor of 0.8 is used.

Table 8. MUEs and RMSEs (ppm) for Combined Eq *11* Fits

	MUE	RMSE
	F = 1.0[a]	
SM8	1.5	1.8
SMD	1.7	2.1
SM8AD	1.5	1.8
	F = 0.8	
SM8	2.2	3.1
SMD	2.7	3.8
SM8AD	2.0	2.6
	F = 1.2	
SM8	1.5	1.9
SMD	1.5	1.8
SM8AD	1.6	2.0

[a] *F* is the scaling factor for the intrinsic Coulomb radii.

Zhan and Chipman,[16] in an approach that motivated the one used here, tested their SVPE and SPE solvent models with the molecular cavity being defined by various isodensity contours, and they found in their best case a root mean squared error of 2.3 ppm for 48 experimental data (we have used 49). Table 8 shows that the SM*x* models have somewhat improved quantitative accuracy compared to SVPE and SPE; in particular they have root mean squared errors of only 1.8−2.1 ppm. Furthermore they correctly predict the inverse dependence of the ^{14}N chemical shift of methyl isothiocyanate on dielectric constant. Zhan and Chipman[16] had concluded that the inclusion of volume polarization is very important for the solvation effects on nitrogen shielding. Therefore it is very encouraging that all three SM*x* methods yield a smaller root mean squared error and exhibit more accurate trends than the SVPE model, which includes this effect. SM8 and SM8AD do not treat volume polarization explicitly, and SMD treats it approximately, like SS(V)PE, because it is based on IEF-PCM.

Mennucci et al. also studied the dependence of nitrogen shielding constants on the cavity size,[21,22] with the concern that cavities that may give qualitatively correct free energies of solvation could lead to inaccurate molecular properties. Various approaches were tested, the most important ones being the following: scaling all the intrinsic Coulomb radii by the same factors, scaling different groups (in the case of a united-atom approach) by different factors, and mixing different kinds of molecular cavities with different scaling factors, i.e., mixing scaled Bondi radii for some atoms with scaled radii for united-atom groups. For acetonitrile in water, they found that scaling the van der Waals radii by 1.4 improved the results in a reasonable manner, while no improvements were seen in the case of cyclohexane.[21] However, in a different paper,[22] an opposite trend was found, since scaling all radii by 1.4 gave rise to improved results for diazines in cyclohexane, but no improvement was seen for water.

Based on these other results and those reported here, it would appear that the radii that have been optimized against free energies of solvation for continuum models that ap-

proximately solve the Poisson equation are somewhat too small for the accurate calculation of ^{14}N chemical shifts, as scaling the radii optimized for free energies by factors ranging from 1.2 to 1.4 tends to lead to improved accuracy (although not in every instance). With the generalized Born models, on the other hand, a single set of radii seems equally suited to either task, although scaling the radii by a factor of 1.2 only degrades chemical shift predictions by a small amount. However, these conclusions are based on a fitting approach that may improve the results in a nonsystematic manner. To understand this situation better, another study was performed to more directly investigate the effects of scaling the radii in generalized Born and NPE models, and the results are discussed in section 6.D.2.

6.D. Solvent Shifts Relative to CCl_4. *6.D.1. Standard Radii.* Equation 11, as used in section 6.C, offers a method to assess the qualitative accuracy of the various solvation models with respect to predicting solvation effects on ^{14}N chemical shifts, but, in the absence of data in enough solvents to do a fit, it is not a practical means for actually predicting chemical shift values. In this section we present a more straightforward protocol for the computation of ^{14}N solvent-to-solvent shifts. Since this protocol calculates solvent-to-solvent shifts, it is convenient to define a reference solvent. We choose CCl_4 as it does not make hydrogen bonds, and it is a commonly used solvent.

In the procedure adopted in this section, one calculates the shielding in a desired solvent using one of the SM*x* models, and one calculates the difference from the result obtained by the same model for the shielding in the reference solvent, CCl_4. If the nitrogen atom of the solute for which the solvent shift is being calculated is known to make strong hydrogen bonds with the chosen solvent, then a cluster calculation is carried out with one explicit solvent molecule (and the rest of the solvent still implicit); otherwise all solvent molecules are implicit. The criterion we adopted for deciding if this situation applies to a specific system is whether the Abraham's hydrogen bond acidity of the solvent in which the solute is immersed is greater than 0.5. In this work, this criterion applied only to CH_3CN in water and TFE, and the cluster geometries for these two cases were optimized in the liquid phase with SM8/M06-L/6-31G(d).

In order to place the performance of the SM*x* models in the context of what can be obtained with other implicit solvation models for solvent-to-solvent shifts, we compared three other models to experiment employing the same basis set (6-31G(d)), the same gas-phase M06-L/6-311++G(2df,2p) geometries, and the same liquid-phase SM8/M06-L/6-31G(d) cluster geometries. The implicit models to which we compared are three IEF-PCM continuum solvent methods present in the popular *Gaussian03* computer package: (i) PCM with UA0[44] radii, (ii) PCM with UAKS[41] radii, and (iii) PCM with Bondi's atomic radii[59] multiplied by the scaling factor 1.2. For (i) we computed the solvent shifts using the B3LYP[95−98] and M06-L[79] density functionals. B3LYP was chosen simply because it is the most popular functional in the chemical literature, and UA0 radii were chosen because they are the default option for PCM calculations in *Gaussian03*. In the case of PCM with the UAKS radii (ii), we chose to use M06-L and PBE0[99] as the density functional since the UAKS parameters were optimized for density functional theory with the PBE0/6-31G(d) method. For (iii) we did calculations only with M06-L. SM8AD calculations were also done with B3LYP and PBE0 in order to assess sensitivity to density functional. Results are presented in Tables 9 and 10 for each solute in the order of increasing solvent dielectric constant. In these tables, as well as later tables, mean unsigned errors were computed from unrounded data.

Large deviations are observed when 1,4-dioxane is chosen as a solvent. This solvent evidently gives rise to interactions that cannot be described by any of the implicit solvent models, all of which are restricted to describing results related to changes in dielectric constant. Benzene is another problematic nonpolar solvent, but this is more understandable as the continuum solvent models cannot reproduce effects such as solvent ring current, which may affect the nuclear shielding constants of solute atoms in aromatic solvents. Continuum solvent models also neglect charge transfer between solute and solvent.

The SM*x* models SM8 and SM8AD provide the lowest MUEs compared to experiment for any of the functional/solvation model combinations tested. However, none of the models show particularly large errors, although use of the UA0 radii in conjunction with PCM cannot be recommended. MUEs from SM8AD with B3LYP and PBE0 are about 0.2 ppm better than those with M06-L. The good performance of SM8 relative to PCM-UA0 is particularly interesting since it has been suggested by three of the authors of *Gaussian03* and co-authors[100] that the reason that the UA0 cavities (which are the default choice in *Gaussian03*) predict less accurate free energies of solvation than SM8 is that they represent a compromise designed to give insight into various solute properties, including magnetic properties and response properties. Although the research underlying the mentioned compromise is apparently unpublished, the inference one could draw from that suggestion is that PCM-UA0 might give more accurate magnetic response properties than SM8. The results here indicate that this is not the case. The SM8 model is more accurate, although the SM8 model was clearly optimized only for free energies of solvation. We should note though that there was an attempt to make the model as physical as possible within the constraints of the chosen functional forms (e.g., no specification of types of atoms is used, and the radii are independent of both overall charge and partial atomic charges).

6.D.2. Solvent Shifts Relative to CCl_4 - Scaled Radii Models. Since the fitting approach indicated that scaling the radii of SMD by a factor of 1.2 improved agreement with experiment, we decided to test this issue for the case of shifts relative to carbon tetrachloride described in section 6.D.1. We examined three models, namely, SMD, PCM-1.2B, and SM8. For each of these models, all of the radii were scaled by a factor of 1.2 (meaning for PCM-1.2B the final radii were 1.44 times larger than Bondi radii), and the nitrogen shielding was calculated with the M06-L functional and 6-31G(d) basis set. The results for these methods are present in Table 11. In no instance is agreement with experiment

Table 9. Predicted Relative Shielding Constants (ppm) Using the M06-L Density Functional

	SM8	SMD	SM8AD	PCM-UA0	PCM-UAKS	PCM-1.2B	experiment
				CH_3CN			
CCl_4-C_6H_{12}	−0.8	−0.1	−0.7	−0.9	−0.6	−0.9	−1.4
CCl_4-1,4-dioxane	−0.1	−0.1	−0.1	−0.1	NA[a]	−0.1	5.0
CCl_4-C_6H_6	0.1	0.1	0.1	0.1	0.0	0.1	2.4
CCl_4-Et_2O	4.1	5.3	3.2	4.3	4.9	4.7	2.5
CCl_4-$CHCl_3$	4.4	6.0	3.3	4.9	3.5	5.4	7.3
CCl_4-CH_2Cl_2	6.6	8.6	5.0	7.1	5.0	7.8	7.4
CCl_4-acetone	8.3	10.6	6.3	8.7	6.1	9.6	6.3
CCl_4-EtOH	8.5	11.0	6.4	8.9	12.0	9.8	11.1
CCl_4-TFE[b]	17.4	18.3	17.5	15.8	NA	17.2	21.6
CCl_4-MeOH	8.7	11.2	6.6	9.2	12.3	10.1	12.4
CCl_4-CH_3CN	8.6	11.3	6.5	9.2	6.5	10.2	8.1
CCl_4-DMSO	9.0	11.5	6.8	9.4	6.6	10.4	6.8
CCl_4-water[b]	14.1	15.7	13.3	12.3	15.1	14.2	18.2
MUE	2.5	2.4	2.8	2.8	NA	2.6	
				CH_3NO_2			
CCl_4-1,4-dioxane	0.0	0.3	0.0	0.0	NA	0.0	−5.3
CCl_4-C_6H_6	−0.1	0.2	−0.1	0.0	−0.0	0.0	−2.7
CCl_4-Et_2O	−2.5	−2.0	−2.7	−1.9	−2.5	−2.3	−3.2
CCl_4-$CHCl_3$	−2.8	−3.3	−3.1	−2.2	−1.7	−2.6	−3.3
CCl_4-CH_2Cl_2	−4.0	−4.2	−4.5	−3.2	−2.5	−3.8	−3.9
CCl_4-acetone	−5.1	−4.7	−5.6	−3.9	−3.0	−4.6	−6.3
CCl_4-EtOH	−5.0	−7.9	−5.5	−4.0	−6.4	−4.8	−4.4
CCl_4-MeOH	−5.1	−8.7	−5.6	−4.1	−6.5	−5.0	−5.1
CCl_4-CH_3CN	−5.6	−5.3	−6.1	−4.1	−3.2	−5.0	−6.9
CCl_4-DMF	−5.2	−4.8	−5.7	−4.1	NA	−5.0	−7.8
CCl_4-DMSO	−5.2	−4.9	−5.8	−4.2	−3.3	−5.0	−9.1
CCl_4-H_2O	−8.8	−9.1	−9.5	−4.3	−6.8	−5.2	−9.1
MUE	1.6	2.3	1.5	2.6	NA	2.0	
				CH_3NCS			
CCl_4-C_6H_{14}	0.5	0.9	0.6	0.7	NA	0.6	3.2
CCl_4-1,4-dioxane	0.0	0.0	0.0	0.0	NA	0.0	−2.7
CCl_4-C_6H_6	−0.1	0.0	−0.1	0.0	0.0	−0.1	−1.1
CCl_4-Et_2O	−1.6	−2.8	−1.8	−2.3	−2.6	−1.9	0.7
CCl_4-$CHCl_3$	−1.6	−3.2	−1.8	−2.6	−1.9	−2.3	−2.1
CCl_4-CH_2Cl_2	−2.5	−4.6	−2.9	−3.8	−2.7	−3.3	−2.0
CCl_4-acetone	−3.2	−5.7	−3.8	−4.7	−3.3	−4.0	−2.8
CCl_4-MeOH	−3.4	−6.0	−4.1	−5.0	−6.6	−4.2	−1.6
CCl_4-DMSO	−3.5	−6.2	−4.2	−5.1	−3.5	−4.3	−6.7
MUE	1.7	2.4	1.8	2.1	NA	1.9	
				CH_3ONO_2			
CCl_4-C_6H_{14}	0.3	0.3	0.4	0.3	NA	0.3	0.3
CCl_4-1,4-dioxane	0.0	0.1	0.0	0.0	NA	0.0	−3.0
CCl_4-C_6H_6	0.0	0.0	0.0	0.0	0.0	0.0	−1.5
CCl_4-Et_2O	−0.1	−1.0	−1.1	−0.9	−1.1	−1.1	−1.5
CCl_4-$CHCl_3$	−1.2	−1.6	−1.4	−1.0	−0.8	−1.2	−1.9
CCl_4-CH_2Cl_2	−1.7	−1.7	−1.9	−1.5	−1.1	−1.7	−2.7
CCl_4-acetone	−1.9	−1.9	−2.2	−1.8	−1.4	−2.1	−3.5
CCl_4-TFE	−1.9	−3.8	−2.2	−1.9	NA	−2.3	−3.9
CCl_4-MeOH	−1.9	−3.8	−2.2	−1.9	−2.7	−2.3	−3.5
CCl_4-1,2-ethanediol	−1.9	−3.9	−2.3	−1.9	NA	−2.3	−3.9
CCl_4-DMSO	−2.0	−2.3	−2.3	−1.9	−1.5	−2.3	−4.8
MUE	1.5	1.0	1.3	1.6	NA	1.4	
subset MUE[c]	1.6	1.9	1.6	2.0	1.9	1.7	
total MUE	1.9	2.0	1.9	2.3	NA	2.0	

[a] NA denotes not applicable because the method is not defined for this solvent. [b] These results were provided by cluster calculations with the solute and a single solvent molecule optimized with SM8/M06-L/6-31G(d). [c] Includes only the rows where all methods are defined.

improved compared to the results provided in Table 9. While this may in part be associated with choosing carbon tetrachloride as a reference solvent, it suggests that scaling of radii cannot be considered to be a general requirement for accurate NPE-based predictions, and the fitting approach embodied in eq 11 is best employed only to assess qualitatively the dielectric dependence of the solvent shifts predicted by each model.

Another way to scale the radii is to scale them differently in each solvent, as in the UAKS model. The original example of this is the work of Luque, Orozco, and co-workers, who called their version of PCM by the name MST, and who scaled the Pauling radii by 1.25 in water,[60] 1.60 in chloroform,[101] and 1.80 in carbon tetrachloride.[102] Another example of this is the UAHF model,[41] upon which the UAKS model is a variation. We

Table 10. Predicted Relative Shielding Constants (ppm) using the PBE0 and B3LYP Density Functionals

	B3LYP		PBE0		
	PCM-UA0	SM8AD	PCM-UAKS	SM8AD	experiment
			CH_3CN		
CCl_4-C_6H_{12}	−0.9	−0.7	−0.6	−0.7	−1.4
CCl_4-1,4-dioxane	−0.1	−0.1	NA[a]	−0.1	5.0
CCl_4-C_6H_6	0.1	0.1	1.0	0.1	2.4
CCl_4-Et_2O	4.8	3.5	5.4	3.6	2.5
CCl_4-$CHCl_3$	5.4	3.7	3.9	3.8	7.3
CCl_4-CH_2Cl_2	7.8	5.6	5.6	5.7	7.4
CCl_4-acetone	9.5	7.0	6.8	7.2	6.3
CCl_4-EtOH	9.7	7.2	13.4	7.3	11.1
CCl_4-TFE[b]	17.2	19.4	NA	20.2	21.6
CCl_4-MeOH	10.0	7.4	13.7	7.5	12.4
CCl_4-CH_3CN	10.1	7.3	7.3	7.4	8.1
CCl_4-DMSO	10.3	7.6	7.4	7.7	6.8
CCl_4-water[b]	13.2	14.7	17.1	15.4	18.2
MUE	2.6	2.4	NA	2.3	
			CH_3NO_2		
CCl_4-1,4-dioxane	0.0	0.0	NA	0.0	−5.3
CCl_4-C_6H_6	0.0	−0.1	0.0	−0.1	−2.7
CCl_4-Et_2O	−2.1	−3.0	0.0	−3.1	−3.2
CCl_4-$CHCl_3$	−2.4	−3.3	−2.8	−3.4	−3.3
CCl_4-CH_2Cl_2	−3.4	−4.8	−7.2	−5.0	−3.9
CCl_4-acetone	−4.2	−6.1	−2.8	−6.3	−6.3
CCl_4-EtOH	−4.3	−6.0	−7.3	−6.1	−4.4
CCl_4-MeOH	−4.4	−6.1	−7.0	−6.3	−5.1
CCl_4-CH_3CN	−4.5	−6.7	−3.4	−6.8	−6.9
CCl_4-DMF	−4.5	−6.2	NA	−6.3	−7.8
CCl_4-DMSO	−4.6	−6.3	−1.6	−6.4	−9.1
CCl_4-water	−4.7	−10.3	−3.6	−10.5	−9.1
MUE	2.3	1.5	2.3	1.5	
			CH_3NCS		
CCl_4-C_6H_{14}	0.8	0.6	NA	0.7	3.2
CCl_4-1,4-dioxane	0.0	0.0	NA	0.0	−2.7
CCl_4-C_6H_6	0.0	−0.1	0.0	−0.1	−1.1
CCl_4-Et_2O	−2.6	−1.8	−2.9	−2.0	0.7
CCl_4-$CHCl_3$	−2.9	−1.8	−2.1	−1.9	−2.1
CCl_4-acetone	−5.2	−3.8	−3.7	−4.0	−2.8
CCl_4-CH_2Cl_2	−4.2	−2.9	−3.0	−3.1	−2.0
CCl_4-MeOH	−5.5	−4.0	−7.4	−4.3	−1.6
CCl_4-DMSO	−5.7	−4.2	−4.0	−4.4	−6.7
MUE	2.2	1.8	NA	1.8	
			CH_3ONO_2		
CCl_4-C_6H_{14}	0.3	0.4	NA	0.5	0.3
CCl_4-1,4-dioxane	0.0	0.0	NA	0.0	−3.0
CCl_4-C_6H_6	0.0	0.0	0.0	0.0	−1.5
CCl_4-Et_2O	−1.0	−1.2	−1.2	−1.3	−1.5
CCl_4-$CHCl_3$	−1.2	−1.6	−0.9	−1.6	−1.9
CCl_4-CH_2Cl_2	−1.7	−2.2	−1.3	−2.3	−2.7
CCl_4-acetone	−2.0	−2.5	−1.6	−2.6	−3.5
CCl_4-TFE	−2.1	−2.5	NA	−2.6	−3.9
CCl_4-MeOH	−2.1	−2.6	−3.1	−2.6	−3.5
CCl_4-1,2-ethanediol	−2.2	−2.6	NA	−2.7	−3.9
CCl_4-DMSO	−2.2	−2.6	−1.7	−2.7	−4.8
MUE	1.4	1.1	NA	1.1	
subset MUE[c]	2.0	1.5	1.9	1.5	
total MUE	2.2	1.7	NA	1.7	

[a] NA denotes not applicable because the method is not defined for this solvent. [b] These results were provided by cluster calculations with the solute and a single solvent molecule optimized with SM8/M06-L/6-31G(d). [c] Includes only the rows where all methods are defined.

Table 11. Prediction of Relative Shielding Constants with the M06-L Functional and Continuum Solvent Models with Radii Scaled by 1.2

	SMD	PCM-1.2B	SM8	experiment
		CH_3CN		
CCl_4-C_6H_{12}	−0.6	−0.6	−0.5	−1.4
CCl_4-1,4-dioxane	0.1	0.0	0.0	5.0
CCl_4-C_6H_6	0.1	0.0	0.1	2.4
CCl_4-Et_2O	3.3	3.0	2.5	2.5
CCl_4-$CHCl_3$	3.8	3.4	2.7	7.3
CCl_4-CH_2Cl_2	5.4	4.9	4.0	7.4
CCl_4-acetone	6.6	6.1	4.9	6.3
CCl_4-EtOH	6.8	6.2	4.5	11.1
CCl_4-TFE[a]	17.4	16.1	16.1	21.6
CCl_4-MeOH	7.0	6.4	4.6	12.4
CCl_4-CH_3CN	7.0	6.4	5.1	8.1
CCl_4-DMSO	7.1	6.5	5.3	6.8
CCl_4-water[a]	13.8	12.3	12.0	18.2
MUE	2.7	3.0	3.7	
		CH_3NO_2		
CCl_4-1,4-dioxane	0.2	0.0	0.0	−5.3
CCl_4-C_6H_6	0.1	0.0	−0.1	−2.7
CCl_4-Et_2O	−1.2	−1.5	−1.4	−3.2
CCl_4-$CHCl_3$	−2.1	−1.7	−2.3	−3.3
CCl_4-CH_2Cl_2	−2.6	−2.4	−2.9	−3.9
CCl_4-acetone	−2.9	−3.0	−3.1	−6.3
CCl_4-EtOH	−5.1	−3.0	−5.1	−4.4
CCl_4-MeOH	−5.8	−3.1	−5.7	−5.1
CCl_4-CH_3CN	−3.3	−3.2	−3.4	−6.9
CCl_4-DMF	−2.9	−3.2	−3.0	−7.8
CCl_4-DMSO	−3.0	−3.2	−3.0	−9.1
CCl_4-H_2O	−6.0	−3.3	−5.9	−9.1
MUE	2.9	3.3	2.8	
		CH_3NCS		
CCl_4-C_6H_{14}	0.6	0.4	0.4	3.2
CCl_4-1,4-dioxane	0.0	0.0	0.0	−2.7
CCl_4-C_6H_6	0.0	0.0	0.0	−1.1
CCl_4-Et_2O	−1.8	−1.2	−1.1	0.7
CCl_4-$CHCl_3$	−2.0	−1.4	−1.2	−2.1
CCl_4-CH_2Cl_2	−2.9	−2.0	−1.7	−2.8
CCl_4-acetone	−3.6	−2.5	−2.1	−2.0
CCl_4-MeOH	−3.8	−2.7	−2.1	−1.6
CCl_4-DMSO	−3.9	−2.7	−2.3	−6.7
MUE	1.7	1.6	1.7	
		CH_3ONO_2		
CCl_4-C_6H_{14}	0.3	0.2	0.2	0.3
CCl_4-1,4-dioxane	0.1	0.0	0.0	−3.0
CCl_4-C_6H_6	0.0	0.0	−0.0	−1.5
CCl_4-Et_2O	−0.6	−0.7	−0.6	−1.5
CCl_4-$CHCl_3$	−1.0	−0.8	−0.9	−1.9
CCl_4-CH_2Cl_2	−1.3	−1.2	−1.1	−2.7
CCl_4-acetone	−1.4	−1.4	−1.5	−3.5
CCl_4-TFE	−2.6	−1.5	−1.0	−3.9
CCl_4-MeOH	−2.6	−1.5	−1.0	−3.5
CCl_4-1,2-ethanediol	−2.7	−1.6	−1.1	−3.9
CCl_4-DMSO	−1.5	−1.6	−1.2	−4.8
MUE	1.5	1.8	2.0	
total MUE	2.3	2.5	2.6	

[a] These results were provided by cluster calculations with the solute and a single solvent molecule optimized with SM8/M06-L/6-31G(d).

test the PCM-MST method in Table 12, which—unlike other tables—shows *errors* in shielding constants, not shielding constants. This approach does much better than other models (except clustered SMD) for the CCl_4-H_2O shift of CH_3CN. Overall, PCM-MST is better than PCM-1.2B for $CC1_4$-H_2O shifts, but it is consistently worse for CCl_4-$CHCl_3$ shifts. Comparing PCM-MST and the SM*x* models on a case by case basis, for the cases in Table 12, excluding the bare solute row, SMD performs the best, and SM8, SM8AD, and PCM-MST perform about equally well. In view of the limited number of solvents for which this test is possible, we will draw our final conclusions based on the mean unsigned errors for the larger and more diverse sets of solvents.

Table 12. Magnitudes of the Errors in Calculated ^{14}N Shielding Constants, Relative to Their Values in CCl_4

	PCM-MST	SM8	SMD	SM8AD	PCM-UA0	PCM-UAKS	PCM-1.2B
			CH_3CN				
CCl_4-$CHCl_3$	3.1	2.9	1.3	4.0	2.3	3.8	1.9
CCl_4-H_2O (bare solute)	2.6	10.2	6.4	12.4	8.6	5.4	7.6
CCl_4-H_2O (cluster)	1.4	4.2	1.1	5.0	5.9	3.1	4.2
			CH_3NO_2				
CCl_4-$CHCl_3$	1.0	0.5	0.0	0.2	1.1	1.6	0.7
CCl_4-H_2O	0.2	0.3	0.0	0.4	4.8	2.3	3.9
			CH_3NCS				
CCl_4-$CHCl_3$	0.5	0.5	1.1	0.3	0.6	0.2	0.2
			CH_3ONO_2				
CCl_4-$CHCl_3$	0.9	0.7	0.4	0.6	0.9	1.2	0.7

6.D.3. Solvent Shifts Relative to CCl_4 - Extended Basis Set Results. In order to test if an increase in the basis set size would give more accurate results for NPE-solvers, we undertook additional calculations for the entire test set at the SMD/M06-L/6-311++G(2df,2p) and SMD/M06-L/111111−31G(d) levels, where the latter is a decontracted version of the 6-31G(d) basis set. Results are presented in Table 13.

Table 13 shows that increasing the basis set size by decontracting core functions does not increase the accuracy of the model tested here. Almost no difference in the mean errors is seen for the methods that used the contracted and decontracted versions of the 6-31G(d) basis set, which illustrates the small effect of core functions in isotropic shielding constants. Indeed, the average change in absolute nuclear shielding upon this decontraction is only 0.3 ppm.

SMD/M06-L/6-311++G(2df,2p) show slight improvements for methyl nitrate and nitromethane, but the opposite is observed for acetonitrile and methyl isothiocyanate, and the total mean unsigned error for this method is 20% larger than when 6-31G(d) is used as the basis set in either its decontracted or standard version. As noted above in section 6.A, we ascribe this decrease in accuracy to the greater charge penetration outside of the solute cavity with the larger basis set. Table 14 lists the errors on total polarization charges (i.e., the amount by which the q_m failed to satisfy eq 7) for each solute in each solvent for the SMD model with the 6-31G(d) and 6-311++G(2df,2p) basis sets; these provide some measure of the amount of outlying charge, and larger errors—up to 50% larger—are found for 6-311++G(2df,2p) in every case. We conclude that the smaller basis set results are likely to be more physical.

7. Additional Discussion

It has been emphasized in previous work[35,103,104] that there is no fundamental way to separate electrostatic and nonelectrostatic contributions to solvation. The long-range effects of bulk solvent can be estimated quite reasonably from the bulk dielectric constant, so the questionable contributions are all localized to the nearby solvent and are dominated by the first solvation shell.[28] Provided, though, that nonelectrostatic terms dependent on the nature and extent of first-solvation-shell effects are optimized to be consistent with a systematic and well-defined scheme for bulk-electrostatic effects, one can obtain good across-the-board accuracy for free energies of solvation in both aqueous and diverse nonaqueous media.[28−34,104] However in SCRF models, only the terms treated self-consistently have an effect on the electronic properties of the solute (for a given solute geometry), and all widely used methods for calculating solvation energies treat only the bulk-electrostatic terms self-consistently.[36,37] Since one can obtain reasonable solvation free energies without carrying out self-consistent calculations even when one underestimates or neglects the electrostatic contributions,[105,106] obtaining accurate solvation free energies does not guarantee accurate response of solute properties to the density,[103,104] and separate tests are necessary. However it is not clear what properties of the solute can provide definitive tests. For example, the partial atomic charges of the solute provide a clear physical picture of solute polarization,[107] but they are not physically observable. The dipole moment of a solute molecule is a physical observable in the gas phase, and it is sensitive to solute response,[103] but it is not precisely defined in solution because there is no unique way to divide electron density between the solute and the solvent. Zhan and Chipman[16] have suggested that the solvent dependence of NMR chemical shielding at nitrogen can be used to evaluate the treatment of solute electron-density response, and we have adopted their suggestion in the present work.

In practical work the extent of predicted solute polarization is primarily controlled by the choice of cavity size and shape, where the solute cavity is defined as the region in which the dielectric constant is unity; the dielectric constant is set equal to the bulk value outside the cavity. In the SM*x* models we have optimized the parameters that control the cavity size and shape primarily to free energies of solvation of ions,[108] both clustered with explicit solvent molecules[109,110] and unclustered; since these quantities are dominated by bulk electrostatics, this provides a realistic way to pin down the partition into bulk electrostatic effects and other effects. The PCM solvation models have used other criteria for cavity parameters.[37,41,44,60−64,101,102] In fact, the multitude of cavity definition protocols in modern continuum solvation models, including some which adjust atomic radii as a function of partial atomic charge (a choice not adopted in the SMx models), attests to the lack of consensus on what is the most realistic way to define the cavities. The present study to confirm the usefulness of the approach adopted in the SMx models is therefore useful.

Table 13. Prediction of Relative Shielding Constants with SMD/M06-L/6-311++G(2df,2p) and SMD/M06-L/111111-31G(d)

	SMD/ 111111−31G(d)	SMD/ 6-311++G(2df,2p)	experiment
	CH_3CN		
CCl_4-C_6H_{12}	−1.0	−2.2	−1.4
CCl_4-1,4-dioxane	−0.1	−0.1	5.0
CCl_4-C_6H_6	0.1	0.1	2.4
CCl_4-Et_2O	5.3	6.6	2.5
CCl_4-$CHCl_3$	6.0	7.5	7.3
CCl_4-CH_2Cl_2	8.6	10.9	7.4
CCl_4-acetone	10.6	13.4	6.3
CCl_4-EtOH	10.9	13.7	11.1
CCl_4-TFE[a]	18.3	19.0	21.6
CCl_4-MeOH	11.2	14.1	12.4
CCl_4-CH_3CN	11.3	14.3	8.1
CCl_4-DMSO	11.5	14.5	6.8
CCl_4-water[a]	15.7	16.7	18.2
MUE	2.5	3.5	
MUE (6-31G(d))	2.4		
	CH_3NO_2		
CCl_4-1,4-dioxane	0.3	0.3	−5.3
CCl_4-C_6H_6	0.2	0.3	−2.7
CCl_4-Et_2O	−2.0	−2.4	−3.2
CCl_4-$CHCl_3$	−3.3	−3.8	−3.3
CCl_4-CH_2Cl_2	−4.3	−5.0	−3.9
CCl_4-acetone	−4.8	−5.6	−6.3
CCl_4-EtOH	−7.9	−8.8	−4.4
CCl_4-MeOH	−8.7	−9.5	−5.1
CCl_4-CH_3CN	−5.3	−6.2	−6.9
CCl_4-DMF	−4.8	−5.7	−7.8
CCl_4-DMSO	−4.9	−5.8	−9.1
CCl_4-H_2O	−9.0	−9.9	−9.1
MUE	2.3	2.3	
MUE(6-31G(d))	2.3		
	CH_3NCS		
CCl_4-C_6H_{14}	0.9	1.0	3.2
CCl_4-1,4-dioxane	0.0	0.0	−2.7
CCl_4-C_6H_6	0.0	0.0	−1.1
CCl_4-Et_2O	−2.8	−3.3	0.7
CCl_4-$CHCl_3$	−3.2	−3.7	−2.1
CCl_4-CH_2Cl_2	−4.7	−5.4	−2.8
CCl_4-acetone	−5.8	−6.7	−2.0
CCl_4-MeOH	−6.1	−7.0	−1.6
CCl_4-DMSO	−6.2	−7.2	−6.7
MUE	2.4	2.8	
MUE(6-31G(d))	2.4		
	CH_3ONO_2		
CCl_4-C_6H_{14}	0.4	0.5	0.3
CCl_4-1,4-dioxane	0.1	0.1	−3.0
CCl_4-C_6H_6	0.1	0.1	−1.5
CCl_4-Et_2O	−1.0	−1.2	−1.5
CCl_4-$CHCl_3$	−1.6	−1.8	−1.9
CCl_4-CH_2Cl_2	−2.0	−2.4	−2.7
CCl_4-acetone	−2.2	−2.7	−3.5
CCl_4-TFE	−3.8	−4.2	−3.9
CCl_4-MeOH	−3.8	−4.3	−3.5
CCl_4-1,2-ethanediol	−3.9	−4.3	−3.9
CCl_4-DMSO	−2.3	−2.7	−4.8
MUE	1.0	0.8	
MUE (6-31G(d))	1.0		
total MUE	2.0	2.4	
total MUE(6-31G(d))	2.0		

[a] These results were provided by cluster calculations with the solute and a single solvent molecule optimized with SM8/M06-L/6-31G(d).

Table 14. Error on Total Polarization Charges (in au) Given by the Methods SMD/M06-L/6-31G(d) and SMD/M06-L/6-311++G(2df,2p)

	SMD/6-31G(d)	SMD/6-311++G(2df,2p)
	CH_3CN	
C_6H_{12}	0.008	0.010
1,4-dioxane	0.009	0.011
C_6H_6	0.009	0.011
CCl_4	0.009	0.011
Et_2O	0.012	0.015
$CHCl_3$	0.012	0.016
CH_2Cl_2	0.014	0.018
acetone	0.014	0.019
EtOH	0.015	0.019
TFE	0.015	0.019
MeOH	0.015	0.019
CH_3CN	0.015	0.019
DMSO	0.015	0.019
H_2O	0.015	0.020
average	0.012	0.016
	CH_3NO_2	
1,4-dioxane	0.004	0.005
C_6H_6	0.004	0.005
CCl_4	0.004	0.005
Et_2O	0.006	0.007
$CHCl_3$	0.007	0.009
CH_2Cl_2	0.007	0.009
acetone	0.007	0.009
EtOH	0.015	0.022
MeOH	0.020	0.028
CH_3CN	0.007	0.009
DMF	0.007	0.008
DMSO	0.007	0.008
H_2O	0.020	0.029
average	0.009	0.012
	CH_3NCS	
C_6H_{14}	0.008	0.009
1,4-dioxane	0.007	0.008
C_6H_6	0.008	0.009
CCl_4	0.008	0.009
Et_2O	0.011	0.013
$CHCl_3$	0.011	0.013
CH_2Cl_2	0.013	0.015
acetone	0.014	0.016
MeOH	0.014	0.016
DMSO	0.014	0.014
average	0.011	0.012
	CH_3ONO_2	
CCl_4	0.004	0.005
C_6H_{14}	0.003	0.004
1,4-dioxane	0.004	0.005
C_6H_6	0.004	0.005
Et_2O	0.005	0.006
$CHCl_3$	0.007	0.009
CH_2Cl_2	0.007	0.009
acetone	0.007	0.008
TFE	0.021	0.029
MeOH	0.021	0.029
1,2-ethanediol	0.022	0.029
DMSO	0.007	0.008
average	0.009	0.012
total average	0.010	0.013

Hydrogen bonding between the solute and solvent is the classic example of a solute−solvent interaction with atomic-scale character that cannot be understood entirely in terms of bulk solvent properties. And yet even hydrogen bonding is often dominated by electrostatics.[111] One particularly informative study of the solvent effect on an NMR nuclear shielding of a hydrogen-bonded system is the study of the ^{17}O chemical shielding in acetone by Aidas et al.[112] They found that the gas-to-water solvent shift is underestimated by 22 pm (30%) by the PCM-UA0 model. They pointed out that it had been suggested[17,21,113] that larger cavities might be employed in the PCM model to improve the calculated shieldings but that larger cavities would make the agreement

worse. To obtain a more accurate value, they instead employed a procedure combining several elements. First they sampled 700 configurations of acetone in water in a molecular dynamics simulation with explicit water molecules. For each configuration they then carried out an SCRF calculation on a supersolute consisting of acetone and the two closest water molecules, which yielded widely distributed values with a standard deviation from the mean of 13 ppm. Averaging over these values yields a shift of 76.5 ppm, which happens to be within 1.1 ppm of experiment. In the standard approach one attempts to find a cavity definition where a calculation at a single representative solute geometry (and often with only implicit water) yields the average result directly. Here we use one explicit water when there is hydrogen bonding of the solvent to the atom at which the chemical shift is measured, but nevertheless the large spread of values found for instantaneous solvent configurations in the work of Aidas et al. shows why it is hard to capture the effect of a specific hydrogen bonding interaction with a calculation for a single solute−solvent configuration.

8. Conclusions

Several combinations of continuum solvation models and density functionals were assessed for their ability to predict the solvent dependence of 49 ^{14}N chemical shifts in four different solutes and seventeen different solvents. We conclude the following from our results:

(1) The continuum solvent models do not give particularly reliable predictions of the solvent shifts; errors range from about 1.7 to 3.0 ppm depending on the model. In part this reflects the importance of specific solute−solvent interactions, but if we nevertheless ask which existing continuum solvation model is best (smallest mean unsigned errors) for computing the solvent dependence of ^{14}N chemical shifts, we find that the generalized Born models SM8 and SM8AD, used with their standard values for the intrinsic Coulomb radii, perform the best-despite the fact that these models were optimized for the computation of molecular free energies of solvation. SMD and PCM-1.2B, which determine the reaction field from approximate solution of the nonhomogeneous Poisson equation, also give better results when their standard radii are used. Although earlier studies[21,22] scaled standard radii by factors greater than one, and the fitting approach presented in section 6.C suggested that this could be a useful way of improving the solvent shifts predicted by NPE-based models, this approach did not improve the accuracy of solvent shifts computed relative to carbon tetrachloride.

(2) Including a specific solvent molecule in a solute−solvent cluster improves predicted ^{14}N chemical shifts when a strong hydrogen bond is made from the solvent molecule to the nitrogen atom of the solute.

(3) All SM*x* models correctly predict the positive or negative dielectric dependence of the solvent shifts in all four solutes studied here, CH_3CN, CH_3NO_2, CH_3NCS, and CH_3ONO_2, in contrast to previous findings with the SVPE and SPE methods.

(4) Increasing the basis set size for PCM-based models does not improve the calculation of NMR chemical shifts in solution, as larger basis sets including diffuse functions lead to larger quantities of outlying charge. In PCM-based models, such outlying charge contributes to unphysical effects on computed nuclear shieldings.

(5) The SM8 and SM8AD solvation models offered the most quantitatively accurate results compared to experiment when solvent-to-solvent shifts were examined. Most calculations were carried out with the M06-L density functional, but substitution of B3LYP or PBE0 for M06-L gave slightly improved accuracy in the case of SM8AD. Solvation models such as PCM that are based on a reaction field determined from an approximate solution of the nonhomogeneous Poisson equation showed somewhat reduced accuracy. The success of the SM*x* models is encouraging because the SM*x* models also provide excellent accuracy for free energies of solvation,[31−34] much better for example than does the default version of PCM in *Gaussian03*.

It thus appears that the Coulomb radii normally used to predict free energies of solvation in the SM*x* models are also the most appropriate ones for predicting NMR shielding constants in solution and that those solvation models that do best for free energies of solvation are also reasonable choices for property calculations. It will be interesting in the future to identify other well-defined solvent response properties that will permit this question to be assessed in more detail.

Acknowledgment. The authors are grateful to Roberto Cammi and Daniel Chipman for many stimulating discussions. This work was supported in part by the National Science Foundation under grant nos. CHE06-10183 and CHE07-04974.

Supporting Information Available: Plots of the dependence on dielectric constant of the relative nitrogen shielding calculated by the scaled radii SMx models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Pople, J. A.; Schneider, W. G.; Bernstein, H. J. *High Resolution Nuclear Magnetic Resonance*. McGraw-Hill: New York, 1959.

(2) Emsley, J. W.; Feeney, J.; Sutcliffe, L. H. *High Resolution Nuclear Magnetic Resonance Spectroscopy*; Pergamon Press: 1965; Vol. 1.

(3) Ando, I.; Webb, G. A. *Theory of NMR Parameters*; Academic Press: New York, 1983.

(4) Buckingham, A. D.; Schaefer, T.; Schneider, W. G. *J. Chem. Phys.* **1960**, *32*, 1227.

(5) Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877.

(6) Begtrup, M.; Taft, R. W.; Kamlet, M. J. *J. Org. Chem.* **1986**, *51*, 2130.

(7) Witanowski, M.; Sitkowski, J.; Biernat, S.; Kamienski, B.; Hamdi, B. T.; Webb, G. A. *Magn. Reson. Chem.* **1985**, *23*, 748.

(8) Witanowski, M.; Sitkowski, J.; Biernat, S.; Sudha, L. V.; Webb, G. A. *Magn. Reson. Chem.* **1987**, *25*, 725.

(9) Witanowski, M.; Sicinska, W.; Webb, G. A. *Magn. Reson. Chem.* **1989**, *27*, 380.

(10) Witanowski, M.; Stefaniak, L.; Webb, G. A. In *Annual Reports on NMR spectroscopy*; Webb, G. A., Ed.; Academic Press: London, 1993, Vol. 25, p. 86.

(11) Witanowski, M.; Sicinska, W.; Biedrzycka, Z.; Grabowski, Z.; Webb, G. A. *J. Magn. Reson. A* **1995**, *112*, 66.

(12) Witanowski, M.; Biedrzycka, Z.; Sicinska, W.; Grabowski, Z. *J. Magn. Reson.* **2003**, *164*, 212.

(13) Alkorta, I.; Elguero, J. *Struct. Chem.* **2003**, *14*, 377.

(14) Bagno, A.; Rastrelli, F.; Saielli, G. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *47*, 41.

(15) Cammi, R. *J. Chem. Phys.* **1998**, *109*, 3185.

(16) Zhan, C.-G.; Chipman, D. M. *J. Chem. Phys.* **1999**, *110*, 1611.

(17) Cammi, R.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1999**, *110*, 7627.

(18) Jaszuński, M.; Mikkelsen, K. V.; Rizzo, A.; Witanowski, M. *J. Phys Chem. A* **2000**, *104*, 1466.

(19) Manalo, M. N.; de Dios, A. C.; Cammi, R. *J. Phys. Chem. A* **2000**, *104*, 9600.

(20) Ksiazek, A.; Borowski, P.; Wolinski, K. *J. Magn. Reson.* **2009**, *197*, 153.

(21) Mennucci, B.; Martinez, J. M.; Tomasi, J. *J. Phys. Chem. A* **2001**, *105*, 7287.

(22) Mennucci, B. *J. Am. Chem. Soc.* **2002**, *124*, 1506.

(23) Fazaeli, R.; Monajjemi, M.; Ataherian, F.; Zare, K. *THEOCHEM* **2002**, *581*, 51.

(24) Manalo, M. N.; de Dios, A. C. *Magn. Reson. Chem.* **2002**, *40*, 781.

(25) Chipman, D. M. *Theor. Chem. Acc.* **2004**, *111*, 61.

(26) Bulo, R. E.; Jacob, C. R.; Visscher, L. *J. Phys. Chem. A* **2008**, *112*, 2640.

(27) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.

(28) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213.

(29) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1996**, *61*, 8720.

(30) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(31) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.

(32) Cramer, C. J.; Truhlar, D. G. In *Trends and Perspectives in Modern Computational Science, Lecture Series on Computer and Computational Sciences Volume 6*; edited by Maroulis, G., Simos,T. E., Ed.; Brill/VSP: Leiden, 2006; pp 112−139. Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760.

(33) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378.

(34) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.,* in press.

(35) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 877.

(36) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(37) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(38) Miertuŝ, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(39) Cancés, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(40) Tomasi, J.; Mennucci, B.; Cancès, E. *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211.

(41) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(42) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43.

(43) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.

(44) Barone, V.; Improta, R.; Rega, N. *Theor. Chem. Acc.* **2004**, *111*, 237.

(45) 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.; and Pople, J. A. *Gaussian03, revision E.01*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(46) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. M.; 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.

(47) *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Politzer, P., Truhlar, D. G., Eds.; Plenum: New York, 1981.

(48) 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.

(49) Hoijtink, G. J.; de Boer, E.; van der Meij, P. H.; Weijland, W. P. *Recl. Trav. Chim. Pays-Bas Belg.* **1956**, *75*, 487.

(50) Peradejordi, F. *Cah. Phys.* **1963**, *17*, 393.

(51) Jano, I. *C. R. Acad. Sci.(Paris)* **1965**, *261*, 103.

(52) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(53) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(54) Cramer, C. J.; Truhlar, D. G. *Rev. Comp. Chem.* **1995**, *6*, 1.

(55) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1995**, *16*, 422.

(56) Grycuk, T. *J. Chem. Phys.* **2003**, *119*, 4817.

(57) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87.

(58) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2046.

(59) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(60) Bachs, M.; Luque, F. J.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 446.

(61) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(62) Mennucci, B.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1995**, *102*, 6837.

(63) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074.

(64) Klamt, A.; Eckert, F. *Fluid Phase Equilib.* **2000**, *172*, 43.

(65) Chipman, D. M. *J. Chem. Phys.* **1997**, *106*, 10194.

(66) Zhan, C.-G.; Bentley, J.; Chipman, D. M. *J. Chem. Phys.* **1998**, *108*, 177.

(67) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293.

(68) Jameson, C. J.; de Dios, A. C. *Nucl. Magn. Reson.* **2003**, *32*, 43.

(69) Facelli, J. C. *Concepts Magn. Reson.* **2004**, *20A*, 42.

(70) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 10754.

(71) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789.

(72) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(73) Kutzelnigg, W. *Isr. J. Chem.* **1980**, *19*, 193.

(74) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, *210*, 223.

(75) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Chem. Phys. Lett.* **1994**, *220*, 299.

(76) Ligabue, A.; Sauer, S. P. A.; Lazzeretti, P. *J. Chem. Phys.* **2007**, *126*, 154111.

(77) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module, version 4.1*; University of Minnesota: Minneapolis, MN, 2008.

(78) *MN-GSM-v2009: Minnesota Gaussian Solvation Module, version 2009*; University of Minnesota: Minneapolis, MN, 2009.

(79) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(80) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6794.

(81) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497.

(82) Bachrach, S. M. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: 1994; Vol. 5. pp 171−228.

(83) Chesnut, D. B.; Rusiloski, B. E. *J. Mol. Struct. THEOCHEM* **1994**, *314*, 19. Schreckenbach, G.; Ziegler, T. *J. Phys. Chem.* **1995**, *99*, 606.

(84) Astrand, P.-O.; Mikkelsen, K. V.; Jorgensen, P.; Ruud, K.; Helgaker, T. *J. Chem. Phys.* **1998**, *108*, 2528.

(85) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.

(86) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.

(87) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982**, *104*, 2797.

(88) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(89) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(90) Davidson, E. R.; Feller, D. *Chem. Rev.* **1986**, *86*, 681.

(91) Jensen, F. *J. Chem. Theory Comput.* **2008**, *4*, 719.

(92) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571.

(93) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045.

(94) Winget, P.; Dolney, D. M.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvent Descriptor Database. http://comp.chem.umn.edu/solvation/mnsddb.pdf (accessed May 1, 2009).

(95) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(96) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(97) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(98) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(99) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1998**, *298*, 113.

(100) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. *Acc. Chem. Res.* **2009**, *42*, 489.

(101) Luque, F. J.; Zhang, Y.; Aleman, C.; Bachs, M.; Gao, J.; Orozco, M. *J. Phys. Chem.* **1996**, *100*, 4269.

(102) Luque, F. J.; Bachs, M.; Aleman, C.; Orozco, M. *J. Comput. Chem.* **1996**, *17*, 806.

(103) Curutchet, C.; Cramer, C. J.; Truhlar, D. G.; Ruiz-Lopez, M. F.; Rinaldi, D.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2003**, *24*, 284.

(104) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2009**, *42*, 493.

(105) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1997**, *101*, 7147.

(106) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257.

(107) Marenich, A. V.; Olson, R. M.; Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2055.

(108) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 408.

(109) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 2493.

(110) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066.

(111) Lommerse, J. P. M.; Price, S. L.; Taylor, R. *J. Comput. Chem.* **1997**, *18*, 757. Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.

(112) Aidas, K.; Møgelhøj, A.; Kjaer, H.; Nielsen, C. B.; Mikkelsen, K. V. *J. Phys. Chem. A* **2007**, *111*, 4199.

(113) Cossi, M.; Crescenzi, O. *J. Chem. Phys.* **2003**, *118*, 8863.

CT900258F

JCTC Journal of Chemical Theory and Computation

Through-Space Effects of Substituents Dominate Molecular Electrostatic Potentials of Substituted Arenes

Steven E. Wheeler* and K. N. Houk*

Department of Chemistry and Biochemistry, University of California, 405 Hilgard Avenue, Los Angeles, California 90095

Received July 8, 2009

Abstract: Model systems have been studied using density functional theory to assess the contributions of π-resonance and through-space effects on electrostatic potentials (ESPs) of substituted arenes. The results contradict the widespread assumption that changes in molecular ESPs reflect only local changes in the electron density. Substituent effects on the ESP above the molecular plane are commonly attributed to changes in the aryl π-system. We show that ESP changes for a collection of substituted benzenes and more complex aromatic systems can be accounted for mostly by through-space effects, with no change in the aryl π-electron density. Only when π-resonance effects are substantial do they influence changes to any extent in the ESP above the aromatic ring. Examples of substituted arenes studied here are taken from the fields of drug design, host−guest chemistry, and crystal engineering. These findings emphasize the potential pitfalls of assuming ESP changes reflect changes in the local electron density. Since ESP changes are frequently used to rationalize and predict intermolecular interactions, these findings have profound implications for our understanding of substituent effects in countless areas of chemistry and molecular biology. Specifically, in many noncovalent interactions there are significant, often neglected, through-space interactions with the substituents. Finally, the present results explain the good performance of many molecular mechanics force-fields when applied to supramolecular assembly phenomena, despite the neglect of the polarization of the aryl π-system by substituents.

I. Introduction

Molecular electrostatic potentials (ESPs) have emerged as powerful predictive and interpretive tools in disparate areas of chemistry, rational drug design, and molecular biology.[1,2] Colorful plots of ESPs have been used to rationalize trends in organic reactivity[3] and binding in host−guest complexes and noncovalent interactions (cation/π, $\pi-\pi$, etc.).[4−9] Information gleaned from ESPs is often utilized in analyses of protein−ligand and protein−protein interactions[10] as well as studies of the folding of model proteins.[11] Quantitative ESP-based reactivity descriptors have also emerged, offering alternatives to traditional substituent constants.[12] Computed ESPs have been correlated with impact sensitivities of explosive compounds[13] and toxicity in polychlorinated biphenyls (PCBs).[14] Moreover, ESPs can be readily computed using standard electronic structure theory packages or derived from experimental X-ray diffraction data,[15] providing a simple, easily accessible tool for understanding numerous phenomena.

The ESP at a given point near a molecule is a measure of the electrostatic energy a positive unit test charge would experience at that point. Negative ESPs correspond to an attractive interaction with this test charge, while positive ESPs indicate repulsion. Nonuniform electrostatic potentials arise in molecular environments from the competing effects of the nuclear charges and the surrounding electrons. The use of ESPs to predict and rationalize reactivity trends was pioneered by Scrocco, Tomasi, and co-workers,[16] who studied electrophilic attack of three-membered rings and nucleic acid bases and proton affinities of amides. Subsequently, ESP plots have been

* Corresponding authors. E-mail: swheele2@chem.ucla.edu (S.W.). E-mail: houk@chem.ucla.edu.

10.1021/ct900344g CCC: $40.75
Published on Web 07/28/2009

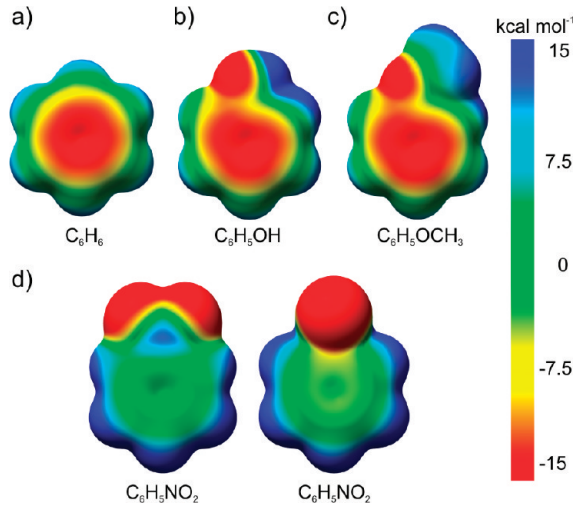


Figure 1. Plots of the electrostatic potential of (a) benzene, (b) phenol, (c) anisole, and (d) planar nitrobenzene (left) and perpendicular nitrobenzene (right) mapped onto electron density isosurfaces (0.001 e/au^3).

applied to sundry chemical systems driven, in large part, by the many reports of Politzer and Murray.[2,3] ESP plots also enjoy wide applicability in the analysis of noncovalent interactions. They have been used in the development of conceptual models of the electrostatic component of prototypical interactions, and they have provided a simple means of approximating the interaction strength and geometry of noncovalent complexes.[4−9] ESPs have been applied to noncovalent interactions of increasingly complex systems, culminating in studies of protein−ligand and protein−protein interactions.[4]

Unfortunately, the literature is peppered with false assumptions regarding the effect of substituents on ESPs. It is common to equate changes in the ESP in a given region with *local* changes in the electron density. For example, in a recent study of noncovalent interactions in mechanically interlocked compounds, Goddard, Stoddart, and co-workers[17] used ESP plots to evaluate electron density differences of the central naphthalene core for a series of substituted systems. Indeed, when presenting ESP plots, many authors explicitly label regions of negative electrostatic potential as "electron-rich" and positive ESP regions as "electron-poor".[18] This connection between ESP values and the local electron density is also advanced in otherwise stellar textbooks on physical organic chemistry[19] and in publications advocating the use of ESP plots in undergraduate education.[20] Politzer and Murray[3] emphasized nearly two decades ago that the ESP is the "net result at a given point of the integrated effects of all of the electrons and nuclei, whereas $\rho(\mathbf{r})$ of course represents only the electronic density at that point." While negative ESPs often do correspond to electron-rich regions, the assumption that changes in ESPs necessarily indicate local changes in the electron density is incorrect.

The gas-phase ESP at a given point, $V(\mathbf{r})$, is defined by eq 1, where Z_A and $\mathbf{R}_A$ are the charge and position of nucleus A, respectively, and $\rho(\mathbf{r}')$ is the electron density at position $\mathbf{r}'$, all in atomic units:

$$V(\mathbf{r}) = \sum_{A}^{\text{nuclei}} \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \qquad (1)$$

The integral in eq 1 runs over all space. Thus, at a given point, the ESP is dependent on the electron density in all surrounding space, though this dependency dies off with distance. Despite this 1/r dependence, seemingly subtle changes in the electron density can have profound effects on the ESP at a given point several angstroms away. For example, a charge of 0.1 *e* contributes more than 10 kcal mol^{-1} to the ESP at a distance of 3 Å.

Substituent effects on aromatic rings have been studied extensively since the pioneering work of Hammett.[21] Generally, the effects of a substituent on an aryl ring are transmitted via numerous potential mechanisms, which are often conceptually divided into π-resonance, inductive (through-σ-bond), and field (through-space) effects.[22] There have been numerous attempts to quantify these often competing effects, leading to the development of a bevy of substituent constants.[23] Among the most popular separations of π- and σ-effects come from the works of Roberts and Moreland,[24] Taft,[25] and Swain and Lupton,[26] who partitioned substituent constants into resonance effects (as quantified by σ_R or R) and inductive/field effects (σ_I or F). In these schemes, more negative numbers indicate stronger electron donating tendency, while more positive values correspond to stronger electron acceptors. Substituent effects arise from some combination of resonance and inductive/field effects, with the relative contribution varying with the substituent. ESP maps of substituted benzenes should similarly reflect both π-resonance and inductive/field effects.

The potentially large contribution of through-space substituent effects is mostly absent in discussions of arene ESPs in the modern literature. Many authors assume changes in arene ESPs reflect donation into or out of the aryl π-system. For example, in their extensive studies of substituent effects on the benzene dimer,[8,9,27−29] Sherrill and co-workers utilized ESP plots to characterize the "degree of π-density" in substituted benzenes. This idea is appealing since it enables simple resonance-based explanations of trends in electrostatic potential plots. That is, the ESP above benzene substituted with an electron-withdrawing group will generally be more positive than that of benzene, rationalized based on resonance forms with a formal positive charge at the *ortho* and *para* positions. More negative ESPs above benzenes substituted with electron donors are often explained by resonance forms with a negative charge on the benzene ring.

Two subtle exceptions to such π-resonance-based explanations of ESPs are phenol and anisole. Based on Hammett substituent constants, OH and OMe are π-electron-donating substituents [R(OCH$_3$) = −0.56, R(OH) = −0.70],[23] so the ESPs above the aryl ring in these two systems should be more negative than in benzene, according to the simple π-resonance picture. However, the ESPs above phenol and anisole are slightly more positive than that of benzene (see Figure 1a−c). Analogously, the interactions of Na^+ with C_6H_5OH and $C_6H_5OCH_3$ are slightly weaker than the $Na^+\cdots C_6H_6$ interaction.[5,30,31] This seemingly counterintuitive phenomenon has been mentioned previously,[27,30,32] most

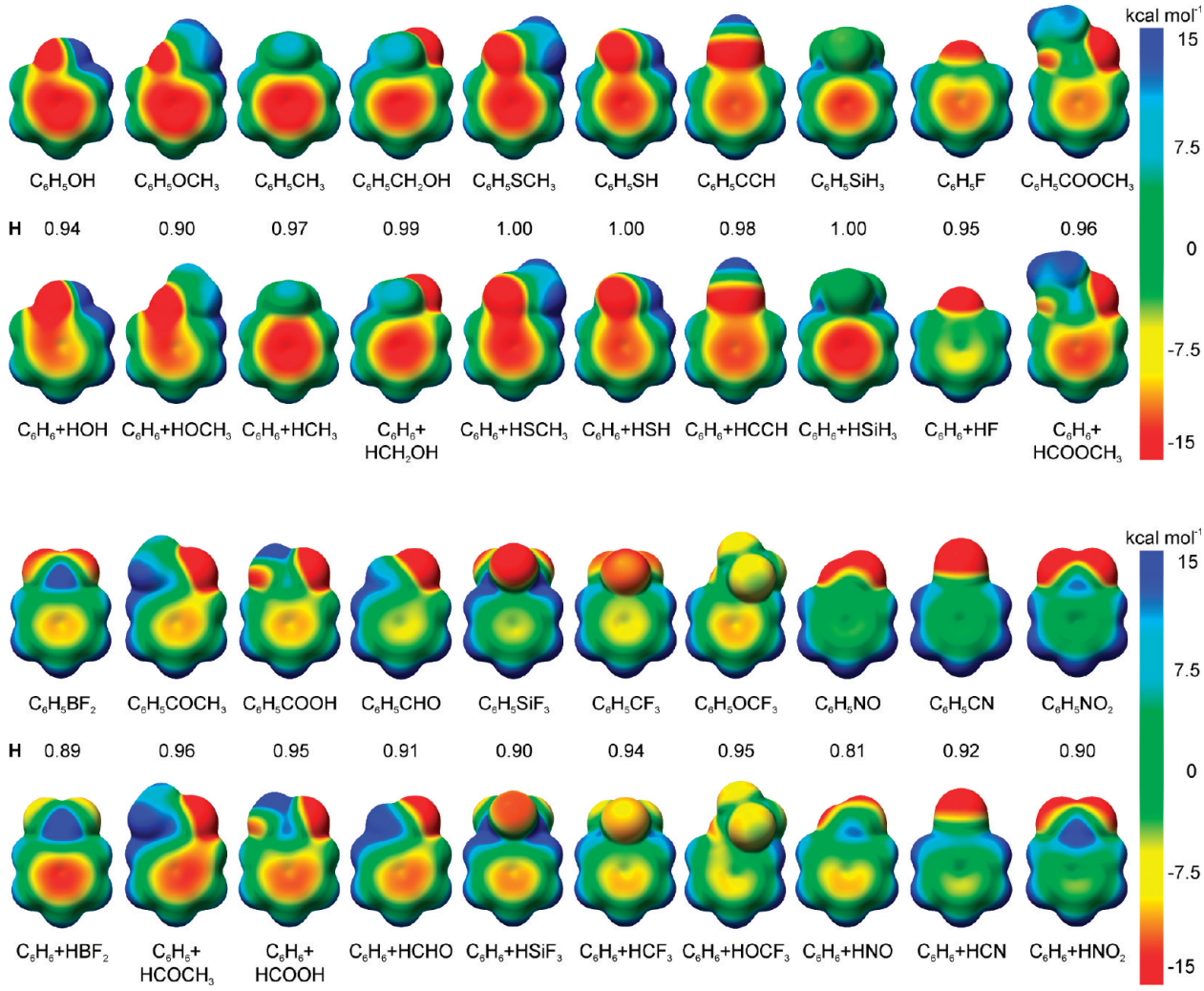


Figure 2. Plots of electrostatic potentials of monosubstituted benzenes (first and third row) and corresponding additive ESPs (second and fourth row). ESPs are mapped on electron density isosurfaces (0.001 e/au^3) for the substituted benzene. The *H* index is computed for the ESP values on the isodensity surfaces for the true and additive ESPs.

prominently by Dougherty and co-workers,[5,30] who noted that the strength of cation$-\pi$ interactions and the ESP above the center of substituted benzenes are more strongly correlated with σ_m constants than σ_p. Hunter and co-workers[7] also reported a strong correlation between σ_m and the ESP at the centroid of substituted aromatic rings. This σ_m dependence is contrary to the prevalent π-resonance-based explanations of ESPs of substituted arenes, since σ_m constants reflect mostly nonresonance effects. In the case of phenol and anisole, the σ-withdrawing effect on the ESP overwhelms the π-donation, as noted by Klärner and co-workers.[32]

A more clear-cut example of the dominant role of inductive/field effects on ESPs was provided by Politzer and co-workers in 1987.[33] The ESP map of nitrobenzene is positive everywhere above the aryl plane, a feature which might naively be attributed to π-electron withdrawal. However, Politzer et al.[33] showed that the ESP of nitrobenzene is essentially unchanged upon a 90° rotation of the nitro group (see Figure 1d). In perpendicular nitrobenzene, there can be no π-resonance between the NO_2 and the aryl π-system, yet the ESP is still positive everywhere above the benzene plane. Clearly, in nitrobenzene, π-resonance has little net effect on the ESP; substituent effects on the ESP must arise from inductive/field effects. This finding is consistent with the typical characterization of NO_2 as a strong inductive electron-withdrawing group but modest resonance acceptor ($F = 0.65$, $R = 0.13$).[23]

Substituent effects on the electrostatic properties of aromatic systems are central to many areas of modern chemistry and molecular biology, and maps of molecular electrostatic potentials constitute a powerful and popular tool for rationalizing and predicting noncovalent interactions. Consequently, a full understanding of chemical and biochemical systems must rest on a sound understanding of the changes in ESPs induced by substituents. Previously, we demonstrated[31] that substituent effects on the ESP at a point approximately 2.4 Å above the center of monosubstituted benzenes arise primarily from direct through-space effects of the substituents, and π-resonance effects play a relatively minor role. We now show that, for a wide range of substituents, changes in ESP maps of substituted aromatic systems are generally dominated by through-space effects of the substituents. More importantly, the widespread, often implicit, assumption that changes in ESPs necessarily indicate *local* changes in electron density is shown to be unfounded and, in many cases, misleading.

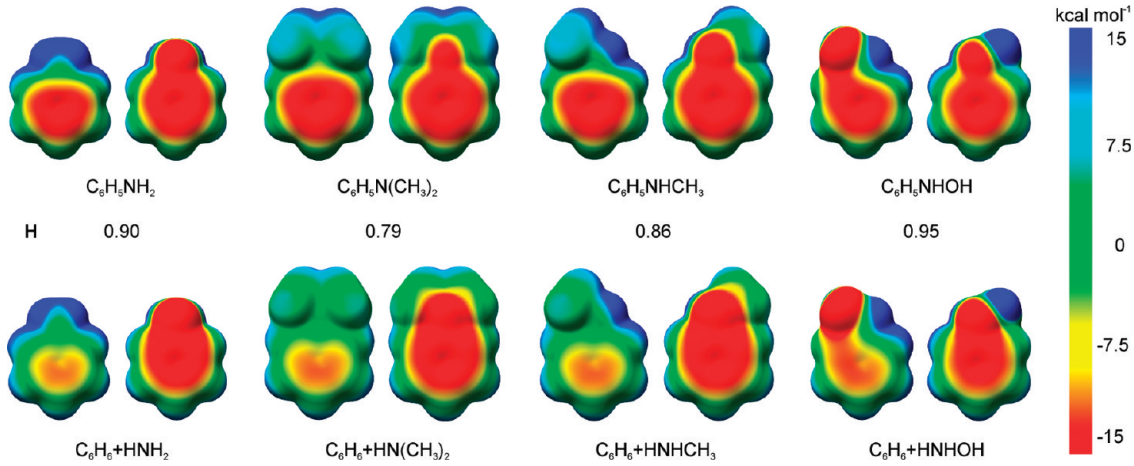


Figure 3. Front and back views of electrostatic potentials of aniline derivatives (top row) and corresponding additive ESPs (bottom row). ESPs are mapped on electron density isosurfaces (0.001 e/au^3) for the substituted benzene. The *H* index is computed for the ESP values on the isodensity surfaces for the true and additive ESPs.

II. Theoretical Methods

The molecular elelectrostatic potential, $V(\mathbf{r})$, was evaluated on a rectangular grid enveloping each molecule according to eq 1 and using electron densities computed at the B3LYP/6-31G(d) level of theory[34] with Gaussian03.[35] These ESP plots are relatively insensitive to the method and basis set employed, as demonstrated for cyanobenzene in Supporting Information (see Figure S1). Graphical representations of these ESPs were generated by mapping the ESP onto a molecular surface corresponding to an isodensity contour at $\rho = 0.005$ or 0.001 e/au^3 using UCSF Chimera.[36] The wide range of systems considered necessitated the use of several different scales for plotted ESPs. The scales utilized are displayed in each figure, and within a given figure, the ESP scale is always the same to facilitate straightforward comparisons of different systems.

An additive ESP model was employed to differentiate between π-resonance and inductive/field effects, constructed as follows for monosubstituted benzenes: for each point on an identical rectangular grid, the ESP was evaluated for C_6H_6, C_6H_5X, and HX, with each system positioned so that conserved atoms were placed identically. For example, for fluorobenzene, all six carbons and the five unsubstituted hydrogens have the same Cartesian coordinates in C_6H_6 as C_6H_5F. The fluorines in C_6H_5F and HF also have identical coordinates. The positions of the substituted atoms were optimized. These constraints result in no discernible difference in ESP plots, as demonstrated in Figure S2 of the Supporting Information. The ESPs of C_6H_6 and HX were then added at each point on this grid, and the resulting additive ESP was mapped onto the electron density isosurface of C_6H_5X.[37] To provide a quantitative measure of the similarity of the additive and true ESPs, the Hodgkin index,[38] *H*, has been computed for the ESP values on the plotted isodensity surfaces. The Hodgkin index for two sets of ESP values ranges from -1.00, for two equal but opposite sets of ESP values, to 1.00 for identical ESPs.

III. Plots of Molecular Electrostatic Potentials

A. Substituted Benzenes. Monosubstituted benzenes serve as models for more complex substituted aromatic systems, and without understanding the effect of substituents in these paradigmatic systems, there is little hope for a sound analysis of more complex substituted arenes. Standard ESP plots are provided for 20 substituted benzenes in Figure 2 (first and third rows). ESP plots for four aniline derivatives are shown in the top row of Figure 3. These ESP plots show the expected qualitative trends: electron-withdrawing substituents generally increase the ESP above the aryl ring, while donors lead to a decrease in the ESP, relative to benzene. These ESP maps reflect both π-resonance and inductive/field effects, the relative contribution of which cannot be discerned from these plots alone.

To assess the role of nonresonance effects, ESPs from an additive model for each of these species are also plotted in Figure 2 (second and fourth row) and the bottom row of Figure 3 and constructed as described in Section II. This primitive model should approximate the polarization of the C−X σ-bond as well as the direct through-space effects of the substituents in the substituted benzene. More importantly, these additive ESPs reflect substituent effects on the ESP *not* due to changes in the aryl π-system, since the π-electron-density is that of unsubstituted benzene.

The similarities of the additive ESPs to the true ESPs in Figure 2 are striking. Computed Hodgkin indices[38] further underscore the similarity of the additive and true ESPs, with most values of *H* exceeding 0.95. For all substituents, the plots of the intact substituted benzenes are qualitatively similar to those derived from this simple additive model. Indeed, for several substituents, the plots are indistinguishable (see, for example, phenylmethanol, thioanisole, benzenethiol, ethynylbenzene, and phenylsilane). For these systems in particular, changes in the ESP relative to that of benzene arise entirely from through-space effects. π-resonance cannot possibly play an appreciable role, since, in the additive ESPs,

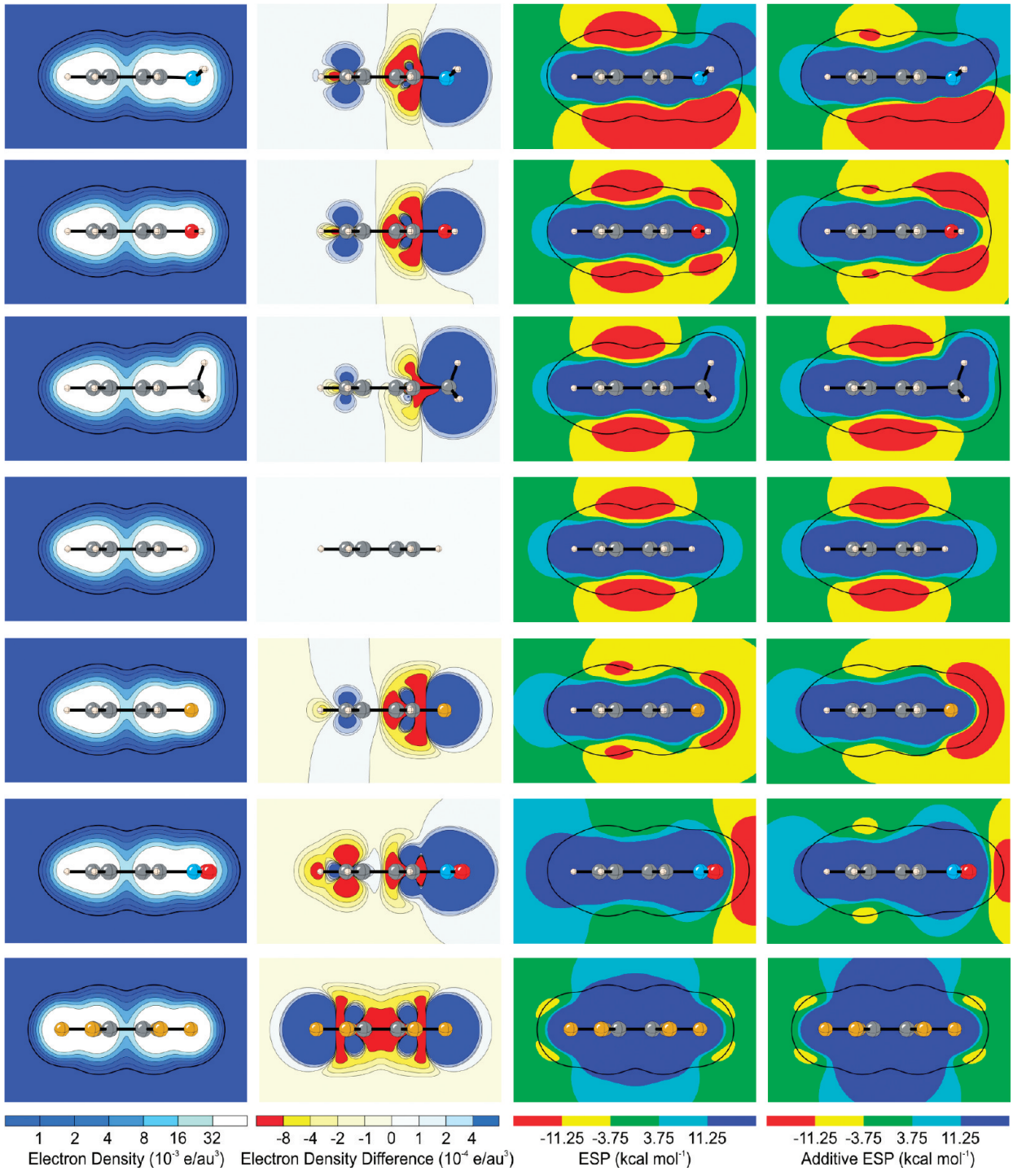


Figure 4. Contour plots of the electron density, electron density difference versus benzene [$\Delta\rho = \rho(C_6H_5X) - \rho(C_6H_6)$], electrostatic potential, and additive ESP for aniline, phenol, toluene, benzene, fluorobenzene, nitrobenzene, and hexafluorobenzene. The thick black line in the density and ESP plots denotes the electron density contour (0.001 e/au^3) used to construct the isodensity surfaces in Figures 2 and 3.

the aryl π-system is identical to that of the unsubstituted benzene by construction.

There is some deviation between the additive and the true ESPs for several of the systems in Figure 2. These deviations occur primarily for strong electron-donating or -accepting substituents (OCH_3, BF_2, SiF_3, NO, and NO_2), suggesting that changes in the aryl π-system influence the molecular ESPs of these substituted systems. Similar behavior was observed previously[31] for the ESP at a single point above the center of substituted benzenes. The observed deviations are in accord with standard resonance parameters for these substituents: for example, OH is a strong π-electron donor ($R = -0.64$), and consequently, the additive ESPs is more positive above the ring than for the true ESP. Conversely, NO is a strong π-electron-withdrawing group ($R = 0.42$), and the additive ESP is more negative than the ESP for C_6H_5NO. The agreement between the additive and true ESPs for the aniline derivatives (Figure 3) is generally slightly

poorer (H = 0.78 to 0.95) than that observed for the substituents depicted in Figure 2, in accord with the very strong π-donating character of substituted amines.

Contour plots of the ESP in a plane perpendicular to the aryl ring and passing through the *ipso* and *para* carbons are shown in Figure 4 for benzene and in five substituted benzenes (NH_2, OH, CH_3, F, and NO_2). These plots, employing the same color-scheme as the surface maps shown in Figures 2 and 3, offer an alternative, complementary view and enable a more complete comparison of the true ESPs with the additive ESPs. To clarify the connection between these contour plots and the ESP maps in Figures 2 and 3, the electron density contour value used to construct the isodensity surfaces is superimposed on the ESP contour plots. Also shown in Figure 4 are contour plots of the additive ESPs for these selected systems. There are small differences between the additive ESPs and the true ESPs. However, the values of the ESP in this plane are mimicked by the additive model, and no changes in the π-electron-density are necessary to recover much of the substituent effect on the ESP. Specifically, for fluorobenzene, the regions of positive and negative ESP are roughly the same between the additive and the true ESP plots. The primary difference is the very small region above the center of the ring where the ESP dips below -11.25 kcal mol^{-1} (red), which is visible in the true ESP but missing in the additive ESP (the additive ESP is -8.9 kcal mol^{-1} in this region). Similarly, for aniline, the additive ESP overestimates the ESP above the ring, as was apparent in Figure 3.

Electron density contour plots for these systems are also shown in Figure 4. The electron densities are similar, as expected, apart from the area immediately surrounding the substituent. However, π-electron densities of aryl rings do change in response to introduced substituents. This can be most readily seen in contour plots of the difference in electron density between the substituted benzenes and the benzene [i.e., $\Delta\rho = \rho(C_6C_5X) - \rho(C_6H_6)$]. Because the scale is chosen to showcase the density differences around C_{para}, the $\Delta\rho$ values immediately surrounding the substituent are far off of the scale.

The changes in the electron density surrounding C_{para} exhibit expected trends: π-donating substituents (NH_2, OH, CH_3, and F) show a net gain in density above and below C_{para}, while π-accepting NO_2 reduces the electron density in this region. Despite these changes in the electron density surrounding C_{para}, the density above and below the center of the ring is essentially unchanged in each of these systems. These density difference plots can be used to rationalize the differences between the additive ESP and the true ESP contour plots. In the additive model, there is no change in the aryl π-system, so the effects of the density differences above and below C_{para} will be neglected. For aniline, the overestimation of the ESP above C_{para} in the additive ESP is due to the neglect of the increase in π-electron density at the *para* carbon in the additive model. Similarly, the additive ESP of nitrobenzene slightly underestimates the ESP above C_{para}, consistent with the decrease in density in that region in the intact system that is not present in the additive model.

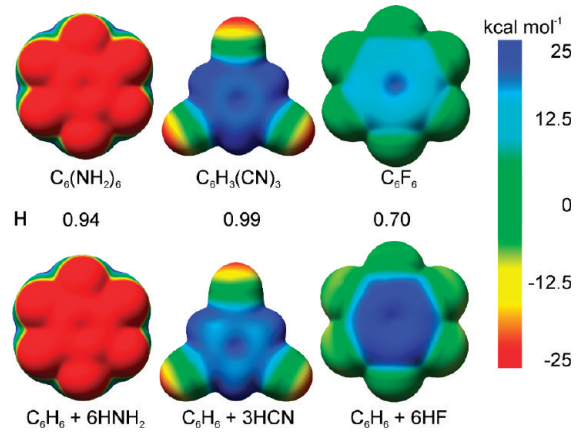


Figure 5. Plots of electrostatic potentials of polysubstituted benzenes (top) and corresponding additive ESPs (bottom). ESPs are mapped onto electron density isosurfaces (0.001 e/au^3) for the substituted benzene. The H index is computed for the ESP values on the isodensity surfaces for the true and the additive ESPs.

Since the additive ESPs are similar to the true ESP in each of these cases, it is clear that the effects of these π-electron density changes on the ESP are relatively minor. This is unsurprising, since these density changes are modest compared to the changes in the electron density surrounding the substituent. The through-space electrostatic effects of the substituents swamp the effect of π-donation and withdrawal, which, in most cases, shows up as a small perturbation of the ESP changes arising from nonresonance effects.

B. Polysubstituted Benzenes. ESPs for three polysubstituted benzenes are presented in Figure 5, along with additive ESPs. Computed Hodgkin indices again indicate a strong similarity between the true ESPs and the additive ESPs. These polysubstiuted benzenes were recently studied by Ringer and Sherrill[9] in the context of the sandwich configuration of the benzene dimer. Ringer and Sherrill argued[9] that the ESP of the pictured rotamer of hexaaminobenzene "confirms an electron-rich π cloud", while $C_6H_3(CN)_3$ and C_6F_6 are similarly shown to have "noticeably depleted electron density in the center of the substituted rings." The negative ESP above hexaaminobenzene and the positive ESP above tricyanobenzene and hexafluorobenzene do not necessarily arise from any change in the π-electron-density, since in the additive ESPs the aryl π-system is identical in each case. Labeling substituted aryl rings "π-electron-rich" or "π-electron-poor" based solely on computed ESP plots is clearly unfounded.

Hexafluorobenzene is of particular importance, since the reversed quadrupole moment of C_6F_6, compared to C_6H_6, is invoked to explain the strong face-to-face interaction of benzene and perfluorobenzene.[4,39−43] Perfluorobenzene also features in discussions of anion/π interactions[44] and in related complexes in which the π-cloud of C_6F_6 purportedly serves as an electron acceptor.[45] The reversal in electrostatics of perfluorobenzene is sometimes attributed to the withdrawal of electron density from the center of the ring by the fluorines. However, Laidig[46] showed in 1991 that the quadrupole moment of C_6F_6 arises primarily from the build-up of electron density along the periphery of the ring (i.e.,

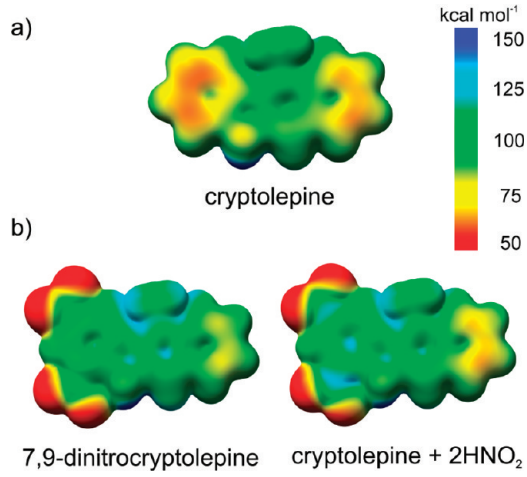


Figure 6. (a) ESP of cryptolepine; (b) ESP of 7,9-dinitrocryptolepine and additive ESP of 7,9-dinitrocryptolepine constructed by adding the ESP of cryptolepine to the ESP of two HNO_2 molecules and mapped onto the electron density isosurface of dinitrocryptolepine. Density isosurfaces correspond to $\rho = 0.005$ e/au^3. The *H* index for the true and the additive ESP plots is 1.00.

on the fluorines) rather than from the drastic changes of the electron density along the C_6 symmetry axis.

Contour plots of the electron density and density difference versus benzene for C_6F_6 are shown in Figure 4, along with contour plots of the ESP and the additive ESP. There is clearly a depletion of electron density above and below the plane of the ring due to the six fluorines. However, the introduction of a large amount of density associated with the fluorines easily swamps the changes above and below the benzene plane. As seen in Figure 4, the additive ESP of perfluorobenzene closely resembles the true ESP, demonstrating that the highly positive ESP above the center of the ring arises primarily from through-space effects of the fluorines and has no effect on the aryl π-system. Thus, while the π-electron-density of C_6F_6 is depleted compared to benzene (See Figure 4), the positive ESP above the ring is not evidence of this but merely of the through-space electrostatic effects of the F substituents.

C. Substituted Cryptolepines. Cryptolepine, a cytotoxic alkaloid from the West African shrub *Cryptolepis sanguinolenta*, is of interest as a lead for the development of both antimalarial and antitumor drugs.[47−49] Cytotoxicity of cryptolepine arises from its intercalation into DNA at nonalternating G-C sequences and inhibition of topoisomerase II.[49,50] The origin of the antimalarial activity is less well understood, though it is thought to involve the inhibition of hemazoin formation, similar to chloroquine.[47] There are ongoing efforts to develop cryptolepine derivatives that offer comparable or exceptional antimalarial activity without the associated cytotoxicity. 7,9-dinitrocryptolepine has been shown to exhibit antimalarial activity in the absence of DNA intercalation and toxicity, though the mode of antimalarial activity might differ from that of the parent compound.[47,51]

Electrostatic interactions are expected to be important in both the DNA intercalation and the inhibition of hemazoin formation for substituted cryptolepines.[52] ESP plots of cryptolepine and dinitrocryptolepine are shown in Figure 6.

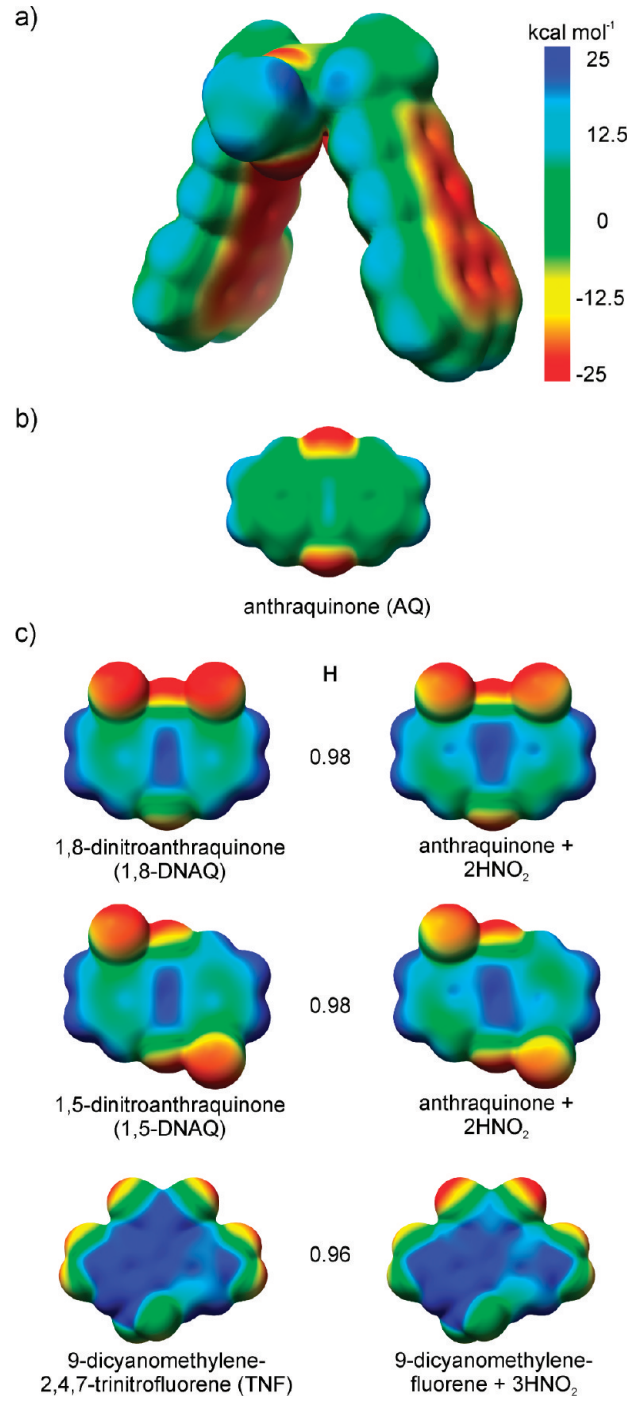
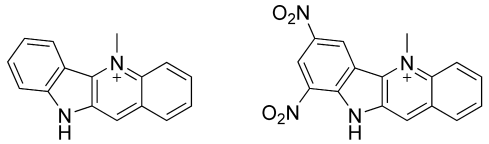


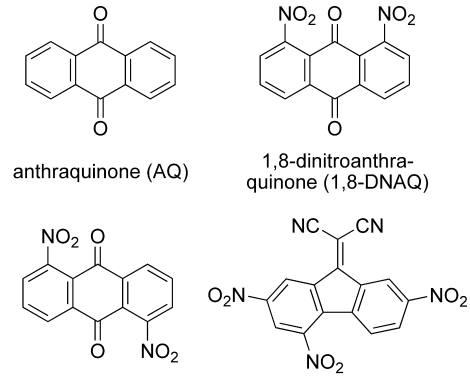
Figure 7. (a) ESPs of the molecular tweezers of Klärner and co-workers;[54,56] (b) ESP of anthraquinone; (c) ESPs (left) and additive ESPs (right) of two dinitroanthraquinones and 9-dicyanomethylene-2,4,5-trinitrofluorene. Density isosurfaces correspond to $\rho = 0.001$ e/au^3. The *H* index is computed for the ESP values on the isodensity surfaces for the true and additive ESPs.

The two nitro groups have a significant effect on the ESP, with the most pronounced changes localized on the substituted ring. An additive ESP for 7,9-dinitrocryptolepine is included in Figure 6b, constructed by adding the ESP of cryptolepine to the ESP of two appropriately placed HNO_2 moieties. Because there are only very minor differences between the true ESP and the additive model ($H = 1.00$), it is clear that the majority of the substituent effect arises from

Scheme 1



Scheme 2



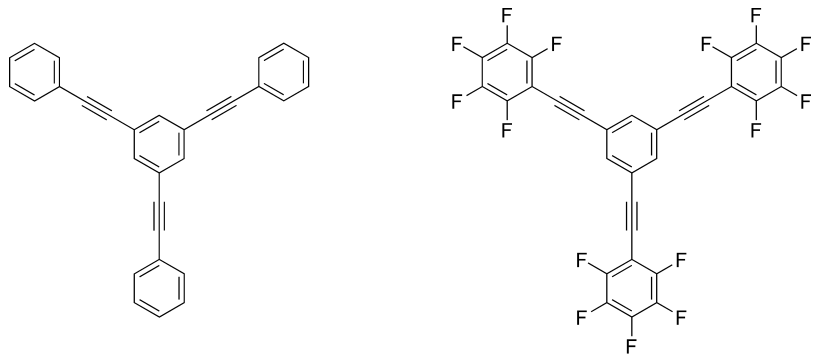
through-space effects. The π-system of cryptolepine plays a very minor role. In general, when considering ESPs of substituted analogs of candidate drugs built on aryl frameworks, the role of direct through-space effects of substituents is potentially significant and must be considered.

D. ESPs in Crystal Engineering and Host−Guest Chemistry. The field of supramolecular chemistry has blossomed in recent years, enabling the construction of complex molecular systems, molecular machines, and materials with novel properties through subtle control over intermolecular interactions.[4,53] Often this control arises from substituent effects on noncovalent interactions. In this regard, plots of molecular electrostatic potentials are valuable tools. One example of host−guest systems for which ESP maps have been employed are the molecular tweezers of Klärner and co-workers.[54−56] Klärner et al. synthesized and characterized a series of molecular tweezers based on bimethylene "hinges" separated by a benzene bridge with polycyclic aromatic "arms".[54] These receptors are powerful binders of what were described as "electron deficient" aryl systems.[56] The preferential binding was rationalized in part based on computed ESPs of the clips and the guest molecules (see Figure 7a−b). It was noted that substituted aryl systems with more positive ESPs (e.g., 1,8-DNAQ, 1,5-DNAQ, and TNF) are bound much more strongly than analogous systems with more negative ESPs (e.g., AQ), due to the favorable electrostatic interactions with the predominantly negative ESPs of the inner walls of the tweezers in the former case. Additive ESP plots of 1,8-DNAQ, 1,5-DNAQ, and TNF are provided in Figure 7c. In each case, the additive ESP is essentially indistinguishable from the true ESPs (H = 0.96−0.98). These significant changes in the ESP arise almost entirely from through-space effects; π-resonance plays no discernible role.

Another example gleaned from the field of supramolecular chemistry exploits the avidity of arenes for perfluorinated arenes,[4,39−43] originally observed by Patrick and Prosser.[40] As mentioned in Section III.B, this favorable interaction results from the opposite sign but comparable magnitude of the quadrupole moments of benzene and hexafluorobenzene.[39,57] This strong attractive stacking interaction has led to the use of the $C_6H_6\cdots C_6F_6$ interaction as a supramolecular synthon,[58] and this interaction has been exploited in countless systems. For example, Grubbs and co-workers[43] utilized perfluoroarene−arene interactions to achieve topological and stereochemical control over the photochemically driven reaction of 1,3-diynes in the condensed phase. Ponzini, Zagha, Hardcastle, and Siegel[59] later demonstrated the utility of such interactions in the generation of highly ordered crystals of 1,3,5-trisphenethynylbenzene and 1,3,5-tris(perfluorophenethynyl)benzene (Scheme 3). In both cases,[59,43] the electrostatic complementarity of the phenyl and perfluorophenyl moieties lead to highly ordered alternating face-to-face stacks in mixed crystals, while the two components on their own formed slipped−stacked arrangements.

ESPs of trisphenethynylbenzene and the fluorinated analog are shown in Figure 8. The complementary nature of the ESPs of the aryl and perfluoroaryl functionalities is immediately apparent. However, the additive ESP of the perfluorinated system (Figure 8, far right) once again shows that the highly positive ESP above the center of perfluorinated aryl rings is reproduced without any changes in the aryl π-system (H = 0.78).

Scheme 3



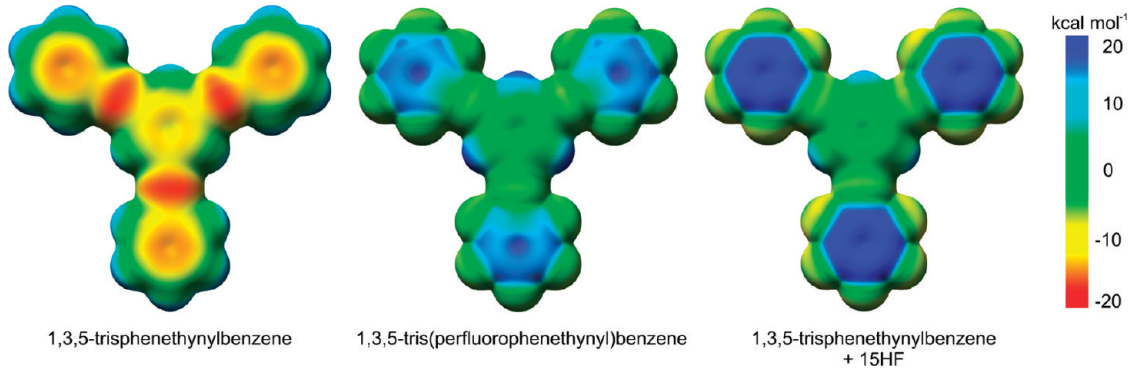


Figure 8. ESP plot of 1,3,5-trisphenethynylbenzene (left) and plot of the true (middle) and additive (right) ESP of 1,3,5-tris(perfluorophenethynyl)benzene, mapped onto electron density isosurfaces (0.001 e/au^3). The *H* index for the true and the additive ESP plots is 0.78.

IV. Implications for Non-Covalent Interactions with Aromatic Systems

Many qualitative models of substituent effects in noncovalent interactions with aromatic rings rest on the assumption that the dominant electrostatic effect arises from the polarization of the aryl π-system. This is most prominent in models of the benzene dimer advocated by Cozzi and Siegel[60] (the polar/π model) and Hunter and co-workers.[7,42,61] The crux of these primarily electrostatic models is that electron donors increase the aryl π-electron density, increasing the electrostatic repulsion with the π-system of the nonsubstituted ring, while electron acceptors enhance the benzene−benzene interaction through the opposite mechanism. While there have been numerous publications criticizing these models,[9,27,28,62] the underlying assumption that substituents modulate the electrostatic properties above the plane of benzene via polarization of the aryl π-system has previously remained largely unaddressed.

We have recently analyzed prototypical noncovalent interactions with substituted benzenes, including the sandwich and edge-to-face configurations of the benzene dimer[63] and the cation−π interaction of Na^+ with C_6H_5X.[31] The primary conclusions were that direct through-space interactions of the substituents were the dominant cause of substituent effects.[31,63] The electrostatic component of these direct interactions is reflected in the current finding that variation in ESPs maps of substituted benzenes is due, in large part, to through-space effects of the substituents. Thus, all models of substituent effects in intermolecular arene interactions based on ESPs should similarly reflect the role of through-space effects. Given the prevalence of electrostatic models, there is a potential for a broad revision of our understanding of the effect of substituents in myriad systems. For example, in the perfluoroarene−arene interactions utilized in crystal engineering, the present results suggest that this strong interaction arises primarily from the direct interaction of the fluorines with the nonfluorinated ring and not from π-polarization. Similarly, in the prototypical anion−π interaction between halide anions and C_6F_6,[44] the favorable interaction is potentially due largely to direct through-space interactions and not from a depleted π-system, as generally assumed.[44] Indeed, Clements and Lewis[64] showed that the attractive interaction between halogenated benzenes and F^- is due to direct interactions with the substituents.

The present work is a clarion call for the reevaluation of models of substituent effects in noncovalent interactions in which ESP arguments are central, since often it was assumed that changes in arene ESPs reflect changes in the aryl π-system. Specifically, the previously underappreciated role of direct through-space interactions of substituents must be reconsidered.

V. Summary and Conclusions

Molecular ESPs are powerful tools for the interpretation and the prediction of chemical phenomena and noncovalent interactions. However, deep-rooted misconceptions regarding the effect of substituents on the ESPs of substituted aromatic systems pervade the literature. Equating changes in ESPs with changes in the local electron density is prevalent, as exemplified by the "π-electron-rich" and "π-electron-poor" monikers assigned to aromatic systems, often based solely on ESP plots. While substituents do perturb the aryl π-system, the effects of these changes on the electrostatic potential surrounding aromatic systems are often swamped by the significant changes in the electron densities associated with the substituents. Several groups[65] have demonstrated that most substituents have no significant effect on the aromaticity of benzene. This resiliency of the benzene π-system shows up again in arene ESPs; for most substituents, the polarization of the aryl π-cloud is modest and does not significantly alter the ESP above the aromatic ring.

The role of through-space effects on ESPs was demonstrated here for a series of substituted benzenes and for more complex substituted arenes taken from disparate areas of research. Specifically, the change in the ESP of cryptolepine, a potent antimalarial and cytotoxic agent, induced by nitro substituents was shown to be independent of the aryl π-system. Similarly, the highly positive ESP values above the face of perfluorinated arenes, which are ubiquitous in supramolecular chemistry, can be reproduced with no alteration of the aryl π-system. Implications for our understanding of noncovalent interactions with substituted aromatic rings are profound, since substituent effects on the electrostatic component of many of these interactions arise primarily from direct interactions with the substituents.

Based on traditional, π-resonance-based models of noncovalent interactions with arenes, one would expect classical molecular mechanics force fields to perform poorly for supramolecular assembly phenomena. This is because MM force fields typically do not explicitly account for the perturbation of aryl π-systems by substituents. However, the present finding that polarization of the aryl π-system has minor affects on arene ESPs explains the often excellent performance of MM force fields for $\pi-\pi$ interactions.[66] The neglect of changes in aryl π-systems by substituents is, therefore, warranted. Treatment of only direct interactions with the substituents should suffice.

A hallmark of chemistry is the development and the widespread employment of qualitative predictive models. ESP plots constitute a powerful tool in this regard, demonstrating utility in many areas of chemistry and molecular biology. Without a sound understanding of substituent effects on ESPs, the utility of these tools is handicapped. A counterintuitive, yet striking, demonstration of the dominance of through-space effects on ESPs of substituted arenes has been provided with far-reaching implications in the understanding of noncovalent interactions in the fields of host−guest chemistry, crystal engineering, and rational drug design, among others. Perhaps most importantly, we have clearly shown that changes in ESPs do not necessarily reflect changes in the local electron density.

Acknowledgment. This work was supported by NIH-1F32GM082114 (S.E.W.) and NIH-GM36700 (K.N.H.). S.E.W. would like to thank H. Lischka for helpful suggestions and H. M. Jaeger and F. A. Evangelista for stimulating discussions regarding this work, which is dedicated to K. M. Williams and N. L. Wheeler. Computer time was provided in part by the UCLA Institute for Digital Research and Education (IDRE).

Supporting Information Available: Plots comparing ESPs at different levels of theory and with and without geometry constraints. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Politzer, P.; Truhlar, D. G., Eds.; Plenum: New York, 1981. *Molecular Electrostatic Potentials: Concepts and Applications*; Murray, J. S., Sen, K., Eds.; Elsevier Science: Amsterdam, The Netherlands, 1996. Naráy-Szabó, G.; Ferenczy, G. G. *Chem. Rev.* **1995**, *95*, 829–847.

(2) Politzer, P.; Murray, J. S. In *Computational Medical Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker, Inc.: New York, 2004; p 213−234.

(3) Politzer, P.; Murray, J. S. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 2.

(4) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.

(5) Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10566–10571.

(6) Gung, B. W.; Amicangelo, J. C. *J. Org. Chem.* **2006**, *71*, 9261–9270.

(7) Cockroft, S. L.; Perkins, J.; Zonta, C.; Adams, H.; Spey, S. E.; Low, C. M. R.; Vinter, J. G.; Lawson, K. R.; Urch, C. J.; Hunter, C. A. *Org. Biomol. Chem.* **2007**, *5*, 1062–1080.

(8) Hohenstein, E. G.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 878–886.

(9) Ringer, A. L.; Sherrill, C. D. *J. Am. Chem. Soc.* **2009**, *131*, 4574–4575.

(10) Naráy-Szabó, G. In *Molecular Electrostatic Potentials: Concepts and Applications*; Murray, J. S., Sen, K., Eds.; Elsevier: Berlin, 1996; p 333−365; RP, A.; B, L.; SL, P.; JG, V. *J. Comput. Aided Mol. Des.* **1995**, *9*, 33–43. Sheinerman, F. B.; Honig, B. *J. Mol. Biol.* **2002**, *318*, 161–177.

(11) Zheng, H.; Comeforo, K.; Gao, J. *J. Am. Chem. Soc.* **2009**, *131*, 18–19.

(12) Galabov, B.; Ilieva, S.; Schaefer, H. F. *J. Org. Chem.* **2006**, *71*, 6382–6387. Suresh, C. H.; Gadre, S. R. *J. Phys. Chem. A* **2007**, *111*, 710–714. Gadre, S. R.; Suresh, C. H. *J. Org. Chem.* **1997**, *62*, 2625–2627. Suresh, C. H.; Gadre, S. R. *J. Am. Chem. Soc.* **1998**, *120*, 7049–7055.

(13) Politzer, P.; Murray, J. S. *J. Mol. Struct.* **1996**, *376*, 419–424. Murray, J. S.; Lane, P.; Politzer, P.; Bolduc, P. R. *Chem. Phys. Lett.* **1990**, *168*, 135–139. Rice, B. M.; Hare, J. J. *J. Phys. Chem. A* **2002**, *106*, 1770–1783.

(14) Chana, A.; Concejero, M. A.; de Frutos, M.; González, M. J.; Herradón, B. *Chem. Res. Toxicol.* **2002**, *15*, 1514–1526.

(15) For recent examples, see Yearley, E. J.; Zhurova, E. A.; Zhurov, V. V.; Pinkerton, A. A. *J. Am. Chem. Soc.* **2007**, *129*, 15013–15021. Zhurova, E. A.; Matta, C. F.; Wu, N.; Zhurov, V. V.; Pinkerton, A. A. *J. Am. Chem. Soc.* **2006**, *128*, 8849–8861.

(16) Scrocco, E.; Tomasi, J. In *Topics in Current Chemistry*; Springer: Berlin, 1973; Vol. 42, p 95−170; Scrocco, E.; Tomasi, J. *Adv. Quantum Chem.* **1978**, *11*, 115–193. Bonaccorsi, R.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1970**, *52*, 5270–5284. Bonaccorsi, R.; Scrocco, E.; Tomasi, J. *Theor. Chim. Acta* **1971**, *21*, 17–27. Bonaccorsi, R.; Pullman, A.; Scrocco, E.; Tomasi, J. *Chem. Phys. Lett.* **1972**, *12*, 622–624. Bonaccorsi, R.; Pullman, A.; Scrocco, E.; Tomasi, J. *Theor. Chim. Acta* **1972**, *24*, 51–60. Berthier, G.; Bonaccorsi, R.; Scrocco, E.; Tomasi, J. *Theor. Chim. Acta* **1972**, *26*, 101–105. Tomasi, J.; Mennucci, B.; Cammi, R. In *Molecular Electrostatic Potentials: Concepts and Applications*; Murray, J. S., Sen, K., Eds.; Elsevier: Berlin, 1996; p 1−85.

(17) Yook, I.; Benítez, D.; Miljanić, O. Š.; Zhao, Y.-L.; Tkatchouk, E.; Goddard, W. A., III; Stoddart, J. F. *Cryst. Growth Des.* **2009**, *9*, 2300–2309.

(18) For recent examples, see: Zhang, Y.; Luo, M.; Schramm, V. L. *J. Am. Chem. Soc.* **2009**, *131*, 4685–4694. Laughrey, Z. R.; Kiehna, S. E.; Riemen, A. J.; Waters, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 14625–14633. Wang, Y.; Stretton, A. D.; McConnell, M. C.; Wood, P. A.; Parsons, S.; Henry, J. B.; Mount, A. R.; Galow, T. H. *J. Am. Chem. Soc.* **2008**, *129*, 13193–13200. Terraneo, G.; Potenza, D.; Canales, A.; Jiménez-Barbero, J.; Baldridge, K. K.; Bernardi, A. *J. Am. Chem. Soc.* **2007**, *129*, 2890–2900. Gorteau, V.; Bollot, G.; Mareda, J.; Perez-Velasco, A.; Matile, S. *J. Am. Chem. Soc.* **2006**, *128*, 14788–14789.

(19) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, CA, 2006; p 14−15.

(20) Shusterman, G. P.; Shusterman, A. J. *J. Chem. Educ.* **1997**, *74*, 771–776.

(21) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125–136. Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, 1940.

(22) Charton, M. *Prog. Phys. Org. Chem.* **1981**, *13*, 120–251.

(23) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165–195.

(24) Roberts, J. D.; Moreland, W. T. *J. Am. Chem. Soc.* **1953**, *75*, 2167–2173.

(25) Taft, R. W. *J. Phys. Chem.* **1964**, *64*, 1805–1815.

(26) Swain, C. G.; Lupton, E. C. *J. Am. Chem. Soc.* **1968**, *90*, 4328–4337.

(27) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2003**, *107*, 8377–8379. Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656–10668.

(28) Ringer, A. L.; Sinnokrot, M. O.; Lively, R. P.; Sherrill, C. D. *Chem.—Eur. J.* **2006**, *12*, 3821–3828.

(29) Arnstein, S. A.; Sherrill, C. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2646–2655.

(30) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303–1324. Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A. *J. Am. Chem. Soc.* **1996**, *118*, 2307–2308.

(31) Wheeler, S. E.; Houk, K. N. *J. Am. Chem. Soc.* **2009**, *131*, 3126–3127.

(32) Klärner, F.-G.; Polkowska, J.; Panitzky, J.; Seelbach, U. P.; Burkert, U.; Kamieth, M.; Baumann, M.; Wigger, A. E.; Boese, R.; Bläser, D. *Eur. J. Org. Chem.* **2004**, *140*, 5–1423.

(33) Politzer, P.; Lane, P.; Jayasuriya, K.; Domelsmith, L. N. *J. Am. Chem. Soc.* **1987**, *109*, 1899–1901.

(34) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(35) *Gaussian 03, Revision C.02*, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; 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.; and Pople, J. A.; Gaussian, Inc.: Wallingford, CT, 2004.

(36) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(37) If these additive ESPs are instead mapped onto an analogously derived additive electron density isosurface the results are indistinguishable from the present results.

(38) Hodgkin, E. E.; Richards, W. G. *Int. J. Quantum Chem.* **1987**, *32*, 105–110.

(39) Battaglia, M. R.; Buckingham, A. D.; Williams, J. H. *Chem. Phys. Lett.* **1981**, *78*, 421–423. Vrbancich, J.; Ritchie, G. L. D. *Chem. Phys. Lett.* **1983**, *94*, 63–68.

(40) Patrick, C. R.; Prosser, G. S. *Nature* **1960**, *187*, 1021–1021.

(41) Dahl, T. *Acta Chem. Scand.* **1988**, *42*, 1–7. West, A. P., Jr.; Mecozzi, S.; Dougherty, D. A. *J. Phys. Org. Chem.* **1999**, *10*, 347–350.

(42) Hunter, C. A.; Lawson, K. R.; Perkins, J.; Urch, C. J. *J. Chem. Soc., Perkin Trans. 2* **2001**, 651–669.

(43) Coates, G. W.; Dunn, A. E.; Henling, L. M.; Dougherty, D. A.; Grubbs, R. H. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 248–251. Coates, G. W.; Dunn, A. E.; Henling, L. M.; Ziller, J. W.; Lobkovsky, E. B.; Grubbs, R. H. *J. Am. Chem. Soc.* **1998**, *120*, 3641–3649.

(44) Ballester, P. *Struct. Bonding (Berlin)* **2008**, *129*, 127–174. Mascal, M.; Armstrong, A.; Bartberger, M. D. *J. Am. Chem. Soc.* **2002**, *124*, 6274–6276. Alkorta, I.; Rozas, I.; Elguero, J. *J. Am. Chem. Soc.* **2002**, *124*, 8593–8598. Quiñonero, D.; Garau, C.; Rotger, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Angew. Chem., Int. Ed.* **2002**, *41*, 3389–3392. Berryman, O. B.; Bryantsev, V. S.; Stay, D. P.; Johnson, D. W.; Hay, B. P. *J. Am. Chem. Soc.* **2007**, *129*, 48–58.

(45) Gallivan, J. P.; Dougherty, D. A. *Org. Lett.* **1999**, *1*, 103–105. Alkorta, I.; Rozas, I.; Elguero, J. *J. Org. Chem.* **1997**, *62*, 4687–4691. Danten, Y.; Tassaing, T.; Besnard, M. *J. Phys. Chem. A* **1999**, *103*, 3530–3534.

(46) Laidig, K. E. *Chem. Phys. Lett.* **1991**, *185*, 483–489.

(47) Wright, C. W.; Addae-Kyereme, J.; Breen, A. G.; Brown, J. E.; Cox, M. F.; Croft, S. L.; Gokcek, Y.; Kendrick, H.; Phillips, R. M.; Pollet, P. L. *J. Med. Chem.* **2001**, *44*, 3187–3194.

(48) Wright, C. W. *J. Pharm. Pharmacol.* **2007**, *59*, 899–904.

(49) Bonjean, K.; De Pauw-Gillet, M. C.; Defresne, M. P.; Colson, P.; Houssier, C.; Dassonneville, L.; Bailly, C.; Greimers, R.; Wright, C.; Quetin-Leclercq, J.; Tits, M.; Angenot, L. *Biochemistry* **1998**, *37*, 5136–5146.

(50) Lisgarten, J. N.; Coll, M.; Portugal, J.; Wright, C. W.; Aymami, J. *Nat. Struct. Biol.* **2002**, *9*, 57–60.

(51) Lisgarten, J. N.; Potter, B. S.; Palmer, R. A.; Pitts, J. E.; Wright, C. W. *J. Chem. Crystallogr.* **2008**, *38*, 815–819.

(52) Bhattacharya, A. K.; Karle, J. M. *J. Med. Chem.* **1996**, *39*, 4622–4629. Bhattacharya, A. K. *J. Mol. Struct. THEOCHEM* **2000**, *529*, 193–201. Medhi, C.; Mitchell, J. B. O.; Price, S. L.; Tabor, A. B. *Biopolymers* **2000**, *52*, 84–93.

(53) Lehn, J. M. *Science* **1993**, *260*, 1762–1763.

(54) Klärner, F.-G.; Kahlert, B. *Acc. Chem. Res.* **2003**, *36*, 919–932.

(55) Kamieth, M.; Klärner, F.-G.; Diederich, F. *Angew. Chem., Int. Ed.* **1998**, *37*, 3303–3306. Klärner, F.-G.; Panitzky, J.; Preda, D.; Scott, L. T. *J. Mol. Model.* **2000**, *6*, 318–327.

(56) Branchi, B.; Balzani, V.; Ceroni, P.; Campañá Kuchenbrandt, M.; Klärner, F.-G.; Bläser, D.; Boese, R. *J. Org. Chem.* **2008**, *73*, 5839–5851.

(57) Williams, J. H. *Acc. Chem. Res.* **1993**, *26*, 593–598.

(58) Thalladi, V. R.; Goud, S.; Hoy, V. J.; Allen, F. H.; Howard, J. A. K.; Desiraju, G. R. *Chem. Comm.* **1996**, 401–402.

(59) Ponzini, F.; Zagha, R.; Hardcastle, K.; Siegel, J. S. *Angew. Chem., Int. Ed.* **2000**, *39*, 2323–2325.

(60) Cozzi, F.; Annunziata, R.; Benaglia, M.; Baldridge, K. K.; Aguirre, G.; Estrada, J.; Sritana-Anant, Y.; Siegel, J. S. *Phys.*

Chem. Chem. Phys. **2008**, *10*, 2686–2694. Cozzi, F.; Cinquini, M.; Annunziata, R.; Dwyer, T.; Siegel, J. S. *J. Am. Chem. Soc.* **1992**, *114*, 5729–5733.

(61) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525–5534.

(62) Müller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143–168. Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697. Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457. Grimme, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 3430–3434.

(63) Wheeler, S. E.; Houk, K. N. *J. Am. Chem. Soc.* **2008**, *130*, 10854–10855. Wheeler, S. E.; Houk, K. N. *Mol. Phys.* **2009**, *107*, 749–760.

(64) Clements, A.; Lewis, M. *J. Phys. Chem. A* **2006**, *110*, 12705–12710.

(65) Campanelli, A. R.; Domenicano, A.; Ramonda, F. *J. Phys. Chem. A* **2003**, *107*, 6429–6440. Krygowski, T. M.; Ejsmont, K.; Stepień, B. T.; Cyrański, M. K.; Poater, J.; Solà, M. *J. Org. Chem.* **2004**, *69*, 6634–6640. Krygowski, T. M.; Stepień, B. T. *Chem. Rev.* **2005**, *105*, 3482–3512. Krygowski, T. M.; Szatylowicz, H. *Trends Org. Chem.* **2006**, *11*, 37–53. Wu, J. I.; Pühlhofer, F. G.; Schleyer, P. v. R.; Puchta, R.; Kiran, B.; Mauksch, M.; Hommes, N. J. R. v. E.; Alkorta, I.; Elguero, J. *J. Phys. Chem. A* **2009**, *113*, 6789–6794.

(66) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944–955.

CT900344G

JCTC Journal of Chemical Theory and Computation

High Accuracy ab Initio Calculations on Reactions of OH with 1-Alkenes. The Case of Propene

Róbert Izsák,*[,†] Milán Szőri,[‡,§] Peter J. Knowles,[†] and Béla Viskolcz[‡]

School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff CF10 3AT, United Kingdom, Department of Chemical Informatics, Faculty of Education, University of Szeged, Boldogasszony sgt. 6, 6725 Szeged, Hungary, and Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo náměstí 2, 16610 Prague 6, Czech Republic

Received March 21, 2009

Abstract: The energetics of terminal, central OH-additions as well as allylic H-abstractions by OH in its reaction with propene was studied as proxies for the 1-alkenes + OH reactions using several single and multireference ab initio techniques with basis set extrapolation where possible. Selection of the localized occupied orbitals forming the active space for multireference methods is discussed. Initial geometries of the reactants, prereaction complex (π-complex), and transition states were determined at the [5,5]-CASPT2/cc-pVTZ level of theory. Frequency analysis was also carried out at this level with the introduction of a scale factor. Analyzing the results, it will be concluded that multireference effects are negligible, and from the various single reference models we will opt for UCCSD(T)/cc-pVTZ for final geometry optimizations and vibrational frequency analysis. These results will be compared with those from approximate models yielding information on the reliability of the latter. Triples contributions are found to be very important, except for the π-complex, which has a UCCSD(T)/CBS relative enthalpy of −10.56 kJ/mol compared to infinitely separated propene + OH. The addition transition states are found to have relative enthalpies of −9.93 kJ/mol for the central and −9.84 kJ/mol for the terminal case. Allylic abstraction mechanisms, although lying significantly higher, still have only slightly positive barriers - a value of 3.21 kJ/mol for the direct and 1.67 kJ/mol for the consecutive case. Conventional transition state theory was used as a rough estimation for determining rate constants and turned out to agree well with experimental data.

1. Introduction

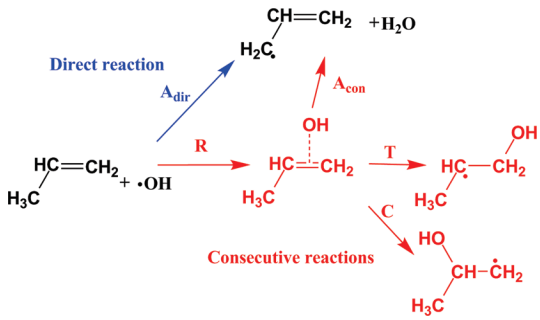
As a result of their importance in a variety of fields, hydrogen transfer reactions are well studied both experimentally and theoretically, making comparisons possible. These reactions are favored model systems for the study of chemical reactivity. In their simplest forms they serve as models for reactive scattering and dynamics[1] or models for heavy + light-heavy atom reaction systems, e.g. symmetric H exchange between halogen atoms and hydrogen halogenids.[2,3] They have served as useful model systems for testing standard theoretical approaches,[4,5] and many attempts have been made to describe similar systems with simple yet reliable approximations.[6,7]

An important special case of asymmetric H-transfer reactions plays an important role among hydrocarbon oxidation mechanisms. These are important in many areas of science, from understanding and reducing pollutant formation in combustion to describing partial oxidation in fuel cells.[8,9] It is widely accepted that the most common initial reaction of hydrocarbons in the atmosphere[10] and in all hydrocarbon frames[11] is the attack by an OH radical. Since propene can be a prototype of 1-alkenes, it is essential to characterize its

* Corresponding author e-mail: izsakr@cf.ac.uk.
† Cardiff University.
‡ University of Szeged.
§ Academy of Sciences of the Czech Republic.

10.1021/ct900133v CCC: $40.75
Published on Web 08/07/2009

Scheme 1. Studied Direct and Consecutive Reactions of the Propene + OH System



relevant reactions to understand the chemical behavior of 1-alkenes with the OH radical.

It is well-known that OH is able to attack the double bond of alkenes in terminal (**T**) and central (**C**) positions (Scheme 1). These addition reactions take place via a van der Waals complex (vdW-complex), a so-called π-complex (**R**). However, the importance of the hydrogen transfers such as the consecutive ($\mathbf{A_{con}}$) and direct ($\mathbf{A_{dir}}$) allylic H-abstractions in the case of alkenes + OH reactions has been recognized only recently.[12]

Although the propene and hydroxyl radical system has been studied previously and reported in several theoretical papers,[13–18] allylic H-abstraction channels were neglected in most cases. Earlier studies[13–16] have focused mostly on the ratio of the terminal and central addition reaction rates. Although Cvetanović reported in his work that 65% of the additions occur at the terminal carbon atom,[19] theoretical calculations at both the MP2/6-31+G(d)[16] and MP4(SDTQ)/6-31G(d,p)//MP2/6-311G(d,p)[13] levels of theory showed that central addition is preferred. However, it is emphasized in both theoretical works that the energy and entropy differences of the terminal and central transition states are quite small.[14] The interest in the kinetic behavior of the 1-alkene + OH system is also shown by papers published very recently[17,18] in this field, in which theoretical calculations offered a mainly kinetic description of the system. These works are mostly based on PMP2/aug-cc-pVQZ//MP2/cc-pVTZ[17] and CCSD(T)/cc-pVDZ//B3LYP/cc-pVTZ[18] methods. We have found relevant discrepancy between these latter two potential energy surfaces, although both state that their results are good descriptions of the overall kinetics. These results will be discussed to some extent later on. All this has led us to determine the accurate energetics of transition states corresponding to the energetically favored reaction channels with small difference in their energetics. Based on this set of calculations we are able to provide a highly accurate framework for kinetic modeling as well as a procedure for the logical choice for the active space in such asymmetric species. On the other hand, our aim was also to provide highly reliable results from benchmark ab initio calculations for further tests with density functional methods for larger alkene homologues.

2. Methodology

The relevant structures of the reaction system were determined by geometry optimizations performed at two different levels of theory. Initial optimizations were carried out at the [5,5]-CASPT2/cc-pVTZ level. This method allows for choosing relevant correlation contributions by selecting the proper active space, and, therefore, it reduces computational requirements compared to more accurate models. On these geometries, various single point calculations were then performed to study some factors such as the effect of multireference treatment, spin contamination, basis sets, and triples contributions. Based on these results, the UCCSD(T)/cc-pVTZ level of theory was chosen for the final geometry optimizations. Harmonic vibrational analysis was carried out at both levels ([5,5]-CASPT2 and UCCSD(T)). Results on these geometries will be compared in the following section.

For the [5,5]-CASPT2/cc-pVTZ level of theory, the active space should involve the SOMO in all cases. For **C** and **T** the π bond must be involved, since it participates in the C−O bond formation, and in the remainder of cases, this orbital corresponds to the most mobile electrons out of doubly occupied orbitals (highest orbital energy in RHF reference). For $\mathbf{A_{dir}}$ and $\mathbf{A_{con}}$ the breaking C−H bond must also be involved. For consistency, a C−H bond is involved in the active space for all cases. This makes the treatment balanced since the active space contains contributions for all non-hydrogen atoms for all species. Reactants (1-propene and OH) were treated in the supermolecular approach with a 1000 Å separation and share the same active space structure. This results in an active space of 5 electrons placed in 5 orbitals, 2 of which are unoccupied in the Hartree−Fock configuration. The occupied orbitals are chosen by first localizing the initial RHF orbitals, and then after analyzing the basis function contributions, the relevant orbitals may be identified. Similar procedures have been discussed in the literature, addressing the difficulty of choosing a balanced active space resulting in a correct correlation treatment.[5,20] Local orbitals simplify the choice of occupied orbitals; however, the difficulty of choosing the right virtual orbitals still remains. Here only the active occupied orbitals are preselected, and the virtuals are chosen purely on the basis of energetic order from the RHF reference. This procedure seems sufficient, since after the MCSCF optimization the active virtuals are the π^*(C−C) and the σ^*(C−H) antibonding orbitals as desired, see Figure 1. For further details see the following section.

The multiconfiguration nature of the wave function assures that the wave function is qualitatively correct, the long-range static correlation effects having been considered - avoiding the dissociation related problems of single reference methods. The choice of the CASPT2 method ensures that the most relevant short-range dynamic electron correlation effects important for geometry optimization are also considered. In the optimizations, the active space described above was used in all cases, for consistency, even in the cases, where the C−H bond remains intact. The removal of this orbital and a corresponding virtual one from the active space or the choice of an alternative C−H orbital, however, does not influence the result of the optimization significantly. Neither does the use of a basis set augmented with diffuse functions results in any relevant change. In both the case of the modified active space and the basis set augmentation, the resulting change

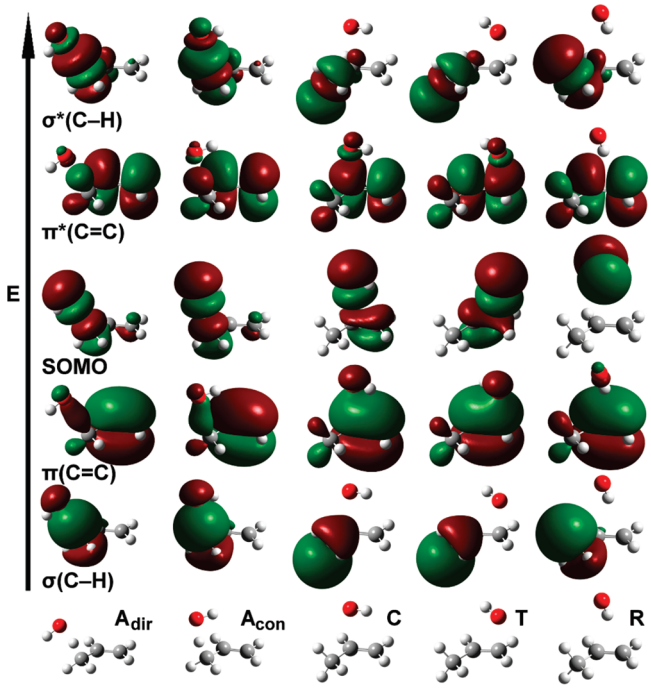


Figure 1. [5,5]-CASSCF/cc-pVTZ active orbitals for the π-complex (**R**) as well as transition state structures of direct ($\mathbf{A_{dir}}$) and consecutive ($\mathbf{A_{con}}$) allylic H-abstraction; terminal (**T**) and central (**C**) OH-addition reaction channels.

is of the order of a few 0.01 Å in bond lengths and a few 0.1 degrees in angles, corresponding to perhaps a few 0.01 kJ/mol in the calculated energies due to reoptimization.

For a further improvement in our results various single point calculations were carried out with different basis sets and high level correlated methods. Various kinds of multireference (MR) methods were used, beginning with CASPT2 and CASPT3 perturbative methods. Although in some ways CASPT3 is an improvement over CASPT2, for barrier heights it definitely seems inferior.[21] These methods are cheaper alternatives of the more expensive MRCI method, namely in our calculations the internally contracted MRCI with singles and doubles (ICMRCISD)[22] was used. With the MRCI results, denoted by Q1 and Q2, the Davidson corrected energies for fixed and relaxed references respectively are given in an attempt to make the wave function size consistent by adding approximate quadruple corrections. As observed in the literature,[23] Q2 usually yields poorer agreement with FCI and should be used only in special cases. Following this, iterative size consistency correction methods follow, namely MRACPF and MRAQCC, which are two variants of an approximate MRCC. Both have a tendency to overshoot the correlation energy, the first one more than the latter.[24] For some further details about these methods, see e.g. ref 24.

Various single reference (SR) methods are also presented, the reliability of which depends on whether the wave function is dominated by one configuration during the calculation. To answer this, one can first take the T1 diagnostics (from CCSD calculations) which indicates the significance of higher excitations and therefore the possibility of a need for a multireference treatment.[25] For all transition states, with cc-pVQZ basis, the T1 values are roughly equivalent or less than the critical 0.02 value (and well below for minima). As will be seen later on, the contribution of connected triple excitations has an important role, and with that included using a perturbative ansatz, a single reference treatment seems sufficient. One can also say based on multiconfiguration calculations that the weight of the ground state configuration is dominating (about 0.97) over all the rest (about 0.02 or less) at all examined geometries, and this dominance shows itself in the occupation numbers as well, those being quite close to the reference state values. This slightly changes with relaxation in MR calculations (see the difference between Q1 and Q2 corrections) and more significantly with the expansion of the active space (the dominance is still conserved although less evident). All this well justifies the use of single reference methods, and a further advantage will be that higher excitations are more feasible to include in the SR case.

The RMP2 values are gained as intermediate values in the coupled cluster procedure. The MP2 model suffers from some artifacts due to its lack of treating single excitations (for a study with FHF see Fox and Schlegel;[4] the arguments should hold for any H transfer with lone pairs close to the radical center). A variety of CC methods were also used, these are as follows: RHF-RCCSD and RHF-UCCSD models, RHF here referring to the reference orbitals. RCCSD is the partially spin restricted coupled cluster method (spin adapted in linear terms, which results in virtually no spin contamination).[26] Triples are treated in a variety of ways: the standard CCSD(T),[27] the simpler CCSD[T] missing the usually important singles contributions and CCSD-T[28] which considers some higher order perturbation terms compared to CCSD(T).

To approximate the nonrelativistic limit, extrapolations were carried out based with the cc-pVXZ bases[29,30] (X=D,T,Q), where the three point exponential formula of Feller[31] was used for HF and MCSCF results, and the two point X^{-3} function form[32] was used for correlation energies with X=T,Q. This latter choice is usually not too different from X=D,T, the most significant difference being with Davidson corrected energies, which show a slower convergence. In the multiconfiguration calculations the choice of the RHF or MCSCF wave functions as a reference in the Davidson correction does not introduce significant differences - which might be the case if there was a significant amount of dynamic correlation in the active space. In the RHF case, the extrapolation was checked against cc-pV5Z results, and it was found that the difference in predicted RHF barrers is less than 0.01 kJ/mol. The effect of augmented bases were also studied using aug-cc-pVXZ bases[30,33] with X=D,T for correlation energies, and for the references an additional X=Q was calculated.

Finally, some additional calculations were carried out in a less systematic fashion with smaller basis sets due to their computational cost. These include calculations with extended active spaces and some UHF-UCCSD(T) calculations for comparison. The explicitly correlated model UCCSD(T)-F12a[34] with the recommended basis (AVTZ) was calculated on the UCCSD(T) geometries. Similarly, the extrapolated UCCSD(T) energies for these structures were determined. In some cases, restricted active space (RASSCF) calculations,

Table 1. Method Dependence of Relative Standard Enthalpy Values (in kJ/mol) Obtained by Extrapolation of cc-pVXZ Basis Sets for the π-Complex (**R**), Transition States of Direct ($\mathbf{A_{dir}}$), and Consecutive ($\mathbf{A_{con}}$) Allylic H-Abstractions As Well As Terminal (**T**) and Central (**C**) OH-Additions

	$\mathbf{A_{dir}}$	$\mathbf{A_{con}}$	C	T	R
[5,5]-CASPT2	4.62	2.44	–9.00	–7.45	–10.37
[5,5]-CASPT3	22.10	20.63	6.66	7.08	–8.61
[5,5]-MRCI	28.88	28.19	11.70	11.50	–6.26
[5,5]-MRCI+Q1	15.36	13.97	2.39	6.44	–8.95
[5,5]-MRCI+Q2	16.64	15.20	3.60	7.22	–9.28
[5,5]-MRACPF	16.67	15.19	1.50	2.11	–8.78
[5,5]-MRAQCC	18.49	17.13	3.05	3.54	–8.41
RMP2	15.94	13.95	–3.91	–3.69	–10.96
RCCSD	16.60	15.29	–0.28	–0.19	–8.32
RCCSD[T]	5.69	3.91	–9.41	–8.60	–10.08
RCCSD-T	7.21	5.44	–8.12	–7.44	–10.00
RCCSD(T)	7.20	5.44	–8.19	–7.54	–10.03
UCCSD	14.35	13.04	–2.05	–1.84	–8.38
UCCSD[T]	2.89	1.15	–11.54	–10.60	–10.26
UCCSD-T	4.60	2.87	–10.16	–9.33	–9.96
UCCSD(T)	4.79	3.06	–10.02	–9.24	–10.01

and the corresponding correlated methods[35] prefixed with RAS were used. This means that we only allow certain excitations in certain regions of the active space. Here in the case of an extended [9,9] active space only double excitations are allowed from the lower two occupied and to the upper two unoccupied orbitals, which reduces computational cost (singles are eliminated due to numerical reasons rather than due to their quantity).

Most of the calculations were carried out with the MOLPRO program package of Werner and Knowles.[36] For the CCSDT and UHF-UCCSD(T) calculations, the MRCC package of Kállay[37] was used. Vibrational frequencies were calculated on the optimized geometries at the [5,5]-CASPT2/cc-pVTZ and UCCSD(T)/cc-pVTZ levels. To ensure a better agreement with experiment, the scale factor 0.958 ± 0.004 was determined for [5,5]-CASPT2 by fitting the calculated frequencies against the experimental values for propene[38] and OH.[39] For the UCCSD(T) frequencies the scale factor used is 0.975 ± 0.0021.[40] For comparative purposes, a variation of G3MP2B3 procedure[41] was also carried out where the B3LYP/6-31G(d) geometry optimization and normal node analysis were replaced by the BH&HLYP/6-31G(d) level of theory due to the fact that the B3LYP functional is not able to characterize the transition state for the consecutive allylic H-abstraction.[12] The BH&HLYP harmonic frequencies were scaled by 0.935.[12] In analogy to G3MP2B3, we term this method G3MP2BH&H, and refer to its earlier use in our publications.[42] All the DFT results were obtained using the Gaussian program package.[43] All enthalpy values are relative to that of the level of propene and OH.

3. Results and Discussion

First, let us discuss the single point results at the [5,5]-CASPT2 geometries. In Table 1 relative enthalpy results extrapolated from the cc-pVXZ basis sets are shown. The first obvious observation is that there is a significant difference between the multireference and the single reference results, especially when comparing the size consistency corrected MR and the triple corrected CC results, the ones that can be considered as the most reliable from the corresponding sets; this difference is approximately 10 kJ/mol.

Let us first analyze the SR results. The RMP2 result agrees with the CCSD results best, which is not surprising. The CCSD models show a considerable difference between results with and without triples corrections in both the restricted and unrestricted cases. Since the triples are important (see Table 1), the most reliable result must be among the corrected results. The CCSD[T] model as described above does not account for some important contributions and is therefore inferior to the others, while the other two methods agree very well, again in accordance with general experience. Therefore the choice of the standard CCSD(T) model is justified. One can still observe some difference between RHF-RCCSD and RHF-UCCSD results. We will return to this later; here we only say that in accordance with previous recommendations,[44,45] we choose the RHF-UCCSD results as the most reliable single reference ones and will use this for comparison.

The MR results appear somewhat divergent. The CASPT2 results agree well with the UCCSD(T) single point ones within 1−2 kJ/mol, which supports the choice of the inexpensive CASPT2 method for geometry optimization. The CASPT3 results seem to overestimate the barrier heights, and so does MRCI because of the size consistency error. It should be noted that in the supermolecular approach size consistency is already approximately dealt with, but the inclusion of higher excitations may still be important. For this reason, the theoretically most reliable results here are the ones with some kind of a correction for the latter error. These (MRCI+Q1, MRCI+Q2, MRACPF, and MRAQCC) give results within a broad 3 kJ/mol range. In all cases the difference between these is significantly smaller compared to that with the MRCI results, indicating the importance of higher excitations, and also the fact that the active space may be too small to involve all significant higher excitations. Indeed, if one compares these with the CCSD results (that is without triples correction), one finds a good agreement, showing that the MR calculations with the present active space is comparable with considering only SD excitations. We will come back to this later.

In Table 2, we present some results coming from extrapolation using augmented basis sets for selected methods. In general, there is a good agreement between the two extrapolations, they mostly differ for $\mathbf{A_{dir}}$ and $\mathbf{A_{con}}$ in MR calculations and for **R** in general. **R** being a weakly bound π-complex, longer range interactions are usually more important, which the augmented basis sets handle better (diffuse functions). The augmented basis sets also show a faster convergence. For all these reasons we will prefer results with augmented bases in the followings and refer to the extrapolations from these as the complete basis set (CBS) limit (see Table 3).

Table 3 begins with the [9,9]-RAS-MRCI+Q1 results. This [9,9] active space is the [5,5] extended with the two C−C bonds and two unoccupied orbitals (and with excita-

Table 2. Method Dependence of Relative Standard Enthalpy Values (in kJ/mol) Obtained by Extrapolation of aug-cc-pVXZ Basis Sets for the π-Complex (**R**), Transition States of Direct ($\mathbf{A_{dir}}$), and Consecutive ($\mathbf{A_{con}}$) Allylic H-Abstractions As Well As Terminal (**T**) and Central (**C**) OH-Additions

	$\mathbf{A_{dir}}$	$\mathbf{A_{con}}$	C	T	R
[5,5]-CASPT2	4.66	2.55	–9.01	–7.58	–10.93
[5,5]-MRCI	28.73	28.07	11.45	11.19	–6.85
[5,5]-MRCI+Q1	16.93	15.59	2.17	11.19	–9.06
[5,5]-MRCI+Q2	18.33	17.01	3.33	4.62	–9.03
UCCSD	14.38	13.16	–2.21	5.56	–9.16
UCCSD[T]	3.05	1.43	–11.51	–10.66	–10.96
UCCSD-T	4.70	3.08	–10.18	–9.43	–10.64
UCCSD(T)	4.90	3.28	–10.02	–9.34	–10.68

tions restricted from/to these extensions). This extension of the active space improves the agreement with single reference methods on the same double-ζ basis, except for **R**. In the next step, we improve the basis set by augmentation and the active space by removing the restriction of double excitations. The resulting [9,9]-MRCI+Q1 values are now even comparable with the extrapolated UCCSD(T) results, but in the case of **R** there is no improvement. If we now take the [5,5]-MRCI+Q1 results (Table 2), it is obvious that the major differences between SR and MR methods are in the case of the transition states, in the case of **R** there is actually a rather good agreement (differing by 1 kJ/mol only). Furthermore, for **R**, the triples contribution yields a contribution of 1 kJ/mol only in CCSD indicating that a consistent treatment of the triples does not change the result significantly. This would explain why [5,5]-MRCI+Q1, which was described above to have an overall SD quality agrees well with UCCSD(T) for the π-complex. In case of the transition state structures, the extension of the active space brings the desired improvement, indicating that the chosen active space gives a consistent treatment of important higher order excitations. With **R** this does not seem to be the case, that is some important higher order contributions are included, whereas others are left out in the extended active space, which causes an unbalanced, inconsistent treatment. To recover consistency, one should change the active space. However, as it was described earlier, this is not an easy task,[5] which seems only necessary for **R**.

How to control which orbitals go into the active space? We have control over the occupied orbitals, but the virtual ones are harder to choose. First, there is no symmetry condition which could help. One could perhaps see that orbitals with large contributions from the transferring H are involved, but, even so, care should be taken to choose such that are only related to the transfer directly and not to other interactions. This is a hard task in the case of a π-complex, where there are several competing noncovalent interactions. If we decide on not manipulating the virtual orbitals, one could try to change the occupied orbitals and hope that the MCSCF optimization will result in the desired virtuals. There are many possibilities to do this, here we only note that a [9,9] active space where the two C−C bonds and the C−H bond is replaced with the O−H bond and the two lone pairs of the oxygen yields no better results (−22.77 kJ/mol for **R**). Since the [9,9] results did not bring improvement, one could try to increase or decrease the active space of **R**. Increasing the active space further is not feasible, neither is a larger basis set. A decrease would take us back to the already discussed [5,5] space, which indeed seems an improvement in consistency, which due to the less emphasized importance of triples contributions with **R** shows itself as a good agreement with UCCSD(T). Since **R** is a minimum structure, it is less likely to have a multiconfigurational nature, so the UCCSD(T) result can be taken as the final word. This seems to be also the case with the transition states, since the above-mentioned not too high T1 value seems to be taken care of by the triples correction and also because the [9,9]-MRCI+Q1 results seem to converge there anyway (if we could allow the use of larger bases). From all this, our conclusion is that consistent MRCI+Q1 values with large enough active space and UCCSD(T) results agree well, and the latter should be chosen for computational and methodological ease.

The remainder of the first section of Table 3 addresses some spin related issues. Besides the RHF based treatments, here some UHF-UCCSD(T) results are also included. As pointed out earlier, there appears to be a roughly 2 kJ/mol difference between restricted and unrestricted CC results based on an RHF reference.[44,45] A somewhat smaller difference (a few tenths of kJ/mols) is observed between methods based on UHF and RHF orbitals.[44,45] It is difficult to reach firm conclusions on the nature of the spin contamination effects arrising from the UHF or the UCCSD(T) procedures from these data. Here, the UCCSD(T) model will be chosen[44,45] with RHF reference[46] as a preferred method for the final geometry optimizations and vibrational frequency calculations. Although the energetic difference between the RCCSD(T) and UCCSD(T) single points remains an open question, the issue may be addressed from a geometrical point of view. As discussed below, **T** (and also **C**) seems to be the most sensitive to correlation methods used for optimization. If one takes this species and optimizes the structure with RCCSD(T) and UCCSD(T) with the 6-31G(d) basis, one gets quite similar geometries: the most sensitive parameter, the C···O distance is 2.13 Å with the unrestricted, and 2.09 Å with the restricted method. If we now perform a UCCSD(T) single point calculation on the RCCSD(T) geometry (or vice versa) and compare it with the UCCSD(T) optimized value, there is only a slight 0.24 kJ/mol difference. On the other hand, the difference between the optimized energies is 2.55 kJ/mol, which corresponds to the above 2 kJ/mol gap between RCCSD(T) and UCCSD(T). This suggests that the choice of restricted or unrestricted CCSD(T) models has only a negligable effect on geometry optimizations in these cases, despite the energetic difference between the two. This is assumed to hold for all species and bases discussed here.

Having chosen the UCCSD(T) method, geometry optimizations and vibrational frequency analysis were performed with the cc-pVTZ basis. It is interesting to compare the [5,5]-CASPT2 and UCCSD(T) geometries to emphasize the good performance of the much cheaper CASPT2 method. In Figure 2, UCCSD(T) results are indicated first, and then in brackets the CASPT2 ones follow. The most significant difference is

Table 3. Method Dependence of Relative Standard Enthalpy Values (in kJ/mol) Obtained at Several Levels of Theory for the π-Complex (**R**), Transition States of Direct ($\mathbf{A_{dir}}$), and Consecutive ($\mathbf{A_{con}}$) Allylic H-Abstractions As Well As Terminal (**T**) and Central (**C**) OH-Additions[a]

	$\mathbf{A_{dir}}$	$\mathbf{A_{con}}$	**C**	**T**	**R**
		[5,5]-CASPT2 geoms&freqs			
[9,9]-RAS-MRCI+Q1/cc-pVDZ	12.30	10.81	–1.64	–1.54	–16.84
[9,9]-MRCI+Q1/aug-cc-pVDZ	5.36	4.33	–10.15	–8.80	–18.22
RCCSD(T)/aug-cc-pVDZ	5.92	4.74	–11.65	–9.99	–12.97
UCCSD(T)/aug-cc-pVDZ	3.96	2.83	–13.13	–11.39	–12.82
UHF-UCCSD(T)/aug-cc-pVDZ	4.36	3.17	–12.59	–10.68	–12.99
		UCCSD(T) geoms&freqs			
UCCSD(T)/cc-pVDZ	15.17	13.67	2.15	3.82	–11.03
CCSDT/cc-pVDZ	14.56	13.10	1.20	2.96	–11.04
UCCSD(T)/CBS	3.21	1.67	–9.93	–9.84	–10.56
UCCSD(T)-F12a/AVTZ	3.03	1.55	–10.54	–10.34	–10.47
		BH&HLYP geoms&freqs			
G3MP2BH&H	0.74	–0.92	–6.60	–5.35	–8.43

[a] [5,5]-CASPT2 (scale factor 0.958), UCCSD(T) (scale factor 0.975), and BH&HLYP (scale factor 0.935) optimized geometries and frequencies are included.

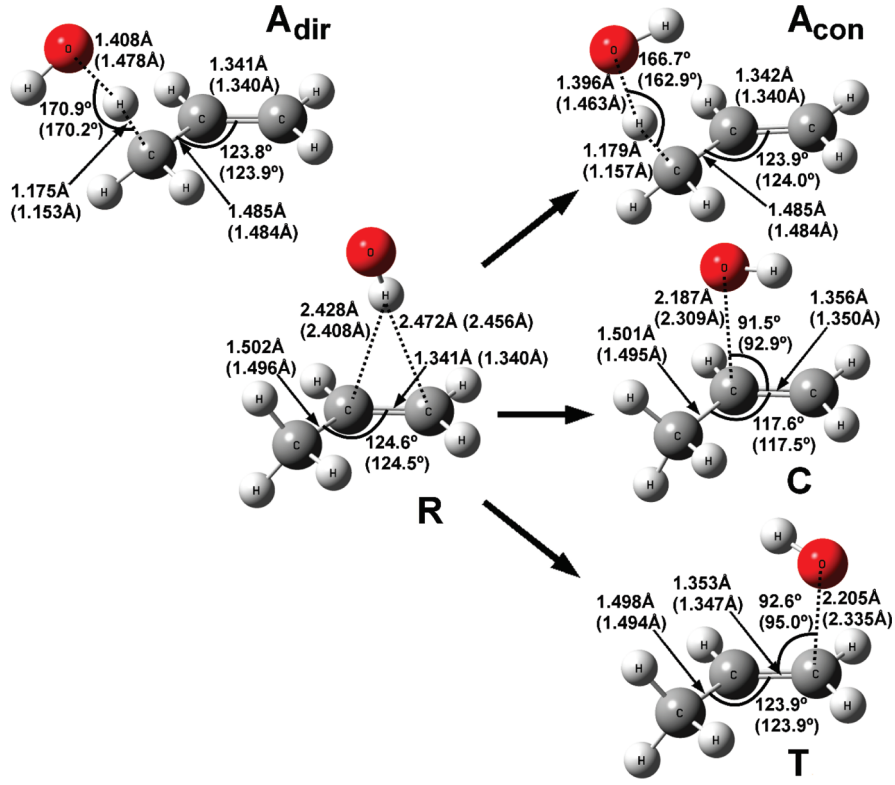


Figure 2. UCCSD(T)/cc-pVTZ geometry parameters followed by [5,5]-CASPT2/cc-pVTZ ones in brackets for the π-complex (**R**) as well as transition state structures of direct ($\mathbf{A_{dir}}$) and consecutive ($\mathbf{A_{con}}$) allylic H-abstraction; terminal (**T**) and central (**C**) OH-addition reaction channels.

the C···O distance in **T** and **C** (about 0.12 Å) which is probably due to some neglected correlation contributions rather than spin contamination effects (see above). However, this only yields a difference of about 0.3 kJ/mol between UCCSD(T) and UCCSD(T)//[5,5]-CASPT2 barriers with cc-pVTZ basis. Comparing the extrapolated UCCSD(T) energies at 0 K, they agree within 1 kJ/mol, which is an excellent agreement. At 298.15 K, the maximum difference in enthalpies is a somewhat larger 1.5 kJ/mol, since in this case differences in frequencies also play a role. Allylic abstraction barriers are 1.5 kJ/mol lower at the UCCSD(T) geometries, and the addition barriers are also much closer to each other compared to the results with CASPT2 (0.09 vs 0.65 kJ/mol difference). From now on, UCCSD(T) geometries will be used by default. Finally, it is also worth mentioning that if we take the extrapolated CASPT2 enthalpy barriers rather than the extrapolated UCCSD(T)//CASPT2 values as above and compare it with optimized UCCSD(T) values, the agreement is still very good (1−2 kJ/mol difference).

In the following, our results will be compared with structural data available in the literature. The UCCSD(T) geometry of the π-complex is in good agreement with the previously published MP2/6-31+G(d) geometry.[16] Most geometry parameters of transition state structures corresponding to addition channels calculated with UCCSD(T)/cc-pVTZ, CASPT2/cc-pVTZ, MP2/6-31+G(d),[16] and MP2/

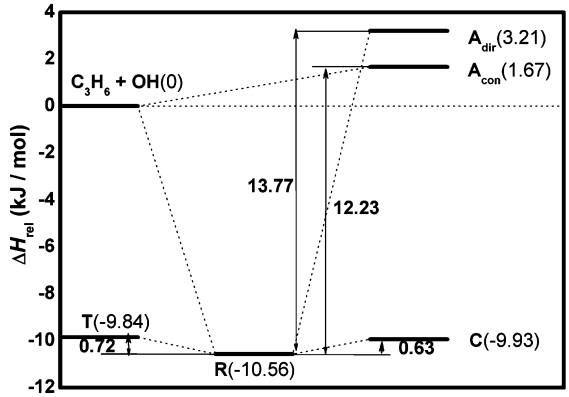


Figure 3. Energetics of the examined structures at the UCCSD(T)/CBS//[5,5]-CASPT2/cc-pVTZ level of theory.

cc-pVTZ[17] are also close to each other. The only exceptions are the bonds being formed (C−O) in **T** and **C**, where the CASPT2 bond lengths are about 0.13 Å larger, whereas the MP2 results are about the same value shorter compared to UCCSD(T) with cc-pVTZ basis. In general CASPT2 predicts earlier transition states than those obtained by single reference methods. Our previous BH&HLYP/6-31G(d) and CCSD/6-31G(d) results[12] on the transition states of allylic hydrogen abstraction channels are consistent with the corresponding UCCSD(T) geometries. Here, while the bonds being broken (C−H) are somewhat larger in the case of the BH&HLYP (1.22 Å) or CCSD (1.23 Å) geometries compared to UCCSD(T) transition states (1.18 Å).

In the second section of Table 3, some single point calculations on UCCSD(T) geometries are shown to further investigate some of the problems which occurred so far. Since it has been concluded above that the triples contribution is of great importance, results gained at the (unrestricted) CCSDT/cc-pVDZ level are included, together with the perturbative CCSD(T) results for comparison. The good agreement (within 1 kJ/mol) between CCSDT and UCCSD(T) results suggests that we can indeed rely on the latter as a good approximation for triples contribution.

Finally, the convergence of the basis set extrapolation using the CCSD(T) results is tested. An explicitly correlated theory, UCCSD(T)-F12a[34] - as implemented in MOLPRO - was utilized with the recommended AVTZ basis set. This improves basis set convergence, so that we can obtain accurate results with relatively small bases (differences from UCCSD(T)/CBS are within 0.7 kJ/mol). This procedure yields very similar results from what we had from extrapolation. It is also worth noting that using the G3MP2BH&H extrapolation scheme also gives quasi-quantitative answers, the differences being in the worst cases around 3−4 kJ/mol.

The UCCSD(T)/CBS result for **R** as shown in Figure 3 is −10.56 kJ/mol relative to the infinitely separated species. The addition barriers - as suggested by earlier theoretical works lie very close to each other; actually our calculations show that they are within 0.09 kJ/mol (the virtual activation enthalpies are −9.93 kJ/mol for **C** and −9.84 kJ/mol for **T**). The allylic H-abstraction enthalpy barriers are (about 12 kJ/mol) higher and have a larger difference, 1.54 kJ/mol, making the consecutive reaction energetically favored (enthalpy barriers: 3.21 kJ/mol for $\mathbf{A_{dir}}$ and 1.67 kJ/mol for $\mathbf{A_{con}}$). Conventional transition state theory (cTST) might provide a rough estimation for the rate constants of these channels by means of our UCCSD(T) results. We assume a 4-fold electronic degeneracy (g) for OH ($g = 4$, ignoring spin−orbit splitting), $g = 2$ for the transition states, and $g = 1$ for propene. A 2-fold degeneracy of reaction paths was considered in the case of the addition transition states, since they have nonsuperimposable mirror images (the OH can come from either side of the propene plain). In the case of direct H-abstraction, a 3-fold rotational degeneracy is assumed (although the conformer with the OH in the propene plain is expected to be energetically a bit different), whereas in the consecutive case the same degeneracy is two. Propene also has a 3-fold conformational degeneracy because of the methyl group (this only affects the abstraction cases). At room temperature, the rate constants are 4.86×10^{-14} (for $\mathbf{A_{dir}}$), 4.49×10^{-14} (for $\mathbf{A_{con}}$), 4.23×10^{-12} (for **C**), and 8.33×10^{-12} (for **T**) in cm^3 molecule^{-1} s^{-1} which in total (1.27×10^{-11} cm^3 molecule^{-1} s^{-1}) agrees within a factor of 2 with the value recommended by IUPAC at this temperature (3.02×10^{-11} cm^3 molecule^{-1} s^{-1}).[47] This latter experimental value is derived from a temperature dependent formula recommended by IUPAC and is partially based on work of Zellner and Lorenz,[48] who suggest a value of $(3.0 \pm 0.5) \times 10^{-11}$ cm^3 molecule^{-1} s^{-1} at 298 K. Similarly, in Atkinson's review,[49] the suggested 2.63×10^{-11} cm^3 molecule^{-1} s^{-1} with a 15% error is again in good agreement with our results. The above-mentioned factor of 2 corresponds to a 1.7 kJ/mol change in the barriers. However, in our estimation, the error of the energy calculations is only around 1 kJ/mol or perhaps even less in some cases (see Table 3 for CCSDT benchmarks for triples errors and also for basis set convergence). The rest of the discrepancy must come from sources like the choice of cTST for our estimations, the quality of the calculated frequencies (e.g., ignoring anharmonicity), and some other issues which a more thorough kinetic study should deal with. Since this was not our goal here, we consider our results in good agreement with experiment. To further support this point, the branching ratio for terminal addition (**T**, 65.8%) was calculated and was found to be in near perfect agreement with Cvetanović's data (65%)[19] with a calculated contribution of 0.4% for direct and consecutive allylic abstraction channels each. These results may also prove the accuracy of our quantum chemical results. It is also worth noting that the cTST overall rate constant obtained from UCCSD(T)/CBS//CASPT2/cc-pVTZ values (2.26×10^{-11} cm^3 molecule^{-1} s^{-1}) is closer to Atkinson's experimental one, whereas the calculated branching ratio for **T** (57.8%) is in a less good agreement with the above data (Cvetanović).

In their recent paper, Zhou et al.[17] have explored the propene + OH potential energy surface with projected MP2 methodology using the PMP2/aug-cc-pVQZ//MP2/cc-pVTZ level of theory and CCSD(T) methodology at the same geometries using the 6-311+G(3df,2p) basis and extrapolation from cc-pVDZ and cc-pVTZ bases. Their aim was to give an overall kinetic description at a broad temperature range, whereas our report focuses on species relevant around room temperature. In at least some of their cases, Zhou finds

that PMP2 results are closer to experimental values than CCSD(T) ones. It might be justified to choose their PMP2 methodology over an elaborate CCSD(T) optimization with so many species examined. However, the inclusion of higher excitations is known to be important with radical transition states,[44,45] which is particularly true for the studied system as we pointed out in the previous discussion. The good results with the PMP2 methodology are probably due to a cancellation of errors,[15] and the inferior behavior of CCSD(T) observed by these authors might be a problem of inadequate extrapolation and simply the fact that CCSD(T) single points are not calculated at their optimized geometries. Both of these issues have been addressed here by using larger bases, using CCSD(T) optimized geometries and by comparing those results with ones from a wider choice of ab initio models. The authors were also able to reasonably reproduce the kinetic behavior of the system based on weak collision master equation/microcanonical variational RRKM theory by lowering the barrier heights of central OH-addition (TS11) and terminal OH-addition (TS12) with 1 kcal/mol. However, in Figure 11 of Zhou's article, the branching ratio for these is around 50−50% at room temperature versus the experimental 65% preference for the terminal case, which is well predicted by our CCSD(T) model (65.8%). If cTST branching is calculated with their results it turns out to be 54% for the central case, indicating that the difference between their calculations and the ones presented here (and also the experimental results) is not due to the choice of the kinetic model but to the fact that the optimized CCSD(T) results yield better and more consistent results compared to PMP2/aug-cc-pVQZ//MP2/cc-pVTZ.

In another recent paper, Huynh et al.[18] describes the kinetics of the enol formation based on CCSD(T)/cc-pVDZ//B3LYP/cc-pVTZ results. In our experience, results with cc-pVDZ basis are still roughly 10 kJ/mol away from the CBS limit. In addition, the pathologic behavior of the B3LYP in relevant cases is also known for a while.[12]

4. Conclusions

Barrier heights for different possible reaction paths were calculated for the propene + OH system with the most accurate models available for general use. The results could be summarized in the following points:

1. The use of single reference methods are sufficient and accurate in this case, and in fact they yield more accurate results than multireference methods due to computational limitations for the latter. On the other hand, the advantage of using a multireference CASPT2 in this case is that if active orbitals are carefully selected, it is able to approximate UCCSD(T) within 1−2 kJ/mol with considerably less computational effort.

2. The RHF-UCCSD(T)/CBS method is expected to yield the most accurate results. Triples contributions are substantial (typically around 10 kJ/mol for barriers). The restricted coupled cluster variant exhibits a slight difference to these results (around 2 kJ/mol), which is, however, unlikely to effect geometry optimizations.

3. G3MP2BH&H yields a result within 3−4 kJ/mol to the extrapolated UCCSD(T). As another way of approximating the complete basis set limit, the explicitly correlated UCCSD(T)-F12a model was utilized giving results within 0.7 kJ/mol maximum difference compared to extrapolated values.

4. Consecutive allylic abstraction and addition mechanisms go through a π-complex (**R**), which lies at −10.56 kJ/mol with respect to the enthalpy level of the infinite separation of the species.

5. The addition mechanisms have negative enthalpy barriers relative to infinite separation (−9.93 kJ/mol for **C** and −9.84 kJ/mol for **T**). There is only a marginal 0.09 kJ/mol energetic difference between the two.

6. The allylic abstraction mechanisms have slightly positive enthalpy barriers relative to infinite separation (3.21 kJ/mol for $\mathbf{A_{dir}}$ and 1.67 kJ/mol for $\mathbf{A_{con}}$), with the consecutive mechanism favored by 1.54 kJ/mol. Although they have significantly higher barriers, they may contribute to the overall reaction system at higher temperatures.

7. Using conventional transition state theory, our UCCSD(T) results were able to reproduce the experimental overall high pressure rate constant within a factor of 2. Calculated branching ratios show the preference of **T** (65.8%) in good agreement with experiment. Allylic abstraction channels have a small contribution of 0.4% each. UCCSD(T)/CBS//CASPT2/cc-pVTZ values show similar good agreement supporting its use as an alternative to more expensive methods.

8. For higher 1-alkene homologues, where UCCSD(T) becomes too demanding to compute, a CASPT2 with similar active space structure may still be an option. Another possibility is to use the G3MP2BH&H method, which is found to be somewhat less accurate compared to CASPT2, but does not require constructing an active space.

Acknowledgment. This work was performed using the computational facilities of the Advanced Research Computing @ Cardiff (ARCCA) Division, Cardiff University. The authors would also like to thank Dr. Massimo Mella for useful discussions.

Supporting Information Available: Optimized geometries and harmonic frequency analysis results with rotational constants at both the [5,5]-CASPT2/cc-pVTZ and UCCSD(T)/cc-pVTZ levels, energies used for extrapolation for our final results, and additional calculated thermochemical data for both geometries. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schatz, G. C.; Sokolovski, D.; Connor, J. N. L. *Advances in Molecular Vibrations and Collision Dynamics*; JAI Press: Greenwich, CT, 1994; Vol. 2B, pp 1−26.

(2) Bondi, D. K.; Connor, J. N. L.; Manz, J.; Römelt, J. *Mol. Phys.* **1983**, *50*, 467.

(3) Dobbyn, A. J.; Connor, J. N. L.; Besley, N. A.; Knowles, P. J.; Schatz, G. C. *Phys. Chem. Chem. Phys.* **1999**, *1*, 957.

(4) Fox, G. L.; Schlegel, H. B. *J. Am. Chem. Soc.* **1993**, *115*, 6870.

(5) Luth, K.; Scheiner, S. *Int. J. Quant. Chem. Chem. Symp.* **1992**, *26*, 817.

(6) Zavitsas, A. A.; Chatgilialoglu, C. *J. Am. Chem. Soc.* **1995**, *117*, 10645.

(7) Isborn, C.; Hrovat, D. A.; Borden, W. T.; Mayer, J. M.; Carpenter, B. K. *J. Am. Chem. Soc.* **2005**, *127*, 5794.

(8) Taatjes, C. A.; Hansen, N.; McIlroy, A.; Miller, J. A.; Senosiain, J. P.; Klippenstein, S. J.; Qi, F.; Sheng, L. S.; Zhang, Y. M.; Cool, T. A.; Wang, J.; Westmoreland, P. R.; Law, M. E.; Kasper, T.; Kohse-Hoinghaus, K. *Science* **2005**, *208*, 1887.

(9) Miller, J. A.; Pilling, M. J.; Troe, J. *Proc. Combust. Inst.* **2005**, *30*, 43.

(10) Poppe, D.; Brauers, T.; Dorn, H.-P.; Karl, M.; Mentel, T.; Schlosser, E.; Tillmann, R.; Wegener, R.; Wahner, A. *J. Atmos. Chem.* **2007**, *57*, 203.

(11) Taatjes, C. A.; Hansen, N.; Miller, J. A.; Cool, T. A.; Wang, J.; Westmoreland, P. R.; Law, M. E.; Kasper, T.; Kohse-Hinghaus, K. *J. Phys. Chem. A* **2006**, *110*, 3254.

(12) Szőri, M.; Fittschen, C.; Csizmadia, I. G.; Viskolcz, B. *J. Chem. Theory Comput.* **2006**, *2*, 1575.

(13) Alvarez-Idaboy, J. R.; Diaz-Acosta, I.; Vivier-Bunge, J. R. *J. Comput. Chem.* **1998**, *8*, 811.

(14) Diaz-Acosta, I.; Alvarez-Idaboy, J. R.; Vivier-Bunge, J. *Int. J. Chem. Kinet.* **1999**, *31*, 29.

(15) Alvarez-Idaboy, J. R.; Mora-Diez, N.; Vivier-Bunge, A. *J. Am. Chem. Soc.* **2000**, *122*, 3715.

(16) Selcuki, C.; Aviyente, V. *J. Mol. Model.* **2001**, *11*, 398.

(17) Zhou, C.-W.; Li, Z.-R.; Li, X.-Y. *J. Phys. Chem. A* **2009**, *113*, 2372.

(18) Huynh, L. K.; Zhang, H. R.; Zhang, S.; Eddings, E.; Sarofim, A.; Law, M. E.; Westmoreland, P. R.; Truong, T. N. *J. Phys. Chem. A* **2009**, *113*, 3177.

(19) Cvetanović, R. J. *12th International Symposium on Free Radicals*; 1976; Laguna Beach, CA.

(20) Tishchenko, O.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1208.

(21) Werner, H.-J. *Mol. Phys.* **1996**, *89*, 645.

(22) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1988**, *89*, 5803.

(23) Werner, H.-J.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2008**, *128*, 034305.

(24) Szalay, P. G. *Chem. Phys.* **2008**, *349*, 121.

(25) Lee, T. J.; Rendell, A. P.; Taylor, P. R. *J. Phys. Chem.* **1990**, *94*, 5463.

(26) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219.

(27) Watts, J.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718.

(28) Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321.

(29) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(30) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358.

(31) Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104.

(32) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

(33) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(34) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.

(35) Olsen, J.; Roos, B. O.; Jørgensen, P.; Jensen, H. J. A. *J. Chem. Phys.* **1988**, *89*, 2185.

(36) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. et al. *MOLPRO, version 2008.1, a package of ab initio programs*; 2008. See: http://www.molpro.net (accessed Jun 24, 2009).

(37) Kállay, M. MRCC, a string-based quantum chemical program suite. See also: Kállay, M.; Surján, P. R. *J. Chem. Phys.* **2001**, *115*, 2945. as well as www.mrcc.hu (accessed Jun 24, 2009).

(38) Silvi, B.; Labarbe, P.; Perchard, J. P. *Spectrochim. Acta A* **1973**, *29*, 263.

(39) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*; Van Nostrand Reinhold Co.: 1979.

(40) NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 14, Sept 2006, Editor: Russell D. Johnson, III. http://srdata.nist.gov/cccbdb (accessed Jun 24, 2009).

(41) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.

(42) Szőri, M.; Abou-Abdo, T.; Fittschen, C.; Csizmadia, I. G.; Viskolcz, B. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1931.

(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, 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.

(44) Peterson, K. A.; Dunning, T. H. *J. Phys. Chem. A* **1997**, *101*, 6280.

(45) Chuang, Y.-Y.; Coitio, E. L.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 446.

(46) Szalay, P. G.; Vázquez, J.; Simmons, C.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 7624.

(47) Atkinson, R.; Baulch, D. L.; Cox, R. A.; Crowley, J. N.; Hampson, R. F.; Hynes, R. G.; Jenkin, M. E.; Kerr, J. A.; Rossi, M. J.; Troe, J. Summary of Evaluated Kinetic and Photochemical Data for Atmospheric Chemistry IUPAC Subcommittee on Gas Kinetic Data Evaluation for Atmospheric Chemistry Web Version February 2006. http://www.iupac-kinetic.ch.cam.ac.uk/ (accessed Jun 24, 2009).

(48) Zellner, R.; Lorenz, K. *J. Phys. Chem.* **1984**, *88*, 984.

(49) Atkinson, R. *Chem. Rev.* **1985**, *85*, 69.

CT900133V

JCTC Journal of Chemical Theory and Computation

Staggered Mesh Ewald: An Extension of the Smooth Particle-Mesh Ewald Method Adding Great Versatility

David S. Cerutti,*,† Robert E. Duke,‡,§ Thomas A. Darden,|| and Terry P. Lybrand†

Center for Structural Biology, Department of Chemistry, Vanderbilt University, 5142 Medical Research Building III, 465 21st Avenue South, Nashville, Tennessee 37232-8725, Department of Chemistry, University of North Carolina, Campus Box 3290, Chapel Hill, North Carolina 27599-0001, Laboratory of Structural Biology, National Institute of Environmental Health Science, Research Triangle Park, 12 Davis Drive, Chapel Hill, North Carolina 27709-5900, and Open Eye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, New Mexico 87508-1338

Received March 3, 2009

Abstract: We draw on an old technique for improving the accuracy of mesh-based field calculations to extend the popular Smooth Particle Mesh Ewald (SPME) algorithm as the Staggered Mesh Ewald (StME) algorithm. StME improves the accuracy of computed forces by up to 1.2 orders of magnitude and also reduces the drift in system momentum inherent in the SPME method by averaging the results of two separate reciprocal space calculations. StME can use charge mesh spacings roughly 1.5 × larger than SPME to obtain comparable levels of accuracy; the one mesh in an SPME calculation can therefore be replaced with two separate meshes, each less than one-third of the original size. Coarsening the charge mesh can be balanced with reductions in the direct space cutoff to optimize performance: the efficiency of StME rivals or exceeds that of SPME calculations with similarly optimized parameters. StME may also offer advantages for parallel molecular dynamics simulations because it permits the use of coarser meshes without requiring higher orders of charge interpolation and also because the two reciprocal space calculations can be run independently if that is most suitable for the machine architecture. We are planning other improvements to the standard SPME algorithm and anticipate that StME will work synergistically will all of them to dramatically improve the efficiency and parallel scaling of molecular simulations.

1. Introduction

With few exceptions,[1,2] the method of choice for computing long-ranged electrostatic interactions in molecular simulations with periodic boundary conditions is the Ewald sum.[3] Whereas simple truncation of long-ranged electrostatic interactions has been shown to give rise to significant simulation artifacts,[4,5] the use of modern Ewald algorithms enables more efficient simulations with effectively no omission of long-ranged electrostatic interactions.

In its original formulation, the Ewald sum for a system of N particles was an $O(N^2)$ computation, but the introduction of particle:mesh methods have reduced the complexity to $O(N \log N)$[6−8] and even to $O(N)$,[9−11] at which point the choice of optimal algorithm falls to the computational constants of the various methods given the problem's size and particle density. Most concisely, particle:mesh methods rephrase the problem of computing the electrostatic potential of a system of point (or otherwise highly localized) charges from solving $O(N^2)$ pairwise interactions to solving Poisson's equation for a highly smoothed version of the system's

* To whom correspondence should be addressed. Phone: (615) 936-3569. Fax: (615) 936-2211. E-mail: david.cerutti@vanderbilt.edu.

† Vanderbilt University.

‡ University of North Carolina.

§ National Institute of Environmental Health Science.

|| Open Eye Scientific Software.

10.1021/ct9001015 CCC: $40.75
Published on Web 08/04/2009

charges and then determining the difference between this smoothed-charge potential and the system's actual point charge density. This approach can be efficient because the smoothed charge density is written as a Gaussian convolution of the point charge density: the interaction of two Gaussian charges rapidly converges to the interaction of two point charges at distances greater than about six times the Gaussian's root-mean-squared deviation. The mesh-based electrostatic potential can then be solved by fast Fourier transforms (FFTs) in $O(N \log N)$ operations or by a finite-difference Poisson solver in $O(N)$ operations, while the modification needed to recover the point-charge potential is computed in $O(N)$ operations, similar to a simple truncation method.

Molecular simulations require accurate force calculations, as well as values of the total system energy. In the primary publications of the many available electrostatic mesh methods[7,8,10,12] there have been analyses of the parameters such as the width of the Gaussian charge smoothing function, direct space truncation length, and mesh density required to obtain a given degree of accuracy in forces acting on each atom or the total system energy. In simulations, the goal is to balance these parameters to maximize efficiency. Mostly, this is a matter of minimizing the total computational effort, but with the availability of highly scalable molecular dynamics codes,[2,13−15] another critical factor in the computational efficiency of an algorithm is the communications requirement. Parallel implementations on many different machine architectures can therefore benefit from algorithms that can obtain a given level of accuracy with the widest possible range of parameters.

In this Article, we draw upon a technique used in the 1970s for improving the accuracy of force calculations in particle: mesh methods. The method, known then as "interlacing," was first applied to plasma simulations by Chen and colleagues[16] and later to molecular simulations by Eastwood.[17] The fundamental improvement is to use two or more meshes staggered such that their points are displaced by some fraction of the mesh spacing, typically 1/2. Averaging the results obtained from each mesh produces significant error cancellation. When it was introduced, the method was viewed as a means for achieving higher levels of accuracy with limited amounts of computer memory, at the expense of speed. After rediscovering the method, however, we observe that it improves the overall computational efficiency on modern computers and may help to improve the parallel scaling of molecular simulations. We apply interlacing to the popular Smooth Particle Mesh Ewald method, and term the extended method "Staggered Mesh Ewald".

2. Summary of Particle Mesh Ewald Methods

If one calculates the electrostatic potential of a periodic system of charges by applying using Coulomb's law over all pairs of charges in a large number of images of the unit cell, the process is cumbersome and the result is only conditionally convergent. The Ewald method employs a mathematical identity to split the Coulomb sum $E^{(\mathrm{coul})}$ into a "direct space" sum $E^{(\mathrm{dir})}$ that converges rapidly (with a short interparticle distance $|\mathbf{r}_{ij}|$) in real or "direct" space, and a "reciprocal space" $E^{(\mathrm{rec})}$ sum that converges absolutely in "reciprocal" space after Fourier transformation.

$$\begin{aligned} E^{(\mathrm{coul})} &= \frac{1}{2}\sum_{\mathbf{n}}\sum_{i}\sum_{j\neq i}\frac{k_c q_i q_j}{|\mathbf{n}\cdot\mathbf{L}+\mathbf{r}_{ij}|} \\ E^{(\mathrm{dir})} &\simeq \frac{1}{2}\sum_{i}\sum_{j\neq i}\frac{k_c q_i q_j(1-\mathrm{erf}(\beta|\mathbf{r}|))}{|\mathbf{r}_{ij}|} \\ E^{(\mathrm{rec})} &= \frac{1}{2}\sum_{\mathbf{n}}\sum_{i}\sum_{j\neq i}\frac{k_c q_i q_j \mathrm{erf}(\beta|\mathbf{r}|)}{|\mathbf{n}\cdot\mathbf{L}+\mathbf{r}_{ij}|} \end{aligned} \quad (1)$$

Above, $\mathbf{n}$ represents all unit cell images, including the primary unit cell, $\mathbf{L}$ is a 3×3 matrix whose columns are the unit cell lattice vectors, i and j run over all charged particles in the system, $\mathbf{r}_{ij}$ is the interparticle distance, k_c is Coulomb's constant, and β is the "Ewald coefficient." (Exclusions of electrostatic interactions between bonded atoms in the primary unit cell are omitted from this discussion for simplicity.) The Ewald method reduces the problem of computing the electrostatic energy (and forces on all particles) to an $O(N^2)$ problem, a double sum over all particles to obtain the reciprocal space sum. (Computing $E^{(\mathrm{dir})}$ is an $O(N)$ problem because interactions can be neglected beyond some direct space cutoff L_{cut}.)

Physically, the Ewald method is equivalent to treating the system of point (or otherwise highly localized) charges as a system of diffuse Gaussian charges, solving the electrostatic potential $E^{(\mathrm{rec})}$ and forces $\partial E^{(\mathrm{rec})}/\partial \mathbf{r}_i$ arising from the Gaussian charge density, and then modifying those quantities with $E^{(\mathrm{dir})}$ and $\partial E^{(\mathrm{dir})}/\partial \mathbf{r}_i$ to recover the interactions of the point charges. Ewald mesh methods take this view of the Ewald reciprocal space procedure so that the reciprocal space sum can be solved on a mesh. (The direct space part is identical to the original Ewald method, and will not be discussed further.)

In general, the procedure with any Ewald mesh method entails four stages: (1) interpolate the charge mesh Q given the positions of particles and the magnitudes of partial charges, (2) smooth the interpolated point charges into Gaussian charges of the desired width, (3) compute the electrostatic potential $\Phi^{(\mathrm{rec})}$ by solving Poisson's equation for the smoothed charge density, and (4) compute the electrostatic potential energy and forces given the derivatives of the charge density in Q and the potential $\Phi^{(\mathrm{rec})}$. Many Ewald mesh methods, including the Smooth Particle Mesh Ewald[8] method that we will focus on during the results, make use of Fast Fourier Transforms (FFTs) to solve Poisson's equation; in those cases it is convenient to combine stages 2 and 3. The charge mesh Q is transformed using the forward three-dimensional FFT to obtain $\hat{Q}$, which is then multiplied element-wise by the transformed reciprocal space pair potential $\hat{\theta}^{(\mathrm{rec})}$. The inverse three-dimensional FFT is then applied to the product to complete the convolution $\Phi^{(\mathrm{rec})} = Q \bigstar \theta^{(\mathrm{rec})}$.

The two FFTs needed to convolute Q with $\theta^{(\mathrm{rec})}$ have $O(N \log N)$ computational complexity, much better than the complexity of the original Ewald method. However, for highly parallel molecular dynamics applications, the FFTs still require global data communication: every processor involved in the FFTs must broadcast its part of the problem to all other processors, and in turn receive similar information

Table 1. Test Cases for the Staggered Mesh Ewald Method[a]

case	cell dimensions (*a*, *b*, *c*), Å	cell dimensions (α, β, γ)	atom count
streptavidin	89.7 × 89.7 × 89.7	90°, 90°, 90°	73 305
protein crystal	91.3 × 81.3 × 91.0	90°, 90°, 90°	73 944
glycerol solution	69.7 × 69.7 × 89.0	60°, 90°, 90°	39 808
cyclooxygenase-2	114.8 × 114.8 × 114.8	109.5°, 109.5°, 109.5°	118 833

[a] The cases presented here span a variety of simulation cell geometries. All systems are in the condensed phase and were pre-equilibrated by molecular dynamics simulations at constant pressure.

from every other processor. This constraint on the ultimate scalability of the calculation has driven the development of real-space methods for solving $\Phi^{(\mathrm{rec})}$.[9−11] However, none of these methods has become widely used on commodity hardware because they are all considerably more expensive than the FFT-based methods: the break-even point comes at very high processor counts, which even today are not widely available. The Staggered Mesh Ewald method presented in this communication offers a way to reduce the total amount of mesh data that must be transformed, which we will show can help to accelerate simulations on a single processor and may help extend the scalability of FFT-based Ewald mesh methods.

3. Summary of Mesh Staggering Methods

Mesh staggering, or "interlacing" as it was originally termed, uses multiple samples of the interpolated charge density of particles on a mesh to suppress errors in the mesh calculation because of "aliasing."[7] Interpolation of a particle to a mesh creates a spectrum of aliases for that particle at each mesh point; because the spectrum is not perfectly smooth, the effects of different aliases on other aliases from the same particle or aliases of a nearby particle can be distorted by their proximity on the mesh. The most basic outcome of aliasing is the fluctuation of forces on particles as a function of their alignment relative to the mesh, which in turn is detrimental to momentum and energy conservation. By sampling multiple spectra of each particle on the mesh, different sets of aliases can be generated. Although each of these spectra contains roughly the same level of error in the interactions of each particle's aliases, the errors from multiple spectra may cancel if the spectra evenly sample the possible alignments of the system's particles relative to the mesh. The simplest and most economical implementation of the mesh staggering technique involves mapping particles to two meshes staggered such that points of one mesh fall exactly halfway in between those of the other.[16]

4. Methods

4.1. Preparation of Primary Test Cases. Anticipating that condensed-phase molecular dynamics simulations will be the primary application of the Staggered Mesh Ewald (StME) method, we selected four test systems: a streptavidin tetramer[18] solvated in a cubic cell, a condensed mixture of 35% v/v glycerol and water in a monoclinic cell, a scorpion toxin protein crystal lattice[19] solvated with water and ammonium acetate in an orthorhombic noncubic cell, and a cyclooxygenase-2 (COX-2) dimer[20] solvated in a truncated octahedral cell. Dimensions and atom counts in all of the simulation cells are provided in Table 1. Together, these four test cases span the available types of periodic simulation cells and encompass a variety of condensed-phase systems.

The SPC/E water model was used in all cases; glycerol parameters were obtained from Chelli and co-workers,[21] and any proteins were modeled with the AMBER FF99SB force field.[22] Prior to electrostatic calculations, all systems were equilibrated with at least 650 ps of molecular dynamics, including position-restrained dynamics if proteins were present and constant-pressure dynamics to reach each system's equilibrium density.

4.2. Accuracy Standards for Ewald Calculations. To compare different Ewald methods, it is necessary to define what parameters determine the accuracy of the calculation and also what is an "acceptable" level of accuracy. We will summarize these parameters here and then assess the efficiency of Ewald methods in terms of acceptable combinations of the parameters in the Results.

The accuracy of the direct space part of any Ewald electrostatics calculation is determined by the direct sum tolerance D_{tol}. Briefly, D_{tol} is the maximum acceptable relative difference between the interaction potential of two Gaussian charges and the interaction potential of two point charges. As we discuss in the Supporting Information, D_{tol} works together with the direct space truncation length L_{cut} to determine the width of the Gaussian charge smoothing function σ.

The accuracy of the reciprocal space part of a Smooth Particle Mesh Ewald (SPME) calculation is primarily a function of the ratio of σ to the mesh spacing μ, but in SPME, there is one other factor involved which is the order of interpolation used to map each point charge to the mesh. For the most generality, we recognize that μ can be different along each of the unit cell dimensions a, b, and c and that the unit cell lattice vectors need not be orthogonal. We therefore discuss results in terms of the number of mesh points in each dimension G_a, G_b, and G_c, in addition to the mesh spacings μ_a, μ_b, and μ_c. Most precisely, the mesh spacings refer to the magnitudes of the bin vectors $\mathbf{v}_a$, $\mathbf{v}_b$, and $\mathbf{v}_c$ as illustrated in Figure 1.

Because there is no single standard for the accuracy of forces in molecular simulations, we chose two based on default SPME parameters from existing molecular dynamics codes. Most codes use the largest μ_a, μ_b, and $\mu_c \leq 1.0$ Å obtainable such that G_a, G_b, and G_c are multiples of 2, 3, and 5; fourth order interpolation (a cubic B-spline) is typically used to map charges to the mesh. The AMBER molecular dynamics modules set $D_{\mathrm{tol}} = 1.0 \times 10^{-5}$ and $L_{\mathrm{cut}} = 8.0$ Å by default, whereas values of $D_{\mathrm{tol}} = 1.0 \times 10^{-6}$ and $L_{\mathrm{cut}} = 12.0$ Å are recommended in the NAMD and CHARMM communities. Because the overall strength of atomic charges differs between systems and slight changes

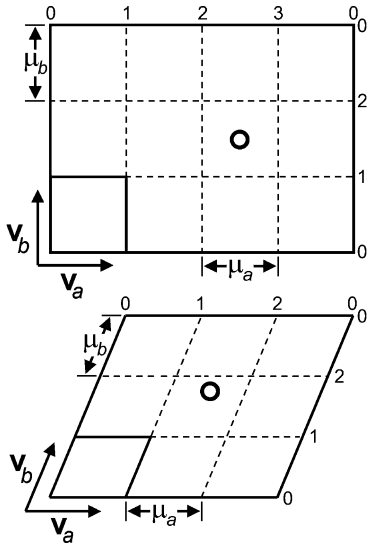


Figure 1. Graphical guide to mesh terminology used in the text. Two examples of a two-dimensional mesh are given. The upper mesh is a rectangular mesh analogous to an orthorhombic three-dimensional unit cell; the lower mesh is analogous to a nonorthorhombic unit cell. We define the bin vectors, $\mathbf{v}_a$ and $\mathbf{v}_b$, as shown for each mesh; note that the magnitudes of the bin vectors correspond to the mesh spacings μ_a and μ_b and that the lattice vectors can be written as $G_a\mathbf{v}_a$ and $G_b\mathbf{v}_b$, where G_a and G_b are the number of mesh cells in each dimension. Each mesh point is indexed from 0 to $G_a - 1$ or $G_b - 1$ and the meshes span a periodic unit cell as shown on the diagram. The mesh bin coordinates, $\mathbf{u} = (u_a, u_b)$ describe the location of a point $\mathbf{r}$ within the mesh: $\mathbf{r} = u_a\mathbf{v}_a + u_b\mathbf{v}_b$. Each mesh contains a small circle to represent a particle; the expression for its bin displacement is given by eq 2 or $\xi = \mathbf{u} - \text{floor}(\mathbf{u})$. In the upper mesh, the particle has mesh bin coordinates $\mathbf{u} = (2.5,1.5)$ and bin displacements $\vec{\xi} = (0.5, 0.5)$; in the lower mesh, the particle's mesh bin coordinates are (1.5, 1.75), and its bin displacements are (0.5, 0.75).

in the size of each system may graduate G_a, G_b, or G_c to the next available integer (i.e., 80 to 90 or 108 to 120), the accuracy of either method is system-specific. We therefore computed forces on all atoms from the four test cases in Table 1 using each set of Ewald parameters and compared them to the results of regular Ewald calculations as described above. We concluded that the AMBER default Ewald parameters can be expected to yield forces accurate to within 7.5×10^{-3} kcal/mol Å, roughly 0.05% relative error, whereas those recommended by the NAMD and CHARMM communities yield forces roughly five times more accurate, to within 1.5×10^{-3} kcal/mol Å or 0.01% relative error. We will refer to these as the "AMBER" and "CHARMM" standards later in this work.

In defining these standards, we emphasize that the default *settings* of a particular molecular dynamics package are separate from the numerical stability of the code itself. The AMBER dynamics engines SANDER and PMEMD both use double precision for all computations and can run simulations with very little energy drift. We also emphasize that the level of accuracy *necessary* to obtain reliable simulation results is not precisely known. The "AMBER" and "CHARMM" standards merely represent two points on a continuum.

4.3. Smooth Particle Mesh Ewald Force Calculations. SPME calculations for this work were performed using the SANDER module of the AMBER software package[13] in debugging mode to print out the forces. Staggered Mesh Ewald calculations, presented in the results, were done by averaging the results of two SPME calculations using the appropriate alignments of the particles and mesh. A high-accuracy regular Ewald sum, in which the forces were converged to a precision of 1.0×10^{-5} kcal/mol Å, was used as the reference for rating the accuracy of any Ewald mesh calculation.

5. Results

Although mesh staggering has been applied to the Particle: Particle Particle:Mesh (P^3M) method,[17] we will first quantify its benefits in the context of the newer Smooth Particle Mesh Ewald (SPME) method for simple cases before moving on to complex molecular systems. We divide the numerical error caused by particle aliasing into two sources: self-image forces that particles exert on themselves and errors in pair interaction forces. As would be expected, we observed that the self-image force errors are proportional to the squares of the individual charges and that pair interaction force errors are proportional to the product of the two charges. However, to simplify the following presentation, we use only $+1e$ and $-1e$ charges, where e is the charge of a proton. We also emphasize that the interpolation order, L_{cut}, and D_{tol} significantly influence the accuracy of SPME calculations, but again to keep the presentation simple, we fix these parameters at $L_{\text{cut}} = 9.0$ Å, $D_{\text{tol}} = 1.0 \times 10^{-6}$, and fourth-order interpolation. The periodic unit cell in the following examples, termed the "test cell," was a 64 Å cube.

5.1. Self-Image Forces in Smooth Particle Mesh Ewald calculations. The first source of numerical error can be observed by computing the SPME force on a single particle. We placed a single charge at a random point in the test cell and computed the electrostatic forces on the particle using a mesh spacing μ of 1.333 or 1.000 Å (corresponding to $G = 48$ or 64 points on a side). Repeating this procedure many times allowed us to plot the error in the force on the particle as a function of its alignment on the mesh, which we describe as its bin displacement ξ

$$\xi_\alpha = \frac{\mathbf{r}_\alpha}{\mu_\alpha} - \text{floor}\left(\frac{\mathbf{r}_\alpha}{\mu_\alpha}\right) \tag{2}$$

where $\mathbf{r}$ is a particle's position relative to the origin of the grid and $\alpha \in \{a, b, c\}$, the three dimensions of the mesh. The concept of a bin displacement is illustrated in Figure 1.

Simply stated, the self-image force error is any deviation from zero as a charge should put no force on itself and forces resulting from the charge's images should perfectly cancel. In Figure 2, the three components of the error are shown to be separable in three dimensions, plotted against the corresponding values of ξ. As has been found with previous investigations on mesh staggering, the self-image force error $\mathbf{F}^{(\text{si})}$ is well described by a Fourier sine series (eq 3)

$$\mathbf{F}_\alpha^{(si)} \simeq \sum_{p=0}^{\infty} W_\alpha^{(p)} q^2 \sin(2p\pi\xi_\alpha) \quad (3)$$

where q is the atomic charge and the $W^{(p)}$ coefficients depend on the direct space cutoff, mesh spacing in each dimension, and the interpolation order. For the cubic test cell, $W_a^{(p)} = W_b^{(p)} = W_c^{(p)}$. Repeating the mesh calculation for a mesh staggered by $\mu/2$ in all dimensions would eliminate all errors associated with sine series terms with odd values of k, most importantly $k = 1$. However, eq 3 raises the possibility of eliminating self-image force errors for all values of k by simply computing the appropriate sine series coefficients.

To see how $\mathbf{F}^{(si)}$ contributes to the total error in a system of multiple charges, we created a sparse set of 200 $+1e$ and 200 $-1e$ charges spaced further than $L_{cut} = 9.0$ Å from one another in the test cell. SPME calculations were carried out as before and were compared to the results from a regular Ewald calculation as described in Methods. The results in Figure 3 show that $\mathbf{F}^{(si)}$ plays a major role in the total error of the SPME calculation for a system of sparse charges; for all particles, the numerical error correlates with $\mathbf{F}^{(si)}$ with Pearson coefficient 0.96 ± 0.01 for $\mu = 1.000$ or 1.333 Å; if we optimize the first two $W^{(p)}$ coefficients to reduce the error in these SPME calculations, the values come very close to those found for the case of solitary charges under similar conditions, as illustrated in Figure 3. Correcting for the self-image forces can improve the accuracy of either of these SPME calculations by a factor of 4 to 5, implying that for simulations of diffuse plasmas the benefits of a second, staggered mesh calculation can be obtained by simply computing the appropriate sine series coefficients and applying a correction force to each particle after each mesh calculation. However, there are clearly other sources of error even with the particles spaced by more than L_{cut}.

As shown in Figure 4, these other sources of error dominate in a condensed system. While the overall error remains weakly correlated with $\mathbf{F}^{(si)}$, we found that simply correcting the self-image error was no longer effective for improving the accuracy of SPME calculations on dense plasmas or solvated biomolecular systems.

5.2. Pair Interaction Force Errors in Smooth Particle Mesh Ewald Calculations. To analyze the pair interactions that seem to be critical for accurate SPME calculations in condensed-phase systems, we set two charges of opposite sign close to one another in the test cell and computed the force between them using the same SPME parameters as before. We then iteratively perturbed the second charge along the x axis and recomputed the forces until the charge had traveled the entire width of the test cell. By repeating this analysis for different fixed positions of the first charge relative to the mesh (sampling the bin displacement ξ for the first charge while sampling all possible x coordinates of the second charge), we were able to plot the error in pair interaction forces between two particles as shown in Figure 5.

As shown in Figure 5, different aliases of each particle interact in complex ways, giving rise to errors that depend both on the interparticle separation in all three dimensions as well as the bin displacement of the first charge. Despite these complexities, however, Figure 5 confirms that mesh staggering eliminates a majority of the pair interaction force errors, particularly if the two meshes are staggered by 1/2 the mesh spacing μ in all directions simultaneously. Chen and co-workers used this multidimensional staggering approach in their simulations of plasmas,[16] although other investigators[7] have suggested that the results of as many as eight meshes, staggered by $\mu/2$ along any and all of the unit cell lattice vectors, should be averaged to obtain the best results. We tried averaging the results of eight such meshes (data not shown), but found this much more expensive approach to give scarcely better results than using only two meshes.

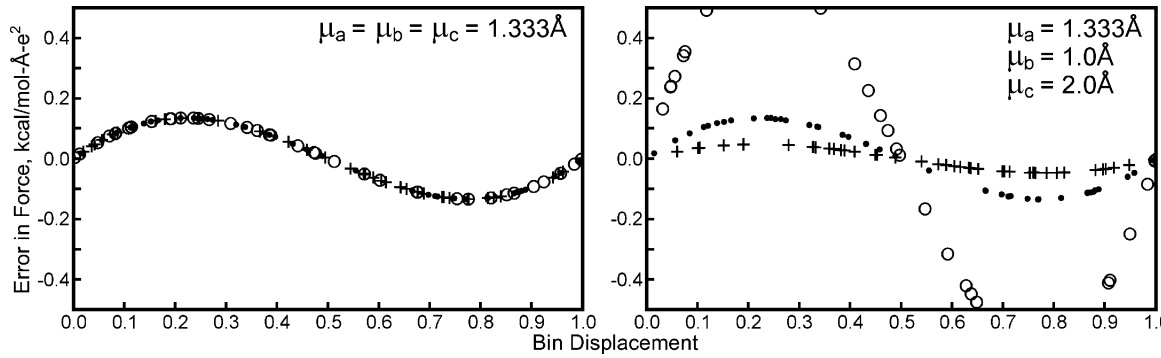


Figure 2. Force errors associated with mapping a single atomic charge to the mesh. When point charges are mapped to a mesh, they suffer an artifactual net force from their own the self-images in other unit cells. These artifactual forces decrease rapidly as the mesh becomes finer but can be significant even for 1.0 Å meshes, a common spacing used in conjunction with direct space cutoffs of ~9 Å. The plots above show the self-image forces on a $+1e$ test charge as it is moved to many random positions inside a 64 Å cubic box, when the mesh spacings given in each panel are used to compute the reciprocal space electrostatics. ●, +, and ○ represent self-image forces in the x, y, and z directions (along $\mathbf{v}_a$, $\mathbf{v}_b$, and $\mathbf{v}_c$ for this mesh). In all cases the self-image forces have a sinusoidal form given in eq 3 with respect to the bin displacement; the amplitude of the error increases rapidly with the mesh spacings μ_a, μ_b, or μ_c but appears to depend only on the charge's bin displacement in each dimension. Molecular dynamics codes typically add a "net force correction" to prevent the reciprocal space calculation from imparting artificial momentum on the system; eliminating the self-image forces would reduce but not obviate the need for such a correction (see also Figure 3).

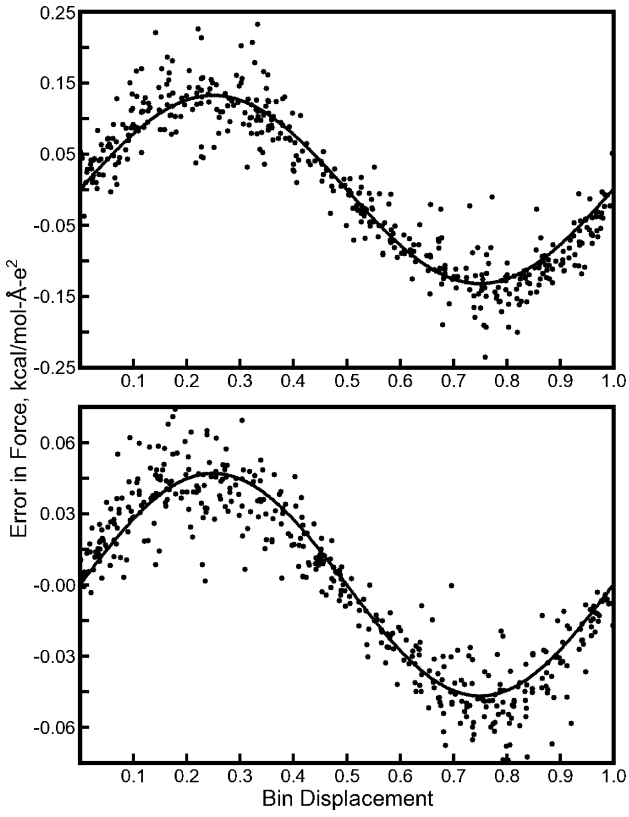


Figure 3. Force errors in a system of many sparse charges. A total of 200 pairs of $\pm 1e$ charges were placed in another 64 Å cubic box similar to the setup in Figure 2. Charges were distributed such that no two came within 9.5 Å of one another (for this example, $L_{cut} = 9.0$ Å, and $D_{tol} = 1.0 \times 10^{-6}$). Black dots in each panel represent the total error in the force on each charge in the *x*, *y*, or *z* directions. The black lines in each panel represent the expected self-image forces, obtained by optimizing the coefficients $W^{(1)}$ and $W^{(2)}$ in eq 3 for each mesh. For $\mu = 1.333$ Å, $W^{(1)} = 0.1335$ and $W^{(2)} = 0.0112$; for $\mu = 1.0$ Å, $W^{(1)} = 0.0476$ and $W^{(2)} = 0.0046$. While the expected self-image forces account for a significant amount of the total error, other sources of error are clearly present even for this sparse system of charges.

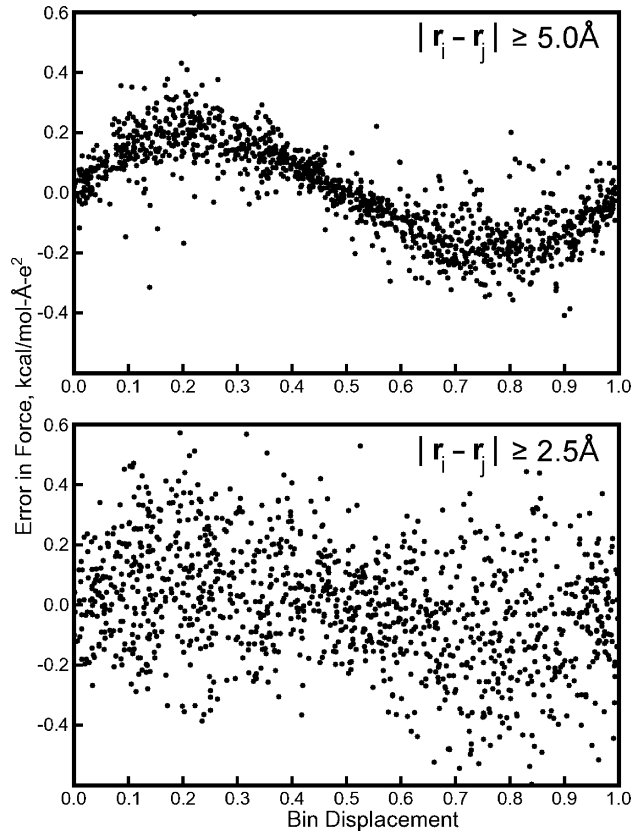


Figure 4. Force errors in increasingly dense systems. Many pairs of $\pm 1e$ charges (1600 in the top panel, 12 800 in the bottom panel) were placed in a 64 Å cubic box in the same manner as in Figure 3 according to the minimum interparticle spacing $|\mathbf{r}_i - \mathbf{r}_j|$ given in each panel. For this example, $\mu = 1.333$ Å, $L_{cut} = 9.0$ Å, and $D_{tol} = 1.0 \times 10^{-6}$. Black dots again represent the total error in the force on each charge in the *x*, *y*, or *z* directions (errors for only 400 charges are shown for clarity). The self-image forces remain a major factor in the total error even at minimum interparticle spacings as low as 5.0 Å, but other sources of error rapidly dominate as the minimum spacing goes below 2.5 Å and thus removing the self-image error is no longer an effective correction for coarse reciprocal space meshes. In a typical MD simulation, interparticle separations of less than 1 Å are common.

5.3. Staggered Mesh Ewald Method. Having confirmed that mesh staggering can eliminate large portions of the self-interaction force error, as well as pair interaction force errors in the SPME method, we sought to quantify the benefits of the mesh staggering in terms of accuracy and overall calculation efficiency when applied to condensed-phase biomolecular systems.

We term the use of two reciprocal space calculations on meshes aligned one-half mesh spacing relative to one another in all three mesh dimensions "Staggered Mesh Ewald" (StME). Because the reciprocal space operations (mapping charges to the mesh, convoluting the density and solving Poisson's equation, and interpolating forces from the smoothed potential) are identical to the procedures in SPME, implementing this method in current molecular dynamics codes can be straightforward. However, we will suggest some additional optimizations later in the Results.

With two meshes to compute but the potential to increase the accuracy by an order of magnitude or more relative to the corresponding SPME calculation, we wanted to thoroughly characterize the numerical error of StME relative to SPME for a variety of simulation parameters. In the 1970s, FFT solvers were efficient with mesh sizes of powers of 2, at the time, "interlacing" typically meant using two coarse meshes with twice the spacing of the equivalent fine mesh, and delivered an intermediate level of accuracy. Modern FFT solvers, however, are able to work efficiently with multiples of 2, 3, 5, and even 7; we therefore have much more freedom in the choice of mesh spacings for maximizing efficiency. Furthermore, since the 1970s, simulations in nonorthorhombic unit cells have become more common; it is important to confirm that mesh staggering is beneficial in these cases as well.

We performed both SPME and StME calculations on all test cases listed in Table 1 and compared them to regular Ewald calculations as described in Methods. The results are plotted in Figures 6 and 7. These tests, which included noncubic and nonorthorhombic cells, demonstrate

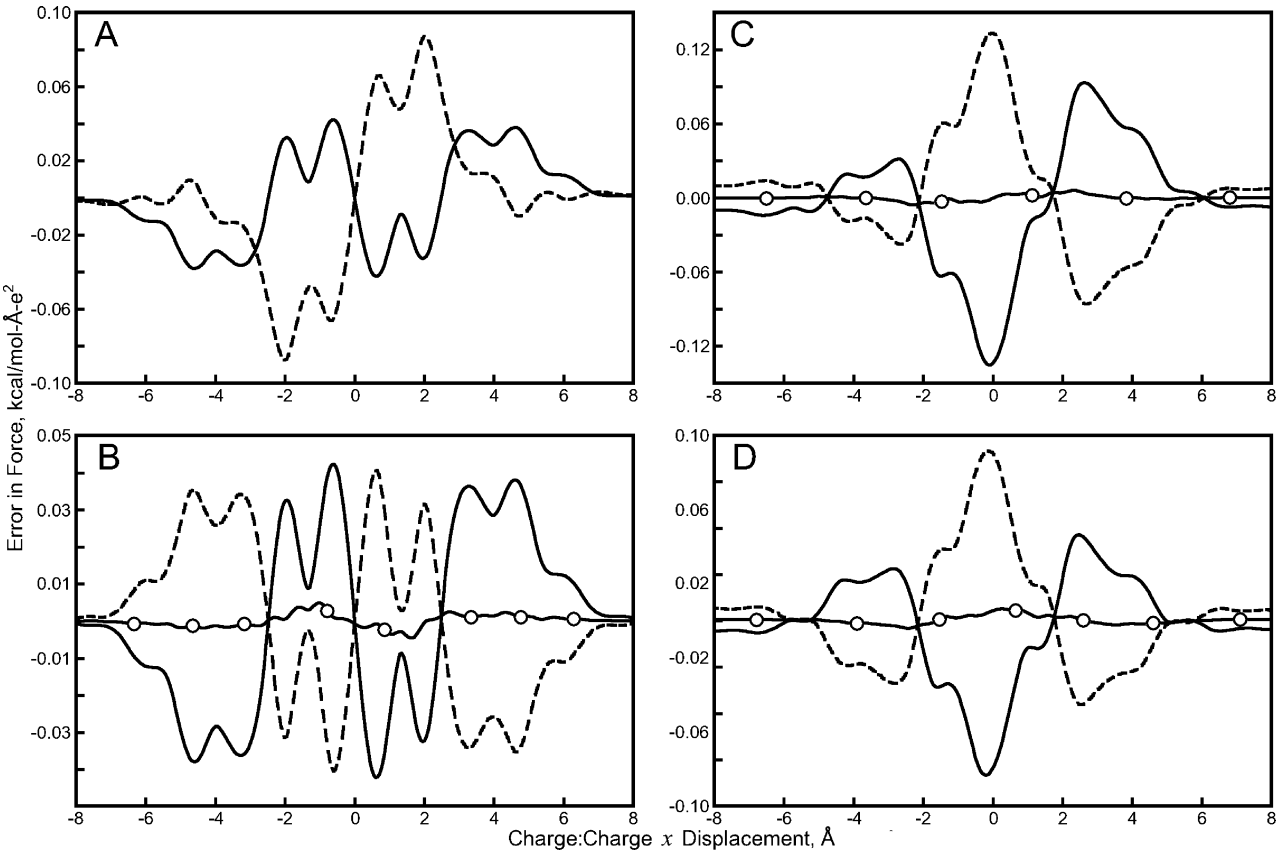


Figure 5. Force errors associated with pair interactions caused by coarse reciprocal space meshes. Significant errors enter the calculation of the force between two $\pm 1e$ charges P_1 and P_2 when a coarse mesh (in this case, $\mu = 1.333$ Å) is used. Error in the x component of the force exerted on P_1 by P_2, excluding any self-image force, is plotted as P_2 is moved parallel to the x axis such that its path intercepts P_1 unless otherwise noted. Results for several different positions of P_1 are shown as a function of the x displacement between P_1 and P_2. Panel A, solid line: P_1 is positioned at the origin. Panel A, dashed line: P_1 is positioned at 0.5 μ along the x axis. In Panel B, the solid line is copied from Panel A, but for the dashed line, P_1 is positioned at (0.5 μ, 0.5 μ, 0.5 μ). The errors are anticorrelated in Panel A and more strongly so in Panel B (the solid line with circles shows the average of the two errors). Panels C and D follow the format of Panel B. Panel C, solid line: P_1 positioned at 0.25 μ on the x axis. Panel C, dashed line: P_1 positioned at (0.75 μ, 0.5 μ, 0.5 μ). Panel D, solid line: P_1 positioned at (0.307 μ, 1.421 μ, 1.804 μ), P_2 moved along the x axis. Panel D, dashed line: P_1 positioned at (0.807 μ, 1.921 μ, 2.804 μ), P_2 moved to sample points (x, 0.5 μ, 0.5 μ).

the applicability of the method to periodic systems in general. Note, however, that the mesh grids are staggered by $(1/2)(\mathbf{v}_a + \mathbf{v}_b + \mathbf{v}_c)$, not simply by the half the mesh spacing in x, y, and z.

Figures 6 and 7 show that, for a given order of interpolation and direct sum tolerance, the StME method can greatly increase the accuracy of computed forces relative to an SPME calculation with similar parameters. In all systems, error reductions exceeding 1 order of magnitude can be obtained for direct space cutoffs of 8 Å to 10 Å with similar mesh sizes and values of D_{tol}; the benefits of StME appear to be highest for direct space cutoffs and mesh densities near those used in typical MD simulations. Furthermore, the StME method continues to produce comparable increases in accuracy, relative to SPME, if the order of interpolation is increased or if D_{tol} is reduced. Although we did not demonstrate that the self-image forces and errors in pair interaction forces could be canceled with staggered meshes in nonorthorhombic unit cells, StME shows equally good performance for other condensed-phase systems in such unit cells.

While electrostatic potentials from two meshes must be computed in StME, the method achieves comparable accuracy to SPME with coarser meshes and smaller values of L_{cut}. For example, when using fourth order interpolation and $D_{\text{tol}} = 1.0 \times 10^{-5}$ for calculations on the streptavidin system, StME achieves nearly the same accuracy with $L_{\text{cut}} = 8.0$ Å and $G_a = G_b = G_c = 60$ as SPME with $L_{\text{cut}} = 9.0$ Å and $G_a = G_b = G_c = 90$. In such a case, the direct space workload is reduced by almost 30% and the overall FFT workload is reduced by more than 40%: each mesh of 60^3 points is 3.375 $\times$ smaller than the mesh of 90^3 points that would become the bottleneck for the SPME calculation. Twice as many charge mapping and force interpolations would be required in the most basic implementation of StME, but as will be discussed, other optimizations can still lead to significant improvements in simulation efficiency with this method. We will make a detailed analysis of the optimal parameters for StME and SPME calculations later in the results.

5.4. Energies and Virials Obtained with Staggered Mesh Ewald. Highly accurate forces are the most important product of a molecular dynamics method, but we also wanted

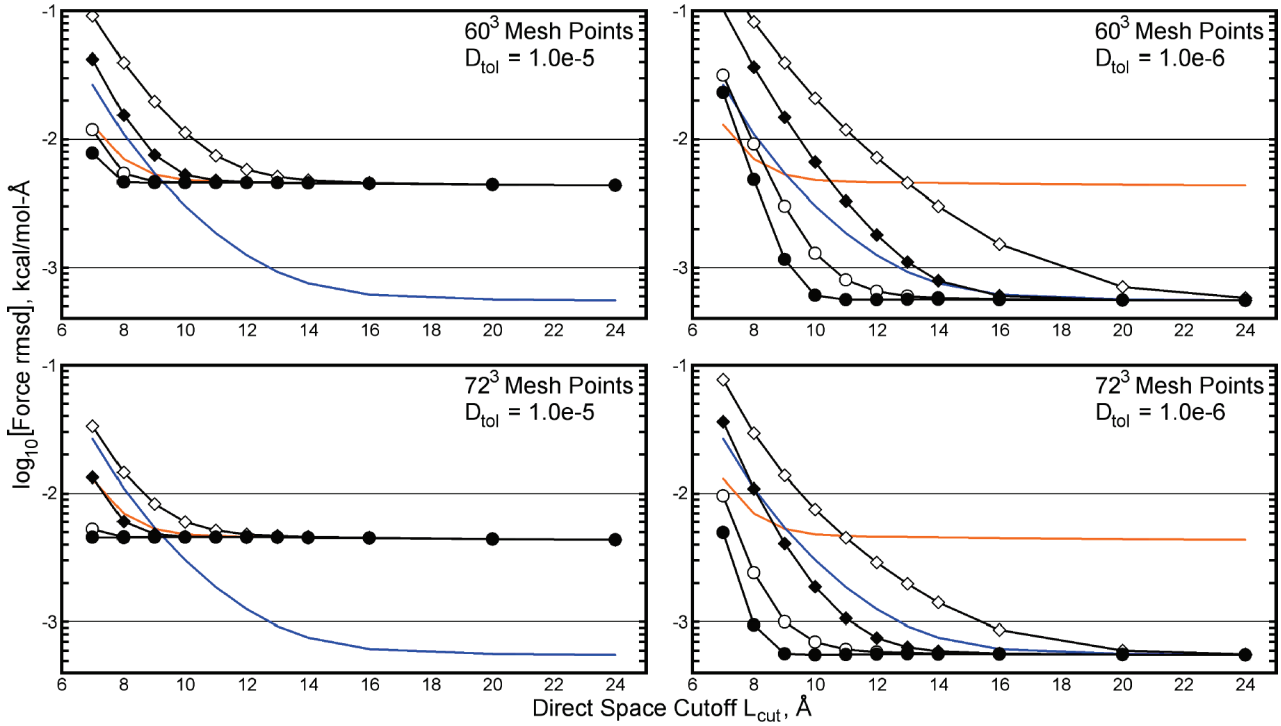


Figure 6. Accuracy of Smooth Particle Mesh Ewald (SPME) and Staggered Mesh Ewald (StME) calculations for the streptavidin test case. Each type of Ewald calculation was run using the parameters given in the top right corner of each panel. Black lines with open or filled symbols represent 4th or 5th order interpolation, respectively; diamonds and circles represent SPME and StME calculations, respectively. In most MD codes, a mesh of 90^3 points would be used, along with $D_{tol} = 1.0 \times 10^{-5}$ or 1.0×10^{-6} and 4th order interpolation; these cases are shown in orange and blue, respectively, for reference. Additional details of the streptavidin system are given in Table 1. Even with a mesh spacing 1.5× the standard value, the StME method offers improved accuracy over the corresponding SPME calculation for nearly all values of L_{cut}. StME maintains its advantage with finer meshes or higher interpolation orders.

to test whether StME could produce energies and virials of comparable accuracy to SPME, particularly when used with coarser meshes. Typically, electrostatics dominates the total potential energy of a molecular system but the energy of the reciprocal space part is fairly small. The reciprocal space calculation also makes fairly minor contributions to the system's virial tensor. Still, errors in these contributions could limit the overall applicability of StME. We also tested whether errors in either of these quantities were systematic or random by repeating the StME and SPME calculations for 20 individual snapshots taken at 100 ps intervals from 2 ns trajectories of each system. These tests were conducted using the SANDER module of the AMBER software.

When the mesh used to compute the reciprocal space electrostatic potential is coarsened, both the energy and elements of the virial tensor trace become increased relative to their values obtained with a very fine mesh. Figure 8 shows that for a mesh spacing μ approaching 1.5 Å, the reciprocal space calculation begins to report energies noticeably different from the values obtained in the limit of a very fine mesh. This behavior holds for the reciprocal space contributions to the virial trace as well; the off-diagonal elements of the virial accumulate very large errors (data not shown).

Despite these limitations, it appears that StME can hold its ground in most constant pressure simulations and in cases when the total system energy is required. In any periodic simulation cell, isotropic position rescaling can be used to adjust the cell size to satisfy a particular external pressure making use of only the trace of the virial tensor. In orthorhombic cells, even anisotropic rescaling can be accomplished without reference to the virial's off-diagonal elements. In our test cases, elements of the virial's trace are consistently biased by nearly the same amount for a given system and a particular set of SPME parameters.

In StME calculations, the strong anticorrelation between the errors of the two coarse meshes carries over into estimates of the energy and virial trace, so that the average of the two results is biased more consistently than either alone. It is likely that the necessary correction factors can be computed at the beginning of a simulation by comparing the results from both StME meshes to the results from a finer mesh or higher interpolation order computed with the same values of D_{tol} and L_{cut}. Periodic updates of the correction factors, perhaps every 10 000 steps or upon significant changes in the unit cell dimensions, appear to be a reliable means of keeping errors in the calculated energy and virial trace within the levels obtained with conventional SPME calculations and accepted parameters. To ensure that the accuracy in estimates of the energy and virial trace (after removing the bias) was comparable to the accuracy of forces in StME, we scanned over a large number of all four Ewald parameters for the streptavidin and COX-2 test cases. The results in Figure 9 confirm that, if StME produces accurate forces, it produces a precise virial trace and energy as well. Further explanation of why the energy and

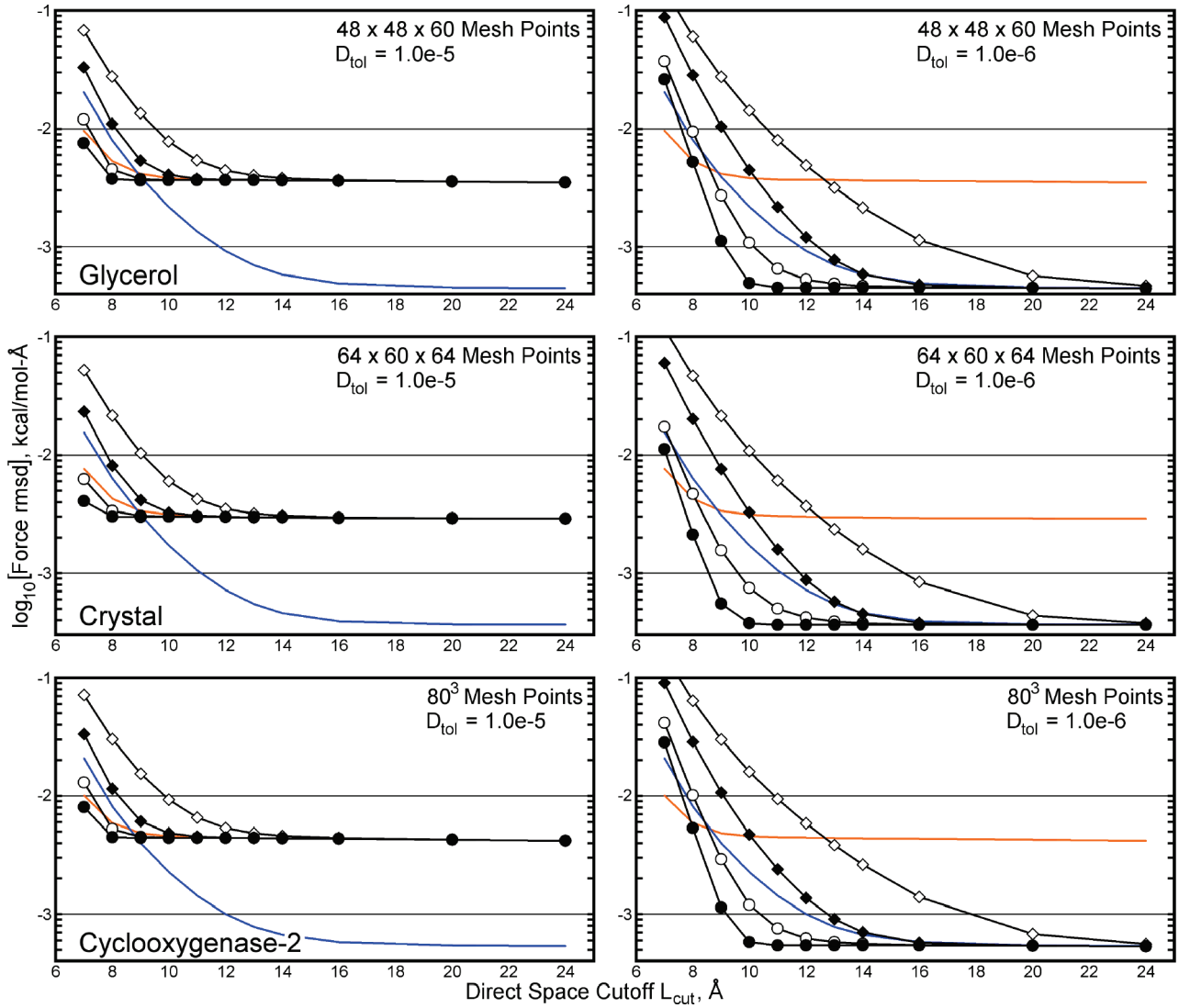


Figure 7. Accuracy of SPME and StME calculations for three other test cases. The format follows Figure 6 but only the case of a mesh spacing 1.5× the default value, that is μ approaching 1.5 Å, is shown. The 35% v/v glycerol:water mixture, protein lattice, and the solvated COX-2 dimer are simulated in monoclinic, orthorhombic, and truncated octahedral cells, respectively. For efficiency, the truncated octahedron is tiled and reshaped into a triclinic unit cell in dynamics simulations. Additional details of all systems can be found in Table 1. The StME method shows comparable performance relative to SPME across all test cases.

virial trace estimates are biased in the manner observed is provided in the Supporting Information.

5.5. Simple Metric for the Accuracy of Ewald Mesh Methods. As in shown in the Supporting Information, it is logical to compute the accuracy of the reciprocal space part of an Ewald mesh calculation as a function of σ/μ, where σ is the width of the Gaussian charge smoothing function defined in Equation S.1 of the Supporting Information and μ, again, is the mesh spacing. We did this for all four of our test cases by computing SPME or StME calculations for σ ranging from 0.5 to 3.0 Å, μ ranging from 0.9 to 2.0 Å, and fourth, fifth, or sixth order interpolation.

Force errors for each test case are plotted as a function of σ/μ in Figure 10 for StME calculations using fourth order interpolation and SPME calculations using fourth, fifth, or sixth order interpolation. In all cases, the accuracy of forces appears to approach a log-linear relationship with σ/μ; in this region, the accuracy of StME is roughly 1.2 orders of magnitude higher than SPME with identical parameters.

Because the accuracy in an Ewald mesh calculation also depends on contributions from the direct space calculation, we assumed that the entire electrostatics calculation could meet "AMBER" or "CHARMM" accuracy if the reciprocal space calculation produced errors up to half the level of either standard (this estimate is conservative, as the direct and reciprocal space forces for any given atom are generally oriented randomly with respect to one another, so the magnitude of the combined error will be at most the sum of the direct and reciprocal space errors). Our findings echo results in Figures 6 and 7: StME calculations using fourth order interpolation can meet the "AMBER" level of accuracy with $\sigma \geq 1.0\ \mu$, whereas SPME run with fourth order interpolation would require $\sigma \geq 1.5\ \mu$. This is reflected in the AMBER default parameters: $L_{cut} = 8.0$ Å and $D_{tol} = 1.0$

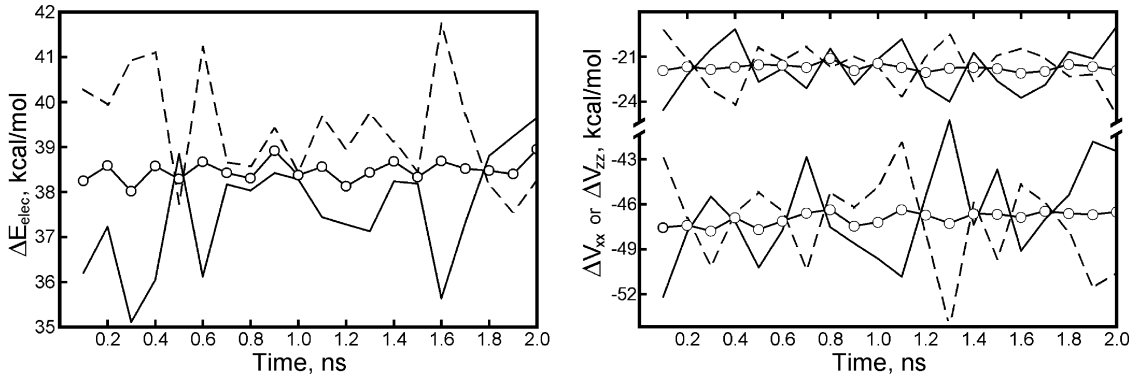


Figure 8. Removable bias in StME estimates of electrostatic energy and elements of the virial tensor trace. In the left panel, StME using 60^3 mesh points, 4th-order interpolation, $D_{tol} = 1.0 \times 10^{-5}$, and $L_{cut} = 8.0$ Å estimates the streptavidin test system's electrostatic energy 38.5 kcal/mol too high relative to a very accurate standard, on average, over the course of a 2 ns simulation. The StME error in the energy estimate is given by the solid line with open circles; plain solid and dashed lines show the error if either of the two StME meshes were used alone. In comparison, a standard SPME calculation run with the AMBER default parameters (a mesh of 90^3 points) delivers an error of only 0.6 ± 0.1 kcal/mol. However, the errors in the reciprocal space energy estimates of each StME mesh are strongly anticorrelated such that the overall error is very consistent: 38.5 ± 0.2 kcal/mol. If the 38.5 kcal/mol bias is removed by measuring against a high accuracy standard occasionally over the course of a simulation, the electrostatic energy can be consistently estimated to within the error of the AMBER default parameters. A similar treatment can be applied to derive the correct reciprocal space virial tensor trace, as shown for the COX-2 test case in the right panel (StME was performed with meshes of 80^3 points; the default AMBER parameters imply a mesh of 120^3 points for SPME). The format of the lines is the same; error in V_{xx} appears in the bottom half of plot, error in V_{zz} in the top half (error in V_{yy} is omitted for clarity).

$\times 10^{-5}$ implies that σ is approximately 1.42 Å to pair with $\mu \leq 1$ Å. Similarly, StME can achieve "CHARMM" accuracy using fourth order interpolation and $\sigma \geq 1.2\ \mu$. SPME with fourth order interpolation would require $\sigma \geq 2.0\ \mu$: this is reflected by the NAMD recommended parameters $L_{cut} = 12.0$ Å and $D_{tol} = 1.0 \times 10^{-6}$, implying $\sigma \approx 1.94$ Å for $\mu \leq 1$ Å. Figure 10 also shows that SPME calculations must use sixth-order interpolation to produce "AMBER" or "CHARMM" accuracy with the σ/μ ratios available to StME with fourth order interpolation.

In the Supporting Information, we show that there is a nearly linear relationship between σ and L_{cut} for a given D_{tol}, which implies that there is then a roughly linear relationship between the acceptable values of L_{cut} and μ for a given value of D_{tol}. In the next section, we will examine how the combination of μ, L_{cut}, and D_{tol} can be used to optimize performance in both SPME and StME.

5.6. Optimal StME and SPME Parameters. To define optimal Ewald parameters, we return to the "AMBER" and "CHARMM" accuracy standards as defined in Methods. Because different computing architectures favor different levels of real-space or reciprocal space calculations, we did intensive scans of L_{cut} between 6.0 and 16.0 Å, μ between 0.7 and 2.0 Å, and D_{tol} between 5.0×10^{-7} and 1.0×10^{-5} for both fourth and fifth order interpolation. Because the choice of D_{tol} does not affect execution time, we sought any value of D_{tol} that could satisfy the accuracy standards for given values of L_{cut} and μ. Results for SPME and StME methods are plotted in Figure 11.

While increasing the interpolation order from 4 to 5 will greatly expand the combinations of L_{cut} and μ that can produce a particular level of accuracy, staggered meshes with fourth order interpolation offer an even wider array of options. Figure 11 also confirms a result evident in Figures 6 and 7 that increasing the interpolation order with staggered meshes is of marginal benefit when seeking the AMBER level of accuracy but offers more significant improvements when seeking the higher CHARMM level of accuracy.

Although we do not have a working version of Staggered Mesh Ewald in an efficient molecular dynamics package, it is not difficult to obtain reasonable estimates of the single-processor efficiency of StME versus SPME. We assume that the costs of the occasional energy and virial bias corrections and the cost of averaging the forces obtained by the two reciprocal space calculations are negligible. If the two reciprocal space calculations share data, the computation of the reciprocal space pair potential $\hat{\theta}^{(rec)}$ need only be done once for both meshes, and there is even the possibility of using "harmonic averaging,"[17] combining the two staggered meshes in Fourier space to eliminate one of the four FFTs and one of the two force interpolation procedures for significant overall savings. However, we assumed that the two calculations must be done independently, because there may be benefits to parallel performance in this regard and independent calculations make the StME implementation trivial. Under these assumptions, the reciprocal space part of an StME calculation takes exactly twice as long as the identical SPME reciprocal space calculation. Tests were conducted on an Intel 2.66 GHz E5430 processor with the serial version of the pmemd module of AMBER10. Similar to findings presented by Crocker and co-workers in the development of their own parameter optimization program MDSimAid,[23] we were unable to significantly improve the performance of single-processor SPME calculations by simply adjusting the parameters. However, Tables 2 and 3, which also provide additional details of the molecular dynamics benchmark, shows that optimized StME parameters can perform somewhat better than optimized SPME param-

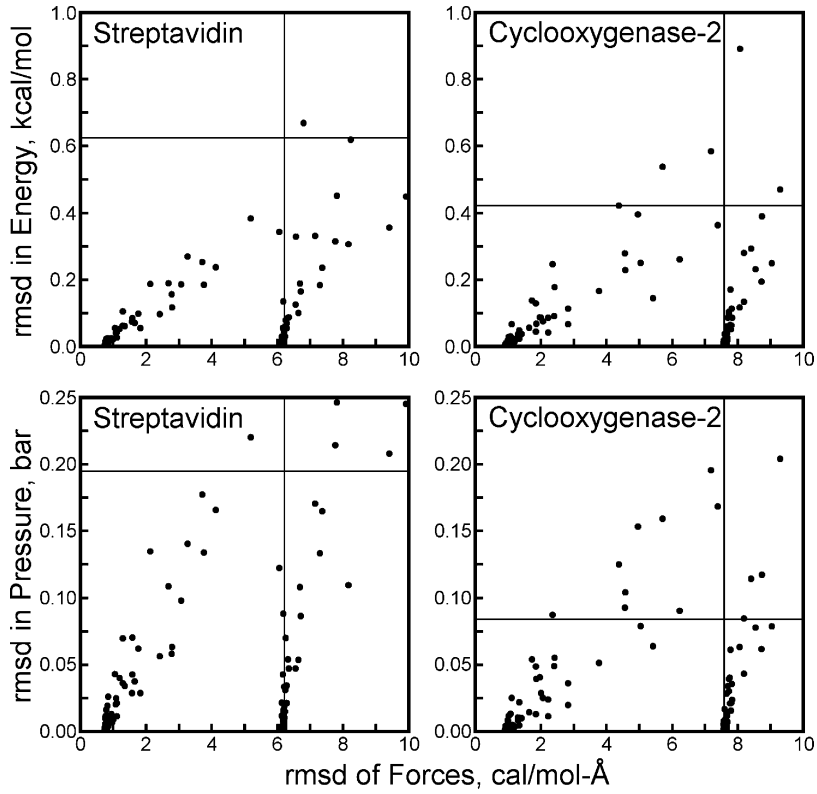


Figure 9. Error in energy and virial trace elements plotted against error in forces obtained by the StME method. Forces, energies, and virials were computed by StME for 20 snapshots of the streptavidin and COX-2 test cases for D_{tol} of 1.0×10^{-5} or 1.0×10^{-6}, μ ranging from 2.0 to 0.9 Å, L_{cut} ranging from 7 to 12 Å, and 4th- or 5th-order interpolation. Each point in the four plots above represents the results for a particular set of Ewald parameters: the root mean squared error in electrostatic energy or instantaneous pressure (after removal of any bias) is plotted against the average force rmsd for all 20 snapshots. The division of the points into two groups stems from the different values of D_{tol}. Crosshairs in each plot intersect at the error in force and error in pressure or energy obtained using the default AMBER parameters ($\mu \leq 1$ Å, 4th order interpolation, $L_{cut} = 8.0$ Å, and $D_{tol} = 1.0 \times 10^{-5}$) for each system. Although the errors in pressure may appear large, only the amplitude is plotted and after removal of bias the error in pressure from StME calculations can be positive or negative. The root mean squared deviations in the instantaneous pressure obtained for streptavidin and COX-2 over the course of each simulation were both 90 bar, fluctuating about an average of 1 bar; in this sense, the errors in pressure are a miniscule amount of extra noise.

eters to obtain either AMBER or CHARMM accuracy on the four test cases.

In contrast to single-processor performance, parallel scaling is difficult to predict. We expect that the performance advantage of StME will carry over into simulations on small numbers of processors and that the ability to reduce the overall FFT workload without increasing the direct space workload will give StME another advantage in highly parallel applications. The optimal parameters will change with the number of processors, and as shown in Figure 11, StME offers many choices. Parallel implementations of the SPME algorithm have been extensively optimized for parallel scaling by multiple independent groups;[13−15] while StME can likely benefit from much of this progress, it will take some effort to devise a parallel StME implementation that is as finely tuned to make a fair comparison with the best SPME implementations.

6. Discussion

In this Article, we have reviewed and renewed an old technique, interlacing, for improving the accuracy of particle:mesh calculations. While the technique was originally used to reduce the memory requirements of such calculations at the expense of simulation speed, our new implementation Staggered Mesh Ewald seems to confer some benefits to overall speed on modern computers. Nowadays, computer memory is plentiful but the ability to use smaller meshes may help to reduce the total communication cost of simulations in parallel applications. We will now discuss how mesh staggering might benefit other Ewald mesh methods, Poisson solvers, and molecular simulations.

6.1. Staggered Meshes for Other Poisson Solvers. Previously, mesh staggering has been shown to be effective with the Particle:Particle Particle:Mesh (P^3M) method,[17] and here, we have shown it to be effective with the Smooth Particle Mesh Ewald (SPME) method. The principal difference between these electrostatic methods is the shape of the charge smoothing function used in the mesh calculation: P^3M uses a spherical hypercone, whereas SPME uses a spherical Gaussian, which confers some advantage in accuracy[9] because the Gaussians are better at conserving the total amount of charge on the mesh. We expect that mesh staggering will also improve the accuracy of Gaussian Split Ewald (GSE) calculations,[9] which are nearly identical to

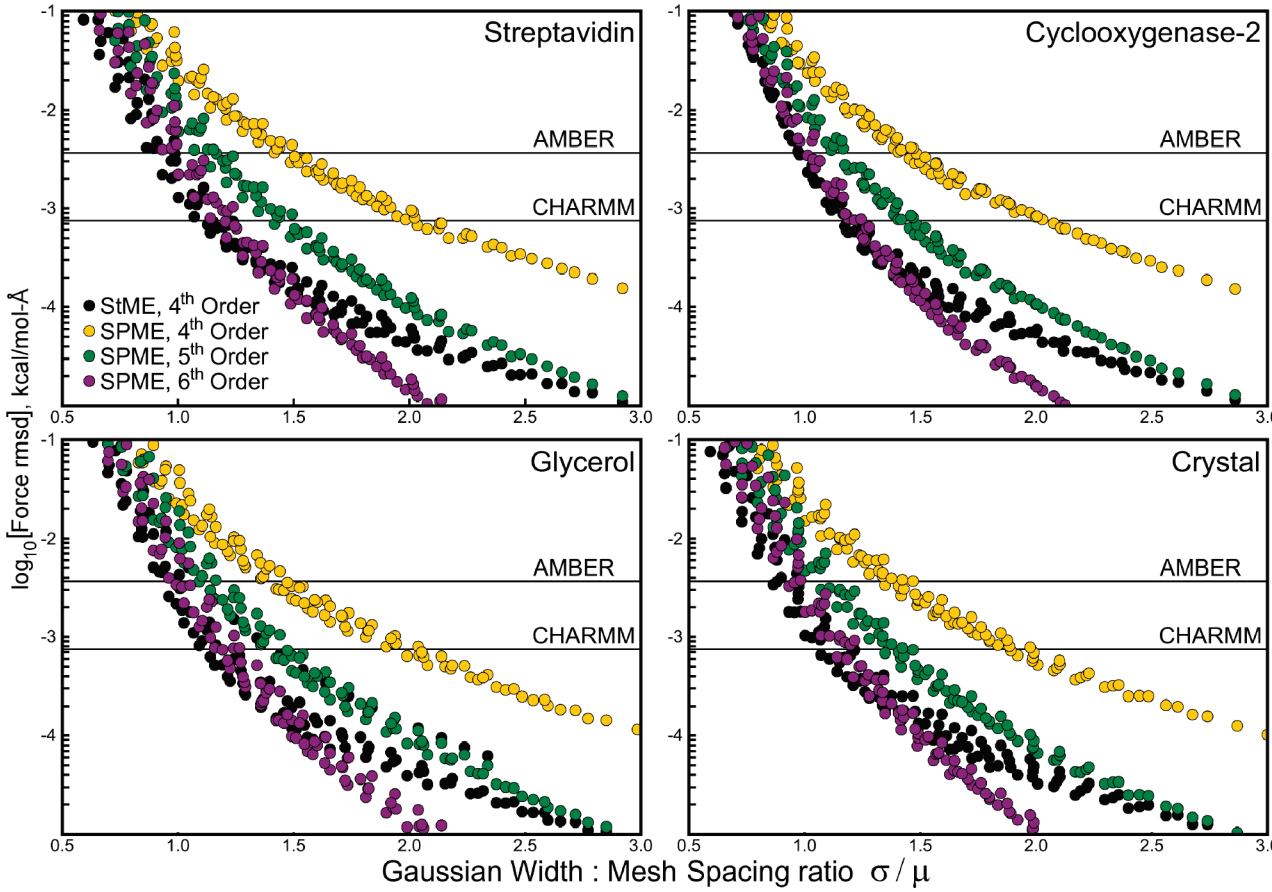


Figure 10. Accuracy of forces as a function of the ratio of charge smoothing to mesh spacing. In each of four systems, the accuracy of forces was computed for a range of values of direct space cutoff L_{cut} and mesh spacing μ. For all calculations, the direct sum tolerance D_{tol} was set to 1.0×10^{-9}, implying that errors in the forces came almost exclusively from the reciprocal space calculation. Equation S.7 of the Supporting Information was used to obtain the width of the Gaussian charge smoothing function in each calculation. Each plot is marked according to the "AMBER" and "CHARMM" accuracy standards, assuming that the reciprocal space calculation must create errors not in excess of half the level of each standard. The accuracy of the reciprocal space calculations as a function of σ/μ is a logical way to compare StME and SPME with different orders of interpolation: it indicates how aggressively the charges must be smoothed and hence provides an indication of how long L_{cut} must be for a particular μ. By this metric, StME with 4th order interpolation performs slightly better than SPME with 6th order interpolation to obtain the levels of accuracy sought in most molecular simulations.

SPME except in the function used to interpolate particles to the mesh (SPME uses a B-spline, whereas GSE uses another Gaussian, but B-splines in fact converge to Gaussians in the limit of high-interpolation order[24]).

Another possible application of mesh staggering is to real-space variants on the SPME method, which use the same particle interpolation and charge smoothing functions but solve Poisson's equation in real space.[10,11] These methods may prove more scalable than FFT-based methods on very high numbers of processors, but the principal drawback of these methods is the cost of smoothing the charge density in real space, a function of the number of mesh points. Staggered meshes can not only reduce the number of mesh points needed but also improve the efficiency of the charge smoothing procedure because the distances between corresponding mesh points are identical on both meshes. Such improvements may help close the gap between real-space and FFT-based Poisson solvers in Ewald calculations.

Fast Multipole methods (FMMs)[25,26] receive attention for the same reasons as real-space based Poisson solvers: the promise of $O(N)$ scaling and also exponentially reduced communication requirements as the interactions become increasingly long-ranged. While FMMs continue to be slower than FFT-based Poisson solvers for condensed phase molecular systems, as with real-space Poisson solvers, considerable progress has been made in recent years. It is possible that staggering the hierarchy of meshes used by FMMs may confer the same benefits as staggering the one mesh used by SPME or P^3M, again helping to close the performance gap between these methods and the standard particle:mesh techniques used in most molecular simulations.

While mesh staggering may have utility in other Poisson solvers, other approximations that have proven useful in standard Poisson solvers may be of utility in Staggered Mesh Ewald. In particular, the use of spherically truncated FFTs,[27] discarding very low-frequency modes in Fourier space much as interactions in the tail of the direct space sum are discarded in standard SPME, can decrease the cost and communication requirements of the FFT needed to take the charge Q mesh into Fourier space. By using spherically truncated FFTs and harmonic averaging[17] in the context of Staggered Mesh

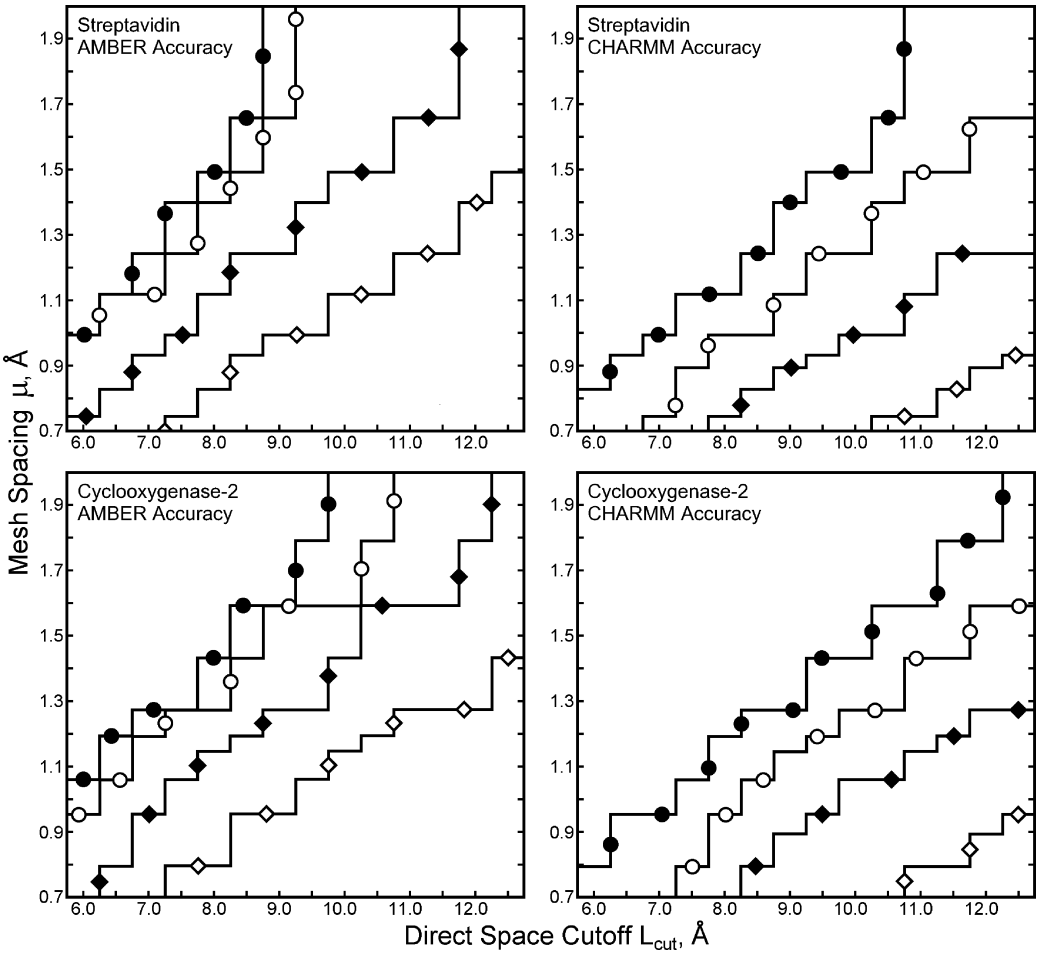


Figure 11. Ewald parameters yielding AMBER or CHARMM levels of accuracy in the streptavidin and COX-2 test cases. By scanning L_{cut} and μ for different orders of interpolation and optimizing D_{tol} for greatest accuracy in each case, we were able to determine the region of the L_{cut} and μ parameter space on which SPME or StME give acceptable levels of accuracy. The format of boundary lines in each panel follows Figure 6: diamonds denote the SPME method and circles denote the StME method, while open and filled symbols denote 4th and 5th order interpolation, respectively. Values of L_{cut} and μ below and to the right of each boundary line produce accurate forces according to the standard listed in each panel.

Table 2. Timings for Optimal Smooth Particle Mesh Ewald (SPME) Parameters and Estimated Timings for Staggered Mesh Ewald (StME) Parameters[a]

	run parameters			timings			
method	L_{cut}, Å	D_{tol}	mesh size	dir.[b]	rec.[c]	Total[d]	Factor[e]
			Streptavidin, AMBER accuracy				
SPME	8.5	4.0×10^{-6}	$96 \times 96 \times 96$	362	123	584	1.00
StME	6.5	4.0×10^{-6}	$80 \times 80 \times 80$	244	156	498	1.17
			Streptavidin, CHARMM accuracy				
SPME	10.0	5.0×10^{-7}	$120 \times 120 \times 120$	482	218	792	1.00
StME	8.5	8.0×10^{-7}	$80 \times 80 \times 80$	360	156	615	1.29

[a] Each system, described in Table 1, was run for 1000 steps in the NVT ensemble using a 1 fs time step, Berendsen thermostat, 10 Å cutoff on Lennard-Jones interactions, 2 Å nonbonded pairlist buffer, and the stated electrostatic parameters. The internal geometry of water molecules was constrained by SETTLE;[36] the lengths of other bonds to hydrogen were constrained by SHAKE.[37] Reciprocal space electrostatics were computed at every time step using the stated parameters and 4th-order interpolation. Timings for StME were estimated by doubling the reciprocal space calculation time of an SPME calculation run with the same parameters. This test used 4th-order interpolation exclusively because higher orders are very rarely used in practice and many codes, including the PMEMD version used for this test, use optimized routines for 4th-order interpolation. While we were able to obtain better overall run times by using higher orders of interpolation in the SPME runs, it would also be possible to improve the efficiency of StME runs if the two reciprocal space calculations were able to share data. This test is meant to offer a basic estimate of the efficiency of StME. [b] Direct space interaction computation time, including van der Waals interactions (all timings are in seconds). [c] Reciprocal space computation time. [d] Total simulation time, including nonbonded pairlist updates and bonded atom force calculations. [e] Overall rate of simulation, relative to SPME.

Ewald, the communication requirements and cost of evaluating the FFTs for the reciprocal space sum might be reduced even further.

6.2. Staggered Mesh Ewald for Highly Parallel Applications. We have not presented results for the performance of StME in the context of parallel molecular dynamics

Table 3. Timings for Optimal Smooth Particle Mesh Ewald (SPME) Parameters and Estimated Timings for Staggered Mesh Ewald (StME) Parameters[a]

	run parameters			timings			
method	L_{cut}, Å	D_{tol}	mesh size	dir.[b]	rec.[c]	total[d]	factor[e]
			glycerol, AMBER accuracy				
SPME	8.5	5.0×10^{-6}	$72 \times 72 \times 90$	219	86	369	1.00
StME	8.0	7.0×10^{-6}	$48 \times 48 \times 60$	200	68	332	1.11
			glycerol, CHARMM accuracy				
SPME	10.0	5.0×10^{-7}	$96 \times 96 \times 120$	284	190	534	1.00
StME	8.5	9.0×10^{-7}	$64 \times 64 \times 80$	222	119	405	1.32
			crystal, AMBER accuracy				
SPME	8.0	7.0×10^{-6}	$90 \times 96 \times 96$	487	117	747	1.00
StME	7.5	8.0×10^{-6}	$64 \times 60 \times 64$	462	100	700	1.07
			crystal, CHARMM accuracy				
SPME	10.0	1.0×10^{-6}	$120 \times 108 \times 120$	588	200	926	1.00
StME	9.0	1.0×10^{-6}	$72 \times 64 \times 72$	540	117	795	1.16
			cyclooxygenase-2, AMBER accuracy				
SPME	8.5	5.0×10^{-6}	$120 \times 120 \times 120$	614	330	1127	1.00
StME	7.0	5.0×10^{-6}	$100 \times 100 \times 100$	458	417	1056	1.07
			cyclooxygenase-2, CHARMM accuracy				
SPME	10.0	8.0×10^{-7}	$160 \times 160 \times 160$	782	719	1696	1.00
StME	9.0	8.0×10^{-7}	$100 \times 100 \times 100$	675	421	1281	1.32

[a] The format, molecular dynamics benchmark protocol, and labeling follows Table 2. StME performs better than SPME across all systems studied, though parallel scaling for an optimized StME implementation has not yet been tested. [b] Direct space interaction computation time, including van der Waals interactions (all timings are in seconds). [c] Reciprocal space computation time. [d] Total simulation time, including nonbonded pairlist updates and bonded atom force calculations. [e] Overall rate of simulation, relative to SPME.

simulations because we do not yet have a working implementation in this respect. We have shown that StME offers moderate performance improvements on single-processor simulations, which can be expected to carry over into parallel applications on small numbers of processors. The scaling of highly parallel algorithms is difficult to predict, but we will address three of the most critical aspects of the reciprocal space calculation with respect to parallel implementations and StME.

On large numbers of processors, the scaling of the Ewald reciprocal space calculation is limited primarily by the number of messages that must be passed between processors to accomplish the FFT operations for convoluting the charge mesh Q with the reciprocal space pair potential $\theta^{(rec)}$. If P nodes are working together to compute an FFT, each node must send its part of the FFT data to all other nodes, and receive FFT data from all other nodes. (It is for this reason that most codes devote a subset of processors to the reciprocal space calculation.) Because StME requires up to 40% less FFT work than regular SPME, the FFTs could be performed on a smaller subset of processors, implying fewer messages to pass.

But, reducing the amount of mesh data can imply other communication costs. A second important factor in the cost of a parallel reciprocal space calculation is the cost of communicating the coordinates and identities of atoms in order to construct the charge mesh Q, before any FFTs take place at all. In SPME with nth order interpolation, each atom influences a rectangular region of the mesh Q that is $n\mu$ points on each side. As n or μ increases, more atoms must be therefore imported from further away in order to consruct Q. As was shown in the results, StME with fourth order interpolation produces similar accuracy to SPME with sixth order interpolation, all other parameters being equal. StME could therefore make use of a large μ such as 1.5 Å with smaller import regions for constructing Q than SPME would require. However, the import regions would still be somewhat larger than those required by an SPME calculation using the typical $\mu = 1$ Å.

A third factor that influences the cost of a parallel reciprocal space calculation is the actual cost of constructing Q and then interpolating forces from the potential $Q \bigstar \theta^{(rec)}$. These particle $\leftrightarrow$ mesh operations, which can be more expensive than the FFTs themselves (data not shown), are typically performed on the same processors that will do the convolution $Q \bigstar \theta^{(rec)}$. Because StME essentially doubles the cost of the particle $\leftrightarrow$ mesh operations, it may be difficult to reduce the number of processors devoted to the FFT operations. However, the particle $\leftrightarrow$ mesh operations can be done on the more numerous processors devoted to the direct space calculation so that the grid data itself could be communicated to a subset of processors for computing the convolution. NAMD[15] is already equipped to run traditional SPME calculations by passing mesh data, not coordinates, to reciprocal space processors when the highest possible scaling is desired; such a reorganization may be necessary to make StME beneficial to highly parallel applications.

6.3. Continued Improvement of Ewald Mesh Methods. Ewald mesh methods will likely remain an important tool for molecular simulations well into the future. Briefly, calculating long-ranged Coulomb electrostatics currently accounts for a majority of the total simulation time, and will continue to do so even as classical models begin to incorporate other charge geometries and explicit polarization effects. While quantum effects are undoubtedly important

for the interaction of charges at very short-range, the Coulomb approximation quickly takes over even on molecular scales. Furthermore, numerous studies have shown that the periodic boundary conditions enforced by Ewald mesh methods are relatively benign, especially in comparison to some alternatives. We will describe the rationale for continued development of Ewald electrostatics in more detail, then outline other avenues to accelerating Ewald calculations which we hope will dramatically accelerate traditional Ewald mesh calculations and serve as a powerful complement to the Staggered Mesh Ewald method.

The majority of the computational effort in molecular dynamics simulations is devoted to electrostatic nonbonded interactions, whereas the calculation of van der Waals dispersion interactions, typically by Lennard-Jones potentials, is minor in comparison. Primarily, this is because electrostatic interactions are very long ranged. Moreover, because of the form of the electrostatic potential, analytic electrostatic force computations require not only a divide operation but also an expensive square root operation to obtain the quantity $[k_{coul}q_iq_j/r_{ij}^3]\mathbf{r}_{ij}$, where k_{coul} is Coulomb's constant, q_i and q_j represent charges of the atoms i and j, and $\mathbf{r}_{ij}$ is the vector between the charges. In contrast, the Lennard-Jones force requires only a divide operation to compute the quantity $[A_{ij}/r_{ij}^8 + B_{ij}/r_{ij}^{14}]\mathbf{r}_{ij}$, where A_{ij} and B_{ij} are constants. A third factor that makes electrostatic calculations dominate the cost of simulations is a peculiarity of current water models, most of which give Lennard-Jones attributes to only the oxygen atoms while placing charges on at least three sites. Because water makes up the majority of the system in most simulations of solvated proteins, there can be many more electrostatic interactions than Lennard-Jones interactions for a given cutoff.

While point charges may not be an adequate representation of atomic charge distributions at close range,[28] Ewald mesh methods are also compatible with other charge geometries. One must merely recall that the direct space interactions are a modification to the electrostatic potential of the smoothed charge distribution computed in the reciprocal space calculation, as discussed in the Introduction and Methods. The direct space modification can just as easily be used to extract interactions of distributed charges, so long as the interaction of two charges in the actual system and the interaction of two Gaussian charges converge at the direct space cutoff. Even if it does not perfectly describe the interaction of subatomic particles at close range, Coulomb's law is still valid for the interaction of charges on the nanometer scale. Therefore, long-ranged electrostatic methods such as the Ewald sum will continue to be essential for molecular simulations, even as new force fields with different local electrostatic approximations and even explicit polarization effects[29] come into use.

There is debate over whether periodicity imposed by Ewald electrostatics is suitable for molecular simulationss,[30,31] and while such a representation may be much more appropriate for crystal lattice simulations,[32] periodic boundary conditions are a very practical solution for simulations of proteins in boxes of water as well. While there are relevant concerns when using periodic boundary conditions with very small systems,[33] finite size effects are by no means limited to periodic systems. Simulations performed in both periodic and nonperiodic unit cells such as droplets[34] or ice shells[35] suggest that periodic boundary conditions are as good or better than numerous alternatives.

Given the importance of Ewald mesh methods to molecular simulations, further developments that permit the use of coarser meshes or reduce the required number of direct space computations are of great interest. In this communication, we have analyzed the errors arising from a coarse mesh in terms of self-image forces and pair interaction force errors. The self-image forces we identified can be corrected on a per-atom basis for low-density plasma simulations, but the pair interaction force errors arising from a coarse mesh require more extensive corrections in condensed-phase simulations. Our solution was to introduce a second mesh calculation, staggered relative to the original. It may also be possible to modify the form of the Ewald "switching" function used to make the transition between the reciprocal space electrostatic potential and the direct space modification. The form used in all Ewald mesh methods to date, most apparent in Equation 1, is dictated by the form of the charge smoothing function, a Gaussian as described in Equation S1 of the Supporting Information. The typical direct space potential satisfies the most important property of an Ewald switching function in that it smoothly vanishes within a reasonable distance, while the associated Gaussian function ensures that charges can be mapped to a mesh with reasonable accuracy. However, it may be possible to design new charge smoothing and potential switching functions that map charges more accurately to coarser meshes or vanish more rapidly. We are pursuing new ways to satisfy these criteria and expect the results to be generally useful for all types of Ewald mesh calculations.

Acknowledgment. D.S.C. thanks Dr. Kristina Furse for the use of her COX-2 trajectory, Peter L. Freddolino and Dr. James C. Phillips for helpful conversations, and Dr. Jessica M.J. Swanson for reading the manuscript. This research was supported by National Institutes of Health Grant GM080214.

Supporting Information Available: Descriptions of the relationships between Gaussian charge smoothing width σ, direct sum tolerance D_{tol}, and the direct space cutoff L_{cut}, detailed description of the Ewald reciprocal space calculation, investigation of the sources of self-image force errors, pair interaction force errors, and biased energy and virial estimates inherent in SPME reciprocal space calculations using coarse meshes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Beck, D. A. C.; Daggett, V. Methods for Molecular Dynamics Simulations of Protein Folding/Unfolding in Solution. *Methods* **2004**, *34*, 112–120.

(2) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(3) DeLeeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of Electrostatic Systems in Periodic Boundary Conditions. I. Lattice Sums and Dielectric Constants. *Proc. R. Soc. Lond. Ser. A* **1980**, *373*, 27–56.

(4) Yonetani, Y. a Severe Artifact in Simulation of Liquid Water Using a Long Cut-off Length: Appearance of a Strange Layer Structure. *Chem. Phys. Lett.* **2005**, *406*, 49–53.

(5) Patra, M.; Karttunen, M.; Hyvönen, M. T.; Falck, E.; Lindqvist, P.; Vattulainen, I. Molecular Dynamics Simulations of Lipid Bilayers: Major Artifacts Due to Truncating Electrostatic Interactions. *Biophys. J.* **2003**, *84*, 3636–3645.

(6) Pollock, E. L.; Glosli, J. Comments on P^3M, FMM, and the Ewald Method for Large Periodic Coulombic Systems. *Comput. Phys. Commun.* **1996**, *95*, 93–110.

(7) Hockney, R. W.; and Eastwood, J. Collisionless Particle Models. In *Computer Simulation Using Particles*; Taylor and Francis Group: New York, 1988; pp. 260−291.

(8) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. H. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(9) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian Split Ewald: A Fast Ewald Mesh Method for Molecular Simulation. *J. Chem. Phys.* **2005**, *122*, 054101.

(10) Sagui, C.; Darden, T. Multigrid Methods for Classical Molecular Dynamics Simulations of Biomolecules. *J. Chem. Phys.* **2001**, *114*, 6578–6591.

(11) Beckers, J. V. L.; Lowe, C. P.; De Leeuw, S. W. An Iterative PPPM Method for Simulating Coulombic Systems on Distributed Memory Parallel Computers. *Mol. Simulat.* **1998**, *20*, 369–383.

(12) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(13) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California, San Francisco: San Francisco, CA, 2008.

(14) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossváry, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; and Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Conference on High Performance Networking and Computing*; Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, Tampa, FL, USA 2006; Association for Computing Machinery: New York 2006.

(15) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Sand Chulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(16) Chen, L.; Langdon, B.; Birdsall, C. K. Reduction of the Grid Effects in Simulation Plasmas. *J. Comput. Phys.* **1974**, *14*, 200–222.

(17) Eastwood, J. Optimal P^3M Algorithms for Molecular Dynamics Simulations. In *Computational Methods in Classical and Quantum Physics*; Hooper, M. B., Ed.; Advance Publications Ltd: London, U.K., 1976; pp 206−228.

(18) Hyre, D. E.; Le Trong, I.; Merritt, E. A.; Green, N. M.; Stenkamp, R. E.; Stayton, P. S. Wildtype Core-Streptavidin with Biotin at 1.4 Å. *Protein Sci.* **2006**, *15*, 459–467.

(19) Smith, G. D.; Blessing, R. H.; Ealick, S. E.; Fontecilla-Camps, J. C.; Hauptman, H. A.; Housset, D.; Langs, D. A.; Miller, R. Ab Initio Structure Determination and Refinement of a Scorpion Protein Toxin. *Acta Crystallogr. D* **1997**, *53*, 551–557.

(20) Kiefer, J. R.; Pawlitz, J. L.; Moreland, K. T.; Stegeman, R. A.; Gierse, J. K.; Stevens, A. M.; Goodwin, D. C.; Rowlinson, S. W.; Marnett, L. J.; Stallings, W. C.; Kurumbail, R. G. Structural Insights into the Stereochemistry of the Cyclooxygenase Reaction. *Nature* **2000**, *405*, 97–101.

(21) Chelli, R.; Procacci, P.; Cardini, G.; Cella; Valle, R. G.; Califano, S. Glycerol Condensed Phases Part I. A Molecular Dynamics Study. *Phys. Chem. Chem. Phys.* **1999**, *1*, 871–877.

(22) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712–725.

(23) Crocker, M. S.; Hampton, S. S.; Matthey, T.; Izaguirre, J. A. MDSIMAID: Automatic Parameter Optimization in Fast Electrostatic Algorithms. *J. Comput. Chem.* **2005**, *26*, 1021–1031.

(24) Unser, M.; Aldourbi, A.; Eden, M. On the Asymptotic Convergence of B-Spline Wavelets to Gabor Functions. *IEEE Trans. Inf. Theory* **1992**, *38*, 864–872.

(25) Lu, B.; Cheng, X.; McCammon, J. A. "New-Version-Fast-Multipole-Method" Accelerated Electrostatic Calculations in Biomolecular Systems. *J. Comput. Phys.* **2007**, *226*, 1348–1366.

(26) Lu, B.; Cheng, X.; Huang, J.; McCammon, J. A. Order *N* Algorithm for Computation of Electrostatic Interactions in Biomolecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 19314–19319.

(27) Fang, B.; Martyna, G.; Deng, Y. A Fine Grained Parallel Smooth Particle Mesh Ewald Algorithm for Biophysical Simulation Studies: Application to the 6D-Torus QCDOC Supercomputer. *Comput. Phys. Commun.* **2007**, *177*, 362–377.

(28) Paricaud, P.; Predota, M.; Chialvo, A. A.; Cummings, P. T. From Dimer to Condensed Phases at Extreme Conditions: Accurate Predictions of the Properties of Water by a Gaussian Charge Polarizable Model. *J. Chem. Phys.* **2005**, *122*, 244511.

(29) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.

(30) Villareal, M. A.; Montich, G. G. On the Ewald Artifacts in Computer Simulations. The Test-Case of the Octaalanine Peptide with Charged Termini. *J. Biomol. Struct. Dyn.* **2005**, *23*, 135–142.

(31) Hünenberger, P. H.; McCammon, J. A. Ewald Artifacts in Computer Simulations of Ionic Solvation and Ion−Ion Interaction: A Continuum Electrostatics Study. *J. Chem. Phys.* **1999**, *110*, 1856–1872.

(32) Cerutti, D. S.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P. Simulations of a Protein Crystal: Explicit Treatment of Crystallization Conditions Links Theory and Experiment in the Streptavidin-Biotin Complex. *Biochemistry* **2008**, *47*, 12065–12077.

(33) Weerasinghe, S.; Smith, P. E. A Kirkwood−Buff Derived Force Field for Mixtures of Urea and Water. *J. Phys. Chem. B* **2003**, *107*, 3891–3898.

(34) Freitag, S.; Chu, V.; Penzotti, J. E.; Klumb, L. A.; To, R.; Hyre, D.; Le Trong, I.; Lybrand, T. P.; Stenkamp, R. E.; Stayton, P. S. A Structural Snapshot of an Intermediate on the Streptavidin−Biotin Dissociation Pathway. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 8384–8389.

(35) Riihimäki, E. S.; Martínez, J. M.; Kloo, L. An Evaluation of Non-periodic Boundary Condition Models in Molecular Dynamics Simulations Using Prion Octapeptides As Probes. *THEOCHEM* **2005**, *760*, 91–98.

(36) Miyamoto, S.; Kollman, P. A. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(37) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C.; Hirasawa, K. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1997**, *23*, 327–341.

CT9001015

JCTC Journal of Chemical Theory and Computation

Capturing the Trans Influence in Low-Spin d^8 Square-Planar Platinum(II) Systems using Molecular Mechanics

Anna. E. Anastasi and Robert J. Deeth*

Inorganic Computational Chemistry Group, Department of Chemistry, University of Warwick, Coventry CV4 7AL, U.K.

Received April 1, 2009

Abstract: Molecular modeling of coordination complexes continues to present challenges for force field methods. Implicit or explicit treatment of the significant d electron effects is mandatory. Ligand field molecular mechanics is designed for coordination complexes by explicitly including the ligand field stabilization energy (LFSE) and it is applied here to model the trans influence in tetracoordinate Pt^{II} complexes of general formulas PtX_4, PtX_3Y, *cis*-PtX_2Y_2, and *trans*-PtX_2Y_2, where X and Y are OH_2, H^-, Cl^-, Br^-, PR_3, SH_2, NR_3, and pyridine. Parameters have been developed within the Merck molecular force field using DFT structures and energies as reference data. Both geometric changes and relative energies are generally well-reproduced although PH_3 and H^- complexes show deviations. However, for phosphine complexes, replacing PH_3 with PMe_3 resolves all bar one of these. The LFSE associated with the low-spin d^8 configuration ensures planar coordination and provides an electronic connection between all the ligands, thus enabling a correct description of the trans influence. The parameters developed for NR_3 and PR_3 with R = H work well for R = Me and Et and, in agreement with experimental and/or DFT structures, display either a tetrahedral distortion or even ligand dissociation.

Introduction

The trans influence was first defined by Pidcock et al. in 1966[1] as "the extent to which a ligand weakens the bond trans to itself in the equilibrium state of a substrate". Trans influences manifest as changes in metal−ligand (M−L) bond lengths, IR frequencies, and/or NMR chemical shifts and are routinely observed in transition metal complexes, particularly planar platinum(II) species.

One of the earliest studies[2] compared Pt^{II}−Cl bond lengths in a number of complexes and proposed the following trans influence series: $R_3Si^- > H^- > PR_3 > C{=}C$, $Cl^- > O(acac)$. The larger trans influence correlates with increased metal−ligand covalency. The series has since been significantly extended.

Pearson rationalized the trans influence in terms of "antisymbiosis".[3] Over a decade earlier, Jorgensen had defined "symbiosis" as the tendency of a hard base to retain its electrons, thus keeping the attached metal hard, too.[4] Conversely, a soft base will transfer charge to the metal, making it soft. Antisymbiosis is the opposite: a hard base bonded to a central atom encourages coordination from a soft base and vice versa.

In molecular orbital terms, the trans influence is most often described as a competition between the two trans ligands for a single, metal-based orbital. One ligand donating strongly into this (initially empty) orbital effectively pre-empts the other, leading to a relative weakening of the latter's bonding. The metal orbital is generally regarded as an sd hybrid utilizing the highest-energy metal d orbital, which, in planar d^8 complexes and assuming the *x* and *y* axes lie long the Pt−L bonds, corresponds to $d_{x^2-y^2}$.[5] Thus, the trans influence is dominated by M−L σ-bonding.[1]

Alternatively, the trans influence can be rationalized by hypervalent valence bond theory.[6] Here, the contribution from the two resonance structures depends on the abilities of X: and Y: to support the lone pair.

* To whom correspondence should be addressed. E-mail: r.j.deeth@warwick.ac.uk. Website: http://warwick.ac.uk/go/iccg.

10.1021/ct9001569 CCC: $40.75
Published on Web 08/20/2009

$$X{:}Pt{-}Y \leftrightarrow X{-}Pt{:}Y$$

Thus, if Y is electronegative and (isolated) Y: is relatively stable, the right-hand structure is favored and the X−Pt bond is short.

The mutual interplay between ligands on opposite sides of the coordination center is clearly an electronic effect. Hence, quantitative theoretical descriptions of the trans influence have largely been based on quantum chemical methods like density functional theory (DFT). However, while DFT has clearly revolutionized the application of quantum mechanics (QM) in transition metal chemistry, current functionals are still not perfect. In addition, DFT is compute-intensive and therefore relatively slow.

In contrast, classical molecular mechanics (MM) is much faster, but conventional MM does not explicitly account for d electron effects and thus may not be generally suited to handle transition-metal complexes. We have therefore developed ligand field molecular mechanics (LFMM), which includes the ligand field stabilization energy (LFSE) directly.[7] LFMM thus describes the metal−ligand coordination better and can deliver DFT-quality results but up to 4 orders of magnitude faster.

Previous applications of LFMM have spanned Jahn-Teller effects in Cu(II) complexes[8,9] and spin state energetics of simple Co(III) species.[10] Here, we present our first complete attempt to describe the trans influence by modeling tetracoordinate Pt^{II} complexes of general formulas PtX_4, PtX_3Y, *cis*-PtX_2Y_2, and *trans*-PtX_2Y_2, where X and Y are OH_2, H^-, Cl^-, Br^-, PR_3, SH_2, NR_3, and pyridine. Since experimental data are available for only a few of the possible complexes, we compare the LFMM results with DFT structures and energies. The LFMM provides a satisfactory description of both the trans influence in $[PtX_3Y]^n$ systems and the relative energies of cis and trans forms of $[PtX_2Y_2]^m$.

Computational Details

Quantum Mechanical Calculations. Density functional optimizations were carried out using the Amsterdam Density Functional suite of programs (version 2006.01).[11] Structures were preoptimized using the local density approximation with a triple-ζ plus polarization basis set (TZP) on all atoms and a scalar ZORA relativistic correction.[12−14] Structures were then fully optimized using a gradient corrected functional, as described later, and include solvent effects based on the conductor-like screening model (COSMO) implemented in ADF.[15] COSMO radii were taken from Allinger's MM3 force field scaled by 0.833.[16,17] Frequencies were calculated numerically.[18,19]

QM charges were computed via the CHelpG method. To maintain compatibility with the way partial atomic charges are implemented in MMFF94, the Hartree−Fock approximation (as apposed to DFT) was used, as implemented in Gaussian 03[20] with 6-31G(d) basis sets on the nonmetal atoms. The LanL2DZ basis set with an extra set of f functions using the exponents determined by Frenking et al. and the accompanying frozen core were used for Pt.[21] Gaussian 03 only has van der Waals radii for elements up to Ar. In the AMBER charge scheme, and in GAMESS-US, heavier atoms are given a radius of 1.8 Å and we have followed this example. Bromine is given a radius of 2.3 Å, also consistent with the GAMESS-US approach.

Table 1. Pt−L Bond Charge Increments

ligand type	bci
OH2	0.2747
N	0.4121
NPYD	0.5531
CL-	0.3631
BR-	0.3939
HYDR	0.4347
S	0.5064
P	0.6683

Molecular Mechanics Parameters. The ligand field molecular mechanics (LFMM) treatment of charges was based on the MMF94 force field.[22] MMFF94 partial atomic charges employ the bond charge increment (bci) scheme, where the base charge q on an atom is modified by the bci value of each attached atom. New bci values for the metal, M, and the ligand donor atoms, L, were determined as follows. The bci for L is a quarter of the decrease in the QM-calculated Pt charge but then scaled relative to the ratio of the QM and LFMM proton charges for HL^+.[23]

Thus, the new bcis are given by

$$\mathrm{bci(L)} = 1/4\Delta q_{\mathrm{Pt}}^{\mathrm{MM}}(\mathrm{PtL_4}) = 1/4\Delta q_{\mathrm{Pt}}^{\mathrm{QM}}(\mathrm{PtL_4})\left(\frac{\Delta q_{\mathrm{H}}^{\mathrm{MM}}(\mathrm{HL^+})}{\Delta q_{\mathrm{H}}^{\mathrm{QM}}(\mathrm{HL^+})}\right)$$

where $\Delta q_{\mathrm{H}}^{\mathrm{MM}}(HL^+)$ is the change in proton charge of the protonated ligand as given by MMFF, $\Delta q_{Pt}^{QM}(PtL_4)$ is the change in the CHelpG Pt^{2+} charge for a $[PtL_4]$ complex, and $\Delta q_{\mathrm{H}}^{\mathrm{QM}}(HL^+)$ is the change in charge on H^+ for a protonated L. The final bci values are collected in Table 1.

In addition to the standard MMFF94 parameters, additional LFMM-specific parameters are required. In particular, and as described elsewhere,[24] the ligand field stabilization energy (LFSE) term is defined in terms of angular overlap model (AOM) parameters e_σ, $e_{\pi x}$, $e_{\pi y}$, and e_{ds}; the M−L stretch employs a Morse function while the angular geometry about the metal center is described via a ligand−ligand repulsion term. LFMM geometry optimizations employed Dommi-MOE,[24] our extended version of the Molecular Operating Environment.[25]

The AOM parameters, e_λ, were estimated by fitting the appropriate expressions to DFT "d" orbital energies from a "spherical configuration" calculation based on the DFT-optimized homoleptic structure. For d^8 Pt(II) species, this involves assigning the molecular orbitals of mainly Pt d parentage equal occupancies of 1.6. For planar PtL_4 systems, the d-orbital splitting yields three degrees of freedom and there are at most three AOM parameters assuming $e_{\pi x} = e_{\pi y}$. The Morse and ligand−ligand repulsion parameters were then roughly optimized using the penalty function approach proposed by Norrby and co-workers,[26] with the DFT-optimized geometries and selected frequencies as target data. All parameters were then further manually refined to give appropriate geometries for mixed ligand complexes. Full

listings of the LFMM parameters are available in the Supporting Information.

Results and Discussion

Despite the absence of an explicit LFSE term, conventional MM has been applied to the trans influence in planar d^8 systems. However, some intervention by the user may be necessary. For example, Rappé et al. have included Pt^{II} parameters in their universal force field.[27] The necessary trans coupling is achieved by varying the bond order of some ligands; for example, carbonyls are generally given a bond order of 2, but if they are trans to a ligand with a high trans influence then a reduced order of $1^1/_2$ is used. For metal−phosphine bonds, the bond order of the M−P bond is varied depending on whether a trans influence should be present.

A generalization of this approach appeared during the course of our study. An extended version of VALBOND,[28] VALBOND-TRANS,[29] incorporates Landis's hypervalence ideas such that the "normal" VALBOND parameters are modified to include contributions from both the ligand and its trans partner. VALBOND-TRANS was applied to octahedral organometallic compounds relevant to various catalytic processes and shows good agreement with DFT and experimental data. However, VALBOND-TRANS explicitly modifies the parameters to reflect the molecule and does not treat deviations in the A−M−B angle from 180°.

Another way of achieving the trans influence is to include a direct ligand−ligand distance term, which spans the intervening metal center. We made a preliminary investigation on model Pt complexes using an early form of LFMM[30] but did not take it any further. The following year, Brandt et al.[31] provided a more complete application of this approach to a series of six-coordinate Ru(II)−polypyridyl complexes, demonstrating that trans influences could be successfully treated with an additional, explicit N−N distance term. In contrast, LFMM tries to capture the trans influence implicitly and makes no assumption that the trans ligands must define a bond angle of 180° at the metal.

Another important feature of planar d^8 systems is their planarity. Landis and co-workers use a Fourier angular potential energy term to accommodate both 90° and 180° L−M−L bond angles in planar Rh(I) complexes,[32] while Cundari et al. have extended their MM2 force field to include three Pt−L atom types (where L = Cl^-, NR_3, CO_2^-) and enforce the planarity of Pt^{II} by including a lone pair on the + and − directions along the *z*-axis.[33] They report some success with their method, with average rms Pt−L differences of 0.08 ± 0.05 Å. However, they generally look at cis complexes and do not assess whether their method is also suitable for trans complexes.

The planar structure associated with the low-spin d^8 configuration can also be rationalized in terms of the LFSE. We have already shown that LFMM automatically generates the correct planar structures for low-spin four-coordinate Ni(II) amine complexes.[34] The observation that the trans influence involves metal-based sd hybrids interacting with ligand orbitals led us to wonder whether the treatment of sd^n hybridization implicit in the angular overlap model (AOM) d−s mixing term included in our model would provide a basis for capturing the trans influence as well.

Choice of Functional. The DFT structures and energies form the target data for the LFMM parameter optimization. Several X-ray crystal structures are available for $[PtCl_4]^{2-}$ in the Cambridge Structural Database, which we access via the EPSRC Chemical Database Service.[35] These provide a reliable estimate of the Pt−Cl distance of 2.30 ± 0.01 Å. We used this value in a preliminary screen of a range of functional/relativistic corrections/solvation combinations (see Supporting Information, Table S1) and eliminated those that gave an error greater than 0.12 Å. At this stage, it was clear that COSMO and relativistic corrections are important, but a range of common functionals survive.

Applying the same selection criterion to $[PtBr_4]^{2-}$ removes all the gradient-corrected functionals except OPBE (Supporting Information, Table S2). The computed Pt−N distances in $[Pt(NH_3)_4]^{2+}$ display a much smaller spread than either halide complex, and the gradient-corrected functionals, including OPBE, all give bond lengths within 0.03 Å of the experimental value of 2.05 Å (Supporting Information, Table S3). As a final test of the OPBE functional, the structures of *cis*- and *trans*-$[PtCl_2(NH_3)_2]$ were compared (Supporting Information, Table S4). The OPBE functional continues to provide a good description of the structure as well as predict that the trans isomer should be more stable than cis. Our favored DFT protocol is therefore OPBE/TZP/ZORA/COSMO.

The structures of all the homoleptic species were therefore optimized at the OPBE/TZP/ZORA level with a COSMO correction to model condensed phase effects.[36] Where more than one orientation of a ligand was possible, e.g., for $[Pt(OH_2)_4]^{2+}$, several geometries were optimized and the lowest energy one was used as the target geometry for the LFMM parameters. All possible ligand combinations of PtX_nY_{4-n} were optimized and added to the set of the target structures. Note that throughout the following the total charge is, for convenience, omitted from the molecular formulas.

A summary of results comparing geometries and energies from DFT and LFMM calculations is presented below.

In evaluating the LFMM data we consider several features: (1) the extent to which some Pt−L bonds change relative to their values in the homoleptic species, (2) the relative cis and trans Pt−X distances for species of the formulas PtX_3Y and the relative Pt−X and Pt−Y distances for both isomers of PtX_2Y_2, and (3) the relative energy of cis and trans PtX_2Y_2 isomers.

PtX_4. The final LFMM parameters utilized the whole target data set of homoleptic and mixed ligand species. Thus, and as shown in Table 2, there are some small discrepancies between the DFT and LFMM Pt−L bond lengths for the former group of complexes. If we had only wanted to study homoleptic systems, obtaining perfect agreement for them would have been trivial.

PtX_3Y. Defining how well LFMM treats the trans influence requires modeling of mixed-ligand systems. In complexes with the formula PtX_3Y we consider two features: (a) the difference between cis and trans Pt−X bond lengths

Table 2. Pt−L Bond Lengths (Å) for DFT- and LFMM-Optimized Geometries of Homoleptic [PtL_4] Species

	Pt-X		
X	DFT	LFMM	Δ(LFMM − DFT)
H^-	1.646	1.594	−0.052
Cl^-	2.310	2.333	0.023
Br^-	2.459	2.459	0.000
NH_3	2.035	2.047	0.012
OH_2	2.024	2.065	0.041
SH_2	2.307	2.290	−0.017
py	2.019	2.029	0.010
PH_3	2.325	2.299	−0.026

and (b) how much the Pt−Y bond length has changed in comparison to the homoleptic species.

Property a is particularly important, as it shows that we can distinguish between cis and trans ligands and correctly mimic the trans influence.

Figure 1 shows the difference between cis and trans Pt−Cl bond lengths in $PtCl_3X$ (mauve). The DFT and LFMM values agree to 0.018 Å or better, and in virtually all instances, the signs of the cis−trans Pt−Cl differences are the same. For those cases where the signs differ (Y = NH_3 or py), the magnitudes of the difference are less than the tolerance in computed bond lengths, i.e., about 0.02 Å. Hence, the cis and trans Pt−Cl distances are essentially the same for Y = Br, NH_3, SH_2, and py, while $Pt-Cl_{trans}$ is greatly lengthened for Y = H, moderately lengthened for Y = PH_3, and moderately shortened for Y = OH_2.

The LFMM description of the Pt−Cl bonds in [$Pt^{II}Cl_3Y$] species is very good. However, large cis−trans Pt−Cl differences are expected to be accompanied by corresponding changes in Pt−Y distance relative to the Pt−Y bond length in the relevant homoleptic system (Figure 1, green), and we begin to discern some apparently larger discrepancies between DFT and LFMM. For Y = H_2O and Y = H^- the models are qualitatively different, but we note that these ligands also have the largest deviations for the homoleptic complexes (see Table 2). The absolute structures compare reasonably well (Figure 2), and had the LFMM Pt−OH_2 distance in $[Pt(OH_2)_4]^{2+}$ been within the "normal" 0.02 Å tolerance of the DFT value, a qualitatively correct picture would have resulted.

A summary of the type of data shown in Figure 1 is shown in Table 3, but for each ligand X in PtX_3Y. Root mean square

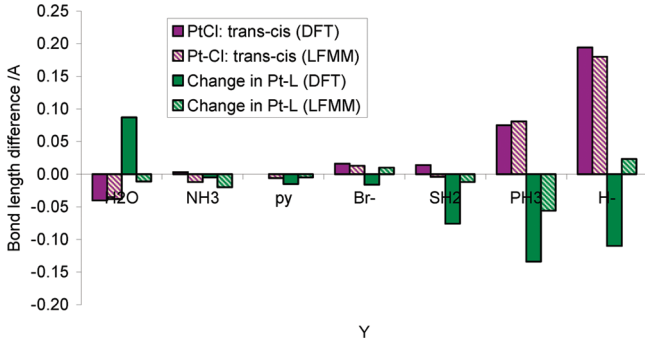


Figure 1. Difference in cis and trans Pt−Cl bond lengths in [$Pt^{II}Cl_3Y$] (in purple) and change in Pt−Y bond length in [$Pt^{II}Cl_3Y$] compared to the homoleptic species [$Pt^{II}Y_4$] (green).

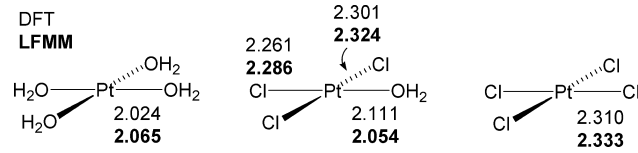


Figure 2. Calculated structural data for selected Pt complexes.

Table 3. RMSD (per Ligand Type) of Absolute Change in Pt−X Bond Lengths (Δ*r*(X)) and Relative *cis*−*trans* Bond Length Differences ($r(X_c) - r(X_t)$, in italics) between DFT and LFMM Structures of [PtX_3Y] and RMSD Values for Change in Pt−Y Bond Lengths (Δ*r*(Y))

X	$\Delta r(X)_{rmsd}$ $\{r(X_c) - r(X_t)\}_{rmsd}$	$\Delta r(Y)_{rmsd}$
H_2O	0.031/*0.030*	0.055
Cl	0.014/*0.011*	0.048
NH_3	0.019/*0.020*	0.037
pyridine	0.019/*0.021*	0.022
Br	0.017/*0.026*	0.058
SH_2	0.027/*0.038*	0.023
PH_3	0.026/*0.035*	0.035
H	0.028/*0.027*	0.058

deviations for the absolute difference in Pt−X and Pt−Y distances are shown and indicate that LFMM is reliably able to predict the Pt−X distances in PtX_3Y complexes. The Pt−Y distances are less accurate, partly due to the effect described above and partly to the fact the LFMM is not always able to predict the full extent of the trans influence. For example, $[PtH_3(NH_3)]^-$ is calculated to have a Pt−N distance of 2.153 Å by DFT, but only 2.082 Å by LFMM. However, both values represent a significant elongation of the Pt−N bond length compared to that found in $[Pt(NH_3)_4]^{2+}$.

Also included in Table 3 are the rms DFT and LFMM deviations between the cis and trans bond lengths. These are all low, indicating that the LFMM gets the balance between the cis and trans influence correct when the cis and trans groups are chemically the same.

Thus, the LFMM gets the right sense but not the full magnitude of the trans influence, at least compared to DFT. A further example is shown in Figure 3 for $[Pt(py)_3Y]$. DFT often gives substantially greater changes in Pt−Y distances than LFMM, particularly for H^-. One potential source of this difference is that each ligand has a single set of LFMM parameters. To the extent that the trans influence is a competition between two ligands such that as one binds more strongly the other weakens, we might anticipate that the AOM parameters in each case should be adjustable to reflect the changing nature of the bonding.[37] This is akin to the idea of a polarizable force field, where the partial atomic charges are variable as opposed to a fixed set. Having multiple sets of AOM parameters also parallels the VALBOND-TRANS idea of explicitly modifying a given ligand's parameters as a function of the trans ligand and would certainly improve matters, as would any increase in the number of parameters, and is an idea for future development. Meanwhile, the current model is qualitatively correct, although we note that increasingly strong trans influences are expected to be increasingly hard for the LFMM model to get right.

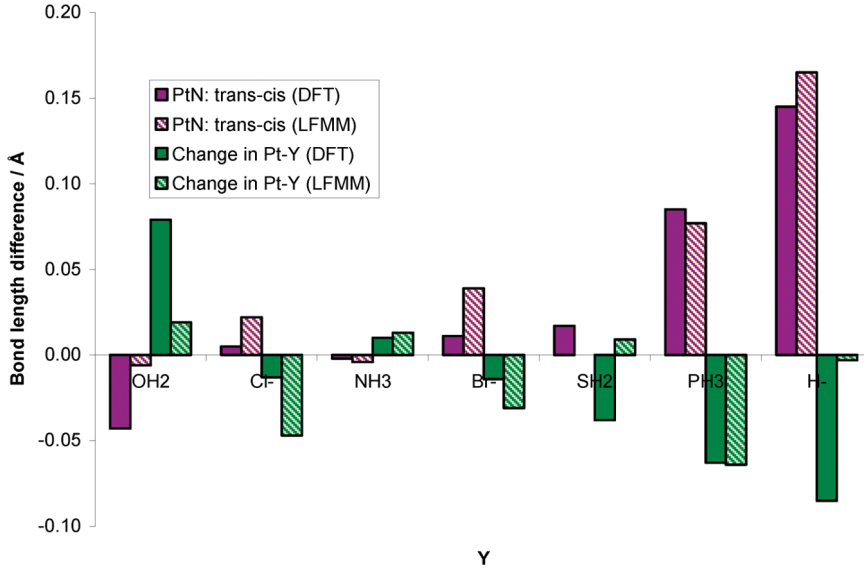


Figure 3. Cis−trans Pt−N difference and change in Pt−Y bond length for [$Ptpy_3Y$].

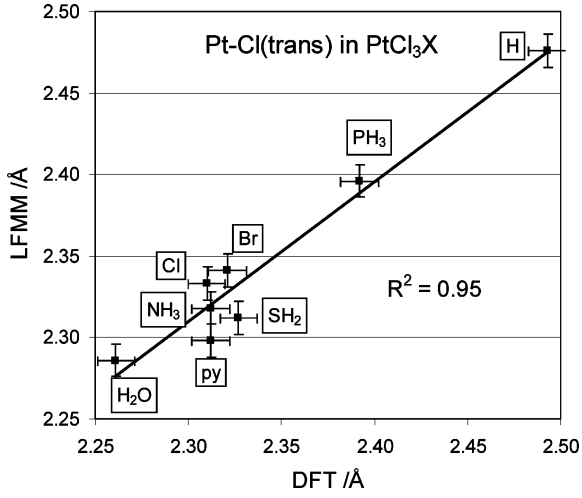


Figure 4. Correlation between computed trans Pt−Cl distances (Å) in [$PtCl_3X$] complexes. Error bars are 0.01 Å.

This issue with the very largest trans influences is not the only problem, since it would not seem to account for H_2O and SH_2, which have apparently fairly modest trans influences and yet some of the biggest errors. The chemical difference between these ligands and the others considered here is their ability to engage in strong intramolecular hydrogen bonding, although how this is related to the errors we observe is unclear.

For the simple ligands considered so far, the following order of decreasing trans influence is obtained:

LFMM: $H^- > PH_3 > SH_2 > Br^- > Cl^-$, pyridine, $NH_3 > H_2O$

DFT: $H^- > PH_3 > SH_2$, Br^-, pyridine, NH_3, $Cl^- > H_2O$

Both the DFT and LFMM data are consistent with experiment, although, as shown graphically in Figure 4, DFT does not significantly differentiate between SH_2, Br^-, pyridine, NH_3, and Cl^-.

PtX_2Y_2. For PtX_2Y_2 complexes we can compare both structural and energetic features: (a) relative Pt−X bond lengths in the cis and trans species, (b) relative Pt−Y bond lengths in the cis and trans species, and (c) relative energies of the cis and trans complexes (compared to the energies determined by DFT).

PtX_2Y_2 Geometries. It is harder to assess the performance of individual parameters for different ligands in PtX_2Y_2 because both the Pt−X and Pt−Y parameters have an effect on the geometry. However, general trends can be observed, and overall there is good agreement between DFT and LFMM.

Again, the chloro complexes $PtCl_2Y_2$ are used to illustrate the performance of the DFT and LFMM calculations. Figure 5 shows how the Pt−Cl and Pt−Y bond lengths differ in $PtCl_2Y_2$ complexes. The qualitative correlation is good, and where there is disagreement ($Y = NH_3$ or SH_2), the difference between cis and trans bond lengths is small. The relative Pt−Y distances are generally also qualitatively correct, but as noted for $PtCl_3Y$ species, the differences between LFMM and DFT are larger for Pt−Y than for Pt−Cl, although here the offending ligand appears to be PH_3, to which we will return.

A summary of the type of information shown in Figure 5 is collected for all PtX_2Y_2 complexes in Table 4. The Pt−PH_3 parameters consistently give poorer rmsd values for the Pt−P cis−trans values, while Pt−H parameters have a large rmsd value for the Pt−Y cis−trans values. However, the latter is in part due to the consistently large trans influence of the hydride ligand such that the apparently bigger deviations are actually a relatively minor proportion of the total change. As shown in Figure 6, the LFMM is always qualitatively correct.

PtX_2Y_2 Energies. The relative energies of the cis and trans isomers depend on several factors. Of course, both forms are usually synthetically accessible, since by exploiting the trans effect, kinetic products can be trapped. The higher energy form is metastable, provided there is a high enough barrier to cis−trans interconversion.

Generally, in the gas phase the trans isomer is by DFT lower in energy, especially if one of the ligands is formally

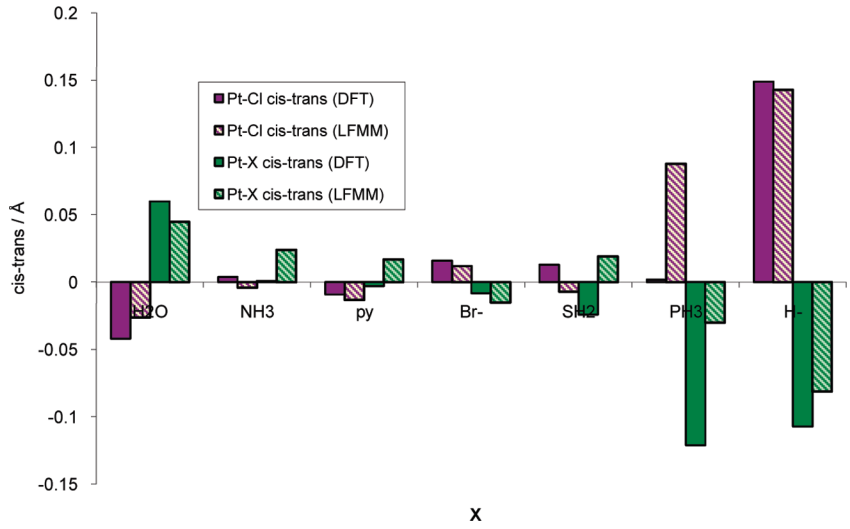


Figure 5. Difference in cis and trans Pt−Cl and Pt−X bond lengths in $[PtCl_2X_2]$.

Table 4. RMSD for Absolute Bond Lengths (Roman Text) and Difference in Cis and Trans Bond Lengths (Italic Text) for $[PtX_2Y_2]$

	rmsd	
X in PtX_2Y_2	$Pt-X_c-Pt-X_t$	$Pt-Y_c-Pt-Y_t$
H_2O	0.038/*0.039*	0.032/*0.039*
Cl	0.026/*0.034*	0.028/*0.041*
NH_3	0.027/*0.022*	0.019/*0.016*
pyridine	0.029/*0.022*	0.016/*0.020*
Br	0.022/*0.030*	0.027/*0.022*
SH_2	0.030/*0.037*	0.030/*0.029*
PH_3	0.044/*0.048*	0.034/*0.038*
H	0.020/*0.019*	0.048/*0.046*

anionic. Given the size of the ligands considered so far, steric effects play only a small role in determining the cis−trans preference. Electrostatics play a much larger role and is the main reason for the trans isomer being preferred, since this minimizes unfavorable ligand−ligand electrostatic repulsions. The trans influence is also important and should favor cis structures as the ligand with the stronger trans influence is opposite the weaker one. Solvation also plays a crucial role in determining the relative stabilities of the isomers. The cis isomer is preferred due to its larger dipole moment (compared to the zero overall dipole moment in the trans isomer).[5] The dipole will be especially large for complexes where X is formally anionic while Y is formally neutral.

The calculated energy differences between cis and trans PtX_2Y_2 species are compared in Figure 7. We have used the lowest energy isomer in each case. The DFT energies may include a solvation correction (DFT^S) or exclude solvation (DFT^{GP}). In either case, the COSMO-optimized structure was employed ($\varepsilon = 78.4$). The LFMM solvation correction ($LFMM^S$) employs the solvation energy computed via the Poisson−Boltzmann scheme as implemented in MOE.

Ideally, we would like a linear correlation between LFMM and DFT energy differences, with a gradient of 1 and an intercept of 0 plus no points in the top left and bottom right quadrants. While this is generally true for the gas phase energies, there are examples where DFT predicts the cis isomer to be more stable and LFMM predicts it to be less stable, and vice versa. When implicit solvent is included, the qualitative predictions of the more stable isomer improve, especially in the "cis" regime, although occasionally LFMM stabilizes the cis isomer too much compared to the DFT results. The worst cases are indicated by the arrows in Figure 7 and involve hydride or PH_3.

Generally, inclusion of solvent stabilizes the cis isomer relative to the trans isomer and the data points tend to move down and to the left. This is due to the presence of a dipole moment in the cis, but not trans, isomer.[5] There are a few anomalous complexes for which this does not happen, i.e., solvation stabilizes trans instead of cis. All of these include water, SH_2, or hydride as ligands. For water and SH_2, the dipole moment of the complex is dependent on the orientation of the hydrogen atoms and this will affect the degree of stabilization afforded by solvation. Additionally, the gas-phase energies are recorded at the solvated geometries. For water and SH_2, the preferred orientation of the ligands may be different at the gas-phase geometries, altering the relative energies. The Pt−H bond lengths may also be affected by the presence of a solvent and be different in the gas phase. In the LFMM calculations more trans structures are stabilized by solvation relative to their cis isomers than for DFT, although some are common to both methods. Again these often include aqua or SH_2 ligands, although not so much hydride ligands. Additionally pyridine ligands show this trend.

Pt−PH_3: A Special Case? The simple phosphine ligand, PH_3, stands out as being particularly poorly treated. On the one hand, this could be considered insignificant, since PH_3 is never used synthetically. On the other hand, PH_3 is a perfectly acceptable computational model and given we are comparing to DFT calculations, we had no a priori reason to expect such a failure. In addition, we have always adopted a "one size fits all" philosophy for am(m)ines; i.e., we use a single set of M−N LFMM parameters for all NR_3 donors.

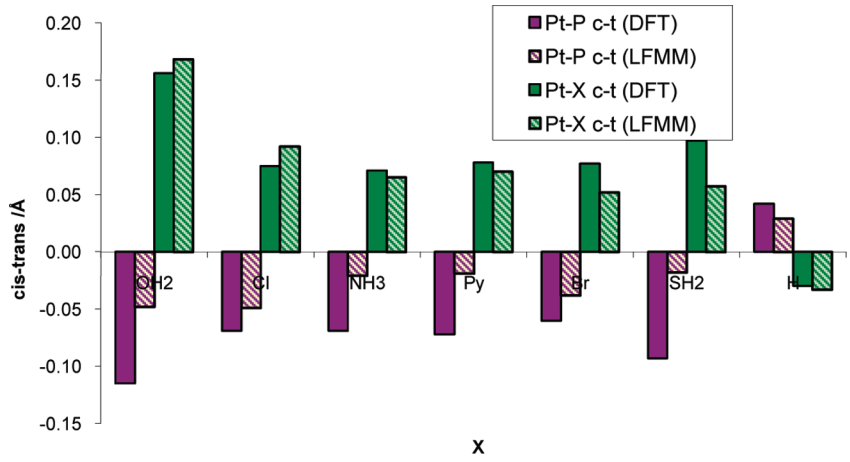


Figure 6. Relative cis and trans geometries of $[PtH_2X_2]$.

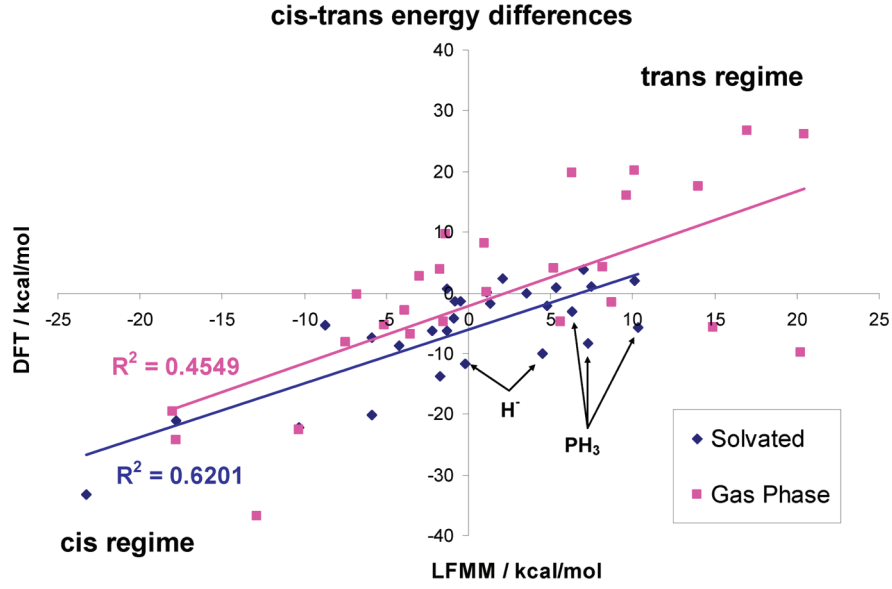


Figure 7. Relative energies of cis and trans isomers/kcal mol^{-1} ($E_{cis} - E_{trans}$). The raw data upon which this figure is based are included in the Supporting Information, Table S5.

We anticipated using the $Pt-PH_3$ LFMM parameters for all $Pt-PR_3$ moieties.

Phosphine does have some qualitatively different features with respect to the other ligands considered here. It is the only π-acceptor ligand in the set, plus its homoleptic complex is one of the few for which the CHelpG analysis suggested a net negative charge on the metal center. We therefore experimented with modifying the partial charges but without significant effect. Finally, we simply replaced PH_3 with PMe_3 and repeated the DFT and LFMM analysis with the same LFMM parameters (Figure 8). Curiously, with solvation corrections added, this "cured" two of the three problem cases such that only one point (dark blue diamonds in Figure 8) is in an incorrect quadrant.

As shown in Figure 9, the geometric comparison of $[Pt(PMe_3)_2X_2]$ is also good. The only complex with cause for concern is $[Pt(PMe_3)_2(py)_2]^{2+}$, where the comparison of both the gas and solution phase relative energies remains less satisfactory, even though the structures appear fine.

$[Pt(PMe_3)_3X]$. The geometries of $[Pt(PMe_3)_3X]$ complexes from LFMM and DFT optimizations are compared in Figure 10. In all cases, except H^-, DFT and LFMM are in agreement. For H^-, if we consider absolute bond lengths, there is only a 0.04 Å difference between the two methods. The large discrepancy in the change in bond length is attributed to the too short Pt−H bond length in the homoleptic species optimized using LFMM.

$[Pt(PMe_3)X_3]$. The geometries of $[Pt(PMe_3)X_3]$ complexes from LFMM and DFT optimizations are compared in Figure 11. In all cases, DFT and LFMM results are in agreement about whether the cis or trans Pt−X bond length should be longer and also about whether the Pt−P bond length has increased or decreased relative to that found in the homoleptic species.

In summary, the LFMM parameters developed on the basis of PH_3 work very well for PMe_3 species.

Other Systems. Having developed a set of LFMM parameters for Pt(II) complexes, we can now explore the

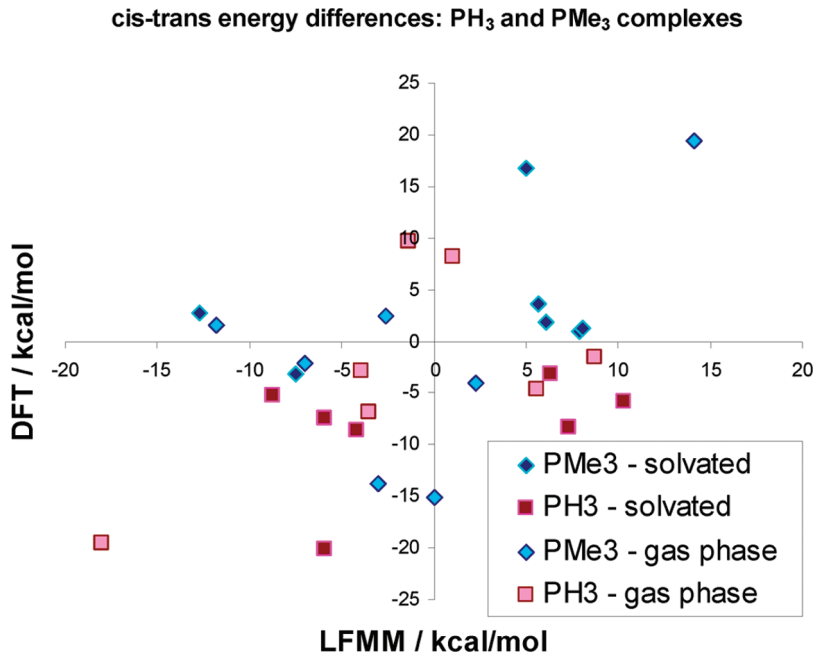


Figure 8. Cis−trans energy difference (kcal/mol) for $[Pt(PR_3)_2X_2]$ complexes, R = H, CH_3. The raw data are included in Supporting Information, Table S6.

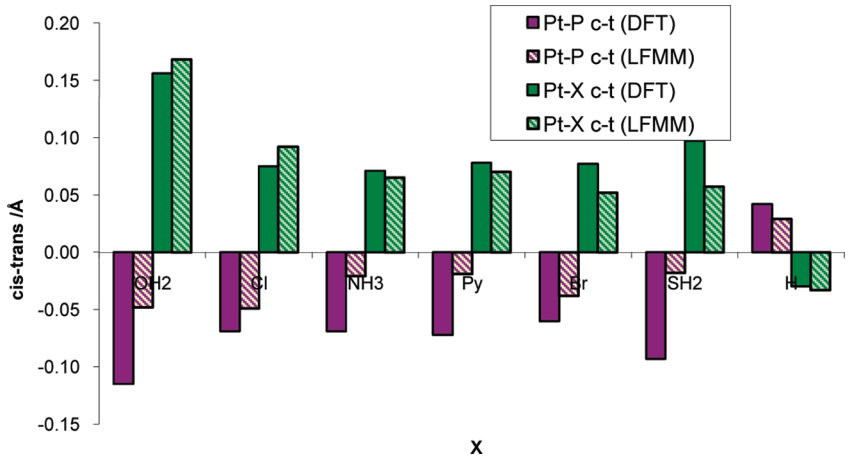


Figure 9. Relative cis and trans geometries of $[Pt(PMe_3)_2X_2]$.

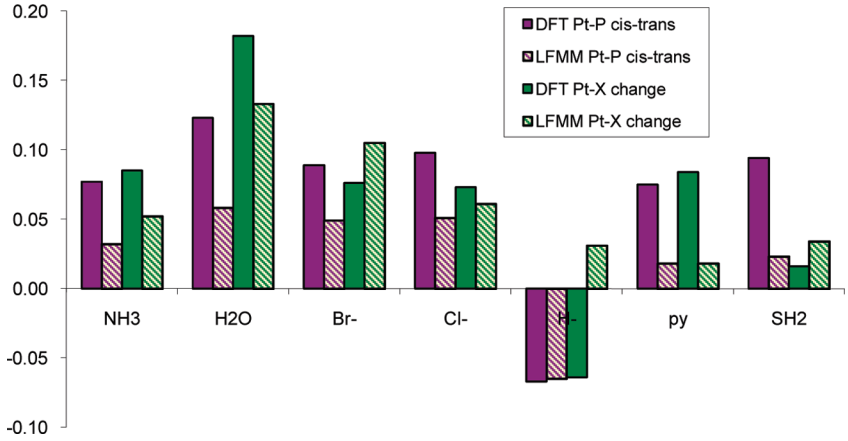


Figure 10. Comparison between DFT and LFMM geometries of $[Pt(PMe_3)_3X]$.

model's performance for species not used in the training set. For example, Rappé et al. have reported UFF calculations for *cis*-dibromo(1,2-diaminocyclohexane)platinum(II), and we compare their results with ours in Table 5.

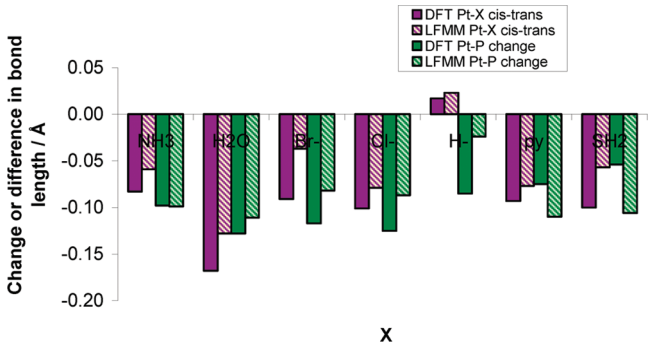
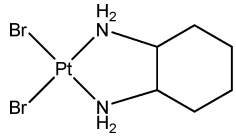


Figure 11. Comparison of DFT and LFMM relative bond lengths (cis−trans) or change in bond length (Pt−P compared to the homoleptic bond length).



LFMM reproduces the crystal structure and DFT structure well, and performs somewhat better than UFF in terms of the Pt−Br distance and the Br−Pt−Br angle.

Further comparisons were made. Four examples of $[PtX_3Y]$ complexes were located in the Cambridge Structural Database[35] and a comparison of the experimentally derived Pt−L distances with those computed by LFMM (and DFT for $[PtH(PMe_3)_3]^+$) is presented in Table 6.

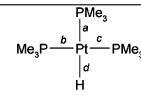
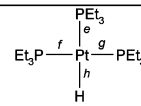
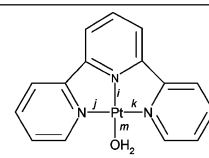
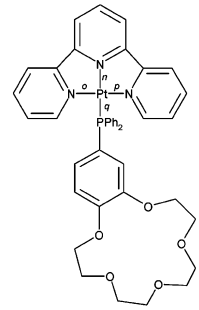
In each case, the relative bond lengths of X_{cis} and X_{trans} are correct, although LFMM tends to generate somewhat longer Pt−P bonds than observed. Note that due to the crystallographic difficulties of estimating H−Pt bond lengths, no comment is offered on the LFMM versus experimental Pt−H bond length comparisons.

A similar comparison with experiment but for $[PtXYZ_2]$ systems is shown in Table 7. Again, LFMM tends to give slightly too long Pt−L contacts, but the qualitative agreement is good.

We next explore the effect of increasing the bulk of the substituents on NR_3 and PR_3 systems. Increasing the steric bulk of the ligands causes the geometry around the metal to distort from square planar toward tetrahedral. The AOM parameters used in the LFMM calculations have to be flexible enough to allow this distortion.

DFT calculations of $[Pt(PR_3)_4]^{2+}$ (R = H, Me, Et) show a progressive distortion from square planar toward tetrahedral. The large L−Pt−L angle changes from 180° to 157.2° to 149°.[51] A crystal structure is available for $[Pt(PEt_3)_4]^{2+}$ that supports the DFT calculations, and LFMM calculations reproduce the trends predicted by DFT.

Table 6. Comparison of Selected Experimental Structures with LFMM Structures for Complexes $[PtX_3Y]$*

Ref	Structure	LFMM (DFT) bond lengths / Å	Experimental bond lengths / Å	
39,40		*a*: 2.423 (2.367) *b*: 2.334 (2.302) *c*: 2.334 (2.302) *d*: 1.637 (1.625)	*a*: 2.323 *b*: 2.287 *c*: 2.293 *d*: *	*a*: 2.315 *b*: 2.289 *c*: 2.299 *d*: 1.922
41		*e*: 2.384 *f*: 2.312 *g*: 2.312 *h*: 1.624	*e*: 2.335/2.291 *f*: 2.297/2.286 *g*: 2.304/2.308 *h*: */*	
42		*i*: 1.917 *j*: 2.046 *k*: 2.053 *m*: 2.085	*i*: 1.914 *j*: 2.009 *k*: 2.021 *m*: 2.040	
43		*n*: 1.999 *o*: 2.097 *p*: 2.073 *q*: 2.287	*n*: 2.003 *o*: 2.081 *p*: 2.073 *q*: 2.232	

* Pt−H distance not reported.

For R = Me and Et, stochastic LFMM conformational searches were carried out and the lowest energy PtL_4 isomers are included in Table 8. The tetrahedral distortion tends to be larger for DFT structures than for LFMM but both show a significant change from planar coordination. For R = Me, there is a DFT structure that is much closer to the LFMM result and only 1.3 kcal mol^{-1} higher than the most stable isomer. The tetrahedral distortion thus seems to be a relatively low-energy mode and relatively large variations in the P−Pt−P angles are associated with relatively small changes in energy.

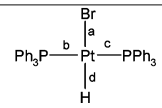
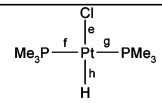
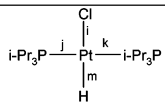
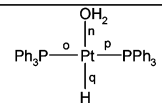
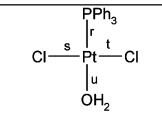
For R = Et, the steric demands are even greater. A crystal structure of the ClO_4^- salt of this complex is available, and both DFT and LFMM are in good agreement with it.[51] An overlay of the LFMM, DFT, and crystal structures is shown in Figure 12. There is some variation in the orientation of some of the ethyl groups, but this is not energetically significant.

A more interesting, if unexpected, result is that the LFMM stochastic searches for $[Pt(PEt_3)_4]^{2+}$ locate a second structural

Table 5. Comparison of Experimental and Computed Geometries of *cis*-Dibromo(1,2-diaminocyclohexane)platinum(II)

	Expt[38]	UFF	LFMM	DFT
Pt−Br/Å	2.434	2.519	2.414	2.456
Pt−N/Å	2.056	2.040	2.036	2.036
Br−Pt−Br/deg	95.61	89.9	97.5	93.2
N−Pt−N/deg	83.5	87.4	86.9	82.3
N−Pt−Br (cis)/deg	89.0/91.9		87.8	92.4/92.1
planarity (Br−Pt−N)/deg	175.4/172.4	underestimated by 6°	174.7/174.7	174.7/174.5

Table 7. LFMM-Optimized and Experimental Bond Lengths for Structures of the Formula $[PtXYZ_2]$*

ref	Structure	LFMM geometry / Å	Experimental Geometry / Å
44-46		a: 2.604 b: 2.361 c: 2.361 d: 1.582	a: 2.523/2.535/2.516 b: 2.283/2.275/2.279 c: 2.283/2.281/2.286 d:1.610/1.592/*
47		e: 2.482 f: 2.307 g: 2.307 h: 1.575	e: 2.423 f: 2.281 g: 2.281 h: *
48		i: 2.497 j: 2.342 k: 2.332 m: 1.586	i: 2.395 j: 2.286 k: 2.287 m: 1.868
49		n: 2.319 o: 2.307 p: 2.310 q: 1.566	n: 2.182/2.191/2.186 o: 2.295/2.291/2.294 p: 2.279/2.285/2.287 q:1.420/1.586/1.612
50		r: 2.257 s: 2.339 t: 2.327 u: 2.159	r: 2.248/2.238 s: 2.329/2.309 t: 2.341/2.358 u: 2.131/2.118

* Pt−H distance not reported.

Table 8. Geometries of $[Pt(PR_3)_4]^{2+}$ for R = H, Me, Et[a]

$[Pt(PH_3)_4]^{2+}$	DFT	LFMM
Pt−P/Å	2.324	2.290
P−Pt−P trans/deg	180.0	180.0
P−Pt−P cis/deg	90.0	90.0

$[Pt(PMe_3)_4]^{2+}$	DFT	LFMM
Pt−P/Å	2.348(2.360)	2.348
P−C/Å	1.824(1.824−1.826)	1.845−1.854
P−Pt−P trans/deg	143.5(157.2)	161.4
P−Pt−P cis/deg	95.6(92.2)	91.5
Pt−P−C/deg	110.0−126.3(113.6−120.7)	118.2−122.0

$[Pt(PEt_3)_4]^{2+}$	DFT	LFMM	experiment[51]
Pt−P/Å	2.365−2.396	2.384−2.409	2.330−2.351
P−C/Å	1.843−1.864	1.771−2.213	1.839−1.931
P−Pt−P trans/deg	145.2−146.9	158.3−158.9	150.4−151.0
P−Pt−P cis/deg	93.6−96.5	91.5−92.4	93.3−94.2
Pt−P−C/deg	105.0−131.2	101.9−129.2	106.5−127.3

[a] For R = Me, the DFT data in parentheses correspond to a structure 1.3 kcal mol^{-1} higher than the overall minimum.

type with one phosphine dissociated. Using DFT, this lies 6.5 kcal mol^{-1} higher in energy than the tetracoordinated structure, but LFMM predicts it to be the more stable form. This arises because the present LFMM parametrization did not include any energetic information that would discriminate these two situations. Our study of $[M(OH_2)_6]^{2+}$ complexes demonstrates that the LFMM can be designed to reproduce such energy differences.[52] Moreover, that study and the current one shows that the use of a Morse function to describe the M−L bond stretching potential allows the LFMM model to support arbitrarily long M−L bond lengths

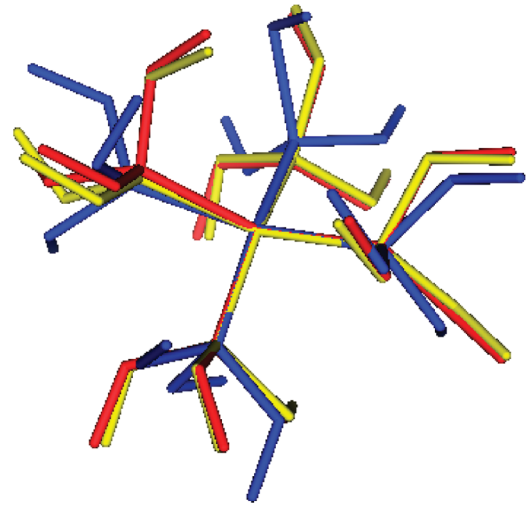


Figure 12. Overlay of optimized geometries of $[Pt(PEt_3)_4]^{2+}$ from DFT (red), LFMM (blue), and a crystal structure (yellow).

Table 9. Selected Bond Lengths (Å) and Angles (deg) for $[Pt(NR_3)_4]^{2+}$ Complexes, R = H, Me

$[Pt(NH_3)_4]^{2+}$	DFT	LFMM
Pt−N	2.035	2.047
N−Pt−N_{trans}	180.0	179.3
N−Pt−N_{cis}	90.0	90.0

$[Pt(NMe_3)_4]^{2+}$	DFT		LFMM	
Pt−N1	2.217	2.258	2.233	2.263
Pt−N2	2.225	2.258	2.232	2.263
Pt−N3	2.260	2.258	2.285	2.263
Pt−N4	2.267	2.258	2.283	2.263
N2−Pt−N3	97.4	95.3	97.1	94.5
N1−Pt−N2	97.4	95.3	96.9	94.5
N1−Pt−N3	138.6	144.4	138.4	147.5
N1−Pt−N4	97.4	95.3	97.1	94.5
N2−Pt−N4	138.0	144.4	138.6	147.5
N3−Pt−N4	96.9	95.3	97.7	94.5

without exacting an infinite energy penalty. That is, the LFMM can effectively describe M−L bond breaking. We will return to this issue in a future publication.

For $[Pt(NR_3)_4]^{2+}$ complexes, the steric influence of R on the geometry around Pt^{II} is also very large, even more so than for the phosphine counterparts. Upon replacing NH_3 with NMe_3 ligands, the steric repulsion is increased so much that, in addition to the tetrahedral distortion found for the PMe_3 system, the Pt−N bond lengths increases by ∼0.2 Å in both DFT and LFMM optimizations, whereas the comparable Pt−P distance was largely unaffected (Table 9). This is consistent with platinum being a soft metal that therefore forms stronger bonds to second-row donors such as phosphines.

Two DFT minima for $[Pt(NMe_3)_4]^{2+}$ have been located, lying only 0.3 kcal mol^{-1} apart. The less stable structure has C_{4h} symmetry and the other is slightly distorted from this. LFMM also finds both these minima and predicts that they lie 1.8 kcal mol^{-1} apart in energy with the symmetric structure being more stable, although this energy difference is small and probably within the error limits of the calcula-

Table 10. Structures and Energies (with respect to symmetric structure in kcal mol^{-1}) of $[Pt(NEt_3)_4]^{2+}$

$[Pt(NEt_3)_4]^{2+}$	DFT	LFMM	LFMM, no bond but charges unadjusted	no bond, charges adjusted
E: isomer 1	(−42.7) −47.7	−8.2 (−7.8)	−22.9 (−23.3)	−15.2 (−7.9)
Pt−N1	2.105	2.178	2.143	2.145
Pt−N2	2.175	2.186	2.202	2.200
Pt−N3	2.148	2.187	2.204	2.201
Pt−N4	5.454	4.702	5.422	4.785
N2−Pt−N3	151.8	156.9	155.1	155.4
N1−Pt−N2	102.8	101.8	102.2	102.0
N1−Pt−N3	105.5	101.2	102.7	102.5
E: isomer 2	(−36.6) −41.9	−8.8 (−7.1)	−9.8 (−9.3)	−17.5 (−9.9)
Pt−N1	2.094	2.122	2.128	2.131
Pt−N2	2.175	2.211	2.205	2.203
Pt−N3	2.149	2.188	2.186	2.179
Pt−N4	5.607	6.251	6.249	5.916
N2−Pt−N3	153.2	156.7	156.9	159.9
N1−Pt−N2	100.8	101.7	101.4	101.5
N1−Pt−N3	106.0	101.3	101.4	101.3
E: isomer 3	(−30.3) −34.8	+4.7 (+4.9)	−11.6 (−10.6)	−2.6 (+4.1)
Pt−N1	2.110	2.186	2.155	2.156
Pt−N2	2.162	2.166	2.179	2.177
Pt−N3	2.152	2.209	2.228	2.225
Pt−N4	5.065	4.642	6.584	4.781
N2−Pt−N3	153.6	158.1	156.8	156.8
N1−Pt−N2	103.6	99.2	100.2	100.1
N1−Pt−N3	102.5	102.1	102.9	102.9

tions. Geometries for both minima are reported in Table 9 and LFMM and DFT are in excellent agreement.

When R = Et, the steric bulk of the ligand is even greater and DFT optimizations show that one ligand spontaneously dissociates from the Pt^{2+} center, leaving an approximately T-shaped $[Pt(NEt_3)_3]^{2+}$ moiety. We have located stationary points for three of these structures, and reoptimized them using LFMM. These results are summarized in Table 10 and the overlay of the lowest energy LFMM and DFT geometries is shown in Figure 13. With care, we were able to optimize a "symmetrical" structure with DFT where all four NEt_3 ligands are coordinated. It has Pt−N bond lengths of 2.371 Å, cis angles of 95.0° and trans angles of 143.9° compared to 2.350 Å, 95.5°, and 145.8° from LFMM, respectively. The symmetrical structure lies ∼43 kcal mol^{-1} higher than the most stable tricoordinate structure in DFT and ∼9 kcal mol^{-1} higher in LFMM.

LFMM predicts slightly longer lengths than DFT, but the "sense" of which bond lengths are longer is retained, along with good comparisons between angles. Moreover, even though there is an explicit Pt−N connection in the LFMM treatment, one Pt−N distance spontaneously lengthens to more than 4.5 Å.

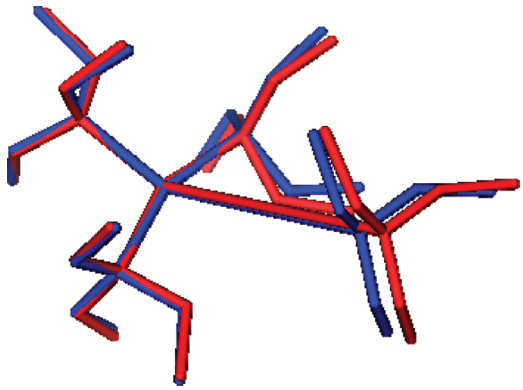


Figure 13. Overlay of LFMM (blue) and DFT (red) optimized geometries for the lowest energy forms of $[Pt(NEt_3)_4]^{2+}$.

Some improvement in the relative Pt−N bond lengths is obtained when this explicit bond between Pt and the "dissociating" amine ligand is deleted, but the change in geometry is not very large.

A search of the Cambridge Structural Database[35] did not reveal any crystal structures of either $[Pt(NEt_3)_3]^{2+}$ or $[Pt(NEt_3)_4]^{2+}$, so we are unable to comment on how this result compares to experiment.

For both the phosphine and amine systems, the LFMM parametrization did not include energetic information and its ability to treat the dissociation found for the R = Et systems is at best qualitative. Removing the explicit Pt−N bond and recalculating charges leads to apparently better energetics and a slight reduction in the $Pt-N^1$ bond length compared to the $Pt-N^{2,3}$ bond lengths. However, a reparameterization is required if an explicit treatment of bond dissociation is desired.

Finally, we return to complexes containing pyridine ligands. Our DFT orbital analysis suggests pyridine acts as a strong σ-donor and a strong π-donor toward Pt^{II}. In $[Pt(py)_4]^{2+}$, the two limiting (i.e., highest symmetry) orientations of the py planes give rise to the same LFSE (Figure 14). Hence, there is no electronic driver to determine the ligand plane orientation, and simple steric considerations suggest that the most stable arrangement is a "propellor" arrangement with the ligand planes tending to lie perpendicular to the PtN_4 coordination plane rather than in the plane where the contacts of the ortho hydrogens would be unfavorable. In practice, both experiment[53] and LFMM give propeller arrangements, although the pitch is 0° for LFMM

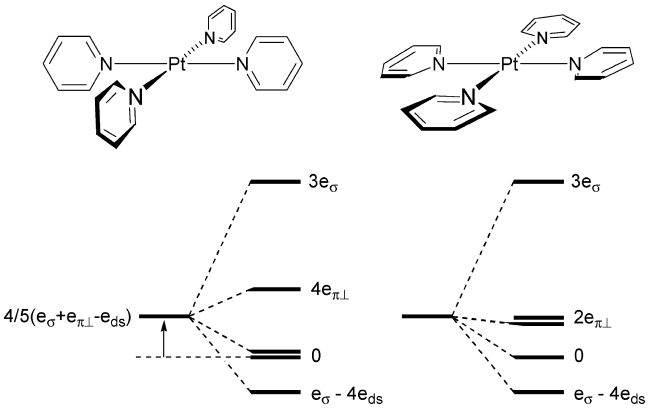


Figure 14. Limiting orientations of pyridine planes for $[Pt(py)_4]^{2+}$ and their associated d orbital energy level diagrams. Left: perpendicular orientation, propeller pitch 0°. Right: parallel orientation, pitch 90°.

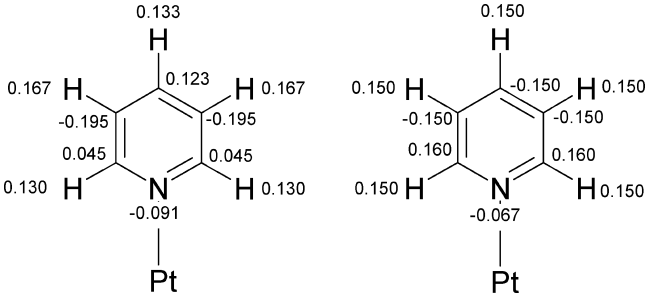


Figure 15. Charges on pyridine ring when coordinated to Pt^{II}.

but ranges from 2 to 28° in the X-ray structure of $[Pt(py)_4]Cl_2 \cdot 3H_2O$ (Figure 15).

In mixed-ligand systems, the pyridine orientation is influenced by both the coligand and solvation. For example, without solvation corrections, DFT and LFMM calculations for *trans*-$[Pt(py)_2H_2]$ and $[Pt(py)H_3]^-$ place the pyridine rings coplanar with the coordination plane due to the favorable electrostatic attraction between the positively charged ortho hydrogen atoms of the pyridine rings and the negatively charged hydrides. For *trans*-$[Pt(py)_2H_2]$, DFT predicts that the coplanar isomer is preferred by 2.7 kcal mol^{-1} and by 0.9 kcal mol^{-1} for $[Pt(py)H_3]^-$. When COSMO solvation corrections are included, the perpendicular py arrangement is preferred by 2.0 and 2.1 kcal mol^{-1} respectively.

LFMM gives a very similar picture. For $[Pt(py)H_3]^-$ without solvation, the coplanar structure is favored by 3.9 kcal mol^{-1} while with a solvation correction, the perpendicular structure is favored by 1.8 kcal mol^{-1}. For *trans*-$[Pt(py)_2H_2]$, attempts to locate the perpendicular isomer by LFMM spontaneously revert to the coplanar form. However, if the Born solvation correction is included in the optimization procedure, as opposed to being added as a single point energy correction, the pyridine rings tilt out of the coordination plane and the energy decreases by 3.8 kcal mol^{-1} relative to the coplanar form.

Conclusions

Ligand field molecular mechanics provides a good description of the trans influence in a wide range of Pt^{II} complexes. Not only can we compute good structures but the relative energies of *cis*- and *trans*-PtX_2Y_2 species virtually always agree with estimates based on DFT.

The ligand field stabilization energy (LFSE) provides a bridge between all the ligands such that variation of any one can lead to changes in all the others. The LFSE is inherently centrosymmetric but, in conjunction with the other LFMM energy terms, we can induce an asymmetric motion such that the stronger trans influenced ligand is more "anchored" and either moves toward the metal or, at the very least, retains its original bond length, while the weaker one moves away. This latter motion can be substantial. For $[PtCl_3H]^{2-}$, the trans Pt−Cl distance is elongated by nearly 0.25 Å relative to that of $[PtCl_4]^{2-}$ (Figure 4).

LFMM calculations give a decreasing trans influence series: $H^- > PH_3 > SH_2 > Br^- > Cl^-$, pyridine, $NH_3 > H_2O$. This compares well with the DFT calculated series: $H^- > PH_3 > SH_2$, Br^-, pyridine, NH_3, $Cl^- > H_2O$. Both the DFT and LFMM data are in agreement with the order reported in the literature.

The LFMM parameters developed for simple ligands also describe the behavior of their bulkier congeners. In both $[Pt(PR_3)_4]^{2+}$ and $[Pt(NR_3)_4]^{2+}$ (R = H, Me, Et), we see progressive tetrahedral distortions and even spontaneous ligand dissociation. Overall, while we are occasionally short of the full magnitude of the trans influence, in virtually every case, the LFMM is in qualitative agreement with DFT. This gives us some confidence to take the model forward to examine platinum-based anticancer agents and their interactions with biomolecules such as DNA.

Supporting Information Available: LFMM parameters, DFT geometry optimization data for $[PdCl_4]^{2-}$, $[PdBr_4]^{2-}$ and $[Pd(NH_3)_4]^{2+}$ using a wide variety of functional/solvation model combinations; cis−trans geometries and energies for $[PtCl_2(NH_3)_2]$ using OPBE with and without COSMO corrections; DFT and LFMM cis−trans energy differences for PtA_2B_2 systems with and without solvation corrections (i.e., the data used to construct Figures 7 and 8). This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgment. A.E.A. acknowledges the award of an EPSRC fellowship and the EPSRC Chemical Database Service for access to the Cambridge Structural Database.

References

(1) Pidcock, A.; Richards, R. E.; Venanzi, L. M. *J. Chem. Soc. A* **1966**, 1707.

(2) Mason, R.; McWeeny, R.; Towl, A. D. C. *Faraday Discuss.* **1969**, *47*, 20.

(3) Pearson, R. G. *Inorg. Chem.* **1973**, *12*, 712–713.

(4) Jorgensen, C. K. *Inorg. Chem.* **1964**, *3*, 1201–1202.

(5) Harvey, J. N.; Heslop, K. M.; Orpen, A. G.; Pringle, P. G. *Chem. Commun.* **2003**, 278–279.

(6) Landis, C. R.; Cleveland, T.; Firman, T. K. *J. Am. Chem. Soc.* **1998**, *120*, 2641–2649.

(7) Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K. *Coord. Chem. Rev.* **2009**, *253*, 795–816.

(8) Bentz, A.; Comba, P.; Deeth, R. J.; Kerscher, M.; Seibold, B.; Wadepohl, H. *Inorg. Chem.* **2008**, *47*, 9518–9527.

(9) Deeth, R. J.; Hearnshaw, L. J. A. *Dalton Trans.* **2006**, 1092–1100.

(10) Deeth, R. J.; Foulis, D. L.; Williams-Hubbard, B. J. *Dalton Trans.* **2003**, 3949–3955.

(11) Baerends, E. J.; Autschbach, J.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; M., D. R.; E., E. D.; van Faassen, M.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; McCormack, D. A.; Michalak, A.; Neugebauer, J.; Nicu, V. P.; Osinga, V. P.; Patchkovskii, S.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L.; Ziegler, T. *ADF2006.01*; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2006.

(12) Guerra, C. F.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.

(13) Van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 9783–9792.

(14) Van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597–4610.

(15) Pye, C. C.; Ziegler, T. *Theor. Chem. Acc.* **1999**, *101*, 396–408.

(16) Allinger, N. L.; Zhou, X. F.; Bergsma, J. *J. Mol. Struct.* **1994**, *118*, 69–83.

(17) Bon, R. S.; van Vliet, B.; Sprenkels, N. E.; Schmitz, R. F.; de Kanter, F. J. J.; Stevens, C. V.; Swart, M.; Bickelhaupt, F. M.; Groen, M. B.; Orru, R. V. A. *J. Org. Chem.* **2005**, *70*, 3542–3553.

(18) Fan, L. Y.; Ziegler, T. *J. Phys. Chem.* **1992**, *96*, 6937–6941.

(19) Fan, L. Y.; Ziegler, T. *J. Chem. Phys.* **1992**, *96*, 9005–9012.

(20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; 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.

(21) Ehlers, A. W.; Bohme, M.; Dapprich, S.; Gobbi, A.; Hollwarth, A.; Jonas, V.; Kohler, K. F.; Stegmann, R.; Veldkamp, A.; Frenking, G. *Chem. Phys. Lett.* **1993**, *208*, 111–114.

(22) Bush, B. L.; Bayly, C. I.; Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 1495–1516.

(23) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

(24) Deeth, R. J.; Fey, N.; Williams-Hubbard, B. J. *J. Comput. Chem.* **2005**, *26*, 123–130.

(25) *MOE*, 2007 ed.; Chemical Computing Group, Montreal: Montreal, 2007.

(26) Norrby, P. O.; Liljefors, T. *J. Comput. Chem.* **1998**, *19*, 1146–1166.

(27) Rappe, A. K.; Colwell, K. S.; Casewit, C. J. *Inorg. Chem.* **1993**, *32*, 3438–3450.

(28) Root, D. M.; Landis, C. R.; Cleveland, T. *J. Am. Chem. Soc.* **1993**, *115*, 4201–4209.

(29) Tubert-Brohman, I.; Schmid, M.; Meuwly, M. *J. Chem. Theor. Comp.* **2009**, *5*, 530–539.

(30) Deeth, R. J.; Munslow, I. J.; Paget, V. J. *NATO ASI Ser.* **1997**, *3/41*, 77–103.

(31) Brandt, P.; Norrby, T.; Akermark, E.; Norrby, P. O. *Inorg. Chem.* **1998**, *37*, 4120–4127.

(32) Allured, V. S.; Kelly, C. M.; Landis, C. R. *J. Am. Chem. Soc.* **1991**, *113*, 1–12.

(33) Cundari, T. R.; Fu, W.; Moody, E. W.; Slavin, L. L.; Snyder, L. A.; Sommerer, S. O.; Klinckman, T. R. *J. Phys. Chem.* **1996**, *100*, 18057–18064.

(34) Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. *J. Am. Chem. Soc.* **1995**, *117*, 8407–8415.

(35) Fletcher, D. A.; McMeeking, R. F.; Parkin, D. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 746–749.

(36) Hocking, R. K.; Deeth, R. J.; Hambley, T. W. *Inorg. Chem.* **2007**, *46*, 8238–8244.

(37) Coe, B. J.; Glenwright, S. J. *Coord. Chem. Rev.* **2000**, *203*, 5–80.

(38) Lock, C. J. L.; Pilon, P. *Acta Crystallogr., Sect. B* **1981**, *37*, 45–49.

(39) Packett, D. L.; Syed, A.; Trogler, W. C. *Organometallics* **1988**, *7*, 159–166.

(40) Adams, R. D.; Barnard, T. S.; Li, Z. Y.; Zhang, L. J. *Chem. Ber. Recl.* **1997**, *130*, 729–733.

(41) Russell, D. R.; Mazid, M. A.; Tucker, P. A. *J. Chem. Soc., Dalton Trans.* **1980**, 1737–1742.

(42) Annibale, G.; Bergamini, P.; Bertolasi, V.; Bortoluzzi, M.; Cattabriga, M.; Pitteri, B. *Eur. J. Inorg. Chem.* **2007**, 5743–5751.

(43) Yam, V. W. W.; Tang, R. P. L.; Wong, K. M. C.; Lu, X. X.; Cheung, K. K.; Zhu, N. Y. *Chem.—Eur. J.* **2002**, *8*, 4066–4076.

(44) Aldridge, S.; Coombs, D.; Jones, C. *Acta Crystallogr., Sect. E* **2003**, *59*, M584−M585.

(45) Habereder, T.; Noth, H. *Appl. Organomet. Chem.* **2003**, *17*, 525–538.

(46) Sivaramakrishna, A.; Su, H.; Moss, J. R. *Acta Crystallogr., Sect. E* **2007**, *63*, M244−M245.

(47) Packett, D. L.; Jensen, C. M.; Cowan, R. L.; Strouse, C. E.; Trogler, W. C. *Inorg. Chem.* **1985**, *24*, 3578–3583.

(48) Robertson, G. B.; Tucker, P. A.; Wickramasinghe, W. A. *Aust. J. Chem.* **1986**, *39*, 1495–1507.

(49) Parkins, A. W.; Richard, C. J.; Steed, J. W. *Inorg. Chim. Acta* **2005**, *358*, 2827–2832.

(50) Rath, N. P.; Fallis, K. A.; Anderson, G. K. *Acta Crystallogr., Sect. C* **1993**, *49*, 2079–2081.

(51) Kozelka, J.; Luthi, H. P.; Dubler, E.; Kunz, R. W. *Inorg. Chim. Acta* **1984**, *86*, 155–163.

(52) Deeth, R. J.; Randell, K. *Inorg. Chem.* **2008**, *47*, 7377–7388.

(53) Wei, C. H.; Hingerty, B. E.; Busing, W. R. *Acta Crystallogr., Sect. C* **1989**, *45*, 26.

CT9001569

JCTC Journal of Chemical Theory and Computation

CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses

Olgun Guvench,[†] Elizabeth Hatcher,[†] Richard M. Venable,[‡] Richard W. Pastor,[‡] and Alexander D. MacKerell, Jr.*,[†]

Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, 20 Penn Street HSF II, Baltimore, Maryland 21201, and Laboratory of Computational Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892

Received May 15, 2009

Abstract: We present an extension of the CHARMM hexopyranose monosaccharide additive all-atom force field to enable modeling of glycosidic-linked hexopyranose polysaccharides. The new force field parameters encompass 1→1, 1→2, 1→3, 1→4, and 1→6 hexopyranose glycosidic linkages, as well as O-methylation at the C_1 anomeric carbon, and are developed to be consistent with the CHARMM all-atom biomolecular force fields for proteins, nucleic acids, and lipids. The parameters are developed in a hierarchical fashion using model compounds containing the key atoms in the full carbohydrates, in particular *O*-methyl-tetrahydropyran and glycosidic-linked dimers consisting of two molecules of tetrahyropyran or one molecule of tetrahydropyran and one of cyclohexane. Target data for parameter optimization include full two-dimensional energy surfaces defined by the Φ/Ψ glycosidic dihedral angles in the disaccharide analogs, as determined by quantum mechanical MP2/cc-pVTZ single point energies on MP2/6-31G(d) optimized structures (MP2/cc-pVTZ//MP2/6-31G(d)). In order to achieve balanced, transferable dihedral parameters for the Φ/Ψ glycosidic dihedral angles, surfaces for all possible chiralities at the ring carbon atoms involved in the glycosidic linkages are considered, resulting in over 5 000 MP2/cc-pVTZ//MP2/6-31G(d) conformational energies. Also included as target data are vibrational frequencies, pair interaction energies and distances with water molecules, and intramolecular geometries including distortion of the glycosidic valence angle as a function of the glycosidic dihedral angles. The model compound optimized force field parameters are validated on full disaccharides through the comparison of molecular dynamics results to available experimental data. Good agreement is achieved with experiment for a variety of properties including crystal cell parameters and intramolecular geometries, aqueous densities, and aqueous NMR coupling constants associated with the glycosidic linkage. The newly developed parameters allow for the modeling of linear, branched, and cyclic hexopyranose glycosides both alone and in heterogeneous systems including proteins, nucleic acids, and/or lipids when combined with existing CHARMM biomolecular force fields.

Introduction

Polysaccharide carbohydrates, composed of individual monosaccharide units that are connected together by glycosidic linkages, have numerous and varied roles in biology, where they serve as energy storage and transport molecules, structural scaffolds, and motifs for molecular recognition. An important subset of these glycosides has as its component monosaccharide hexopyranoses, such as glucose (Figure 1, compound **1**), galactose, and mannose. Members of this subset include cellulose and starch, which are both composed exclusively of glucose, yet have dramatically different

* Corresponding author: Telephone: 410-706-7442. Fax: 410-706-5017. E-mail: alex@outerbanks.umaryland.edu.

[†] University of Maryland School of Pharmacy.

[‡] National Institutes of Health.

10.1021/ct900242e CCC: $40.75
Published on Web 08/20/2009

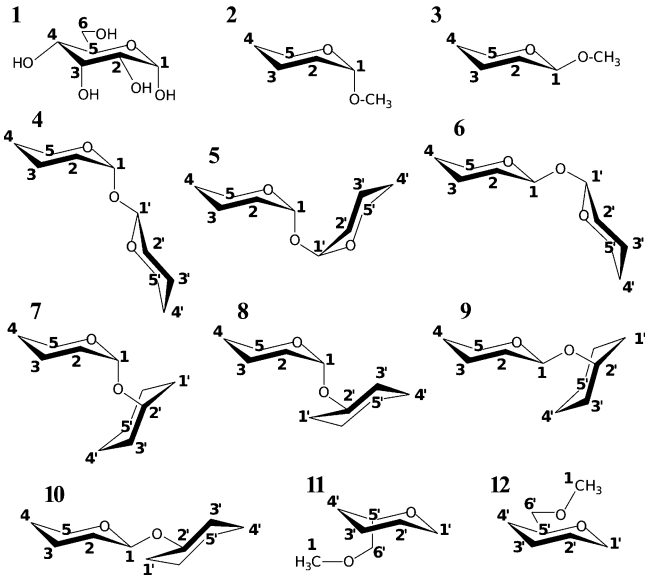


Figure 1. α-D-glucopyranose (**1**) and model compounds (**2**−**12**). Model compound atoms are numbered to correspond with the standard hexopyranose atom numbering, as shown on α-D-glucopyranose.

properties and functions arising from cellulose being composed exclusively of one epimer of glucose and starch exclusively of the other epimer.[1] In addition to existing as isolated biological polymers, hexopyranose polysaccharides are commonly found covalently linked to other biomolecules. Attachment of hexopyranose polysaccharides to proteins and lipids yields glycoproteins and glycolipids, respectively, and these heterogeneous biomolecules play crucial roles in protein folding and molecular recognition.

Significant efforts spanning several decades have been made toward the development of molecular mechanics force field models for investigating the structure and energetics of polysaccharides.[2−16] The large variety of component monosaccharides, the inherent conformational complexity of monosaccharides, and the additional degrees of freedom that need to be considered in parametrizing glycosidic linkages all present challenges to the development of comprehensive carbohydrate force fields. As such, the range of force field applicability (i.e., types of carbohydrates) is often limited, and the study of heterogeneous systems (e.g., glycolipids, glycoproteins, protein: carbohydrate complexes) is complicated by inconsistent treatment of 1,4 nonbonded interactions and by differing force-field development protocols. Presently, the availability of high-quality experimental data and the ability to probe a larger variety of conformational energies on larger model systems using more accurate quantum mechanical methods allow for the inclusion of more and better target and validation data in the carbohydrate force field development process, as reflected in recent efforts.[17−20] It is anticipated that such advances will lead to both more complete and accurate carbohydrate force field models.

To enable the atomic-level modeling of a wide variety of carbohydrates in an aqueous environment and interacting with other biopolymers such as proteins, nucleic acids, and lipids, our laboratory has undertaken the development of a comprehensive force field for carbohydrates[19,20] in the context of the CHARMM additive biomolecular force fields.[21−33] The long-term goal of these efforts is to enable atomically detailed investigation of carbohydrates in molecular recognition, including a detailed understanding of the molecular basis of and the physical insights into the interaction of carbohydrates with other biomolecules such as proteins and lipids. The present work builds on previously published work on model compounds, including linear and cyclic ethers,[34] ethylene glycol,[35] and 2-ethoxy tetrahydropyran,[36] as well as all-atom pairwise additive force field development efforts for hexopyranose monosaccharides[19] and acyclic polyols, acyclic carbohydrates, and inositol.[20] Among the findings of these prior efforts is that, using the relatively simple functional form of the CHARMM all-atom additive force field equation, it is possible to develop carbohydrate force field parameters that give good agreement with gas phase, crystalline, and aqueous conformational properties as well as aqueous phase densities of very dilute to very concentrated aqueous solutions. In the present study, the parameters developed for hexopyranoses have been extended to a comprehensive collection of glycosidic linkages, i.e., 1→1, 1→2, 1→3, 1→4, and 1→6 glycosidic links with all possible combinations of chiralities for the two ring-carbon atoms in each glycosidic link as well as to both α- and β-anomers of 1-*O*-methyl hexopyranoses. Therefore, these parameters allow for the modeling of all possible unsubstituted linear, cyclic, and branched hexopyranose polysaccharides.

The present work represents a departure from previous parametrizations of hexopyranose glycosidic linkages in that extensive high-level quantum mechanical data on relatively large model compound systems are used in the parametrization procedure. Specifically, two molecules of tetrahydropyran or one molecule each of tetrahydropyran and cyclohexane connected by a glyclosidic linkage are employed as model compounds for hexopyranose disaccharides. Optimized two-dimensional scans at the MP2/6-31G(d) level with a resolution of 15° for the full range of the Φ/Ψ dihedral angles of a glycosidic linkage are performed, and MP2/cc-pVTZ single point energies are computed for all of these optimized structures. In sum, over 5 250 MP2/cc-pVTZ//MP2/6-31G(d) conformational energies are used in the parametrization process, and they include consideration of all possible chiralities at the ring carbon atoms involved in the glycosidic linkages. Particular care is taken in the parametrization of the $C_1-O_{link}-C_n'$ valence angle distortion, where C_1 corresponds to the anomeric carbon in the first ring, O_{link} the oxygen in the glycosidic linkage, and C_n' the glycosidic ring carbon atom in the second ring, such that the force field accurately reproduces the quantum mechanical distortion of this angle, ranging nearly 30°, over the full range of Φ/Ψ dihedral values. Extensive validation of the parameter set is done by comparing results from molecular dynamics (MD) simulations of disaccharide crystals and

aqueous solutions to experimental disaccharide crystal geometries, NMR *J*-coupling values, and aqueous solution densities.

Methods

All molecular mechanics calculations were performed with the CHARMM program,[21,37,38] using the same potential energy function as for the CHARMM protein,[23−25] nucleic acid,[26−28] and lipid all-atom additive force fields:[29−33]

$$\begin{aligned} U(r) &= \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \\ &\quad \sum_{\text{Urey-Bradley}} K_{\text{UB}}(S - S_0)^2 \\ &+ \sum_{\text{dihedrals}} K_\chi(1 + \cos(n\chi - \delta)) + \sum_{\text{impropers}} K_{\text{imp}}(\varphi - \varphi_0) \\ &+ \sum_{\text{nonbonded}} \varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{r_{ij}} \end{aligned} \quad (1)$$

In eq 1, K_b, K_θ, K_{UB}, K_χ and K_{imp} are bond, valence angle, Urey−Bradley, dihedral angle, and improper dihedral angle force constants, respectively. b, θ, S, χ and φ are the bond distance, valence angle, Urey−Bradley 1,3-distance, dihedral angle, and improper dihedral angle values, and the subscript 0 represents an equilibrium value. Additionally, for the dihedral term, n is the multiplicity, and δ is the phase angle as in a Fourier series. The nonbonded interaction energy between pairs of atoms i and j consists of the Lennard-Jones (LJ) 6−12 term and the Coulomb term. ε_{ij} is the LJ well depth, $R_{\min,ij}$ is the interatomic distance at the LJ energy minimum, q_i and q_j are the partial atomic charges, and r_{ij} is the distance between atoms i and j. The Lorentz−Berthelot combining rules are applied to determine LJ parameters between different atom types.[39]

A modified version of the rigid three-site TIP3P model was used to represent water,[40,41] and the SHAKE algorithm[42] was applied to keep water molecules rigid and to constrain covalent bonds between hydrogen and heavy atoms to their equilibrium values. Gas-phase energies were calculated using infinite nonbonded cutoffs, and gas-phase energy minimizations were done to a tolerance of 10^{-6} kcal·mol^{-1}·Å^{-1}. Aqueous and crystal simulations were performed using periodic boundary conditions,[39] with a force-switched smoothing function[43] applied to the LJ interactions in the range of 10−12 Å and particle mesh Ewald[44] with a real-space cutoff of 12 Å to compute the Coulomb interactions. A time step of 1 fs was used for integration of the equations of motion with the "leapfrog" integrator,[45] and all MD employed Nosé−Hoover thermostatting,[46,47] Langevin piston barostatting,[48] and a long-range correction to the pressure to account for LJ interactions beyond 12 Å.[39] Condensed-phase simulations were done at experimental temperature and pressure, with aqueous simulations done using a cube as the periodic unit cell and crystal simulations employing the appropriate experimental unit cell geometries. For the aqueous simulations with a cubic unit cell, the cell edge lengths were varied isotropically to maintain the target pressure during simulation, whereas unit cell edge lengths in the crystal simulations were allowed to vary independently. Angular crystal cell parameters of 90° were constrained to this value, while those not 90° were allowed to vary independently. Data for the crystal simulations were collected for 4 ns following 1 ns of equilibration.

QM calculations were performed using the Gaussian 03 program.[49] Geometry optimizations and vibrational calculations were performed using the MP2/6-31G(d) model chemistry,[50,51] with tight tolerances applied when optimizing structures for vibrational calculations. A scale factor of 0.9434 was applied to the QM frequencies, as required to account for limitations in the level of theory and to reproduce experimental frequencies.[52] Potential energy decomposition analysis was performed using the MOLVIB utility in CHARMM using internal coordinates as per Pulay et al.[53] Relaxed potential energy scans were obtained by optimizing the geometry at the MP2/6-31G(d) level followed by MP2/cc-pVTZ single point calculations (MP2/cc-pVTZ//MP2/6-31G(d)).[50,51,54] All potential energy scans were performed in 15° increments with only the scanned dihedral angles (e.g., Φ/Ψ or Ψ/Ω for the 2D surfaces) constrained.

QM calculations for interaction pairs consisting of a water molecule with a particular model compound followed the standard procedure for the CHARMM force field, thereby ensuring consistency of the nonbonded parameters with the remainder of the CHARMM additive biomolecular force fields.[22] Per this procedure, the solute−water interaction distance was optimized at the HF/6-31G(d) level, with constraints on all other degrees of freedom. Following optimization, HF/6-31G(d) interaction energy target data were calculated as $1.16*(E_{\text{pair}} - E_{\text{solute}} - E_{\text{water}})$, with no basis-set superposition error correction, and the empirical scaling factor of 1.16 introduced to yield parameters appropriate for a condensed-phase force field.[24,55] Target data for interaction distances were the QM-optimized distances minus 0.2 Å, again to yield parameters appropriate for a condensed-phase force field. The water intramolecular geometry in both the QM and the MM calculations of pair interaction data was that of the TIP3P water model,[40] and the model compound geometry was one that was previously gas-phase optimized in the MP2/6-31G(d) or the CHARMM representation, respectively.

Relaxed QM potential energy surfaces at the MP2/cc-pVTZ//MP2/6-31G(d) level of theory were used as target data for fitting the dihedral force constants in an automated manner using the freely available Monte Carlo simulated annealing (MCSA) dihedral parameter fitting program fit_dihedral.py[56] (available for downloading at http://mackerell.umaryland.edu). For each dihedral being fit, three multiplicities n of 1, 2, and 3 were included, and the corresponding K_χ values (eq 1) were optimized to minimize the root-mean-square (rmse) error between the empirical and QM energies. In the MCSA approach, the adiabatic empirical energy surfaces are initially obtained with the force constants K_χ on the dihedral parameters being parametrized set to zero; restraining potentials on dihedral angles are used to maintain the values of the dihedral angles that are being scanned. The energy difference between the resulting empirical surface and the target QM surface is then determined, and the dihedral

parameters fit to reproduce that energy difference. K_χ values were constrained to be no more than 3 kcal/mol, and phase angles δ were limited to 0 and 180° to maintain symmetry of the dihedral potentials about $\chi = 0°$, allowing for applicability of the parameters to both enantiomers of a chiral species.

An extension to the MCSA fitting program fit_dihedral.py was introduced to allow for independent rms alignment of multiple MM surfaces to their corresponding QM surfaces during the simultaneous fitting of these surfaces. Previously, the target function to be optimized was the rms error *RMSE* between the QM conformational energies E_i^{QM} and the MM conformational energies E_i^{MM}:

$$RMSE = \sqrt{\frac{\sum_i w_i (E_i^{QM} - E_i^{MM} + c)^2}{\sum_i w_i}} \quad (2)$$

where w_i is a weight factor for conformation i and the constant c, defined by

$$\frac{\partial RMSE}{\partial c} = 0 \quad (3)$$

optimally aligns the QM and MM data.[56] When simultaneously fitting the same dihedral parameters in two or more different molecules (e.g., configurational isomers), it is sometimes desirable to allow for independent rms alignment of the QM and MM data on a per-molecule basis. Thus the target function for MCSA fitting was expanded to

$$RMSE_{sum} = \sum_g w_g RMSE_g \quad (4)$$

where $RMSE_g$ is the rmse error for a grouping of data and is defined by eqs 2 and 3 for that grouping g. Each grouping of data is aligned independently, that is the c values are computed separately for each grouping, and w_g is a weight factor applied to a particular grouping. Thus, for example, data for configurational isomers can be placed in separate groupings so as to optimally fit the conformational energies of the individual configurational isomers with a single set of dihedral parameters without requiring fitting of the QM ΔE values between configurational isomers.

The aqueous solution density ρ was calculated from the average volume $\langle V \rangle$, obtained from the simulations, using eqs 5 and 6:

$$\rho = \frac{N}{\langle V \rangle} \quad (5)$$

$$N = \frac{(N_{water} + N_{solute})\langle MW \rangle}{N_{Avogadro}} \quad (6)$$

Here, N_{water}, N_{solute}, and $N_{Avogadro}$ are the number of water molecules, solute molecules, and Avogadro's number, respectively, and $\langle MW \rangle$ is the average molecular weight of the solution. For all aqueous density simulations, $N_{water} = 1\,100$, while N_{solute} was adjusted to give the appropriate density for each simulated system. These aqueous simulations were equilibrated for 0.5 ns followed by 5 ns of data collection, with the volume of the system calculated every 10 ps and averaged over the simulation time to give $\langle V \rangle$.

NMR heteronuclear three-bond proton−carbon coupling constants (in Hz) for the glycosidic angle rotation $^3J_{COCH}$ were computed from the simulations using the modified Karplus equation developed by Tvaroska et al.[57]

$$^3J_{COCH} = 5.7\cos^2\phi - 0.6\cos\phi + 0.5 \quad (7)$$

as well as that by Cloran et al.[58]

$$^3J_{COCH} = 7.49\cos^2\phi - 0.96\cos\phi + 0.15 \quad (8)$$

where ϕ is the C−O−C−H dihedral angle of interest. Similarly, for the three-bond carbon−carbon coupling constants $^3J_{COCC}$, the modified Karplus equation developed by Bose et al. [59]

$$^3J_{COCC} = 3.70\cos^2\phi + 0.18\cos\phi + 0.11 \quad (9)$$

was used, where ϕ is the C−O−C−C dihedral angle of interest. Bose et al. also developed a second simpler equation in which the cos ϕ term was excluded[59]

$$^3J_{COCC} = 3.49\cos^2\phi + 0.16 \quad (10)$$

and this second equation was also used to analyze the MD data. To compute J values from MD simulations, the value of the dihedral angle of interest was tabulated for each snapshot of each disaccharide in the simulation. Each dihedral value was then used as ϕ in one of the above equations to get a J value for that disaccharide in that snapshot. The resultant J values for the dihedral angle of interest were then averaged over all disaccharides in the system and all snapshots from the simulation to get an ensemble averaged value of J for comparison with previously published experimental data. Simulations consisted of four disaccharides in 1 100 molecules of water, corresponding to a concentration of 200 mM, which is approximately 10-fold more concentrated than that of the comparison NMR experiments though still sufficiently dilute so as to minimize the direct solute−solute interactions. Of the four disaccharide molecules, the chirality at the reducing end anomeric carbon was α for one molecule and β for the remaining three, so as to reflect the anomeric ratio for glucopyranose in water (α:β = 1:2).[60] In accord with the comparison NMR experiments performed using pure D_2O, simulations were done at 298 K and 1 atm and with no ions added. Simulations were equilibrated for 0.5 ns, followed by 20 ns of data collection during which a snapshot was taken every 1 ps.

Results and Discussion

Parameter Optimization. The focus of the present work was the development of a transferable set of parameters for the glycosidic linkages between hexopyranoses. Applying parameters developed for hexopyranose monosaccharides and associated model compounds[19] left as parameters to-be-determined only those involving the glycosidic linkage oxygen atom. Starting values for bond, angle, dihedral, LJ, and electrostatic parameters involving this atom were

transferred from published ether parameters[34] and further optimized, as described below, using the model compounds in Figure 1, and, to a limited extent, the data from disaccharide crystals. Parametrization was done in a self-consistent fashion such that whenever one parameter was changed, properties were recomputed and additional parameters reoptimized, if necessary;[21−23] all data presented throughout reflect the final set of self-consistently optimized parameters. The optimized parameters were subsequently applied to full disaccharides for validation by comparison of MD results to available experimental data.

Conformational energies for all compounds were studied using the MP2/cc-pVTZ//MP2/6-31G(d) model chemistry, which provides a reasonable compromise between accuracy and computational expediency. Previously, it was seen that energies for hexopyranose monosaccharides using this model chemistry were in excellent agreement compared to MP2/cc-pVTZ optimized structures.[19] Additional studies on ethylene glycol[35] and 2-ethoxy tetrahydropyran[36] using a variety of model chemistries support the use of the MP2/cc-pVTZ model chemistry for energy calculations on carbohydrates.

1-O-Methylation. *O*-methylation at the C_1 position in hexopyranoses leads to formation of an acetal, thereby preventing spontaneous isomerization between the α- and β-anomers, and is a common chemical modification in synthetic hexopyranose polysaccharides. Force field parameters were developed for this moiety because of its common occurrence and its similarity to glycosidic linkages between hexopyranoses. Compounds **2** and **3**, which are *O*-methyl derivatives of tetrahydropyran, were used as model compound analogs for the α- and β-anomers of 1-*O*-methyl hexopyranoses. The location of *O*-methylation and the geometry of the tetrahydropyran ring were chosen to mimic a 1-*O*-methyl-D-hexopyranose molecule having the energetically favored $^4C_1$ ring conformation.

Tetrahydropyran parameters were those previously developed in the context of hexopyranoses[19] and bond, angle, dihedral, LJ, and partial charge parameters involving the *O*-methyl group were transferred from existing ether parameters.[34] Following this transfer of parameters, those that remained missing included the $O_{ring}-C_1-O_{methyl}$ valence angle and the $O_{ring}-C_1-O_{methyl}-C_{methyl}$, $C_2-C_1-O_{methyl}-C_{methyl}$, and $C_5-O_{ring}-C_1-O_{methyl}$ dihedral parameters. Also in question was the ability of the transferred nonbonded parameters on the O_{ring} and O_{methyl} atoms to capture the energetics of hydrogen bonding in the context of these molecules.

QM MP2/cc-pVTZ//MP2/6-31G(d) scans of the $O_{ring}-C_1-O_{methyl}-C_{methyl}$ dihedral in both the α-anomer **2** and β-anomer **3** were done to characterize the potential energy surfaces and determine the location of the minimum energy conformations of the two anomers (Figure 2a). The global minimum for both surfaces was at $O_{ring}-C_1-O_{methyl}-C_{methyl} = 60°$ for the α-anomer. The minimum on the β-anomer surface was 1.55 kcal/mol higher in energy and located at $O_{ring}-C_1-O_{methyl}-C_{methyl} = -60°$. Although the *O*-methyl moiety is located in the axial position in the α-anomer, it is energetically more favorable than that of the equatorial substituted β-anomer. This well-known result is counter to the general trend that equatorial substitutions on saturated six-membered rings are more favorable than axial

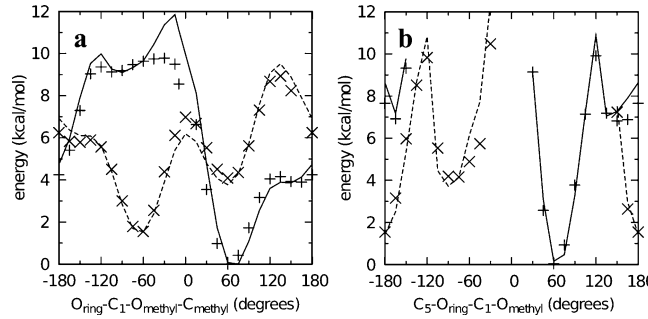


Figure 2. Dihedral potential energy scans in the QM (points) and MM (lines) representation for model compounds **2** and **3**. Scans for the α- (crosses and solid lines) and β- (x's and dashed lines) anomers (**2** and **3,** respectively) have all been offset to the same global minimum.

ones, a phenomenon referred to as the "anomeric effect." The energy difference between the two anomers in the molecular mechanics framework is determined by the $C_5-O_{ring}-C_1-O_{methyl}$ dihedral, and ring deformation potential energy scans of this dihedral were also done for inclusion as target data during dihedral parameter fitting (Figure 2b).

An initial set of parameters was developed by transferring existing ether C−C−O valence angle parameters to the $O_{ring}-C_1-O_{methyl}$ valence angle and fitting the $O_{ring}-C_1-O_{methyl}-C_{methyl}$, $C_2-C_1-O_{methyl}-C_{methyl}$, and $C_5-O_{ring}-C_1-O_{methyl}$ dihedral parameters to reproduce the $O_{ring}-C_1-O_{methyl}-C_{methyl}$ and $C_5-O_{ring}-C_1-O_{methyl}$ QM potential energy scans. Using these initial parameters, both anomers were then fully geometry optimized in the QM MP2/6-31G(d) and MM representations starting from $O_{ring}-C_1-O_{methyl}-C_{methyl} = 60°$ for the α-anomer and −60° for the β-anomer. Comparison of the MM geometries to the QM geometries showed errors in the C_1-O_{methyl} bond and the $O_{ring}-C_1-O_{methyl}$ and $C_2-C_1-O_{methyl}$ angles. Additionally, pair interaction energies with water for these two conformations using the transferred partial charge of −0.34 *e* on O_{methyl} were consistently underestimated in the MM representation compared to that of the QM representation. Because of the errors in geometries and water pair interaction energies using transferred parameters from the linear ethers, further parameter optimization was undertaken.

A second iteration of parameter optimization was done to correct the QM water pair interaction energies, while ensuring the molecular geometries and conformational energies were well reproduced. The C_1-O_{methyl} geometry was improved by decreasing the transferred equilibrium bond length by 0.020 Å, and the $O_{ring}-C_1-O_{methyl}$ and $C_2-C_1-O_{methyl}$ geometries were improved by decreasing the transferred equilibrium value parameters by 1.5° to 110.0° and by 2.5° to 109.0°, respectively. Reproduction of the target vibrational frequencies was improved by increasing the $O_{ring}-C_1-O_{methyl}$ force constant from 45 to 90 kcal·mol^{-1}·radian^{-2}. The partial charge on O_{methyl} was adjusted to −0.36 *e*, with +0.01 *e* added to the partial charges on C_1 and C_{methyl} to maintain charge neutrality, and the dihedral parameters were refit. With optimized partial charges, water pair interaction energies and distances in the MM representation compared very favorably to the QM target data, both for O_{methyl} and for O_{ring}, which retained its initial partial

Table 1. Solute−water Pair Interaction Energies and Distances for Model Compounds **2** and **3**

			energy (kcal/mol)			distance (Å)		
		water orientation[a]	1.16*HF[b]	MM	MM−QM	HF-0.20[b]	MM	MM−QM
2 (α-)	O_{methyl}	ai	−4.87	−4.50	0.37	1.90	1.90	0.00
		aii	−4.64	−4.75	−0.11	1.91	1.89	−0.02
		bi	−4.20	−3.32	0.88	1.90	1.87	−0.03
		bii	−4.35	−4.55	−0.20	1.90	1.83	−0.06
	O_{ring}	ai	−5.88	−5.70	0.18	1.81	1.72	−0.09
		aii	−5.53	−5.79	−0.26	1.84	1.72	−0.12
3 (β-)	O_{methyl}	ai	−5.99	−6.09	−0.09	1.83	1.74	−0.09
		aii	−5.40	−5.56	−0.16	1.86	1.75	−0.11
		bi	−4.20	−3.38	0.82	1.90	1.85	−0.05
		bii	−4.62	−4.83	−0.21	1.89	1.81	−0.08
	O_{ring}	ai	−5.58	−6.16	−0.58	1.88	1.85	−0.03
		aii	−4.78	−4.86	−0.08	1.91	1.88	−0.03
		average			0.05			−0.06
		standard deviation			0.44			0.04

[a] Molecular geometries are as illustrated in Figure 3. [b] HF target energies have been scaled by 1.16, and distances have been shortened by 0.20 Å.

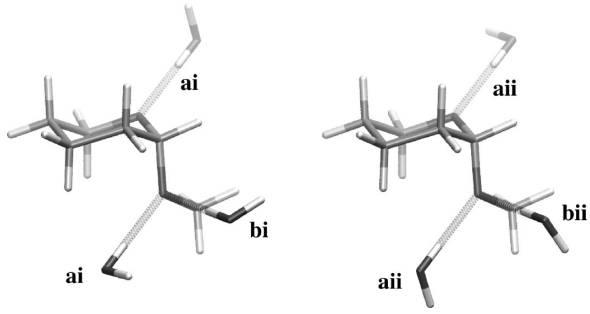


Figure 3. Water pair interaction geometries with model compound **2**. For convenience, three molecules of water have been illustrated simultaneously interacting with the model compound, though in actuality, all data (Table 1) are for a single water molecule interacting with a single model compound molecule. Interactions labeled "a" have the water H−O bond vector in the C−O−C angle plane and along the C−O−C bisector, while interactions labeled "b" have the H−O bond vector in an ideal tetrahedral geometry (109.5°) with respect to the C−O−C angle. The water molecule in type "i" interactions lies perpendicular to the C−O−C plane. Type "ii" interactions differ from type "i" interactions by the rotation of the noninteracting water hydrogen atom by 90° around the H−O bond vector. Molecular graphics prepared with VMD.[66]

charge of −0.40 *e* transferred from tetrahydropyran (Table 1, Figure 3). In order to use the same bonded parameters for both anomers, it was necessary to have the force field slightly underestimate $O_{ring}-C_1-O_{methyl}$ and overestimate $C_2-C_1-O_{methyl}$ for the α-anomer, while doing the reverse for the β-anomer (Table 2). The valence angle parameters also yielded good agreement between the QM and MM vibrational frequencies having contributions from these degrees of freedom (Table 3). Conformational energies as well as the energy difference between the two anomers were very well reproduced with the fit dihedral parameters (Figure 2). In the conformational energy scans, the force field gives a ΔE of 1.53 kcal/mol between the minimum energy conformations of the two anomers, in comparison to the reference MP2/cc-pVTZ//MP2/6-31G(d) value of 1.55 kcal/mol. In this case, the force field representation gives a significantly more accurate ΔE than the MP2/6-31G(d) value (2.25 kcal/mol), highlighting the importance of the choice of QM model chemistry for the generation of target data.

Table 2. Fully-Optimized QM and MM Geometries for Model Compounds **2** and **3**

		QM	MM	MM−QM
2 (α-)	$O_{ring}-C_1-O_{methyl}$[a]	112.0	110.3	−1.6
	$C_2-C_1-O_{methyl}$	107.1	108.8	1.8
	$C_1-O_{methyl}-C_{methyl}$	112.3	112.0	−0.3
	$O_{ring}-C_1-O_{methyl}-C_{methyl}$	61.1	67.1	6.0
	C_1-O_{methyl}	1.411	1.404	−0.008
	$O_{methyl}-C_{methyl}$	1.427	1.424	−0.003
3 (β-)	$O_{ring}-C_1-O_{methyl}$	108.3	109.9	1.5
	$C_2-C_1-O_{methyl}$	108.6	106.8	−1.8
	$C_1-O_{methyl}-C_{methyl}$	112.7	111.9	−0.7
	$O_{ring}-C_1-O_{methyl}-C_{methyl}$	−62.6	−65.3	−2.8
	C_1-O_{methyl}	1.392	1.401	0.009
	$O_{methyl}-C_{methyl}$	1.427	1.423	−0.005

[a] Valence and dihedral angles are in degrees, bond lengths are in Å.

1→1 Glycosidic Linkages. Four different 1→1 glycosidic linkages can be formed between two hexopyranose monosaccharides depending on which anomers are linked together: α(1→1)α, α(1→1)β, β(1→1)α, and β(1→1)β. This number is reduced to three if the identities of the two component monosaccharides are the same, as the resultant α(1→1)β and β(1→1)α linked disaccharides constitute identical molecules. Accordingly, **4**, **5**, and **6** were used as model compound analogs of 1→1 linked hexopyranose disaccharides. These model compounds are dimers of tetrahydropyran connected by a glycosidic linkage and are analogous to α(1→1)β, α(1→1)α, and β(1→1)β hexopyranose disaccharides, respectively. Parameters to be determined for these linkages were those involving the glycosidic oxygen O_{link}. These included $O_{link}-C_1$ bond parameters, $C_1-O_{link}-C_1$, $O_{link}-C_1-O_{ring}$, and $O_{link}-C_1-C_2$ valence angle parameters, $C_1-O_{link}-C_1-X$ dihedral parameters (X = C_2, or O_{link}), and nonbonded O_{link} parameters.

An attractive possibility for minimizing computer time and parametrization effort is to transfer analogous parameters from **2** and **3** to the full disaccharides, and this approach has been taken in other work. To test this approach, the QM

Table 3. Vibrational Frequencies for Model Compound **2** Having Contributions from the $O_{ring}-C_1-O_{methyl}$ and $C_2-C_1-O_{methyl}$ Valence Angles

	frequency (cm^{-1})		% XCZ contribution[a,b]		% YCZ contribution[a,b]	
frequency no.	QM	MM	QM	MM	QM	MM
2	125.8	131.4	13	14		
3	170.2	187.3	7	7	7	9
4	221.1	245.4		10		
5	297.6	286.4	18	31		
6	310.8	337.2	5	10	12	15
7	380.6	366.5	8			
8	415.0	439.8				9
9	504.7	492.3	7	6		
10	553.2	558.8	6		37	29
11	661.2	651.0	5		12	9
12	797.5	797.0				
13	805.0	827.0				
14	855.0	881.6				
15	869.6	893.8				
16	890.9	905.6				
17	946.3	949.5	7			
18	1001.3	995.0				
19	1028.0	1041.1				
20	1032.9	1068.6				
21	1060.1	1075.4				
22	1107.9	1108.8	5			
23	1123.7	1131.7			6	

[a] XCZ and YCZ are methine-associated internal coordinates per the definition of Pulay et al.[53] such that X = C_2, Y = O_{ring}, C = C_1, and Z = O_{methyl}. [b] Contributions from XCZ/YCZ of less than 5% to a particular frequency are not shown.

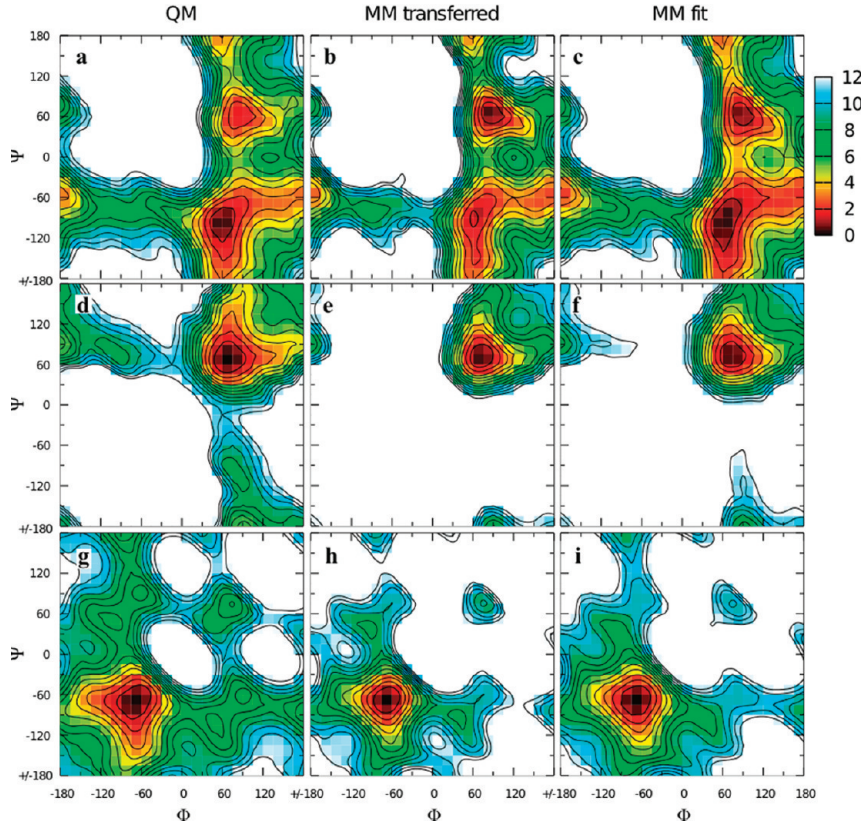


Figure 4. Φ/Ψ potential energy surfaces for model compounds **4** (top row), **5** (middle row), and **6** (bottom row) in the QM (first column), MM using transferred parameters (middle column), and MM using fit parameters (last column) representations. Energies are in kcal/mol with contours every 1 kcal/mol.

Φ/Ψ two-dimensional relaxed potential energy surface for the α(1→1)β analog **4** (Φ = $O_{ring}-C_1-O_{link}-C_1'$, Ψ = $C_1-O_{link}-C_1'-O_{ring}'$) was compared to the MM surface calculated using transferred parameters (Figure 4a,b). The obvious shortcoming of this approach is the incorrect ordering of the two minima on the QM surface at Φ/Ψ = 60°/−90° and 90°/60°; on the QM surface the global minimum is as 60°/−90°, while on the MM surface it is at 90°/60°. Additional QM and MM Φ/Ψ scans done on **5** and **6** showed better agreement between the QM and the MM

surfaces, which is not surprising as they only have one low-lying minimum each (Figure 4d,e,g,h). However, in both instances the minima had notably steeper walls in the MM representation. Starting at the global minimum of 75°/75° for **5** (Figure 4d,e), the energy rises significantly faster moving along the valley toward 180°/90° in the MM as compared to that of the QM surface. Likewise for **6**, starting at the global minimum of −75°/−75° and moving up the valley toward −180°/−60°, the energy increases significantly more quickly in the MM as compared to that of the QM representation (Figure 4g,h). These problems reveal deficiencies in the transferred parameters when applied to disaccharide analogs.

Given the shortcoming of transferring parameters developed for **2** and **3** to the disaccharide analogs, parameters were fit directly to target QM Φ/Ψ data. The $O_{ring}-C_1-O_{link}-C_1$ and $C_2-C_1-O_{link}-C_1$ dihedral parameters that determine Φ/Ψ energetics were fit to **4**, with extra weighting given to points in the two minima and ignoring points greater than 12 kcal/mol above the global minimum (in eq 2, $w_i = 20$ for points in the minima, 0 for points with energy greater than 12 kcal/mol, and 1 otherwise). The optimized parameters based on **4** were then directly transferred to **5** and **6**. Additionally, the $C_1-O_{link}-C_1$ valence angle parameters were optimized to reproduce distortion of this angle as a function of Φ/Ψ for **4**−**6**. This required reduction of the force constant transferred from $C_1-O_{methyl}-C_{methyl}$ from 95 kcal·mol^{-1}·radian^{-2} to a value of 50 kcal·mol^{-1}·radian^{-2} and increasing the equilibrium value from 109.7° to 111.5°.

Very good energies and $C_1-O_{link}-C_1$ valence angle geometries as a function of Φ/Ψ were achieved by the self-consistent optimization of the dihedral and valence angle parameters. The Φ/Ψ energy surface for **4** using these optimized parameters gives the correct ordering of the two minima (Figure 4a,c). The MM model also satisfactorily reproduces the energy surfaces for **5** and **6**, with the walls of the minima being less steep than with the transferred parameters, in better agreement with the QM data (Figure 4d,f,g,i). The MM model also does an excellent job of reproducing the change in $C_1-O_{link}-C_1$ valence angle geometry as a function of Φ/Ψ for all three model compounds (Figure 5). The $C_1-O_{link}-C_1$ geometry varies by more than 20° across the Φ/Ψ surface for each model compound. Quite remarkably, the simple form of the potential energy function, with a harmonic term for valence angle distortion, is able to correctly reproduce such large magnitude valence angle distortions.

Water pair interactions energies and distances, intramolecular geometries of minimum energy conformations, and vibrational frequencies are also all well reproduced using the optimized parameter set. The partial charge and LJ parameters for O_{link}, transferred from the O_{methyl} atom in **2** and **3**, capture the hydrogen-bond energies and distances for **4**−**6** (Table 4). Of note is the fact that for the α(1→1)α analog **5**, the hydrogen-bond length is longer, and the energy is significantly less favorable than for the α(1→1)β and β(1→1)β analogs **4** and **6**. This reflects the fact that water access to the O_{link} atom in the minimum energy geometry of **5** is sterically blocked by the two rings, whereas having

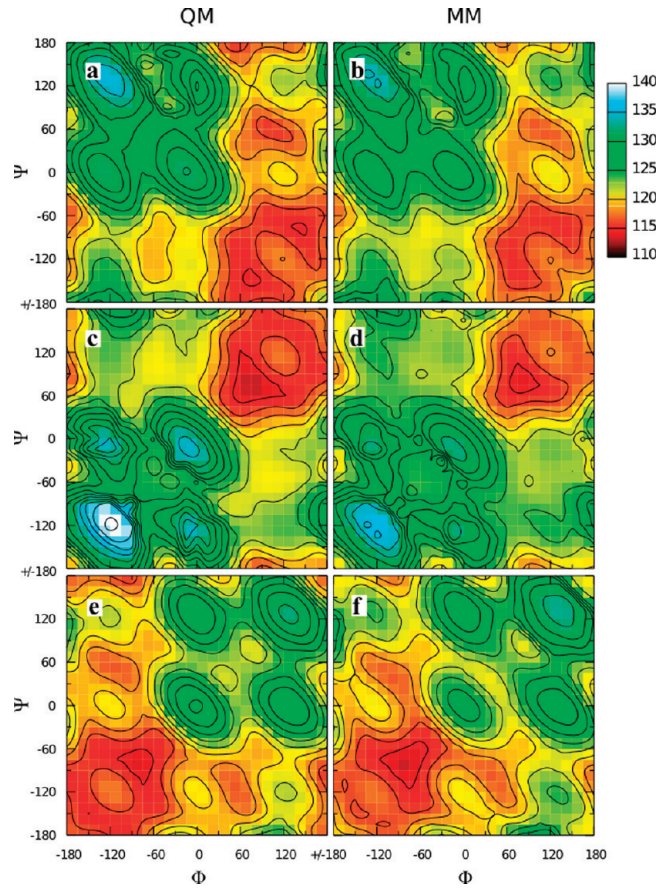


Figure 5. $C_1-O_{link}-C_1'$ valence angle as a function of Φ/Ψ in relaxed potential energy scans for model compounds **4** (top row), **5** (middle row), and **6** (bottom row) in the QM (first column) and MM using fit parameters (last column) representations. Angles are in degrees with contours every 2 degrees.

at least one β-anomer in the linkage opens up access to this atom. Minimum energy structures in the MM representation do a good job of reproducing bond, valence angle, and dihedral angle geometries relative to the QM MP2/6-31G(d) representation (Table 5); we note that the equilibrium length for the C_1-O_{link} bonds in **4**−**6** was set to 1.415 Å, as is done for general linear ethers, compared to the value of 1.395 Å for compounds **2** and **3**. The Φ and Ψ angles are particularly well represented by the MM model (Table 5), even though the same dihedral angle parameters are used for **4**−**6**. Finally, the good agreement between the MM and QM vibrational frequencies (Supporting Information, Figure S1) serves as additional confirmation of the appropriateness of the force field parameters.

1→2, 1→3, and 1→4 Glycosidic Linkages. Using two molecules of tetrahydropyran linked by a bridging ether as model compounds for the 1→2, 1→3, and 1→4 glycosidic linkages leads to a prohibitively large number of possibilities for all combinations of both linkage type and stereochemistry. Because the O_{ring} atom in the second ring in these types of linkages is at least two atoms removed from O_{link}, a cyclohexane molecule was used to represent the second ring for the respective model compounds. For 1→2 linked pyranose disaccharides, the resultant model compounds, with all possibilities for stereochemistry at the C_1 and C_2'

Table 4. Solute−water Pair Interaction Energies and Distances for Model Compounds **4**−**6**

model compound	water orientation[a]	energy (kcal/mol) 1.16*HF[b]	energy (kcal/mol) MM	energy (kcal/mol) MM−QM	distance (Å) HF-0.20[b]	distance (Å) MM	distance (Å) MM−QM
4	−90	−4.82	−4.42	0.40	1.91	1.88	−0.03
	90	−4.92	−5.59	−0.68	1.96	1.85	−0.11
5	90	−1.93	−2.72	−0.79	2.51	2.49	−0.02
6	90	−5.75	−5.64	0.10	1.85	1.77	−0.09
	average			−0.24			−0.06
	standard deviation			0.58			0.04

[a] Dihedral angle (degrees) defined by $C_1-O_{link}\cdots O_{water}-H_2$, where H_2 is the noninteracting water hydrogen atom. In the cases of **5** and **6**, 90 and −90 are equivalent due to the internal symmetry of these two model compounds. In all cases, the water H_1-O_{water} bond vector is in the $C_1-O_{link}-C_1'$ angle plane and along the $C_1-O_{link}-C_1'$ bisector. [b] HF target energies have been scaled by 1.16, and distances have been shortened by 0.20 Å.

Table 5. Optimized Geometries for Model Compounds **4**−**6** in the MP2/6-31G(d) QM and MM Representations

model compound	valence angle/ dihedral angle/ bond length[a]	QM	MM	MM−QM
4	$O_{ring}-C_1-O_{link}$	112.7	110.9	−1.8
	$O_{ring}'-C_1'-O_{link}$	107.6	109.7	2.1
	$C_2-C_1-O_{link}$	106.2	108.7	2.6
	$C_2'-C_1'-O_{link}$	108.8	106.8	−2.0
	$C_1-O_{link}-C_1'$	114.4	115.0	0.6
	Φ	58.1	60.3	2.2
	Ψ	−97.1	−95.9	1.2
	C_1-O_{link}	1.428	1.419	−0.009
	$C_1'-O_{link}$	1.403	1.418	0.014
5	$O_{ring}-C_1-O_{link}$	111.2	109.7	−1.5
	$C_2-C_1-O_{link}$	106.7	108.7	1.9
	$C_1-O_{link}-C_1'$	112.7	113.2	0.6
	Φ (=Ψ)	68.6	73.1	4.4
	C_1-O_{link}	1.425	1.417	−0.008
6	$O_{ring}-C_1-O_{link}$	107.5	109.4	1.9
	$C_2-C_1-O_{link}$	108.1	106.9	−1.3
	$C_1-O_{link}-C_1'$	113.6	113.0	−0.6
	Φ (=Ψ)	−68.8	−67.9	0.9
	C_1-O_{link}	1.405	1.414	0.009

[a] Valence angles and dihedral angles are in degrees, bond lengths are in Å.

positions, are compounds **7**−**10**. Additionally, simply by renumbering the atoms on the cyclohexane ring, these four molecules correspond to all possible stereochemistries for the 1→3 and 1→4 glycosidic linkages. Thus, energies and geometries of these four molecules were used to develop parameters for 1→2, 1→3, and 1→4 linked hexopyranose disaccharides. As with the 1→1 linkages, a single set of parameters was developed for all model compounds in the series.

Parameters, except for the $O_{ring}-C_1-O_{link}-C_2'$, $C_2-C_1-O_{link}-C_2'$, $C_1-O_{link}-C_2'-C_1'$, and $C_1-O_{link}-C_2'-C_3'$ dihedral parameters, were transferred from the 1→1 linkages, and cyclohexane parameters were those previously developed in the context of hexopyranoses.[19] Automated fitting of these four dihedrals, with the $C_1-O_{link}-C_2'-C_1'$ and $C_1-O_{link}-C_2'-C_3'$ parameters equivalenced (i.e., constrained to be the same),[56] was undertaken by including in the fit all points having QM energies less than 12 kcal/mol relative to the respective global minima on the four two-dimensional MP2/cc-pVTZ//MP2/6-31G(d) scans. While the fits were generally good, two issues arose: systematic deviation of the MM C_1-O_{link} and $O_{link}-C_2'$ bond and $C_1-O_{link}-C_2'$ angle values from the QM data, and difficulty reproducing the global

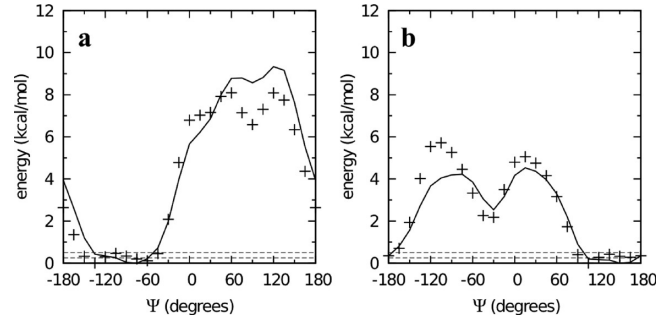


Figure 6. Φ = −60° slices from the two-dimensional Φ/Ψ potential energy surfaces for model compounds **9** (a) and **10** (b) in the QM (crosses) and MM using fit parameters (solid lines) representations. Dashed horizontal lines are at 0.25 and 0.50 kcal/mol.

energy minima on the surfaces for **9** and **10**, where the energy is nearly constant over a large range of Ψ for a value of Φ = −60° (Φ = $O_{ring}-C_1-O_{link}-C_2'$, Ψ = $C_1-O_{link}-C_2'-C_1'$).

To overcome these limitations, a special weighting scheme was applied to the dihedral fitting, and the relevant bond and angle parameters were optimized to target the QM data. With regard to dihedral fitting, the automated dihedral fitting algorithm was extended to allow separate rms alignment of each of the four MM surfaces to its corresponding QM surface during the simultaneous fitting of all four surfaces, as described in the Methods Section. This allowed the fitting to overcome any discrepancies in the MM ΔE values between any pair of surfaces as compared to the QM ΔE. Additionally, points with an energy value within 0.25 kcal/mol of the global minimum for each surface were given a weight w_i of 1 000, compared to a weight of 1 for other points below the 12 kcal/mol cutoff.[56] This emphasis on a few low-lying points was helpful in fitting the flat global minima for compounds **9** and **10** (Figure 6). There were sufficiently few points with this high weighting such that, combined with optimized bond and angle parameters, the full shapes of the energy surfaces were all well reproduced (Figure 7). Bond optimization entailed adjusting the equilibrium values of the C_1-O_{link} and $O_{link}-C_2'$ bonds to 1.395 and 1.435 Å, respectively. Additionally, the final optimized $C_1-O_{link}-C_2'$ parameters (force constant 50 kcal·mol^{-1}·radian^{-2} and equilibrium value of 109.2°) gave impressive reproduction of the angle distortion as a function of Φ/Ψ (Figure 8).

Intramolecular geometries and vibrational frequencies demonstrated the appropriateness of the final bonded pa-

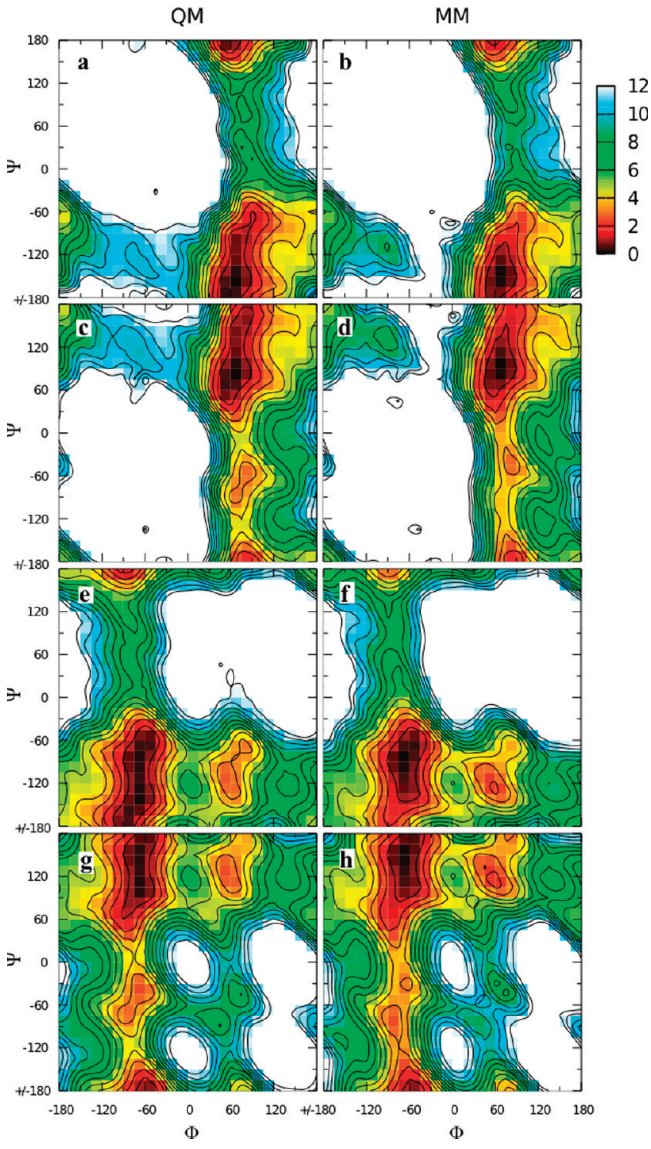


Figure 7. Φ/Ψ potential energy surfaces for model compounds **7** (top row), **8** (second row), **9** (third row), and **10** (bottom row) in the QM (first column) and MM using fit parameters (last column) representations. Energies are in kcal/mol with contours every 1 kcal/mol.

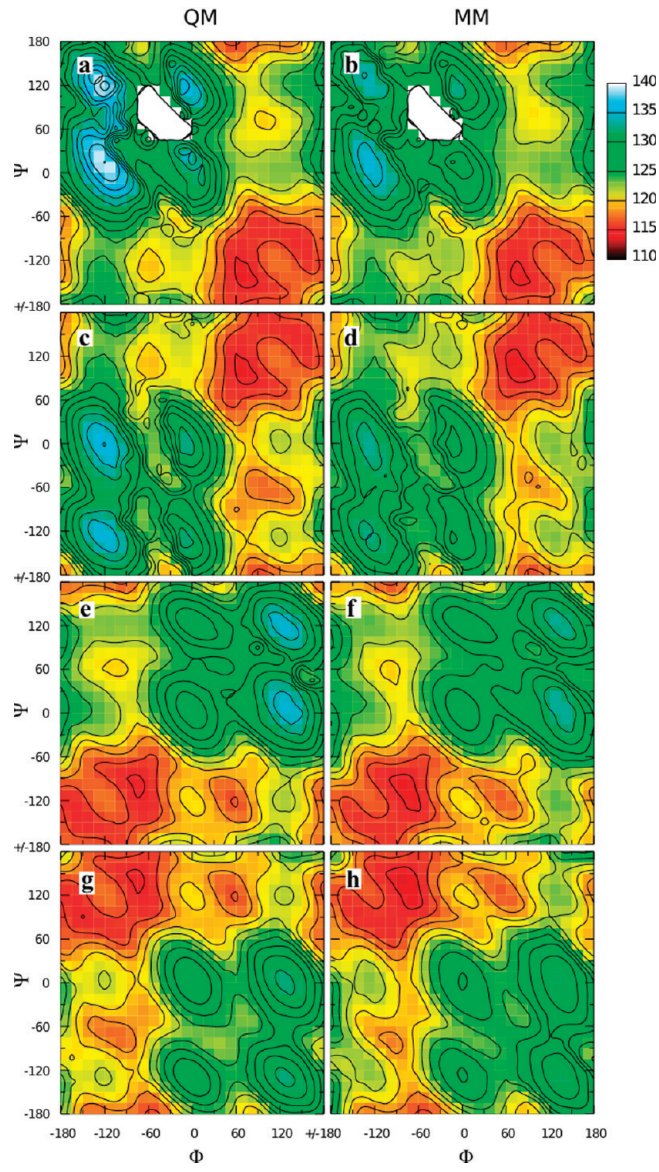


Figure 8. $C_1-O_{link}-C_2'$ valence angle as a function of Φ/Ψ in relaxed potential energy scans for model compounds **7** (top row), **8** (second row), **9** (third row), and **10** (bottom row) in the QM (first column) and MM using fit parameters (last column) representations. Angles are in degrees with contours every 2 degrees.

rameters, and water pair interaction distances and energies confirmed the transferability of the O_{link} partial charge and LJ nonbonded parameters from model compounds **4**−**6**. Fully unconstrained MP2/6-31G(d) geometry optimizations done on the global minimum energy conformations from the two-dimensional Φ/Ψ scans showed little change in the value of the dihedral angles relative to the constrained values from the scans. Using the geometries from fully unconstrained QM optimizations, MM optimizations on compounds **7** and **8** gave structures in excellent agreement with the optimized QM structures (Table 6). However, for model compounds **9** and **10**, the Ψ dihedral needed to be restrained to the QM value during the MM geometry optimization to prevent changes in this angle by +49.3° and +46.2°, respectively. It is important to note that this simply reflects the very flat minimum energy well that these conformations exist in (Figures 6 and 7e−h) and not a significant deficit in the force field. From Figure 6, it is apparent that Ψ values spanning 60° for **9** and **10** in the global minimum energy wells correspond to conformations within 0.5 kcal/mol of each other, with a majority of these being within 0.25 kcal/mol of each other. Because of the differences in the QM- versus unrestrained MM-optimized conformations for **9** and **10**, vibrational frequencies for comparison were computed only for **7** and **8**, and these had the same degree of accuracy as for model compounds **2**−**6** (Supporting Information, Figure S2).

Water pair interaction energies and distances for the QM and MM optimized conformations showed excellent agreement for **8** and **10**, while hydrogen bonds were shorter and stronger in the MM representation relative to the QM for **7** and **9** (Table 7). Importantly, the hydrogen bonds for **8** and **10** are significantly shorter and more favorable in both representations than for **7** and **9**, which have relatively long

Table 6. Optimized Geometries for Model Compounds **7**−**10** in the MP2/6-31G(d) QM and MM Representations

model compound	valence angle/ dihedral angle/ bond length[a]	QM	MM	MM−QM
7	$O_{ring}-C_1-O_{link}$	111.8	110.2	−1.7
	$C_2-C_1-O_{link}$	106.7	108.4	1.7
	$C_1'-C_2'-O_{link}$	106.3	107.2	0.9
	$C_3'-C_1'-O_{link}$	109.7	109.4	−0.3
	$C_1-O_{link}-C_2'$	113.7	114.0	0.3
	Φ	66.3	68.0	1.7
	Ψ	−151.9	−149.2	2.7
	C_1-O_{link}	1.414	1.400	−0.014
	$C_2'-O_{link}$	1.446	1.441	−0.004
8	$O_{ring}-C_1-O_{link}$	111.9	110.2	−1.7
	$C_2-C_1-O_{link}$	106.8	108.5	1.8
	$C_1'-C_2'-O_{link}$	110.5	109.3	−1.2
	$C_3'-C_1'-O_{link}$	107.1	107.3	0.2
	$C_1-O_{link}-C_2'$	113.6	113.7	0.0
	Φ	66.3	67.6	1.3
	Ψ	88.2	91.5	3.3
	C_1-O_{link}	1.414	1.401	−0.013
	$C_2'-O_{link}$	1.441	1.440	−0.001
9	$O_{ring}-C_1-O_{link}$	108.6	110.3	1.7
	$C_2-C_1-O_{link}$	108.5	106.6	−1.9
	$C_1'-C_2'-O_{link}$	106.5	107.6	1.1
	$C_3'-C_1'-O_{link}$	109.4	108.8	−0.7
	$C_1-O_{link}-C_2'$	114.4	114.1	−0.3
	Φ	−64.6	−61.0	3.7
	Ψ	−136.8	[b]	
	C_1-O_{link}	1.391	1.398	0.007
	C_2'-O_{link}	1.447	1.442	−0.006
10	$O_{ring}-C_1-O_{link}$	108.6	110.1	1.5
	$C_2-C_1-O_{link}$	108.5	106.7	−1.9
	$C_1'-C_2'-O_{link}$	110.0	108.8	−1.1
	$C_3'-C_1'-O_{link}$	107.5	107.8	0.2
	$C_1-O_{link}-C_2'$	114.2	113.7	−0.5
	Φ	−63.9	−62.7	1.2
	Ψ	104.2	[b]	
	C_1-O_{link}	1.391	1.398	0.008
	$C_2'-O_{link}$	1.443	1.440	−0.003

[a] Valence and dihedral angles are in degrees, bond lengths are in Å. [b] Value restrained to the QM geometry during optimization.

and weak hydrogen-bonding interactions with the water molecule regardless of its orientation. Thus, the MM model is properly reproducing the relative hydrogen-bond strengths for the different anomers. The disagreement for the weaker hydrogen-bonding interactions may be ascribed to the inability of the water molecule to approach the O_{link} atom for **7** and **9** due to steric hindrance, and the MM model predicts a shorter interaction distance because it properly includes favorable dispersion interactions via the LJ force field term, whereas the Hartree−Fock level of theory used in the QM water pair interaction calculations does not model dispersion.

As discussed at the beginning of this section, the use of two molecules of tetrahydropyran linked by a bridging ether as model compounds for all possible 1→2, 1→3, and 1→4 glycosidic linkages would have been prohibitive. To confirm the appropriateness of the dihedral parameters developed using **7**−**10**, MM scans using these parameters were done on analogs of **7**−**10** ($\mathbf{7}_O$, $\mathbf{8}_O$, $\mathbf{9}_O$, $\mathbf{10}_O$) in which cyclohexane was replaced by a second molecule of tetrahydropyran and compared to QM data. The scans were done at fixed values of Φ, chosen based on the global minima on the Φ/Ψ surfaces for **7**−**10** (Figure 9). In all four cases, the MM scans capture the features of the QM scans, and for $\mathbf{8}_O$, $\mathbf{9}_O$, and $\mathbf{10}_O$ (Figure 9b−d), the global minimum is correctly captured. In the case of $\mathbf{7}_O$ (Figure 9a), the global minimum is at Ψ = −90° in the QM representation as compared to that of Ψ = −150° in the MM. More than anything else, this is a reflection of the flat well surrounding the global minimum, also seen for compound **7** (Figure 7a). In the case of $\mathbf{7}_O$, energies for conformations in this well ranging from Ψ = −180° to Ψ = −75° are within 1.5 kcal/mol of each other. The MM representation correctly captures the span of this wide well and also the adjacent peak at Ψ = +15°. Thus, the dihedral parameters developed on **7**−**10** appear also to be appropriate when the cyclohexane ring is replaced by a second tetrahydropyran ring.

1→6 Glycosidic Linkages. Compared to 1→1, 1→2, 1→3, and 1→4 linkages, 1→6 glycosidic linkages have an additional dihedral degree of freedom making full QM characterization of an analogous model compound too computationally expensive. The three dihedral degrees of freedom in a 1→6 linkage (Φ = $O_{ring}-C_1-O_{link}-C_6'$, Ψ = $C_1-O_{link}-C_6'-C_5'$, Ω = $O_{link}-C_6'-C_5'-O_{ring}'$) were, thus, parametrized based on the results for **2** and **3** as models for the Φ dihedral angle and **11** and **12** as models for the Ψ and Ω dihedral angles. A single set of dihedral parameters gave excellent results for the Ψ/Ω surfaces for both **11** and **12** (Figure 10). These dihedral parameters along with optimization of the $O_{link}-C_6'-C_5'$ angle geometry also gave excellent minimum energy geometries (Table 8) and good agreement with vibrational frequencies (Supporting Information, Figure S3), completing the parameter development for all the model compounds.

Further Parameter Optimization in Full Disaccharides. The parameters developed in the model compound analogs of disaccharides, along with existing parameters for hexopyranose monosaccharides,[19] enabled the construction and simulation of disaccharides in the crystalline state (Table 9). Analysis of intramolecular geometries obtained from MD simulations of various disaccharide crystals showed several systematic differences in the bond and valence angle geometries, as determined by the model compound parameters, and these parameters were further optimized to correct for these systematic deviations. For $C_1-O_{link}-C_1'$ linked compounds, the model compound vs full disaccharide parametrized equilibrium valence angle values were as follows: $C_2-C_1-O_{link}$ 109.0° vs 105.0° and $O_5-C_1-O_{link}$ 110.0° vs 112.0°, with the model compound values followed by the full disaccharide values. For $C_1-O_{link}-C_n'$ (n = 2, 3, 4), the equilibrium bond and angle values were: $O_{link}-C_n'$ 1.435 vs 1.415 Å, $C_2-C_1-O_{link}$ 109.0° vs 105.0°, and $O_5-C_1-O_{link}$ 110.0° vs 112.0°. Of note is that the changes to the $C_2-C_1-O_{link}$ and $O_5-C_1-O_{link}$ angle parameters were the same for both $C_1-O_{link}-C_1'$ and $C_1-O_{link}-C_n'$ (n = 2, 3, 4) linked disaccharides relative to the model compounds. Perhaps it is not surprising that these two angle parameters required special attention; even in the case of model compounds **2** and **3**, the changes in QM angle geometries going from the α- to the β-anomer resulted in MM errors having magnitudes of 1.5° to 1.8° but of differing signs depending on the anomer (Table 2), attesting to the challenge

Table 7. Solute−water Pair Interaction Energies and Distances for Model Compounds **7**−**10**

		energy (kcal/mol)			distance (Å)		
model compound	water orientation[a]	1.16*HF[b]	MM	MM−QM	HF-0.20[b]	MM	MM−QM
7	−180	−1.19	−2.48	−1.29	3.10	2.84	−0.25
	−90	−1.28	−2.60	−1.33	2.96	2.81	−0.15
	0	−0.88	−2.19	−1.32	3.03	2.84	−0.18
	90	−1.24	−2.27	−1.03	2.99	2.84	−0.15
	average			−1.24			−0.19
	standard deviation			0.14			0.05
8	−180	−3.88	−4.19	−0.32	2.00	2.07	0.06
	−90	−4.16	−4.17	−0.01	1.99	2.06	0.07
	0	−3.27	−3.49	−0.21	2.24	2.10	−0.14
	90	−3.61	−3.75	−0.14	2.11	2.09	−0.03
	average			−0.17			−0.01
	standard deviation			0.13			0.10
9	−180	−1.87	−2.98	−1.11	2.82	2.64	−0.17
	−90	−2.13	−3.26	−1.13	2.60	2.62	0.02
	0	−1.56	−2.57	−1.02	2.53	2.67	0.14
	90	−1.97	−2.45	−0.49	2.67	2.67	0.00
	average			−0.94			−0.01
	standard deviation			0.30			0.13
10	−180	−5.56	−6.13	−0.57	1.87	1.75	−0.12
	−90	−6.37	−6.56	−0.19	1.82	1.74	−0.08
	0	−5.46	−5.42	0.04	1.87	1.77	−0.10
	90	−6.05	−5.44	0.62	1.83	1.76	−0.06
	average			−0.02			−0.09
	standard deviation			0.50			0.03

[a] Dihedral angle (degrees) defined by $C_1-O_{link}\cdots O_{water}-H_2$ where H_2 is the noninteracting water hydrogen atom. In all cases, the water H_1-O_{water} bond vector is in the $C_1-O_{link}-C_1'$ angle plane and along the $C_1-O_{link}-C_1'$ bisector. [b] HF target energies have been scaled by 1.16, and distances have been shortened by 0.20 Å.

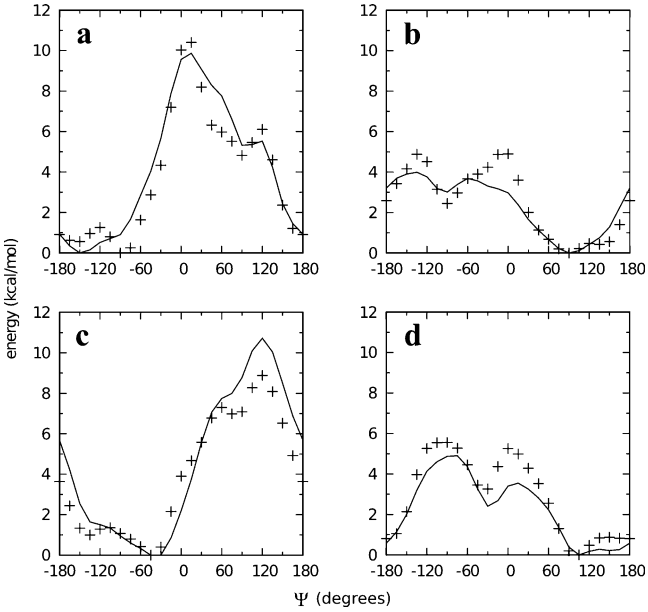


Figure 9. Energy as a function of Ψ for analogs of compounds **7**−**10** (**7**$_O$, **8**$_O$, **9**$_O$, **10**$_O$) in which the cyclohexane ring has been replaced with a second tetrahydropyran ring. Φ = +60° for compounds **7**$_O$ (a) and **8**$_O$ (b), and Φ = −60° for compounds **9**$_O$ (c) and **10**$_O$ (d). QM energies are represented as crosses and MM energies (using dihedral parameters developed on compounds **7**−**10**) as solid lines.

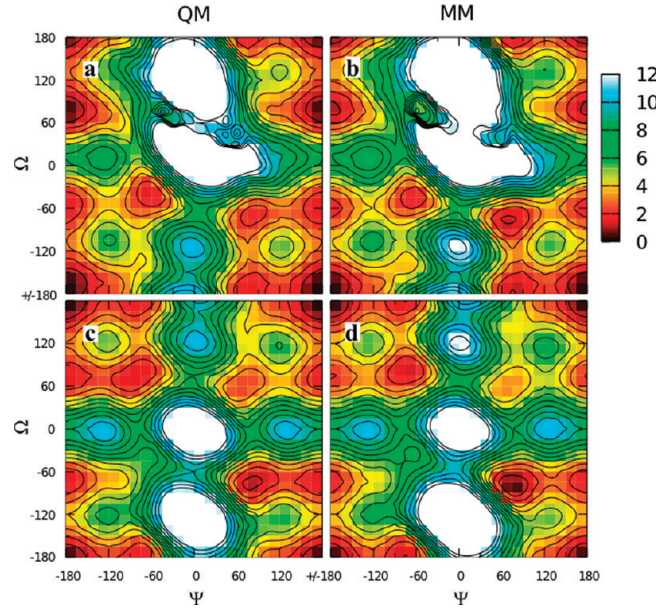


Figure 10. Ψ/Ω potential energy surfaces for model compounds **11** (top row) and **12** (bottom row) in the QM (first column) and MM using fit parameters (last column) representations. Energies are in kcal/mol with contours every 1 kcal/mol.

of developing general parameters involving the anomeric oxygen. The final set of parameters for all linkages between hexopyranoses is listed in Supporting Information, S4−S9.

Validation. A number of MD simulations were undertaken to validate the development of bonded and nonbonded parameters for glycosidic linkages and included a variety of disaccharides both in the aqueous phase and the crystalline phase. Crystal simulations were used primarily to confirm

Table 8. Optimized Geometries for Model Compounds **11** and **12** in the MP2/6-31G(d) QM and MM Representations

model compound	valence angle/ dihedral angle/ bond length[a]	QM	MM	MM−QM
11	$C_1-O_{link}-C_6'$	111.5	111.2	−0.3
	$O_{link}-C_6'-C_5'$	107.1	107.6	0.5
	$C_6'-C_5'-O_{ring}$	110.5	110.4	−0.1
	$C_6'-C_5'-C_4'$	113.5	114.3	0.8
	Ψ	178.9	179.3	0.4
	Ω	−170.8	−169.7	1.1
	$O_{link}-C_6'$	1.420	1.424	0.004
	$C_6'-C_5'$	1.525	1.510	−0.015
12	$C_1-O_{link}-C_6'$	111.4	111.1	−0.3
	$O_{link}-C_6'-C_5'$	107.6	107.8	0.2
	$C_6'-C_5'-O_{ring}$	105.1	106.8	1.7
	$C_6'-C_5'-C_4'$	113.2	113.1	−0.1
	Ψ	179.3	179.5	0.2
	Ω	177.7	177.8	0.1
	$O_{link}-C_6'$	1.418	1.424	0.006
	$C_6'-C_5'$	1.515	1.504	−0.011

[a] Valence and dihedral angles are in degrees, bond lengths are in Å.

the ability of the force field to reproduce intramolecular geometries, thereby testing bond, angle, and dihedral parameters, with additional analysis of crystal cell volumes and their relevance to the nonbonded parameters. Aqueous phase simulations served to test the conformational properties of the various glycosidic linkages, as determined by the dihedral parameters, as well as the ability of the force field model to capture aqueous solution density as a function of concentration. The results of these extensive simulations compared favorably to the available experimental data, thereby serving to validate the force field parametrization.

Crystalline Disaccharide Intramolecular Geometries and Unit Cell Parameters. A comprehensive list of disaccharides representative of the parametrized linkages was simulated as infinite crystals using MD at constant temperature and pressure to test the ability of the force field to maintain intramolecular geometries and unit cell parameters. This list of 16 compounds, taken from the Cambridge Structural Database (CSD) of small molecules,[61] included disaccharides containing not only glucose as a component hexopyranose monosaccharide but also allose, galactose, and mannose (Table 9). Additionally, a number of the disaccharides were *O*-methylated at the reducing end pyranose and/or contained water molecules in the crystal unit cell, both of which further tested the robustness of the force field. Finally, in all cases the simulated unit cell contained at least two independent disaccharide molecules, and a number of the simulations had four independent disaccharides. Overall, MD results of intramolecular geometries and unit cell parameters tabulated by averaging over snapshots from the trajectories and over the independent disaccharide molecules, in the case of intramolecular geometries, were quite consistent with the experimental values.

Errors in the disaccharide intramolecular geometries over the entire set of 16 crystals were impressively small for bonds and valence and dihedral angles involving the glycosidic oxygen O_{link} (Table 10). Both bonds involving this atom had average errors no larger than 0.01 Å, and the spread in errors, as judged by the standard deviation in the errors, was also quite small, reflecting the fact that these bond lengths were well reproduced across the entire set of crystals. The situation was similar for valence angles, with all average errors for valence angles being less than 1.5° and having similarly small standard deviations in the errors. Of particular note is the $C_1-O_{link}-C_n'$ valence angle with an average error of only −0.3°. Parameters for this angle were from careful parametrization aimed at reproducing not only the geometries of minimum energy conformations of the disaccharide analog model compounds **4**−**10** but also the distortion of this angle with changing Φ/Ψ, including high-energy regions in the MP2/cc-pVTZ//MP2/6-31G(d) QM surfaces (Figures 4, 5, 7, and 8). These parameters were directly used in the simulations of the full disaccharides and, along with $C_1-O_{link}-C_6'$ parameters transferred from linear ethers, yielded good reproduction of this glycosidic link angle across the full set of 1→1, 1→2, 1→3, 1→4, and 1→6 linked disaccharides (Figure 11). Finally, Φ, Ψ, and Ω dihedral angle values were well reproduced in the disaccharide crystal simulations using the dihedral parameters developed on model compounds **2**−**12** (Table 10).

Disaccharide crystal unit cell geometries and volumes were also tabulated from the MD simulations of the systems in Table 9 and were in good accord with the experimental values (Table 11). The crystal unit cell length parameters *A*, *B*, and *C* were well maintained for 15 of the 16 simulated systems. In those systems where the crystal unit cell angle parameter β was not 90°, it was allowed to vary during the simulation, and in all of the 9 total systems for which this was the case, β stayed close to the crystallographic value. The outlier with regard to the crystal unit cell parameter data was α,α-allo-trehalose trihydrate (YOXFUG). This case is the most challenging in the entire set in that it is a trihydrate with six molecules of water and two of the disaccharide in the unit cell. Furthermore, it is the only crystal in the set to have P1 symmetry, thus, all three angular unit cell parameters α, β, and γ were allowed to vary unconstrained during the simulation, whereas all other crystals had at most one unconstrained angular parameter (β) (Table 11). It is important to note that unit cell volumes are systematically overestimated across the entire set of compounds, a trend that has been noted in prior parametrizations of cyclic[19] and linear carbohydrates.[20] As noted in these prior works, this likely is a reflection of the difficulty in applying nonbonded parameters developed from simulations of neat liquid alcohols[19] to the highly directional hydrogen-bonding networks in crystals. Also as noted in these prior works, when used in aqueous simulations, the parameters give excellent reproduction of solution densities, as discussed next.

Solution Densities. Aqueous solutions at various concentrations of different disaccharides were simulated under the experimental conditions of 298 K and 1 atm to calculate densities (eqs 5 and 6), which compared favorably with experimental values across all solutes. For all disaccharides and densities, the calculated solution densities are in good agreement with the experimental solution densities, with all errors within 1.5% (Table 12). For the 15 aqueous solutions that were simulated, the average total error in the densities is 0.85%, and the average total absolute error is 1.10%.

Table 9. Disaccharide Crystals

CSD accession code	common name	nonreducing end pyranose	reducing end pyranose	linkage	linkage geometry[a]	no. of disacs[b]	no. of water molecules[b]
DEKYEX	α,α-trehalose	α-glucose	α-glucose	α(1→1)α	aa	2	0
TREHAL01	α,α-trehalose dihydrate	α-glucose	α-glucose	α(1→1)α	aa	4	8
YOXFOG	α,α-allo-trehalose trihydrate	α-allose	α-allose	α(1→1)α	aa	2	6
YOXFUM	α,α-galacto-trehalose	α-galactose	α-galactose	α(1→1)α	aa	4	0
TIQDUS	α,β-trehalose monohydrate	α-glucose	β-glucose	α(1→1)β	ae	2	2
FABYOW10		α-mannose	α-1-*O*-methyl-mannose	α(1→2)	aa	2	0
RESMOR		α-mannose	β-1-*O*-methyl-glucose	α(1→2)	ae	4	0
MOGLPR	methyl α-nigeroside	α-glucose	α-1-*O*-methyl-glucose	α(1→3)	ae	2	0
MALTOS11	β-maltose monohydrate	α-glucose	β-glucose	α(1→4)	ae	2	2
MMALTS	methyl β-maltoside monohydrate	α-glucose	β-1-*O*-methyl-glucose	α(1→4)	ae	4	4
MELIBM10	α-melibiose monohydrate	α-glucose	α-glucose	α(1→6)	ae	4	4
SOPROS	α-sophorose monohydrate	β-glucose	α-glucose	β(1→2)	ee	4	4
WAGBOV	methyl β- laminarabioside monohydrate	β-glucose	β-1-*O*-methyl-glucose	β(1→3)	ee	4	4
BLACTO	β-lactose	β-galactose	β-glucose	β(1→4)	ee	2	0
CELLOB02	β-cellobiose	β-glucose	β-glucose	β(1→4)	ee	2	0
GENTBS01	β-gentiobiose	β-glucose	β-glucose	β(1→6)	ee	4	0

[a] Indicates the configuration at the ring carbons involved in the glycosidic linkage. For example, "ae" indicates that the glycosidic linkage ring carbon on the nonreducing end pyranose has the glycosidic linkage directed axially, while the ring carbon on the reducing end pyranose has it directed equatorially. [b] Number of molecules in the unit cell used for MD simulation.

Table 10. Crystalline Disaccharide Internal Geometries[a]

	C_1-O_{link}	$O_{link}-C_n'$	$C_1-O_{link}-C_n'$	$O_{ring}-C_1-O_{link}$	$C_2-C_1-O_{link}$	$O_{link}-C_n'-C_{n+1}'$[b]	$O_{link}-C_n'-X$[c]	Φ	Ψ	Ω[d]
average error	0.008	0.002	−0.3	0.7	0.9	0.6	1.3	−0.8	−3.9	−3.1
standard deviation of errors	0.008	0.012	1.4	1.2	1.2	1.5	1.2	6.0	8.2	n/a

[a] Data are from simulations of crystals listed in Table 9; MD data for a particular crystal were averaged over the independent disaccharide molecules in the crystal as well as over the MD trajectory; bonds are in Å and valence and dihedral angles are in degrees. [b] Data exclude MELIBM10 and GENTBS01 because these are 1→6 linked disaccharides. [c] $X = O_{ring}'$ for 1→1 linked disaccharides and C_{n-1}' for all others. [d] Data include only MELIBM10 and GENTBS01.

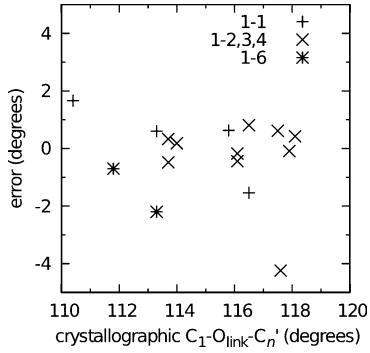


Figure 11. $C_1-O_{link}-C_n'$ valence angle errors in MD simulations of disaccharide crystals. Errors are calculated as the MD value (averaged over the simulation and over all independent disaccharides in the unit cell) minus the crystallographic value.

Moreover, the force field model is able to reproduce experimental results both for dilute and concentrated solutions. Specifically, for simulations of dilute aqueous trehalose, cellobiose, gentiobiose, and melibiose ranging in concentration from 0.1 to 0.3 mol/kg, errors range from 0.87 to 1.48%. For simulations of concentrated maltose, with concentrations between 1.5 to 2.6 mol/kg, errors range from −0.26 to −1.07%. For the most concentrated system, 2.6 *m* maltose, nearly one-third of the atoms in the system belong to solute molecules. As a result, there is significant direct contact between solute molecules, and it is not entirely surprising that the density is somewhat underestimated given that in the limiting case in which there are no water molecules, i.e., crystalline disaccharides, densities are systematically underestimated. More important though is that for all the aqueous systems, errors in density lie within a relatively small window, and the results are quite favorable given the 25-fold difference in concentrations between the most dilute versus the most concentrated systems studied.

Disaccharide Conformational Properties in Aqueous Solution. Karplus-type equations have previously been developed for both vicinal $^{13}C-^{1}H$ and $^{13}C-^{13}C$ spin couplings to relate the observed three-bond coupling constant values ^{3}J to dihedral angle values in the glycosidic linkage (see Methods Section).[57−59] Such equations provide a means of comparing the conformational properties of the glycosidic linkage in the MD simulations of disaccharides in water to the NMR observables. Consequently, they can be used to help validate the dihedral parameters developed using gas-phase MP2/cc-pVTZ//MP2/6-31G(d) QM data from model compounds when applied to full disaccharides in aqueous solution.

The selection of disaccharides for MD *J*-coupling studies was based on the availability of experimental data as well as the desire to test all the different dihedral parameters. As described above, separate glycosidic dihedral parameters were developed for 1→1 linkages (model compounds **4**−**6**; "category 1"), for 1→2, 1→3, and 1→4 linkages (model compounds **7**−**10**; "category 2"), and for 1→6 linkages (model compounds **2**−**3** and **11**−**12**; "category 3"). Accordingly, disaccharides were chosen that fell into each of these three categories: α,α-trehalose was used to test category 1

Table 11. Crystalline Disaccharide Unit Cell Geometries and Volumes

	A (Å)			*B* (Å)			*C* (Å)			*β* (degrees)[a]			volume (Å^3)		
	expt	MD	% error	expt	MD	% error	expt	MD	% error	expt	MD	% error	expt	MD	% error
DEKYEX	12.97	13.16	1.5	8.23	8.15	−0.9	6.79	7.05	3.9	98.12	99.62	1.5	717.4	746.0	4.0
TREHAL01	12.23	12.36	1.1	17.89	17.94	0.3	7.60	7.85	3.3	90.00	90.00		1661.7	1740.7	4.8
YOXFOG	9.40	8.56	−8.9	11.03	11.26	2.0	9.15	10.35	13.1	96.8	96.5	−0.4	853.8	931.0	9.0
YOXFUM	11.07	11.13	0.5	18.28	18.91	3.4	6.87	6.85	−0.2	90.00	90.00		1389.6	1442.0	3.8
TIQDUS	9.77	10.05	2.8	8.58	8.13	−5.2	10.10	10.50	4.0	111.68	109.39	−2.1	786.0	808.1	2.8
FABYOW10	8.09	8.17	1.0	9.78	9.88	1.0	9.98	10.16	1.8	104.58	103.15	−1.4	763.0	797.5	4.5
RESMOR	9.38	9.32	−0.6	10.65	10.76	1.0	15.83	16.57	4.7	90.00	90.00		1580.4	1661.9	5.2
MOGLPR	6.59	6.90	4.6	13.24	13.73	3.7	9.45	9.23	−2.3	108.00	101.40	−6.1	784.1	853.2	8.8
MALTOS11	4.87	5.06	3.9	15.08	15.30	1.4	10.70	10.76	0.5	97.07	97.80	0.8	779.2	823.6	5.7
MMALTS	22.64	24.22	7.0	4.84	4.90	1.3	17.32	17.01	−1.8	117.3	120.7	2.9	1687.2	1736.9	2.9
MELIBM10	8.90	9.08	2.0	10.89	11.03	1.2	15.87	16.03	1.1	90.00	90.00		1538.5	1604.8	4.3
SOPROS	22.05	22.14	0.4	14.44	14.38	−0.4	4.87	5.03	3.5	90.00	90.00		1548.5	1603.3	3.5
WAGBOV	14.55	14.63	0.6	24.25	24.23	−0.1	4.94	5.07	2.7	90.00	90.00		1742.2	1798.0	3.2
BLACTO	10.84	11.03	1.8	13.35	13.55	1.5	4.95	5.06	2.2	91.31	86.43	−5.3	716.6	752.5	5.0
CELLOB02	10.97	11.07	0.9	13.05	13.35	2.3	5.09	5.14	0.9	90.83	91.22	0.4	728.8	758.6	4.1
GENTBS01	8.87	9.01	1.5	22.85	23.31	2.0	7.20	7.22	0.3	90.00	90.00		1459.1	1516.6	3.9
average			1.3			0.9			2.4			−1.1			4.7
standard deviation			3.3			2.1			3.5			3.0			1.8

[a] Constrained to 90° in the simulation if equal to 90° in the experimental crystal, otherwise allowed to vary independently during the simulation.

Table 12. Comparison of the Calculated and Experimental Densities for Different Molal Concentrations of Trehalose, Cellobiose, Gentiobiose and Mellibiose[a]

molality (mol/kg)	N_{solute}	expt	MD	% error
	trehalose + water[64]			
0.106	2	1.011	1.024	1.29
0.159	3	1.018	1.029	1.08
0.210	4	1.024	1.034	0.98
	cellobiose + water[64]			
0.148	3	1.016	1.030	1.38
0.202	4	1.022	1.035	1.27
0.271	5	1.031	1.040	0.87
	gentiobiose + water[64]			
0.146	3	1.016	1.030	1.38
0.205	4	1.023	1.036	1.27
0.270	5	1.030	1.041	1.07
	melibiose + water[64]			
0.144	3	1.015	1.030	1.48
0.200	4	1.022	1.036	1.37
0.255	5	1.029	1.041	1.17
	maltose + water[65]			
1.465	29	1.144	1.141	−0.26
1.717	34	1.163	1.156	−0.60
2.626	52	1.216	1.203	−1.07
Average				0.85

[a] In 1 100 TIP3P waters at $T = 298$ K and $P = 1$ atm.

linkage parameters; maltose and cellobiose was used to test category 2 linkage parameters; and melibiose and gentiobiose was used to test category 3 parameters. In addition to the availability of recent NMR *J*-coupling data for these molecules,[62,63] the tested category 2 and 3 disaccharides are representative of both α and β linkages, which is important given that the same dihedral parameters are used for a given type of linkage regardless of the chiralities at the ring carbon atoms involved in the linkage. While category 2 encompasses 1→2 and 1→3 linkages as well as 1→4 linkages, the 1→4 linked molecules maltose and cellobiose were chosen to represent this category because of their biological significance. Although the set of disaccharides does cover all of the dihedral parameters developed for glycosidic linkages based on the conformational properties of model compounds **2**−**12**, it is worth noting that the configuration at the carbon on the reducing-end monosaccharide involved in the glycosidic linkage in all of these disaccharides is equatorial, reflecting a lack of available experimental data.

The referenced experimental glycosidic $^3J_{COCH}$ coupling constants for trehalose, maltose, cellobiose, melibiose, and gentiobiose span the range of 2.36−4.75 Hz (Table 13). Values for $^3J_{COCH}$ calculated from the MD simulations are representative of this range and correlate well with the experimental data. Interestingly, depending on whether eqs 7 or 8 were used to convert the MD conformational data to $^3J_{COCH}$, the calculated values for a given dihedral in a given simulation differ by as much as 0.86 Hz (cellobiose

Table 13. $^3J_{COCH}$ Coupling Constants (Hz)

		$^3J(H_1C_1O_{link}C_n')$			$^3J(C_1O_{link}C_n'H_n')$		
disaccharide	linkage	expt[62a]	MD eq 7	MD eq 8	expt[62a]	MD eq 7	MD eq 8
trehalose	α(1→1)α	3.19 (0.05)	2.54	2.72			
maltose	α(1→4)	4.12 (0.2)	3.87	4.44	4.68 (0.1)	4.35	5.08
cellobiose	β(1→4)	4.08 (0.04)	3.15	3.51	4.75 (0.25)	4.83	5.69
melibiose	α(1→6)	3.1 (0.4)	2.39	2.52	2.9 (0.4)	1.78/2.04[b]	1.75/2.10[b]
gentiobiose	β(1→6)	2.68 (0.08)	2.50	2.67	2.36 (0.06)	1.87/2.01[b]	1.88/2.05[b]

[a] Experimental errors are in parentheses. [b] Data are for both protons on C_6'.

Table 14. $^3J_{COCC}$ Coupling Constants (Hz)

		^3J ($C_2C_1O_{link}C_n'$)			^3J($C_1O_{link}C_n'C_{n-1}'$)			^3J ($C_1O_{link}C_n'C_{n+1}'$)		
disaccharide	linkage	expt[63]	MD eq 9	MD eq 10	expt[63]	MD eq 9	MD eq 10	expt[63]	MD eq 9	MD eq 10
trehalose	α(1→1)α		3.40	3.43		n/a	n/a			
maltose	α(1→4)		2.57	2.62		0.54	0.57		2.18	2.23
cellobiose	β(1→4)	3.0	3.04	3.08	≤0.5	0.70	0.76	2.0	1.63	1.70
melibiose	α(1→6)		3.42	3.45		3.35	3.37		n/a	n/a
gentiobiose	β(1→6)	3.0	3.40	3.42	2.6	3.30	3.33	n/a	n/a	n/a

$^3J(C_1O_{link}C_n'H_n')$). Furthermore, the variability of the calculated values can lead to underestimation when using one equation and overestimation when using the other, as is the case for both $^3J(H_1C_1O_{link}C_n')$ and $^3J(C_1O_{link}C_n'H_n')$ for maltose. Readers interested in further detail are directed to ref 61 (Table 1), which summarizes both older and more recent experimental and computed $^3J_{COCH}$ values and shows that for e.g., maltose, experimental data can vary by as much as 2.2 Hz and computed data by 2.4 Hz depending on the source. In the case of $^3J_{COCC}$ values, the experimental data are limited to cellobiose and gentiobiose (Table 14). The reference data range from ≤0.5 up to 3.0 Hz, and both the low and the high $^3J_{COCC}$ are faithfully reproduced by the simulations. The overall closeness of the calculated data to the experimental data across all compounds, especially taking into account the inherent uncertainties introduced by using empirical equations to convert observed MD conformational data to NMR ^3J values, suggests the force field parameters are properly reproducing the true aqueous solution conformational behavior of this variety of glycosidic linkages.

Conclusions

The present work extends the developing CHARMM all-atom additive force field for carbohydrates[19,20] to hexopyranose glycosides. The newly developed and validated parameters allow for the modeling of all possible 1→1, 1→2, 1→3, 1→4, and 1→6 linkages between hexopyranoses as well as *O*-methylation at the C_1 position. The parameter development protocol and the force field functional form are consistent not only with recent CHARMM carbohydrate parametrizations[19,20] but also with the CHARMM all-atom additive force fields for proteins, nucleic acids, and lipids; the parameter set is, therefore, an extension of the CHARMM all-atom additive biomolecular force field. Given the comprehensive coverage of glycosidic linkages and the compatibility with other CHARMM biomolecular force fields, the presented force field is expected be of utility for the simulation of linear, cyclic, and branched hexopyranose glycosides in solution, including in heterogeneous systems that include proteins, lipids, and/or nucleic acids.

The present work has focused on the development of a highly optimized force field for the glycosidic linkages between hexopyranoses, with validation focusing on disaccharide crystalline intramolecular geometries and unit cell parameters, solution densities, and conformational properties in aqueous solution. Preliminary data on dynamic properties such as diffusion coefficients and NMR relaxation rates show promise with regard to experimental results, and a thorough analysis of simulated dynamic properties and existing experimental data is under preparation.

Acknowledgment. This work was supported by NIH GM070855 (ADM) and F32CA119771 (OG). The authors acknowledge computer time and resources from the National Cancer Institute Advanced Biomedical Computing Center, Department of Defense High Performance Computing, and the Pittsburgh Supercomputing Center.

Supporting Information Available: Supporting information includes three figures and force field parameters. This material is available free of charge via the Internet at http://pubs.acs.org. The force field topologies and parameters in CHARMM readable format may be obtained from the Web site of ADM at http://mackerell.umaryland.edu.

References

(1) Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*; Harwood Academic Publishers: Amsterdam, The Netherlands, 1998.

(2) Melberg, S.; Rasmussen, K. *Carbohydr. Res.* **1979**, *69*, 27.

(3) Ha, S. N.; Madsen, L. J.; Brady, J. W. *Biopolymers* **1988**, *27*, 1927.

(4) Homans, S. W. *Biochemistry* **1990**, *29*, 9110.

(5) Grootenhuis, P. D. J.; Haasnoot, C. A. G. *Mol. Simul.* **1993**, *10*, 75.

(6) Glennon, T. M.; Zheng, Y. J.; Legrand, S. M.; Shutzberg, B. A.; Merz, K. M. *J. Comput. Chem.* **1994**, *15*, 1019.

(7) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraserreid, B. *J. Phys. Chem.* **1995**, *99*, 3832.

(8) Kouwijzer, M.; Grootenhuis, P. D. J. *J. Phys. Chem.* **1995**, *99*, 13426.

(9) Reiling, S.; Schlenkrich, M.; Brickmann, J. *J. Comput. Chem.* **1996**, *17*, 450.

(10) Ott, K. H.; Meyer, B. *J. Comput. Chem.* **1996**, *17*, 1068.

(11) Senderowitz, H.; Still, W. C. *J. Org. Chem.* **1997**, *62*, 1427.

(12) Momany, F. A.; Willett, J. L. *Carbohydr. Res.* **2000**, *326*, 194.

(13) Eklund, R.; Widmalm, G. *Carbohydr. Res.* **2003**, *338*, 393.

(14) Lii, J. H.; Chen, K. H.; Allinger, N. L. *J. Comput. Chem.* **2003**, *24*, 1504.

(15) Vergoten, G.; Mazur, I.; Lagant, P.; Michalski, J. C.; Zanetta, J. P. *Biochimie* **2003**, *85*, 65.

(16) Lins, R. D.; Hunenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400.

(17) Lii, J. H.; Chen, K. H.; Johnson, G. P.; French, A. D.; Allinger, N. L. *Carbohydr. Res.* **2005**, *340*, 853.

(18) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622.

(19) Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2008**, *29*, 2543.

(20) Hatcher, E. R.; Guvench, O.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2009**, *5*, 1315.

(21) MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The energy function and its paramerization with an overview of the program. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 1; pp 271.

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

(23) Guvench, O.; MacKerell, A. D., Jr. Comparison of protein force fields for molecular dynamics simulations. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Humana Press Inc.: New Jersey, 2008; pp 63.

(24) 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.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

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

(26) MacKerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946.

(27) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.

(28) MacKerell, A. D., Jr.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105.

(29) Schlenkrich, M.; Brinkman, J.; MacKerell, A. D., Jr.; Karplus, M. An empirical potential energy function for phospholipids: criteria for parameter optimization and applications. In *Membrane Structure and Dynamics*; Merz, K. M., Roux, B., Eds.; Birkhauser: Boston, MA, 1996; pp 31.

(30) Feller, S. E.; Yin, D. X.; Pastor, R. W.; MacKerell, A. D., Jr. *Biophys. J.* **1997**, *73*, 2269.

(31) Yin, D. X.; MacKerell, A. D., Jr. *J. Comput. Chem.* **1998**, *19*, 334.

(32) Feller, S. E.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2000**, *104*, 7510.

(33) Feller, S. E.; Gawrisch, K.; MacKerell, A. D., Jr. *J. Am. Chem. Soc.* **2002**, *124*, 318.

(34) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2007**, *3*, 1120.

(35) Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. A* **2006**, *110*, 9934.

(36) Woodcock, H. L.; Moran, D.; Pastor, R. W.; MacKerell, A. D., Jr.; Brooks, B. R. *Biophys. J.* **2007**, *93*, 1.

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

(38) Brooks, B. R.; Brooks, C. L., III.; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.

(39) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U.K., 1987.

(40) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(41) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. *J. Phys. Chem.* **1994**, *98*, 2198.

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

(43) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.

(44) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(45) Hockney, R. W. The potential calculation and some applications. In *Methods in Computational Physics*; Alder, B., Fernbach, S., Rotenberg, M., Eds.; Academic Press: New York, 1970; Vol. 9; pp 136.

(46) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(47) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.

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

(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven Jr., 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, K.; Kitao, O.; Nakai, H.; Klene, M.; Li, T. W.; 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.04 ed.; Gaussian, Inc.: Pittsburgh, PA, 2003.

(50) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(51) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(52) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.

(53) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.

(54) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(55) MacKerell, A. D., Jr.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 10559.

(56) Guvench, O.; MacKerell, A. D., Jr. *J. Mol. Model.* **2008**, *14*, 667.

(57) Tvaroska, I.; Hricovíni, M.; Petráková, E. *Carbohydr. Res.* **1989**, *189*, 359.

(58) Cloran, F.; Carmichael, I.; Serianni, A. S. *J. Am. Chem. Soc.* **1999**, *121*, 9843.

(59) Bose, B.; Zhao, S.; Stenutz, R.; Cloran, F.; Bondo, P. B.; Bondo, G.; Hertz, B.; Carmichael, I.; Serianni, A. S. *J. Am. Chem. Soc.* **1998**, *120*, 11158.

(60) Capon, B. *Chem. Rev.* **1969**, *69*, 407.

(61) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2002**, *58*, 380.

(62) Cheetham, N. W. H.; Dasgupta, P.; Ball, G. E. *Carbohydr. Res.* **2003**, *338*, 955.

(63) Olsson, U.; Serianni, A. S.; Stenutz, R. *J. Phys. Chem. B* **2008**, *112*, 4447.

(64) Galema, S. A.; Hoiland, H. *J. Phys. Chem.* **1991**, *95*, 5321.

(65) Lourdin, D.; Colonna, P.; Ring, S. G. *Carbohydr. Res.* **2003**, *338*, 2883.

(66) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.

CT900242E

JCTC Journal of Chemical Theory and Computation

An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware

M. J. Harvey*,† and G. De Fabritiis*,‡

High Performance Computing Service, Information and Communications Technologies, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom and Computational Biochemistry and Biophysics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain

Received May 28, 2009

Abstract: The smooth particle mesh Ewald summation method is widely used to efficiently compute long-range electrostatic force terms in molecular dynamics simulations, and there has been considerable work in developing optimized implementations for a variety of parallel computer architectures. We describe an implementation for Nvidia graphical processing units (GPUs) which are general purpose computing devices with a high degree of intrinsic parallelism and arithmetic performance. We find that, for typical biomolecular simulations (e.g., DHFR, 26K atoms), a single GPU equipped workstation is able to provide sufficient performance to permit simulation rates of ≈50 ns/day when used in conjunction with the ACEMD molecular dynamics package[1] and exhibits an accuracy comparable to that of a reference double-precision CPU implementation.

I. Introduction

Atomistic molecular mechanics simulations are widely used in the study of a range of biomolecular and inorganic molecular systems. However, the $O(N^2)$ scaling of the electrostatic interactions, combined with the explicit time step scheme, make it challenging to access the microsecond time scale. To reduce the computational cost, the contribution of longer-range interactions is typically approximated by using a scheme with more favorable scaling properties, such a derivative of the Ewald summation method.[2] The smoothed particle mesh Ewald summation (SPME),[3] which scales with $O(N\log N)$ is perhaps the most widely used of these methods in popular MD packages.

Production molecular dynamics (MD) simulations are typically performed on highly parallel multiprocessor or cluster supercomputers. Parallelization of the directly calculated nonbonded electrostatic interactions may be efficiently achieved by using spatial decomposition methods, such as the neutral territory schemes of Bowers et al.[4] as only local neighbor communication is required. Parallelization of the PME method is more challenging because of the long-range communications pattern required by the 3D Fourier transform step, and consequently, the development of efficient implementations remains an active topic of research.[5,6]

Due to the computing resources required to perform MD simulations, there has been considerable research into accelerating the computation by using, for example dedicated, specialized hardware[7] or commodity, high-performance accelerator processors.[8] In recent years, commodity graphics processing units (GPUs) have acquired nongraphical, general purpose programmability and have undergone a doubling of computational power every 12 months. Of the devices currently available on the market, those produced by Nvidia offer the most mature programming environment, the so-called compute unified device architecture (CUDA),[9] and have been the focus of the majority of investigation in the computational science field.

Several groups have lately shown results for MD codes which are accelerated by use of CUDA capable GPUs, for example,[10−14] and it has been recognized that high-performance, low-cost GPU accelerated molecular dynamics simulations could have a significant practical impact in the

* Corresponding authors. E-mail: m.j.harvey@imperial.ac.uk and gianni.defabritiis@upf.edu.

† Imperial College London.

‡ Universitat Pompeu Fabra.

10.1021/ct900275y CCC: $40.75
Published on Web 08/14/2009

field of in silico drug discovery.[15] In recent work, we presented ACEMD, a MD package designed to run on high-performance Nvidia GPUs.[1,16] ACEMD exhibits performance comparable to clusters of ≈100−200 CPUs and enables microsecond scale simulations on a single GPU equipped workstation, representing a significant improvement in the accessibility and cost-efficiency of MD simulation. A principal feature of ACEMD is its SPME treatment of long-range electrostatic terms as this is generally considered a requirement for production quality biomolecular simulations.

In this paper, we focus on the GPU implementation of the SPME scheme used in ACEMD and discuss the resulting performance.[1] The SPME calculation is performed on a single GPU but utilizes the device's massive intrinsic parallelism, necessitating significant modification of the implementation in comparison to a serial or parallel block decomposed version for conventional CPUs.

II. Algorithm and Implementation

A. The SPME Method. The SPME method of Essman and co-workers[17] is an efficient scheme for computing long-range electrostatic forces which exhibits improved $O(N\log N)$ scaling in comparison to the $O(N^2)$ scaling of standard Ewald summation, differing from the latter in the use of interpolation and the 3D FFT in computing the reciprocal space summation.

The electrostatic interaction energy of an N particle system is given by the sum of the pairwise Coulombic interactions:

$$E = \sum_{i=1}^{N-1}\sum_{j>i}^{N}\frac{q_i q_j}{r_{ij}} \tag{1}$$

where q_i is the charge of the ith particle and r_{ij} the distance between the ith and jth particles. The traditional Ewald summation method re-expresses E as the sum of three potentials $E = E_{\mathrm{dir}} + E_{\mathrm{rec}} + E_{\mathrm{self}}$, where E_{dir} is summed over all pairs in real space, E_{rec} is summed in reciprocal space, and E_{self} is a corrective term representing the self-energy of the system. These three potentials are defined as

$$E_{\mathrm{dir}} = \sum_{i=1}^{N-1}\sum_{j>i}^{N}\frac{q_i q_j \mathrm{erfc}(\beta r_{ij})}{r_{ij}} \tag{2}$$

$$E_{\mathrm{rec}} = \frac{1}{2\pi V}\sum_{\vec{m}\neq 0}\frac{\exp(-\pi^2\vec{m}^2/\beta^2)}{\vec{m}^2}S(\vec{m})S(-\vec{m}) \tag{3}$$

$$E_{\mathrm{self}} = -\frac{\beta}{\sqrt{\pi}}\sum_{i}^{N} q_i^2 \tag{4}$$

where $\mathrm{erfc}(x)$ is the complementary error function $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$, β is the Ewald parameter, $\vec{m} = (m_1, m_2, m_3)$ are the reciprocal space lattice vectors, and V is volume of the unit cell in reciprocal space. $S(\vec{m})$ is the lattice structure factor, given by

$$S(\vec{m}) = \sum_{j} q_j \exp(\vec{m}\vec{r}_j) \tag{5}$$

The SPME method approximates this with

$$S(\vec{m}) \approx \tilde{S}(\vec{m}) = b_1(m_1)b_2(m_2)b_3(m_3)F(Q)(m_1, m_2, m_3) \tag{6}$$

where $b_i(m_i)$ are Euler exponential splines given by

$$b_\alpha(m_\alpha) = \exp(2\pi i(n-1)m_\alpha/K_\alpha)\times \left[\sum_{k=0}^{n-2} M_n(m+1)\exp(2\pi i m_\alpha k/K_\alpha)\right]^{-1} \tag{7}$$

for $\alpha = 1, 2, 3$. $\mathrm{F}(Q)$ is the discrete Fourier transform of the 3D matrix Q of dimension $K_1 \times K_2 \times K_3$ which is given by

$$\begin{aligned} Q(k_1,k_2,k_3) &= \sum_{i=0}^{N}\sum_{p_\alpha=0}^{K_\alpha-1} q_i M_n(u_i^1 - k_1 - p_1K_1) \\ &\times M_n(u_i^2 - k_2 - p_2K_2) \\ &\times M_n(u_i^3 - k_3 - p_3K_3) \end{aligned} \tag{8}$$

where $\vec{u}_i$ is the scaled fractional coordinate of particle i within the bounds $0 \le u_i^\alpha \le K_\alpha$. M_n is the cardinal B-spline of order n. The summations are over all integers p_α in the range $0 \le p_\alpha < K_\alpha$. Cardinal B-splines have the following recursive form:

$$M_n(u) = \frac{u}{n-1}M_{n-1}(u) + \frac{n-u}{n-1}M_{n-1}(u-1) \tag{9}$$

for $n \ge 2$. For $n = 2$, $M_2(u) = 1 - |u - 1|$ for $0 \le u \le 2$ else $M_2(u) = 0$. B-splines have compact support, and the effect of applying Equation 8 is to 'spread' each charge q_i out over a cubic volume of n^3 elements of Q.

The approximate reciprocal space energy sum $\tilde{E}_{\mathrm{rec}}$ can be re-expressed as the convolution:

$$E_{\mathrm{rec}} \approx \tilde{E}_{\mathrm{rec}} = \frac{1}{2}\sum_{m_\alpha=0}^{K_\alpha-1} Q(m_1,m_2,m_3)\times(\theta_{\mathrm{rec}}{*}Q)(m_1,m_2,m_3), \tag{10}$$

where

$$\theta_{\mathrm{rec}} = F(B\times C) \tag{11}$$

and B and C are the arrays given by

$$B(m_1,m_2,m_3) = \prod_{i=1}^{3}|b_i(m_i)|^2 \tag{12}$$

$$C(m_1,m_2,m_3) = \frac{1}{\pi V}\frac{\exp(-\pi^2\vec{m}^2/\beta^2)}{\vec{m}^2} \tag{13}$$

From which the per particle forces may be derived by differentiation with respect to $\vec{r}_i$. As Q is continuously differentiable $n-2$ times with respect to the particle positions, and θ_{rec} is independent of the particle positions, the atomic force terms are given by

$$\frac{\delta\tilde{E}_{\mathrm{rec}}}{\delta r_i^\alpha} = \sum_{m_\alpha=0}^{K_\alpha-1}\frac{\delta Q}{\delta r_i^\alpha}(m_1,m_2,m_3)\times(\theta_{\mathrm{rec}}{*}Q)(m_1,m_2,m_3) \tag{14}$$

B. Nvidia GPU Architecture. Contemporary GPUs produced by Nvidia Corp. are general purpose, single program

multiple data (SPMD) processing units with a high degree of intrinsic parallelism and arithmetic performance, for example, over 30K execution threads may run concurrently on the Tesla C1060 device, with a peak arithmetic rate of ≤933 single precision GFLOPS. These devices are composed of a set of independent *multiprocessors*, each of which contains eight scalar cores. A scalar program fragment, known as a *kernel* may be run concurrently in a *block* of *threads* in parallel *warps* of 32 threads. Each block is restricted to executing on a single multiprocessor. Multiple independent blocks may be executed concurrently across the device in a *grid*.

Threads within a common block may intercommunicate via a small 16 kB region of in-core shared memory. Each multiprocessor has a large register file of which each thread receives a private, static allocation. The total number of threads, which may execute on a multiprocessor, and thus, the degree of parallelism, is dependent on the register resources required by each individual thread and also on the shared memory required by the block. Current devices support up to 1 024 threads and up to 8 blocks per multiprocessor.

Multiprocessors are grouped together on a single package and share a high-bandwidth link to external DRAM, known as *global memory*. Access to this memory is uncached and so is particularly costly, although the interleaved execution of warps acts to mitigate the cost of this latency.

The current programming model for these GPUs is CUDA, a C-like language with type qualifier extensions for indicating data locality and a special function call syntax for specifying the parallelism of a kernel invocation. An associated API provides functions for managing GPU host memory allocation and transfer. CUDA code is compiled to an intermediate byte code which is interpreted at runtime, providing forward compatibility with future device architectures.

For further details on device capabilities and programming model, the reader is referred to the CUDA SDK documentation.[9]

C. SPME Implementation. The implementation of SPME has two distinct components: the real space evaluation of the pairwise E_{dir} term and the reciprocal space evaluation $\tilde{E}_{rec}$. The evaluation of the former is trivially implemented into an existing nonbonded electrostatic force computation program with a change of potential function,[1] and we do not discuss this further. $\tilde{E}_{rec}$, however, involves several different computational steps, and its calculation frequently becomes the limiting step in parallel molecular dynamics simulations.[6,18]

In this section, we describe in detail the implementation of the computation of $\tilde{E}_{rec}$ using CUDA. Neither part of the computation is performed by the host CPU nor is it necessary to transfer any state between the GPU and host memory. In this way the performance of the code is entirely dependent upon the specification of the GPU.

The calculation of E_{rec} may be divided into five steps, which we summarize below before describing their implementation:

(1) Charge spreading: Spreading of charges on to the array Q (eq 8). The scaled fractional coordinates of each particle are calculated, and the Q array is populated with n^3 terms computed from the B-spline coefficients $M_n(u_i^\alpha - j)$, where $i = 1, ..., N$, $\alpha = 1, 2, 3$ and $j = 0, ..., n$.

(2) 3D fast Fourier transform (FFT): Real-to-complex 3D FFT in place transformation of Q into reciprocal space.

(3) Energy computation: Application of eq 10 in reciprocal space to compute terms for $\tilde{E}_{rec}$. Q is replaced by the product of itself with array B and C, as defined in eqs 12 and 13.

(4) 3D FFT: Computation of the convolution θ_{rec} (eq 11) through a complex-to-real 3D FFT of the array resulting from step 3.

(5) Force computation: real space computation of per atom force terms $\delta\tilde{E}_{rec}/\delta r_i^\alpha$ by multiplication with $\delta Q/\delta r_i^\alpha$ (eq 14).

Charge Spreading. In this step, each charge q_i is mapped to a site on the real space PME grid at a location determined by particle scaled fractional coordinates $\vec{r}_i$. The charge is then distributed over points in a surrounding volume according to eq 8. The volume of the spreading region is dependent on the order n of the cardinal B-spline. For $n = 4$, typically used in production simulations, each charge is spread over $n^3 = 64$ grid points.

This spreading is straightforward to do in a serial implementation but poses a significant challenge when performed with fine-grained parallelism. A naïve implementation, which used one thread to map the spread of each individual charge onto the grid, would encounter synchronization problems when different threads attempt to accumulate charge on the same grid location, thus, necessitating the use of thread safe atomic memory operations.[23] However, as only integer atomic operations are supported on current hardware, either the floating point atomic operations must be emulated using an atomic compare-and-set loop construct or the charge spreading must be altered to use fixed precision arithmetic. Furthermore, in order to achieve acceptable performance from the uncached GPU memory subsystem, it is necessary for the threads within a block to perform memory accesses to contiguous address ranges. The naïve approach has an essentially unordered memory access pattern, leading to poor performance.

Rather than performing a charge spreading, a per grid point gather is used instead. This is conducted in three steps which we term *placement*, *accumulation*, and *overflow*. First, each particle is mapped to a grid location, and its charge and position are recorded in a three-dimensional array. At the PME grid sizes and system densities typical of biomolecular simulation, this array is 90% sparse. Nevertheless, because the placement is performed in parallel, one particle per thread, the setting of elements within this array must still be performed atomically using the global memory atomic operation primitives in order to prevent access conflicts. Each grid site is permitted to hold a single charge; additional charges are placed in a simple list referred to as the overflow list.

Second, a separate kernel is used to sum the charge at each grid point by finding the contributions from all charges within the surrounding n^3 points (Figure 1). This kernel is executed in a block of K_1 threads which operates on a single full grid width stencil along the m_x vector. Each thread i accumulates the charge for the ith grid point in the row. Each

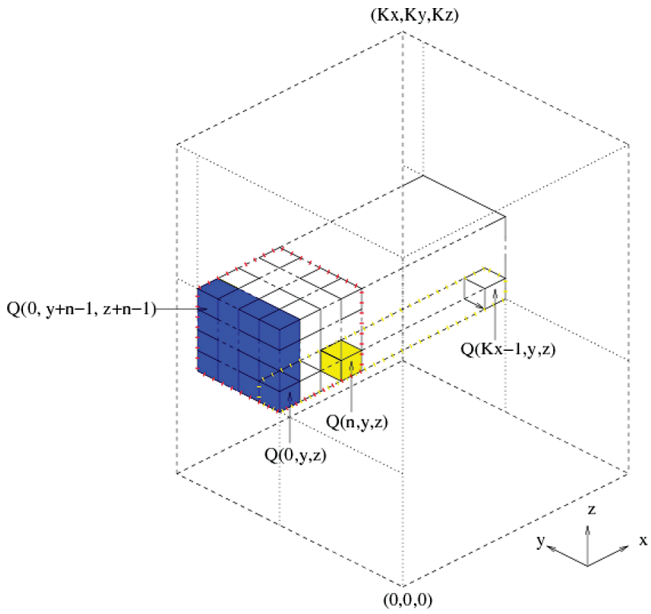


Figure 1. Charge-spreading phase in the real space charge grid. For efficient implementation, the charges are gathered to each grid point (shown as cells for clarity) for a spline order of n, each grid point receives charge terms from n^3 neighboring points (red dashed region indicates the cubic charge import region for the yellow (light-gray) point $Q(n, y, z)$. The charge gathering computational kernel operates in (K_y, K_z) blocks of K_x threads, each block calculating charge terms for a whole x row of the grid. Each thread loops over the n^2 neighbor cells in the y, z plane (blue/dark-gray region), calculating in turn the $3n$ B-spline terms M_n, corresponding to any charge located in the cell and accumulating the charge contribution for its cell. The spline terms are held in shared memory, allowing each thread to reuse the values computed by its neighbors for computing charge contributions from the adjacent $(n - 1)$ y, z planes.

thread loads in turn the n^2 surrounding points in the y, z plane, and, if containing a particle, the associated $3n$ spline terms $M_n(u_i^\alpha - j)$ (for $\alpha = 1, 2, 3$, $j = 0, ..., n$) are computed and placed into shared memory where they may be accessed by other threads.

This use of shared memory reduces the number of global memory accesses by a factor of n and has a memory access pattern favorable for the hardware. Although this method requires each set of spline terms to be computed n^2, the arithmetic cost of the calculation is sufficiently low to make their recomputation preferable to loading them from a precomputed table in global memory.

Lastly, the charges in the overflow list are spread onto the grid using both the naïve method described initially and the floating point atomic memory operations synthesized with an atomic integer compare-and-set loop construct. Although this processing is slow, it is much faster than a second iteration of the gathering kernel.

One consequence of the use of the per-grid-point gather is that the scaling characteristic of the charge griding phase changes from $O(N)$ to $O(K^3)$.

3D FFT. 3D FFTs are performed using the CUDA cuFFT library.[19] Although this library provides 3D FFT routines, the current version (version 2.2) shows poor performance.

Table 1. Performance of the PME Code in the Parallel MD Program ACEMD (Running on Three GTX280 GPUS) for a Range of Model Sizes Using Production Parameters of B-spline Order $n = 4$ and Timestep = 4 fs[a]

model	atoms	PME grid size	10^6 timesteps/day	ns/day
DHFR	26K	65	12	49
ApoA1	92K	108	2.7	11
STMV	1M	220	0.17	0.69

[a] STMV performance estimated from timings given in Figure 3.

Consequently, we implement the multidimensional FFT explicitly using three sets of orthogonal 1D transforms. The transforms are interleaved with in-place transpositions of the charge array Q, which is necessary to satisfy the contiguous memory layout of the input data required by the FFT library.

Energy Computation. The $b(m_i)$ and exponential terms for the Euler exponential splines (eq 7) are precomputed once and stored in the GPU's *constant* memory. This is a small (64 kB), read-only region of global memory that has cached read access for reduced access latency. Calculation of the product of the transformed Q with $B \times C$ (eq 11) is then performed in place using a kernel that computes one column of values along m_z per thread. When required for diagnostic output, energy terms can be accumulated in a separate buffer, and the sum transferred back to the host.

Force Computation. The computation of $\delta\tilde{E}_{rec}/\delta r_i^\alpha$ is performed using a kernel which operates on a single particle per thread. The $3n$ spline terms M_n are again computed, along with the first derivatives M'_n, once the per particle and n^3 force terms are computed as per eq 14. Because the particle distribution is essentially unordered, memory accesses are uncoalesced, but the kernel is sufficiently simple to permit a large number of threads per block, effectively hiding the memory access latency.

III. Discussion

A. Performance. The performance was also measured on production simulations of models of dihydrofolate reductase (DHFR) ($62 \times 62 \times 62$ Å^3, 23 558 atoms) and apoA1 ($108 \times 108 \times 78$ Å^3, 92 224 atoms), two protein systems commonly used for benchmarking exercises.[1] The DHFR and apoA1 models are typical of the scale of system routinely simulated using all-atom biomolecular mechanics with 1 million atom simulations of complete virions (e.g., the satellite tobacco mosaic virus, STMV)[20] at the far extreme, representing a range of simulation cell sizes from $\approx 60^3$−220^3 Å^3. We determine the practical performance in timesteps/day achievable with production simulation parameters and ACEMD, as shown in Table 1.

ACEMD is a parallel code able to distribute all other aspects of an MD simulation (bonded, nonbonded force terms, etc.) over multiple GPUs and, in practice, the PME computation represents the performance-limiting critical path in simulations over this range of sizes, making the optimization of the PME implementation essential for best performance.

To more closely assess the performance characteristics of the code, it was tested over a range of synthetic cubic input systems constructed with linear dimension in the range of

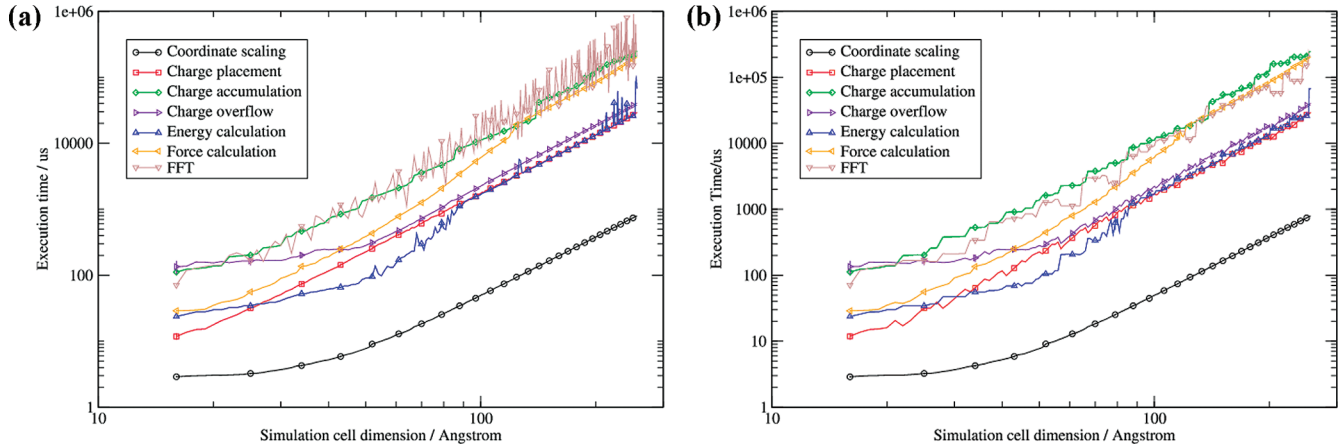


Figure 2. Execution times for the individual GPU kernels given as a function of the linear dimension of a cubic simulation cell for a constant system number density of 0.1Å^{-3}. FFT timing includes both forward and reverse 3D FFT steps. (a) Shows timings for the case of the PME grid having linear dimension $K = \text{int}(l_{\text{sim}})$. (b) Timings are given for optimized K, where $K \geq \text{int}(l_{\text{sim}})$, as determined by the method given in the main text. Error bars indicate σ^2.

$16 \leq l_{\text{sim}} \leq 256$ Å and with a random uniform distribution of particles with number density ρ_N of 0.1 atoms Å^{-3}, representative of the sizes and density of biomolecular systems of interest.

Execution times, T_{kernels}, for each individual kernel were recorded using the standard CUDA runtime profiling tool. The total execution wall time, T_{PME}, was measured from start to finish and so includes all additional host(CPU)-side overhead.

K_α were initially chosen to be equal to $\text{int}(l_{\text{sim}})$, as is commonly used in production MD simulations. For each synthetic system size, 100 iterations were performed, each with a different random particle distribution in order to characterize the sensitivity of the implementation to fluctuations in the distribution (Figure 2a).[24]

For large K, T_{PME} exhibits cubic scaling with respect to the PME grid dimension (dashed lines in Figure 3), corresponding to a broadly linear scaling with respect to particle number. However, for small grid sizes and particle counts, $K < 50$, there is insufficient available parallelism to fully occupy the GPU resources. This can be seen more clearly in the component kernel timings in Figure 2 (a and b), most prominently in the coordinate scaling, and in the energy calculation and overflow list kernels. In the case of the first two, the kernels are computationally simple, and a high thread occupancy may be achieved, e.g., the coordinate scaling kernel may run 15 360 threads (512 threads/multiprocessor) simultaneously, corresponding to a test system size of $K = 53$. For smaller systems, there is insufficient parallelism to fully occupy the GPU, and the execution time of the kernel is weakly dependent on system size; all blocks may be spread out across multiprocessors and executed completely in parallel. At $K = 53$, the GPU is fully subscribed, and a step in the execution time can be seen, indicating that some multiprocessors have now to process a second block. After this point, the execution time scales approximately linearly with particle count as subsequent step changes are much less significant owing to divergence between the work executing on the separate multiprocessors.

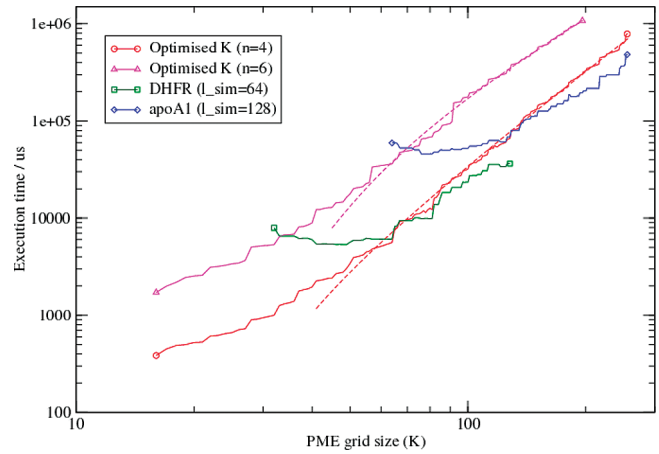


Figure 3. T_{PME} timing showing the effect of varying the PME grid size for a fixed system size K and the B-spline order n. Magenta (circles) and red (triangles) lines show timings for optimized PME grid dimensions with a spline order of 4 (red) and 6 (magenta) for the DHFR model. Dashed lines show cubic fits for $l_{\text{sim}} > 40$. For the green (squares) and blue (diamonds) lines, the simulated systems are of constant size (26K atoms DHFR and 92K apoA1 benchmarks, respectively), and the ordinal indicates the PME grid dimension K.

In the case of the overflow and charge placement kernels, a significant variation in runtime is observable; this arises from collisions between atomic operations of different threads. The synthesized floating point atomic operations used in the overflow kernel are much more sensitive to collisions as they use a looping construction that may make several accesses to global memory. Because a warp of threads is executed synchronously, it requires only one thread to suffer an unfavorable memory access for the execution time of the entire block to increase.

For the charge accumulation and energy calculation kernels, the number of threads per block is directly related to K_1. Initially, several blocks may run concurrently per multiprocessor, but as K_1 increases, the occupancy falls.

We use the Nvidia-supplied CUDA FFT library. As is characteristic of the FFT algorithm, the computational cost is highly sensitive to the prime factorization of the input

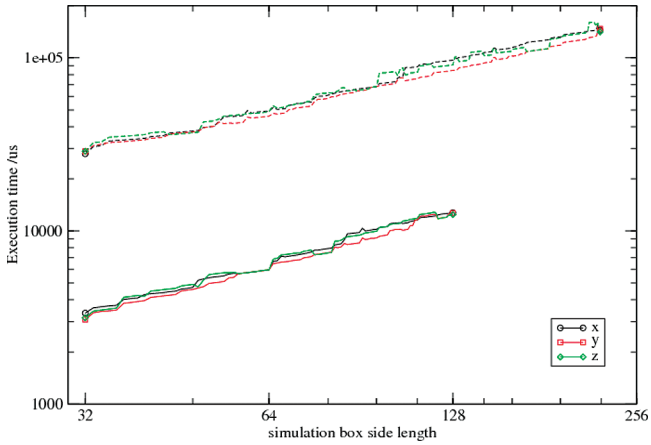


Figure 4. T_{PME} timing showing the minor sensitivity of the implementation to the orientation of the input for cuboid systems. The length of each simulation box dimension l_α is varied in turn within the range $l_{sim}/2 \leq l_\alpha \leq 2\ l_{sim}$ with the remaining sides held at a length of l_{sim}, where $l_{sim} = 64$ (solid lines) and 128 (dashed lines). A constant system density of $\rho_N = 0.1$ Å^3 is maintained.

dimension (Figure 2a). For this reason, some PME implementations, such as that of Desmond[21] are designed to use only specific grid sizes (2^n in that case). However, for the relatively small systems (O(100K) atoms) that ACEMD is intended to simulate, this would be quite restrictive. Therefore, for each l_{sim}, we select a modified K in the range $l_{sim} \leq K \leq 1.5\ l_{sim}$ such that T_{PME} is minimized. By selecting for the minimum T_{PME}, rather than just minimum FFT cost, we compensate for the increased runtime of components of the implementation which are sensitive to the ρ_N and K of the input. $T_{kernels}$ for these optimized dimensions are shown in Figure 2b, and T_{PME} is shown in Figure 3. The sawtooth pattern of the charge placement and overflow kernels are indicative of atomic operation collisions, the probability of them occurring dropping as K becomes larger than int(l_{sim}).

Because the accumulation and energy kernels operate preferentially along one dimension of the PME grid, we test the sensitivity of the implementation to the orientation of the input for cuboid systems. Figure 4 shows the execution time for synthetic input systems of constant ρ_N, varying each dimension l_α (for $\alpha = 1, 2, 3$) in turn in the range $0.5\ l_{sim} \leq l_\alpha \leq 2\ l_{sim}$ for $l_{sim} = 64\ 128$ and using the predetermined 'optimized' grid dimensions. While there are differences between the three variations, reflecting differing grid sizes and multiprocessor occupancies, we conclude that these are insufficiently large to require care over the orientation of the input system.

Finally, we consider the effect of the spline interpolation order on performance. When optimizing the SPME parameters for production MD simulations, it is not unusual to reduce grid size and increase the spline interpolation order to optimize performance or control accuracy. We tested the performance impact of these changes using two commonly used biomolecular benchmarks − DHFR (26K atoms) and apoA1 (92K atoms). In each case, the PME grid size was varied between $0.5\ l_{sim} < K_\alpha < 2\ l_{sim}$, and the timings were taken (Figure 3, green (squares) and blue (diamonds) lines). It can be seen that T_{PME} is minimal at approximately $K_\alpha = 0.6 l_{sim}$, increasing for smaller grids as the overflow list processing becomes significant. However, increasing the interpolation order to six significantly increases the execution time of the charge accumulation (Figure 3, magenta line) and no performance increase is yielded.

B. Accuracy. The GPU code exclusively uses single precision floating point arithmetic and, thus, operates at a reduced precision in comparison to a typical double precision CPU code. To quantify any reduction in accuracy, we compared the force terms produced by the GPU code (using ACEMD)[1] with a reference double precision CPU code (NAMD)[22] when performing the joint Amber−Charmm DHFR benchmark. The relative error of the reciprocal force terms, $|F_{NAMD}| - |F_{ACEMD}|/|F_{NAMD}|$, was found to be 10^{-5}, significantly below the 10^{-3} generally considered the acceptable maximum for relative error in the force terms for biomolecular simulation.[21] As such, we consider single precision arithmetic to be adequate for production simulation.

For long simulations in the NVE ensemble in which energy conservation is important, the code could be converted to use the double precision arithmetic capabilities of the latest GPUs. As this would incur a significant performance penalty (up to ≈ 8 times slower) and is not presently necessary for our work. This remains an area for future investigation. Furthermore, a double precision version of the cuFFT library is not yet available (as of CUDA 2.2).

Further data on the accuracy of benchmark MD simulations performed with PME in conjunction with ACEMD are given in Section IV of ref 1.

IV. Conclusions

We have implemented the smooth particle mesh Ewald method on Nvidia GPU hardware using the CUDA programming language. The implementation has been integrated into ACEMD,[1] permitting all-atom biomolecular simulations with long-range electrostatics to be fully accelerated on GPU devices with an accuracy comparable to a reference double precision CPU implementation. On contemporary hardware, a high level of performance is observed, commensurate with molecular dynamics simulation rates between 100 ns/day for systems of a few thousand atoms and 1 ns/day for one million atom systems, when used within ACEMD.

Acknowledgment. This work was partially funded by the HPC-EUROPA project (R113-CT-2003-506079). G.D.F. acknowledges support from the Ramon y Cajal scheme and the EU Virtual Physiological Human Network of Excellence. We gratefully acknowledge Nvidia Corporation (http://www.nvidia.com) for their hardware donations.

References

(1) Harvey, M. J.; De Fabritiis, G.; Giupponi, G. *J. Chem. Theor. Comp.* **2009**, *5* (*6*), pp 1632−1639.

(2) Ewald, P. *Ann. Phys.* **1921**, *369*, 253–287.

(3) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *19*, 8577–8593.

(4) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Phys.: Conf. Series* **2005**, *16*, 300–304.

(5) Oh, K. J.; Deng, Y. *Comp. Phys. Commun.* **2007**, *177*, 426–431.

(6) Fang, B.; Martyna, G.; Deng, Y. *Comp. Phys. Commun.* **2007**, *177*, 362–377.

(7) Shaw, D. E. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Proc. 34th Int. Symp. Comput. Architecture* **2007**, 1−12.

(8) De Fabritiis, G. *Comput. Phys. Commun.* **2007**, *176*, 600.

(9) Technical Report, *NVIDIA CUDA Compute Unified Device Architecture Programming Guide 2.0*; Nvidia Corp **2008**, http://developer.download.nvidia.com/compute/cuda/2_0/docs/NVIDIA_CUDA_Programming_Guide_2.0.pdf; (Accessed August 4, 2008).

(10) Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comp. Phys.* **2008**, *22*, 5342–5359.

(11) Meel, J. A. V.; Arnold, A.; Frenkel, D.; Portegies, S. F.; Belleman, R. G. *Mol. Simul.* **2008**, *34*, 259–266.

(12) Phillips, J. C.; Stone, J. E.; Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters *Proc. 2008 ACM/IEEE Conf. Supercomputing* **2008**; Article no. 8.

(13) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; LeGrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864–872.

(14) Liu, W.; Schmidt, B.; Voss, G.; Müller-Wittig, W. *Comput. Phys. Commun.* **2008**, *179*, 634–641.

(15) Giupponi, G.; Harvey, M. J.; de Fabritiis, G. *Drug Discov. Today* **2008**, *13*, 1052.

(16) ACEMD home page: http://multiscalelab.org/acemd. Accessed August 4, 2008.

(17) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(18) Toukmaji, A.; Paul, D.; Board, J., Jr. Technical Report: Distributed Particle Mesh Ewald: A Parallel Ewald Summation Method. *Proc. Int. Conf. Parallel Distributed Processing Techniques Applications*; Arabnia, H. R., Ed.; CSREA Press, **1996**; pp 33−43.

(19) Technical Report, *CUDA CUFFT Library, Document PG-00000−003 V2.0*; Nvidia Corp., **2008**, http://developer.download.nvidia.com/compute/cuda/2_0/docs/CUFFT_Library_2.0.pdf; (Accessed August 4, 2008).

(20) Freddolino, P. L.; Arkhipov, A. S.; McPherson, S. B.; Schulten, K. *Structure* **2006**, *14*, 437–449.

(21) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters; *Proc. Supercomputing 2006*; Tampa, FL, November 11−17, 2006.

(22) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(23) Atomic functions are memory operations that are designed to ensure that concurrent accesses to the same memory location by competing threads always leave the contents of that location in a consistent state by preventing race conditions. These functions require dedicated hardware support and have limited set of operations, as detailed in Appendix C of ref 9.

(24) The test hardware was a single Nvidia Tesla C1060 an HP xw660 workstation (dual Xeon 5430, 4GB RAM, Red Hat Enterprise Linux 5.3, CUDA 2.1, Nvidia driver 180.22).

CT900275Y

JCTC Journal of Chemical Theory and Computation

Theoretical Study of the X_2NO Systems (X = F, Cl, Br, I): Effects of Halogen Substitution on Structural and Spectroscopic Properties

Cristina Puzzarini[†] and Vincenzo Barone*[,‡]

Dipartimento di Chimica "G. Ciamician", Università di Bologna, Via F. Selmi 2, 40126 Bologna, Italy, and Scuola Normale Superiore di Pisa, Piazza dei Cavalieri 7, 56126 Pisa, Italy, IPCF-CNR, via Moruzzi 1, 56124 Pisa, Italy

Received April 11, 2009

Abstract: Structural and spectroscopic properties of the X_2 NO series of radicals, with X = F, Cl, Br, I, have been computed by the coupled cluster ansatz in conjunction with hierarchical series of basis sets, accounting for, in most cases, core correlation effects and extrapolation to the complete basis set limit. Namely, equilibrium structures, vibrational frequencies, and hyperfine coupling constants have been considered. Methods rooted into the density functional theory have been used to estimate anharmonic and, in conjunction with the polarizable continuum model, environmental effects. The remarkable agreement with the available experimental data, limited to the lighter member of the series, confirms the reliability of our computational approach and suggests that the data for heavier species represent reliable benchmarks for future experimental data and/or cheaper computational methods devised for larger systems.

I. Introduction

Because of their remarkable stability and strong localization of the unpaired electron on the NO moiety, nitroxides are among the most widely employed and carefully studied classes of organic free radicals (see for example refs 1 and 2). In particular, substituent effects on the geometry, electronic structure, and reactivity of nitroxides have been investigated in deep detail in a number of experimental and theoretical studies.[3−6] However, a systematic study of halogen substituents is still lacking in spite of the remarkable interest it could have from both fundamental and application points of view.

Substituent effects on the magnetic properties of nitroxide radicals can be either direct (i.e., related to polarization or electron transfer) or indirect (i.e., related to changes of hybridization issuing from pyramidalization of the nitrogen environment). Sometimes these two effects influence the value of the magnetic properties (especially isotropic hyperfine couplings) in opposite directions, making more difficult the interpretation of the experimental results. For instance, the presence of halogen substituents on the nitrogen atom increases the nitrogen isotropic hyperfine coupling constants because of the effect of the electron-acceptor power of halogen groups, but this is more than compensated by the pyramidalization of the NO moiety. Analogous effects can be observed on infrared spectra since delocalization of the singly occupied molecular orbital (SOMO) leads to an increase of the NO force constant because of its antibonding nature. It is thus clear that a systematic theoretical study employing reliable quantum mechanical (QM) approaches can provide invaluable information to complement and interpret spectroscopic results.

Concerning specifically halo-nitroxides, only F_2NO is experimentally well characterized,[7−9] which is much more pyramidal than its hydrogen (H_2NO) analogue. Also from a computational point of view, only F_2NO has been analyzed in detail and at high level of theory till now (see ref 10 and references therein). We report in the following a thorough computational study of the whole series of dihalogen substituted nitroxides, which, together with providing reference data for forthcoming experimental studies, has allowed

* To whom correspondence should be addressed. E-mail: vincenzo.barone@sns.it.

[†] Università di Bologna.

[‡] Scuola Normale Superiore di Pisa.

10.1021/ct9001762 CCC: $40.75
Published on Web 07/30/2009

also the unraveling of a number of interesting magneto-structural relationships.

II. Methodology

A. Equilibrium Structure. In the present investigation, we have thoroughly investigated the electronic ground state of the X_2NO systems, which is $^2A'$ for all the species considered, and where X is an halogen atom, fluorine, chlorine, bromine, or iodine.

The coupled cluster (CC) method with single and double excitations with a perturbative treatment of connected triples [CCSD(T)][11] has mainly been used in the present study. The variant denoted R/UCCSD(T)[12] has been employed, which is based on restricted open-shell Hartree−Fock (ROHF) orbitals but spin unrestricted in the solution of the CCSD equations. The coupled cluster model has been found to be adequate for studying these open-shell systems as test computations, carried out at the multiconfiguration self-consistent field (MCSCF)[13] level, showed that all the species considered are well described by a single reference wave function. Furthermore, nondynamical electron correlation seems not to be particularly relevant; in fact, the coupled cluster T_1 diagnostic[14] has been calculated to be in the range of 0.019−0.024 for all radicals. This is also confirmed by the D_1(ROCCSD) diagnostic,[15] which has been found lower than 0.04 in all cases. For the lighter member of the series, the full CC single, doubles and triples (CCSDT)[16] model, as well as the CCSDT(Q) approximation[17] with a perturbative treatment of quadruples on top of a CCSDT calculation, have also been considered to test the influence of higher-order excitations.

Correlation consistent-type basis sets have been used in the present investigation. More precisely, the standard cc-pVnZ (n = T, Q, 5) bases[18] have been employed for nitrogen, oxygen, and fluorine, the tight-d augmented valence cc-pV(n+d)Z basis sets[19] (n = T, Q, 5) for the chlorine atom, and the cc-pVnZ-PP sets[20] (n = T, Q, 5) for bromine and iodine. The latter ones are series of correlation consistent basis sets in conjunction with small-core relativistic pseudopotentials that leave 25 electrons to be handled explicitly for both Br and I. On the whole, these basis sets will be denoted as VnZ (n = T, Q, ...) in the text. The frozen core (fc) approximation has been adopted in conjunction with the above-mentioned series of bases.

For all the species considered, geometry optimizations have been performed using numerical gradients, as implemented in MOLPRO,[21] except those at the CCSDT and CCSDT(Q) levels, which have been carried out with the MRCC package.[22] For CCSD(T) geometry optimizations, the step-sizes used were 0.0005 Å for bond lengths and 0.1 degrees for bond angles. A convergence criterion stronger than the default one has been employed: both the maximum component of the gradient and the maximum component of the step have been constrained to be lower than 1.0×10^{-6} a.u.

To account for basis set truncation effects, since a hierarchical sequence of bases has been considered, the systematic trend of the optimized geometrical parameters can be exploited to estimate the complete basis set (CBS) limit. Making the assumption that the convergence behavior of the structural parameters mimics that of the energy, the consolidated $1/n^3$ extrapolation form[23] has been used to describe the convergence of the correlation contribution. To obtain extrapolated structures, the CBS limit value of the correlation contribution has then been added to the HF-SCF CBS limit, which is assumed to be reached at the HF-SCF/V6Z level

$$r(\mathrm{CBS}) = r_\infty^{\mathrm{SCF}} + \Delta r_\infty^{\mathrm{corr}} \quad (1)$$

where

$$\Delta r^{\mathrm{corr}}(n) = \Delta r_\infty^{\mathrm{corr}} + An^{-3} \quad (2)$$

Even if this procedure is only empirically based, a few papers available in the literature show its reliability.[24]

To take into account the effects of core-valence (CV) electron correlation, which are expected to be important, geometry optimizations have been carried out also including all electrons in the correlation treatment. The weighted core-valence correlation consistent cc-pwCVnZ (n = T, Q) basis sets[25] (cc-pwCVnZ-PP[26] for Br and I) have been used in this step. These basis sets will be denoted as wCVnZ (n = T, Q) in the following. Then, making use of the additivity approximation, the core correlation corrections have been added to the CBS limit of geometrical parameters

$$r(\mathrm{CBS} + \mathrm{CV}) = r(\mathrm{CBS}) + r(\mathrm{wCV}n\mathrm{Z,all}) - r(\mathrm{wCV}n\mathrm{Z,valence}) \quad (3)$$

where r(wCVnZ,all) and r(wCVnZ,valence) are the geometries optimized at the R/UCCSD(T)/wCVnZ level correlating all (except the 1s electrons of Cl) and only valence electrons, respectively.

B. Spectroscopic Properties. Since electron spin resonance (ESR) spectroscopy is one of the most important experimental techniques for characterizing radicals, the magnetic hyperfine coupling has been investigated. The hyperfine interaction contains an isotropic contribution, the so-called Fermi contact term, related to the spin density at the nucleus,[27] and an anisotropic contribution, denoted as dipolar hyperfine coupling, which can be derived from the classical expression of interacting dipoles.[28] In particular, the former determines the positions of ESR lines, whereas the latter tunes spectral shapes depending on different dynamical conditions. Therefore, we focus our attention on the former.

The essential quantities to be calculated are the spin densities at the nucleus of interest, and thus, the isotropic hyperfine coupling constants (hcc) have been evaluated as expectation values of the corresponding one-electron operator[28] at the CCSD(T) level of theory in conjunction with core-valence basis sets of triple- and quadruple-ζ quality, possibly augmented by diffuse functions.[29] More precisely, computations at the CCSD(T)/CVnZ (n = T, Q) and CCSD(T)/aCVnZ (n = T) levels (where aCVnZ denotes aug-cc-pCVnZ) have been performed at our best estimated equilibrium structures (CBS + CV). For I, only pseudopotential-based basis sets have been used, whereas for Br all-electron bases,[30] as well as pseudopotentials, have been

employed. In all calculations, all electrons have been correlated (except the 1s electrons of Cl and Br). Only for the CVTZ basis, the frozen core approximation has also been used in order to figure out the extent of CV effects. The CFOUR program package[31] has been employed for these computations, and the unrestricted Hartree−Fock (UHF) wave function has been used as reference in CCSD(T) calculations.

As clear from the type of bases used, core correlation effects have been directly taken into account. On the other hand, to estimate the effect caused by basis set truncation, the two-parameter CBS extrapolation formula proposed by Bartlett and co-workers in ref 32 has been employed

$$A_n^{(K)} = A_\infty^{(K)} + Be^{-(n-1)} \tag{4}$$

where $A_n^{(K)}$ and $A_\infty^{(K)}$ denote the hcc of the K-th nucleus,

$$A_n^{(K)} = \frac{8\pi}{3}\frac{g_e}{g_0}g_K\beta_K\sum_{\mu,\nu}P_{\mu,\nu}^{\alpha-\beta}\langle\phi_\mu|\delta(r_{nK})|\phi_\nu\rangle \tag{5}$$

computed with the CVnZ basis and extrapolated to the CBS limit, respectively.

According to the notation of ref 32, this extrapolation will be denoted as CBS2. This extrapolation function, which is an approximation of the exponential/Gaussian 3-parameter equation by Peterson, Woon, and Dunning,[33] has been chosen as well-tested for the property under consideration by Bartlett and co-workers (see ref 32).

Since vibrational and environmental effects may significantly affect the hcc's,[5] to provide reliable predictions for experimental values, they have also been estimated in the present investigation. The vibrational effects have been accounted for by adding to the CCSD(T) results the difference between equilibrium and zero-point averaged values computed at the B3LYP/EPRIII level. The vibrational averaging is based on a perturbative approach, and we refer interested readers to ref 34 for all computational details. It is noteworthy that the EPRIII basis set (12s8p2d1f/8s5p2d1f)[35] has been chosen because it was purposely developed and widely validated for calculations of hyperfine coupling constants.

Environmental (argon matrix) effects have been estimated by means of the polarizable continuum model (PCM)[36] and a dielectric constant of 1.43 at the B3LYP/EPRIII level, as well. This approach is expected to provide at least semiquantitative results as, while non electrostatic contributions play a significant role in determining thermodynamic characteristics in nonpolar solvents, local spectroscopic properties are sensitive only to electrostatic contributions.

The Gaussian03 program package[37] has been used for calculations of both vibrational and environmental effects.

Other useful information on radicals come from IR spectroscopy. The theoretical prediction of infrared harmonic and anharmonic vibrational frequencies requires the evaluation of harmonic and anharmonic force fields, respectively. These have been computed for the main isotopic species of the X_2NO radicals. More precisely, the harmonic part of the force field has been obtained at the CCSD(T)/VTZ as well as CCSD(T)/aVTZ (aVTZ denoting aug-cc-pVTZ[29]) levels of theory (in the fc approximation) by means of analytic second derivatives of the energy, as described in ref 38 and as implemented in the CFOUR program package.[31] To account for the anharmonic part, a mixed approach has been considered, i.e., anharmonic effects have been analyzed by a second-order perturbative treatment based on third and fourth energy derivatives computed at the B3LYP/EPRIII level,[34] as implemented in Gaussian03.[37] A number of studies have shown that addition of DFT anharmonic contributions to CCSD(T) harmonic frequencies allows a reliable reproduction of experimental fundamentals.[10,39−42]

III. Results and Discussion

A. Molecular Structure. The equilibrium geometries and energies of the X_2NO systems, with X = F, Cl, Br, and I, obtained at the CCSD(T) level using different basis sets are summarized in Table 1. The extrapolation to the CBS limit has been carried out as explained in the methodology section (eqs 1−2), and the results obtained are listed in Table 1, together with best estimated equilibrium structures provided by inclusion of the core correlation corrections (eq 3). It has to be noted that the results for F_2NO come from a previous study by the same authors.[10]

As evident from Table 1, the convergence to the CBS limit is practically reached at the CCSD(T)/V5Z level for bond distances involving only first-row atoms, whereas for bond lengths involving heavier atoms deviations as large as ∼0.01 Å between the CCSD(T)/VQZ and CCSD(T)/V5Z levels are observed. Unfortunately, computations in conjunction with the V6Z basis have been found to be too costly. For Cl_2NO, a non-monotonic trend has been observed for the NO distance and the angle; since we essentially perform a 2-point extrapolation and the SCF trend is monotonic, the CBS limit for them can be obtained. As far as CV corrections are concerned, it should be noted that in general they are relevant for bond lengths, whereas they are less important for angles. As pointed out in ref 10 and as is evident from the results of Table 1, these corrections are well estimated using both the wCVQZ and wCVTZ bases. For this reason, they have been computed at the CCSD(T)/wCVTZ level for the species with X = Cl, Br, and I, for which the wCVQZ basis is hardly applicable. Concerning scalar relativistic effects, they are expected to be relevant only for Br_2NO and I_2NO, and they are assumed to be completely recovered by the use of pseudopotentials.

On the basis of equilibrium structure evaluations for radicals performed following an approach analogous to that carried out in the present work (see, for instance, refs 10, 24, 41, and 42), an accuracy of about 0.001−0.005 Å is expected for the CBS + CV equilibrium distances derived in the present investigation, where the smaller uncertainty is referred to bond lengths involving first-row atoms and the larger to those for which the convergence to the CBS limit is slower.

From the results collected in Table 1, one may note that the NO distance decreases along the series, that is, when going from F_2NO to I_2NO, by about 0.03 Å. This is clearly related to the decrease of electronegativity from fluorine to

Table 1. Equilibrium Structure and Energy of X_2NO (X = F, Cl, Br, I)

RHF/UCCSD(T)[a]	X-N (Å)	N-O (Å)	∠XNO (deg.)	∠XNOX (deg.)	Energy (E_h)
F_2NO[b]					
aVTZ	1.4470	1.1693	117.14	121.24	−329.0874340
aVQZ	1.4379	1.1674	117.18	121.46	−329.1728545
VTZ	1.4349	1.1730	117.24	121.83	−329.0560601
VQZ	1.4345	1.1684	117.20	121.64	−329.1604408
V5Z	1.4345	1.1673	117.17	121.50	−329.1964277
CBS	1.4345	1.1667	117.15	121.40	
wCVTZ(fc)	1.4358	1.1711	117.20	121.69	−329.0788772
wCVTZ(all)	1.4330	1.1699	117.24	121.79	−329.2921954
wCVQZ(fc)	1.4339	1.1683	117.19	121.65	−329.1701498
wCVQZ(all)	1.4309	1.1669	117.21	121.70	−329.4087740
CBS+CV(CQ)[c]	1.4315	1.1653	117.17	121.45	
CBS+CV(CT)[d]	1.4317	1.1655	117.19	121.50	
Cl_2NO					
aVTZ	1.9967	1.1513	117.69	133.89	−1049.1366472
aVQZ	1.9764	1.1501	117.75	133.22	−1049.2086088
VTZ	1.9968	1.1514	117.62	134.11	−1049.1164283
VQZ	1.9766	1.1497	117.74	133.34	−1049.2020942
V5Z	1.9658	1.1506	117.72	132.94	−1049.2292431
CBS	1.9565	1.1514	117.71	132.60	
wCVTZ(fc)	1.9952	1.1494	117.61	134.16	−1049.1322261
wCVTZ(all)	1.9866	1.1495	117.65	133.94	−1049.7523524
CBS+CV	1.9479	1.1515	117.75	132.38	
Br_2NO					
aVTZ	2.2184	1.1432	118.11	144.07	−961.0783449
VTZ	2.2521	1.1392	117.87	149.98	−961.0479344
VQZ	2.2178	1.1384	118.06	144.05	−961.1351502
V5Z	2.2073	1.1380	118.03	143.30	−961.1609242
CBS	2.1963	1.1377	118.02	142.51	
wCVTZ(fc)	2.2463	1.1378	117.89	149.45	−961.0738369
wCVTZ(all)	2.2193	1.1375	117.81	148.07	−962.6724172
CBS+CV	2.1693	1.1374	117.94	141.16	
I_2NO					
aVTZ	2.4571	1.1463	119.33	154.84	−723.3562812
VTZ	2.5105	1.1418	118.85	180.02	−723.3348060
VQZ	2.4672	1.1404	119.11	158.43	−723.4136048
V5Z	2.4605	1.1394	119.08	160.51	−723.4413458
CBS	2.4545	1.1387	119.07	162.59	
wCVTZ(fc)	2.5004	1.1404	118.97	169.39	−723.3081629
wCVTZ(all)	2.4773	1.1389	118.87	169.10	−724.8908209
CBS+CV	2.4266	1.1362	118.94	162.30	

[a] According to the text, the standard cc-pV*n*Z basis sets for first-row elements, the cc-pV(*n*+d)Z bases for Cl, and the cc-pV*n*Z-PP pseudopotential-based sets for Br and I have been used. [b] Ref 10. [c] CV(CQ) means CV corrections at the CCSD(T)/wCVQZ level. [d] CV(CT) means CV corrections at the CCSD(T)/wCVTZ level.

iodine, and thus to corresponding decrease of σ-withdrawing ability. At the same time the nominally singly occupied orbital (SOMO) is more delocalized for larger and more polarizable substituents and this leads to a shortening of the NO bond in view of the antibonding character of this orbital (π^*) in the NO moiety. On the other hand, steric effects on the XNO angle are not so marked; in fact, it increases only by less than 1°. It is more interesting to observe how the dihedral XNOX angle varies from X = F to X = I. It is evident that this angle enlarges by about 10 degrees from one element of the series to the another, a little bit more when going from Br_2NO to I_2NO. In fact, while F_2NO is strongly pyramidal, I_2NO is close to being planar with a dihedral INOI angle of about 168°. Actually, the geometry optimization employing the VTZ basis fails in finding the minimum structure and converges to an approximately planar structure close to the transition state for the inversion motion. Transition states governing nitrogen inversion have been optimized for all X_2NO species employing the aVTZ basis set. The geometries, the corresponding equilibrium and ground-state (at the harmonic approximation) energy barriers are collected in Table 2. The energy barriers clearly reflect the decreasing along the series of the displacement of nitrogen out of the plane defined by the atoms directly bonded to it.

Table 2. Structure of the Planar Transition States and Barriers to Planarity

UCCSD(T)/aVTZ[b]	X−N (Å)	N−O (Å)	∠XNO (deg)	ΔE_e (kcal/mol)	ΔE_0[a] (kcal/mol)
X = F	1.3624	1.2119	126.52	11.30	12.70
X = Cl	1.7359	1.2290	123.07	11.67	12.61
X = Br	1.9100	1.2246	122.67	6.52	7.54
X = I	2.4794	1.1435	119.42	0.07	0.39

[a] ZPV corrections computed at the UCCSD(T)/aVTZ level within harmonic approximation. [b] According to the text, the aug-cc-pVTZ basis set for first-row elements, the aug-cc-pV(T+d)Z basis for Cl, and the aug-cc-pVTZ-PP pseudopotential-based set for Br and I have been used.

As mentioned in the Methodology section, the effect of high-order excitations on molecular structure has been investigated for the lighter member of the series, F_2NO. More

Table 3. Equilibrium Structure of F_2NO: Higher-Order Excitations and Other Contributions

	X−N (Å)	N−O (Å)	∠XNO (deg)	∠XNOX (deg)
V5Z	1.4345	1.1673	117.17	121.50
Δr(CBS)[a]	0.0	−0.0006	−0.02	−0.10
Δr(CV)[b]	−0.0030	−0.0014	+0.02	+0.05
Δr(*diff*)[c]	+0.0034	−0.0010	−0.02	−0.18
Δr(full-T)[d]	−0.0003	−0.0001	−0.02	−0.02
Δr((Q))[e]	+0.0033	+0.0012	−0.01	−0.03
Best estimate	1.4379	1.1654	117.12	121.22

[a] r(CBS) − r(V5Z). [b] r(wCVQZ, all) − r(wCVQZ, valence). [c] r(aVQZ) − r(VQZ). [d] According to eq 6. [e] According to eq 7.

Table 4. Equilibrium Structure of X_2NO (X = F, Cl, Br, I) at the B3LYP/EPRIII Level

	X−N (Å)	N−O (Å)	∠XNO (deg)	∠XNOX (deg)
X = F	1.4581	1.1616	117.01	128.99
X = Cl	2.0806	1.1468	117.95	141.99
X = Br	2.2655	1.1318	118.29	147.85
X = I	2.4914	1.1304	118.84	151.70

precisely, full triples corrections have been obtained with the VTZ basis as

$$\Delta r(\text{full} - \text{T}) \cong r(\text{CCSDT}) - r(\text{CCSD(T)}) \qquad (6)$$

whereas the VDZ set has been used for quadruples corrections

$$\Delta r(\text{Q}) \cong r(\text{CCSDT(Q)}) - r(\text{CCSDT}) \qquad (7)$$

From the results collected in Table 3, it is first of all evident that higher-order corrections are quite small, being on the order of 0.0001−0.001 Å for distances and 0.01−0.03° for angles. In particular, as expected, full triples corrections are almost negligible, while those caused by quadruples excitations are larger. From Table 3 it is furthermore clear that the effect of diffuse functions is non-negligible for F_2NO; this finding is essentially related to the strong electronegativity of fluorine. In Table 3, an equilibrium structure that account for all the contributions considered is provided: this should be considered as the best estimates obtainable at the moment. Since higher-order excitations are found to be quite small, we can claim that the CBS + CV structures given in Table 1 might be considered as the best estimated structures for X_2NO, with X = Cl, Br, and I.

As far as the comparison with literature values is concerned, to the best of our knowledge, this is restricted to F_2NO; therefore, we refer interested readers to ref 10. We briefly recall that the investigation by the present authors is the only systematic study at high level of theory reported in the literature. With respect to experiment, as far as we know, there are no data for comparison.

Finally, it has to be noted that, although the B3LYP/EPRIII geometries (Table 4) are not quantitatively accurate, they are sufficiently good to estimate anharmonic corrections to vibrational frequencies as well as vibrational effects on hyperfine couplings.

B. Spectroscopy. Isotropic hyperfine coupling constants of all X_2NO species, as obtained at the CCSD(T) level of theory by different basis sets, are summarized in Tables 5 and 6. In particular, in Table 5, we focus our attention on nitrogen and oxygen, while in Table 6, we report the results for halogens. In this way, from Table 5, we can point out how hcc's vary for N and O, whereas from Table 6, we can address the evolution of such constants along the halogen series. Concerning the latter, hcc's are given only for F, Cl, and Br, that is, only for those halogens for which all-electron basis sets can be used. For bromine, it has to be noted that results are affected by the missing account for relativistic effects which are expected to be important for a nucleus as heavy as Br and that the effect of correlating inner core electrons of Br actually does not justify the additional computational effort (with respect to keeping 1s2s2p electrons of Br frozen). Test computations on other Br-containing radicals showed that this effect is at the most 5%. It should be recalled that some results for F_2NO were previously reported and discussed in ref 10.

Both Tables 5 and 6 allow us to investigate the basis-set effects on isotropic hyperfine coupling constants. First of all, in all cases the absolute values increase by enlarging the basis set. Concerning the convergence to the CBS limit, it is evident that the values are nearly converged at the CCSD(T)/CVQZ level; in fact, the differences between this level and the CBS2 limit are generally of the order of 1−3%. We only note a larger discrepancy for chlorine (about 7%), but this finding was actually expected as the convergence for energy and properties is known to be slower for second-row elements and for heavy atoms in general. From the comparison between frozen core and all electron calculations, the effect resulting from core correlation can be pointed out. We note that a general conclusion cannot be drawn as CV corrections are small for N, that is, lower than 1%, while they are relevant for O (∼10%), F (<3%) and Cl (>10%). Furthermore, in all cases but Cl, CV corrections enlarge the absolute value of the hcc. With respect to the effect of diffuse functions (from the comparison between aCVTZ and CVTZ results), it can be noticed that this is not negligible, being in most cases on the order of 2−4%. Furthermore, it is surely worth noting the changes along the X_2NO series. First, it is evident from Table 5 that the hcc of N largely decreases when going from F_2NO to I_2NO; in fact, for instance at the CCSD(T)/aCVTZ level hcc varies from 89.1 G for F_2NO to 78.9 G for Cl_2NO, 70.0 G for Br_2NO, and 59.3 G for I_2NO. A similar trend is observed for oxygen; in fact, the hcc is negative for F_2NO, less negative for Cl_2NO, and positive for Br_2NO, and even more positive for I_2NO. For both constants these changes are related to the structural modifications observed along the series of radicals investigated (direct effect), as well as to the increased polarization of halogen atoms involved (indirect effect).

To further investigate structural effects on hcc's, they have also been computed at the planar structures given in Table 2, and collected in Table 7. To gain proper hints, in Table 7 they are compared to hcc's obtained at the same level of theory (CCSD(T)/CVTZ, all electrons correlated) but calculated at the CCSD(T)/aVTZ minimum structures. Let us now concentrate our attention on nitrogen hyperfine couplings, which represent one of the most widely used experimental probes for stereoelectronic and environmental effects.[5] The nitrogen isotropic hyperfine couplings computed for planar structures allow to compare different radicals in

Table 5. Isotropic Hyperfine Coupling Constants (Gauss) of X_2NO: Nitrogen and Oxygen Atoms

		B3LYP/EPRIII	CCSD(T)/aCVTZ	CCSD(T)/(fc)CVTZ	CCSD(T)/CVTZ	CCSD(T)/CVQZ	CCSD(T)/CBS2	exptl[a]
F_2NO	N							
	vacuum	95.40	89.13[b]	88.62	88.93	90.69[b]	91.72	
	Δ_{vib}[c]	0.35	0.35	0.35	0.35	0.35	0.35	
	Δ_{matrix}[d]	0.79	0.79	0.79	0.79	0.79	0.79	
	total	96.54	90.27	89.76	90.07	91.83	92.86	93 93.635(3)
	O							
	vacuum	−10.77	−14.03[b]	−12.87	−13.98	−14.60	−14.95[b]	
	Δ_{vib}[c]	0.44	0.44	0.44	0.44	0.44	0.44	
	Δ_{matrix}[d]	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	
	total	−10.34	−13.60	−12.44	−13.55	−14.17	−14.52	
Cl_2NO	N							
	vacuum	72.41	78.90	79.52	79.72	81.01	81.76	
	Δ_{vib}[c]	−2.64	−2.64	−2.64	−2.64	−2.64	−2.64	
	Δ_{matrix}[d]	0.68	0.68	0.68	0.68	0.68	0.68	
	total	74.37	76.94	77.56	77.76	79.05	79.80	
	O							
	vacuum	−0.67	−5.48	−4.80	−5.35	−5.69	−5.89	
	Δ_{vib}[c]	0.42	0.42	0.42	0.42	0.42	0.42	
	Δ_{matrix}[d]	0.06	0.06	0.06	0.06	0.06	0.06	
	total	−0.19	−5.00	−4.32	−4.87	−5.21	−5.41	
Br_2NO	N							
	vacuum	60.67	69.99	71.23	71.55	71.69[e]	71.77	
	Δ_{vib}[c]	−1.53	−1.53	−1.53	−1.53	−1.53	−1.53	
	Δ_{matrix}[d]	0.38	0.38	0.38	0.38	0.38	0.38	
	total	59.52	68.84	70.08	70.40	70.54	70.62	
	O							
	vacuum	0.56	1.33	1.38	1.41	1.50[e]	1.56	
	Δ_{vib}[c]	0.46	0.46	0.46	0.46	0.46	0.46	
	Δ_{matrix}[d]	0.04	0.04	0.04	0.04	0.04	0.04	
	total	1.06	1.83	1.88	1.91	2.00	2.06	
I_2NO	N							
	vacuum	44.89	59.32	53.45	53.45	60.57	70.53	
	Δ_{vib}[c]	−0.50	−0.50	−0.50	−0.50	−0.50	−0.50	
	Δ_{matrix}[d]	0.15	0.15	0.15	0.15	0.15	0.15	
	total	44.54	58.97	53.10	53.10	60.22	70.18	
	O							
	vacuum	2.90	7.75	7.47	8.11	8.55	8.80	
	Δ_{vib}[b]	0.85	0.85	0.85	0.85	0.85	0.85	
	Δ_{matrix}[c]	0.05	0.05	0.05	0.05	0.05	0.05	
	total	3.80	8.65	8.37	9.01	9.45	9.70	

[a] Ar-matrix: ref 8, upper line. SF6-matrix: ref 7, lower line. [b] Ref 10. [c] Vibrational corrections computed at the B3LYP/EPRIII level of theory. See text. [d] Environmental corrections computed at the B3LYP/EPRIII level of theory. See text. [e] Pseudopotential basis (cc-pVQZ-PP) has been used for Br.

Table 6. Isotropic Hyperfine Coupling Constants (Gauss) of X_2NO at the CCSD(T) Level: Halogens

		aCVTZ	(fc)CVTZ	CVTZ	CVQZ	CBS2	exptl[a]
F_2NO	F						
	vacuum	139.27[b]	126.61	129.81	137.72[b]	142.33	
	Δ_{vib}[c]	−3.84	−3.84	−3.84	−3.84	−3.84	
	Δ_{matrix}[d]	2.04	2.04	2.04	2.04	2.04	
	total	137.47	124.81	128.01	135.92	140.53	143 143.235(5)
Cl_2NO	Cl						
	vacuum	13.04	14.10	11.99	13.86	14.95	
	Δ_{vib}[c]	−0.23	−0.23	−0.23	−0.23	−0.23	
	Δ_{matrix}[d]	0.31	0.31	0.31	0.31	0.31	
	total	13.12	14.18	12.07	13.94	15.03	
Br_2NO	Br						
	vacuum	36.71	32.44	34.88	—	—	
	Δ_{vib}[c]	−0.55	−0.55	−0.55	—	—	
	Δ_{matrix}[d]	0.96	0.96	0.96	—	—	
	total	37.12	32.85	35.29	—	—	

[a] Ar-matrix: ref 8, upper line. SF6-matrix: ref 7, lower line. [b] Ref 10. [c] Vibrational corrections computed at the B3LYP/EPRIII level of theory. See text. [d] Environmental corrections computed at the B3LYP/EPRIII level of theory. See text.

an unbiased way since only spin polarization contributions remain operative, which are roughly proportional to the π spin density on the nitrogen atom.[5] Then, the nitrogen hyperfine coupling parallels the contribution of nitrogen to the SOMO, which increases in going from fluorine to bromine and iodine. The anomalous behavior of chlorine is

Table 7. Isotropic Hyperfine Coupling Constants (Gauss) of X_2NO: Comparison of Planar and Equilibrium Structure Results

		CCSD(T)/VTZ	
		TS[a]	equilibrium[a]
F_2NO	N	20.24	89.93
	O	−19.74	−13.59
Cl_2NO	N	15.46	79.05
	O	−18.48	−2.83
Br_2NO	N	39.13	67.20
	O	2.66	0.75
I_2NO	N	54.73	55.18
	O	12.23	9.48

[a] Optimized at the CCSD(T)/aVTZ level, where according to the text, aVTZ is the standard aug-cc-pVTZ basis set for first-row elements, the aug-cc-pV(T+d)Z basis for Cl, and the aug-cc-pVTZ-PP pseudopotential-based set for Br and I.

related to the "anomalous" behavior of the NO distance for planar structures, that is, it increases going from X = F to X = Cl and then decreases from X = Cl to X = I. Going to pyramidal structures, the direct effect issuing from involvement of nitrogen *s* orbitals in the SOMO become operative: while this implies, of course, an increase of the hyperfine coupling constant, a direct comparison between different substituents is impaired by their different pyramidalization.

A rigorous comparison with or prediction of experiment needs to account for vibrational and environmental corrections. The former have been obtained as previously explained and have been found to be small but in general not entirely negligible, being of the order of 1−3%. Concerning the environmental effects, they are very small and thus mostly negligible for O, whereas they are larger, that is, on the order of 0.5−1%, for N. Furthermore, the latter show a monotonic trend along the series: they are always positive and decrease from F_2NO to I_2NO. Because the B3LYP/EPRIII level of theory has been employed for evaluating vibrational as well as environmental corrections, the corresponding results (given in Table 5), obtained at the corresponding optimized geometries, deserve to be mentioned. It is interesting to note that on average they are fairly good, that is, this level of theory is able to qualitatively well reproduce the hcc's.

As outlined above, some results for F_2NO were already published in ref 10. This point deserves to be briefly commented on. First, it has to be noted that some of the vibrational corrections reported in ref 10 were given with the wrong sign. In the second place, the matrix corrections of ref 10 differ from those here collected because in the present investigation the gas-phase equilibrium structure has been employed instead of that reoptimized in condensed phase. Finally, in ref 10, a little bit more accurate results were obtained from the extrapolation to the CBS limit of the aCVTZ and aCVQZ values. These are not here reported because we chose to deal only with those basis sets affordable for the other radicals considered.

Since rather accurate experimental data are available for F_2NO, we can take the opportunity of their comparison to our theoretical results for estimating the accuracy reachable by CCSD(T) computations. On the basis of this comparison and of the extent of the effects previously discussed (especially convergence to CBS and core correlation), we can claim that for first row elements our best estimated hcc's are accurate to a few percent, that is, 1−2%. For heavier atoms, we expect a lower accuracy because of the slower convergence to the CBS limit, as well as to the neglect of relativistic effects (mostly for bromine). On the whole, we may conclude that our results can provide useful predictions for future experimental investigations.

Moving to IR spectroscopy, computed frequencies at both the harmonic and anharmonic levels are reported in Table 8 for all X_2NO species. The comparison with previous theoretical calculations and the available experimental results is also reported.

On the basis of the results and discussions reported in refs 10, 41, and 42, we can point out that the typical accuracy of 15−20 cm^{-1} (refs 43 and 44) for vibrational frequencies of closed-shell molecules calculated at the CCSD(T) level in conjunction with triple-ζ quality basis sets also applies to open-shell species. Therefore, such a level of theory is able to either check reliability of experiment or provide reliable predictions for experimental determinations. In ref 10, we already showed that such a level of theory was able to cast doubts on the assignment of the experimental vibrational spectra of F_2NO recorded in argon matrix and allowed us to suggest some revisions.

For improving our predictive capabilities, we went beyond the harmonic approximation as explained in the computational section, that is, anharmonic frequencies have been estimated by adding the anharmonic corrections at the B3LYP/EPRIII level to the CCSD(T) harmonic frequencies; the resulting values are reported in Table 8. This mixed approach has been validated quite convincingly for closed-shell systems,[39] and very recently, it has been successfully used for a number of free radicals.[10,41,42] On the basis of the results presented in refs 10, 41, and 42, our anharmonic vibrational frequencies are expected to have un accuracy better than 15 cm^{-1}.

The results collected in Table 8 allow us to discuss two interesting points: first, how the NO stretching (i.e., ν_1) changes along the series considered; second, how the frequency of the rocking mode (i.e., ν_4) involving halogens varies from F_2NO to I_2NO and how this change is related to the extent of pyramidality and to the barrier to planarity. Concerning the former effect, we note that the frequency parallels the trend of the N−O distance from F_2NO to I_2NO. In fact, as we note that the N−O bond length decreases from X = F to X = Br (by about 0.03 Å) and slightly increases from X = Br to X = I (by about 0.001 Å), the frequency increases from F_2NO to Br_2NO (by ∼200 cm^{-1}) and only slightly decreases from Br_2NO to I_2NO (by about 6 cm^{-1}). Therefore, we note that as the N−O distance is very similar in Br_2NO and I_2NO, in the same way ν_1 is very similar in the two species. With respect to the rocking mode, which correlates with the out-of-plane bending mode of planar structure, it is evident that reduction of the pyramidality is paralleled by a lowering of the frequency value up to a nearly vanishing value from 413 cm^{-1} for F_2NO to 117 cm^{-1} for Cl_2NO,

Table 8. Harmonic and Anharmonic Vibrational Frequencies (cm^{-1}) of X_2NO

	ν_1 (a′) NO str	ν_2 (a′) sym NX_2 str	ν_3 (a′) sym bend	ν_4 (a′) rock	ν_5 (a″) asymm NX_2 str	ν_6 (a″) NX_2 sciss
F_2NO[a]						
harmonic						
B3LYP/EPRIII	1632.2	738.2	582.0	383.3	812.9	429.4
CCSD(T)/VTZ	1592.3	769.7	614.0	429.3	873.9	470.6
CCSD(T)/aVTZ	1594.8	751.4	595.1	413.1	846.1	458.1
anharmonic						
B3LYP/EPRIII[b]	1612.4 (−30)	717.0 (−21)	557.1 (−25)	360.0 (−23)	807.9 (−5)	412.2 (−17)
VTZ+ anharm contr (DFT)	1562	749	589	406	869	454
aVTZ+anharm contr (DFT)	1565	730	570	390	841	441
experiment	1572.7	761	552.7		813	
Cl_2NO						
harmonic						
B3LYP/EPRIII	1868.9	386.9	259.7	92.5	596.5	123.8
CCSD(T)/VTZ	1738.5	477.2	290.3	91.7	635.6	136.0
CCSD(T)/aVTZ	1675.4	492.2	301.3	117.0	648.8	195.5
anharmonic						
B3LYP/EPRIII[b]	1845.2 (−24)	325.1 (−62)	226.2 (−34)	71.9 (−21)	588.3 (−8)	69.5 (−54)
VTZ+ anharm contr (DFT)	1715	415	256	71	628	82
aVTZ+anharm contr (DFT)	1651	430	267	96	641	142
Br_2NO						
harmonic						
B3LYP/EPRIII	1882.6	292.3	214.9	58.1	557.9	117.2
CCSD(T)/VTZ	1821.6	282.0	235.6	37.7	551.5	120.4
CCSD(T)/aVTZ	1779.8	313.1	213.2	49.4	566.0	107.6
anharmonic						
B3LYP/EPRIII[b]	1860.8 (−22)	216.2 (−76)	179.9 (−35)	40.4 (−18)	540.4 (−18)	59.7 (−57)
VTZ+ anharm contr (DFT)	1800	206	201	20	534	63
aVTZ+anharm contr (DFT)	1758	237	178	31	548	51
I_2NO						
harmonic						
B3LYP/EPRIII	1838.4	228.7	180.8	44.3	514.0	117.6
CCSD(T)/aVTZ	1779.1	204.8	161.3	31.3	519.4	80.4
anharmonic						
B3LYP/EPRIII[b]	1811.5 (−27)	158.5 (−70)	163.5 (−17)	34.4 (−10)	496.3 (−18)	91.0 (−27)
aVTZ+anharm contr (DFT)	1752	135	144	21	511	53

[a] Ref 10. [b] Anharmonic contributions reported in parentheses.

49 cm^{-1} for Br_2NO, and 31 cm^{-1} for I_2NO. As a matter of fact, the heavier species, I_2NO, is characterized by a nearly planar structure with very large amplitude out-of-plane bending.

IV. Conclusion

The present paper analyzes the structure, vibrational spectrum, and hyperfine couplings for the X_2NO series of free radicals, with X being an halogen atom. In most cases, CCSD(T) calculations have been carried out in conjunction with hierarchical series of bases and, when possible, accounting for extrapolation to the CBS limit and core correlation as well as relativistic effects. The estimated accuracy of our results is such that we are confident they may provide suitable benchmarks for the more approximate methods to be used for larger systems as well as reliable predictions for experiments, and allow the definition of more accurate magneto-structural relationships.

For spectroscopic properties an effective combination of coupled-cluster equilibrium values and harmonic frequencies, together with vibrational corrections and anharmonic contributions obtained by hybrid density-functional methods, is reported.

Substitution of fluorine by larger halogen atoms leads to a progressive decreasing of the pyramidal character and, in parallel, to a progressive decreasing of the semirigidity of the radical. This is clearly reflected in the properties investigated. For instance, in addition to the lowering of the rocking mode frequencies discussed a few paragraphs above, we may recall that the nitrogen isotropic hyperfine coupling decreases along the series because of the reduction of the contribution of its 2s orbitals to the nominally single occupied orbital up to its complete vanishing for a planar structure.

Acknowledgment. This work has been supported by the Italian Research Council, CNR (PROMO funds), as well as by University of Bologna (RFO funds). The authors also thank the CINECA supercomputing center for a grant of computer time on the IBM SP5 machine. The authors gratefully thank Prof. K. A. Peterson for useful discussions and reading the manuscript.

References

(1) Eaton, S. S.; Eaton, G. R. In *Biological Magnetic Resonance*; Berliner, L. J., Eaton, S. S., Eaton, G. R., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; Vol, 19, p 2.

(2) Tsvetkov, Yu. D. In *Biological Magnetic Resonance*; Berliner, L. J., Bender, C. J., Eds.; Kluwer Academic/Plenum Publishers: New York, 2004; Vol. 21, p 385.

(3) Khramtsov, V. V.; Weiner, L. M.; Grigoriev, I. A.; Volodarsky, J. B. *Chem. Phys. Lett.* **1982**, *91*, 69–72.

(4) *Substituent Effects in Radical Chemistry*; NATO ASI Series; Vieche, H. G., Janousek, Z., Merenyi, R., Eds.; D. Reidel Publisher: Dordrecht, The Netherlands, 1986; p 189.

(5) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231–1254.

(6) Carlotto, S.; Cimino, P.; Zerbetto, M.; Franco, L.; Corvaja, C.; Crisma, M.; Formaggio, F.; Toniolo, C.; Polimeno, A.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 11248–11258.

(7) Morton, J. R.; Preston, K. F. *J. Chem. Phys.* **1980**, *73*, 4914–4916.

(8) Misochko, E. Ya.; Akimov, A. V.; Goldschleger, I. V.; Wight, C. A. *J. Am. Chem. Soc.* **1998**, *120*, 11520–11521.

(9) Misochko, E. Ya.; Akimov, A. V.; Goldschleger, I. V.; Boldyrev, A. I.; Wight, C. A. *J. Am. Chem. Soc.* **1999**, *121*, 405–410.

(10) Puzzarini, C.; Barone, V. *J. Chem. Phys.* **2008**, *129*, 084306/1−7.

(11) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(12) Scuseria, G. E. *Chem. Phys. Lett.* **1991**, *176*, 27–35. Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733. Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1994**, *99*, 5219–5227.

(13) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053–5063. Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259–267.

(14) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem. Symp.* **1989**, *23*, 199–207. Lee, T. J.; Scuseria, G. E. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Ed.; Kluwer: Dordrecht, The Netherlands, 1995, p 47.

(15) Leininger, M. L.; Nielsen, I. M. B.; Crawford, T. D.; Janssen, C. L. *Chem. Phys. Lett.* **2000**, *328*, 431–436.

(16) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 7041–7050. Scuseria, G. E.; Schaefer, H. F., III. *Chem. Phys. Lett.* **1988**, *152*, 382–386. Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *93*, 6104–6105.

(17) Bomble, Y. J.; Stanton, J. F.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 054101/1−8.

(18) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(19) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

(20) Peterson, K. A. *J. Chem. Phys.* **2003**, *119*, 11099–11112. Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113–11123.

(21) MOLPRO (version 2006. 1) is a package of ab initio programs written by the following: Werner, H.-J., Knowles, P. J., with contribution from Almlöf, J., Amos, R. D., Berning, A., Deegan, M. J. O., Eckert, F., Elbert, S. T., Hampel, C., Lindh, R., Meyer, W., Nicklass, A., Peterson, K., Pitzer, R., Stone, A. J., Taylor, P. R., Mura, M. E., Pulay, P., Schütz, M., Stoll, H., Thorsteinsson, T., Cooper, D. L.

(22) MRCC, a generalized CC/CI program by M. Kállay, see http://www.mrcc.hu.

(23) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(24) Puzzarini, C. *J. Chem. Phys.* **2005**, *123*, 024313/1−14. Puzzarini, C. *Chem. Phys.* **2008**, *346*, 45–52.

(25) Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

(26) Peterson, K. A.; Shepler, B. C.; Figgen, D.; Stoll, H. *J. Phys. Chem. A* **2006**, *110*, 13877–13883.

(27) Fermi, E. *Z. Phys.* **1930**, *60*, 320–333.

(28) Frosch, A.; Foley, H. M. *Phys. Rev.* **1952**, *88*, 1337–1349.

(29) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(30) Peterson, K. A. unpublished results.

(31) CFOUR (development version). A quantum chemical program package written by Stanton, J. F.; Gauss, J.; Harding, M. E.; Szalay, P. G. with contributions from Auer, A. A.; Bartlett, R. J.; Benedikt, U.; Berger, C.; Bernholdt, D. E.; Christiansen, O.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Klein, K.; Lauderdale, W. J.; Matthews, D.; Metzroth, T.; O'Neill, D. P.; Price, D. R.; Prochnow, E.; Ruud, K.; Schiffmann, F.; Stopkowicz, S.; Tajti, A.; Varner, M. E.; Vázquez, J.; Wang, F.; Watts, J. D. and the integral packages MOLECULE (Almlöf, J.; Taylor, P. R.), PROPS (Taylor, P. R.), ABACUS (Helgaker, T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.), and ECP routines by Mitin, A. V.; van Wüllen, C. For the current public version, see http://www.cfour.de.

(32) Al Derzi, A. R.; Fau, S.; Bartlett, R. J. *J. Phys. Chem. A* **2003**, *107*, 6656–6667.

(33) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410–7415.

(34) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108/1−10.

(35) Barone V. In *Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p, 287.

(36) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43–54. Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

(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, S.; 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.; Gonzelez, C.; Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Wallingford, CT, 2004.

(38) Gauss, J.; Stanton, J. F. *Chem. Phys. Lett.* **1997**, *276*, 70–77.

(39) Carbonniere, P.; Lucca, T.; Pouchan, C.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384–388. Begue, D.; Carbonniere, P.; Pouchan, C. *J. Phys. Chem. A* **2005**, *109*, 4611–4616. Begue, D.; Benidar, A.; Pouchan, C. *Chem. Phys. Lett.* **2006**, *430*, 215–220.

(40) Barone, V.; Carbonniere, P.; Pouchan, C. *J. Chem. Phys.* **2005**, *122*, 224308/1−8.

(41) Puzzarini, C.; Barone, V. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6991–6997.

(42) Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2009**, *467*, 276–280.

(43) Begue, D.; Carboniere, P.; Barone, V.; Pouchan, C. *Chem. Phys. Lett.* **2005**, *416*, 206–211.

(44) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. *J. Phys. Chem. A* **2007**, *111*, 11242–11248.

CT9001762

JCTC Journal of Chemical Theory and Computation

Molecular Dynamics and Room Temperature Vibrational Properties of Deprotonated Phosphorylated Serine

A. Cimas,[†] P. Maitre,[‡] G. Ohanessian,[§] and M.-P. Gaigeot*,[†]

Laboratoire Analyse et Modélisation pour la Biologie et l'Environnement, UMR8587 CNRS, Université d'Evry val d'Essonne, boulevard F. Mitterrand, Bat. Maupertuis, 91025 Evry Cedex, France, Laboratoire de Chimie Physique, Université Paris Sud 11, UMR8000 CNRS, Faculté des sciences, bâtiment 350, 91405 Orsay Cedex, France, and Laboratoire des Mécanismes réactionnels, Département de Chimie, Ecole Polytechnique, CNRS, 91128 Palaiseau Cedex, France

Received April 13, 2009

Abstract: The local structure of phosphorylated residues in peptides and proteins may have a decisive role on their functional properties. Recent IRMPD experiments have started to provide spectroscopic signatures of such structural details; however, a proper modeling of these signatures beyond the harmonic approximation, taking into account temperature and entropic effects, is still lacking. In order to bridge this gap, DFT-based Car−Parrinello molecular dynamics simulations have been carried out for the first time on a phosphorylated amino acid, gaseous deprotonated phosphoserine. It is found that all vibrational signatures are successfully reproduced, and new deconvolution techniques enable the assignment of the vibrational spectrum directly from the dynamics results and the comparison of vibrational modes at several temperatures. The lowest energy structure is found to involve a strong hydrogen bond between the deprotonated phosphate and the acid with relatively small free energy barriers to proton transfer; however, we find that proton shuttling between the two sites does not occur frequently. Anharmonicities turn out to be important to reproduce the frequencies and shapes of several experimental bands. Comparison of room temperature and 13 K, effectively harmonic dynamics, allows insight to be obtained into vibrational anharmonicities. In particular, a significant blue-shift and broadening of the C=O stretching frequency from 13 to 300 K can be ascribed to intrinsic anharmonicity rather than to anharmonic coupling to other modes. On the other hand, significant couplings are found for the stretching motions of the hydrogen bonded P−O bond and of the free P−OH bond, mainly with modes within the phosphate group.

1. Introduction

Reversible protein phosphorylation of the side chain alcohol group of serine (S), threonine (T), or tyrosine (Y) residues, is a very common post-translational modification (PTM) of proteins. It has a strong impact on protein function, as it influences metabolic pathways, membrane transport, gene transcription, etc.[1,2] The impact of phosphorylation on protein function is at least in some cases related to the conformational changes it induces.[3] It is thus of great interest to unravel the mechanism by which a single phosphorylation may significantly alter the structure of a macromolecule. Studies on peptides have revealed that phosphorylation can promote high helical content.[4] It has been recognized in several cases that the presence of nearby arginine (R) residues is likely to lead to strong hydrogen bonding between the negatively charged phosphate and the positively charged guanidinium of R.[3] Small models can therefore be of high value to quantify such individual interactions.[5]

* Corresponding author phone: +33-1-69470140; e-mail: mgaigeot@univ-evry.fr.

[†] Université d'Evry val d'Essonne.

[‡] Université Paris Sud 11.

[§] Ecole Polytechnique.

10.1021/ct900179d CCC: $40.75
Published on Web 08/13/2009

Phosphate groups are expected to have rather recognizable vibrational signatures;[6] therefore, IR spectroscopy is in principle a valuable tool to identify phosphates and their environment. In particular, IR spectroscopy of organic molecules is known to be highly sensitive to hydrogen bonding, a feature that makes this technique very appealing for structural characterization of phosphorylation. In previous work, we have shown that the recently developed InfraRed Multiple Photon Dissociation (IRMPD) spectroscopic technique can be used to detect the occurrence of phosphorylation in amino acids and peptides in the gas phase.

IRMPD is carried out by irradiating gaseous ions trapped in high vacuum cells of mass spectrometers with IR photons.[7−10] Because the number density of gaseous ions is necessarily small, absorption spectroscopy cannot be carried out. The photons are used to generate ion fragmentation, which is efficiently monitored with the high sensitivity of mass spectrometers. The fragmentation ratio as a function of photon energy generates an IR action spectrum. This technique requires high laser power and wide tunability, which are made available in the 200−2000 cm^{-1} region by free electron lasers (FEL).[7−10] IRMPD has the potential for distinguishing between isomers, and also between conformers, although the latter can be much more difficult. This new spectroscopic tool has already found many applications, including the structural characterization of small biological molecules such as peptides.[11−23]

In a previous paper, we obtained the IRMPD signatures of the three protonated, phosphorylated amino acids: phospho-serine ($[pS+H]^+$), phospho-threonine ($[pT+H]^+$), and phospho-tyrosine ($[pY+H]^+$).[24] The results indicated that phosphate specific bands exist as expected and that they are easily detectable in IRMPD conditions, in the vibrational range accessible with a FEL. Detailed band assignment based on quantum chemical calculations established that some features were common to all three species, while others were specific. Furthermore, a second distinction was established, in which it is the hydrogen bonding capability of the environment that is distinctive. This conclusion was confirmed by a subsequent study of the phosphorylated dipeptide $[GpY+H]^+$, showing the potential of the method.[25] The proof of principle for extension to larger biomolecules was given on a 12-residue fragment of the protein stathmin.[24] It is clear however that the most likely form of phosphates at physiological pH is deprotonated, either singly or doubly. We have therefore initiated a research program in which we generate and interpret the IRMPD spectra of deprotonated phosphorylated amino acids[26] and peptides.

Since IRMPD spectroscopy is a relatively new technique, extensive molecular modeling is required to interpret the bands. This has been done with success in the past with harmonic calculations (see, e.g. refs 11−23 for applications to peptides), i.e. optimizing the geometry of several possible structures for each species and generating harmonic vibrational spectra at the optimized geometries. A match between experimental and calculated IR signatures is subsequently sought to identify populated isomers and/or conformers. However modeling of band breadth has not been yet tackled, and there are a number of cases for which discrepancies between experimental and computed frequencies remain unexplained.

Because IRMPD experiments are often carried out at room temperature, a theoretical approach closer to the experimental conditions and applied in the present work consists in simulating the dynamical behavior of the molecule through molecular dynamics (MD) simulations conducted at the average experimental temperature, for instance through DFT-based Car−Parrinello MD (CPMD),[27] and calculating the IR spectrum directly from the dynamics. MD is essential for including temperature, conformational dynamics, in particular the interconversion between different conformers[28] or isomers such as those connected by proton transfers,[12,14] with the advantage that entropic effects are directly taken into account.

IR spectra calculations through MD simulations are based on a dipole time correlation function.[29] Within the past few years, we have shown that DFT-based MD is the proper tool for the calculation of IR spectra of DNA and peptide building blocks, in the gas phase or immersed in liquid water,[12,14,28,30−32] at room temperature. We have in particular demonstrated the role of conformational dynamics at room temperature in the interpretation of finite temperature spectroscopy of peptides,[28] which is relevant for IRMPD.

The main advantage of IR spectra calculations through finite temperature MD over static calculations is that all anharmonic effects are naturally described. This is to be opposed to the two successive harmonic approximations usually adopted for the determination of IR spectra from static *ab initio* calculations, i.e. the harmonic approximation of the potential energy surface at the optimized geometries and the electrical harmonic approximation for the transition dipole moments. Both approximations are relaxed in *ab initio* molecular dynamics, simply because they are not needed. In fact, the finite temperature dynamics takes place on all accessible parts of the potential energy surface (be they harmonic or anharmonic), provided that time propagation is long enough. As the calculation of IR spectra with molecular dynamics is related only to the time-dependent dipole moment of the molecule, it does not require any harmonic expansion of the transition dipole moments. Therefore, if the dipole moments and their fluctuations are accurately calculated along the trajectory, the resulting IR spectrum should be reliable too. The quality of the potential energy surface is entirely contained in the force field, calculated at the DFT/BLYP level in the present work, as in our previous investigations.[12,14,28,30−32] The very good reproduction of the relative positions (and intensities, when they are directly comparable to experiment[31,32]) of the different active bands in our previous works indicates that this level of theory is satisfactory, at least on weakly interacting floppy peptide building blocks (gas phase or immersed in liquid water). The B3LYP hybrid functional has been recently implemented in the CPMD[33] and CP2K MD packages,[34] but the CPU cost is reported to be ∼40−100 times larger than that of a GGA functional MD. Considering our previous results together with the extra-CPU cost for a hybrid DFT-based MD, we chose to keep to local BLYP MD simulations in the present work. New functionals have recently been implemented to

Table 1. Relative Energy Values (kcal/mol) between the Optimized Configurations of $[pSer\text{-}H]^-$ [b]

	B3LYP/all electron						BLYP/all electron							
	6-31+G(d)		6-311++G(d,p)		aug-cc-pVTZ		6-31+G(d)		6-311++G(d,p)		aug-cc-pVTZ		CPMD/BLYP	
isomer	without ZPE	including ZPE	without ZPE	including ZPE	without ZPE	including ZPE	without ZPE	including ZPE	without ZPE	including ZPE	without ZPE	including ZPE	90 Ry	110 Ry
pSer-H_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00
pSer-H_2	1.46	0.64	1.75	0.94	1.19	0.59	1.30	0.47	1.55	0.75	1.03	0.47	0.84	0.92
pSer-H_3	3.89	3.71	5.69	5.38	-	-	2.54	2.37	4.18	3.91	4.59[a]	-	4.45	4.42
pSer-H_4	3.94	3.69	5.79	5.38	-	-	2.60	2.37	4.31	3.94	4.74[a]	-	4.46	4.44
pSer-H_5	11.89	10.80	-	-	-	-	11.45	10.47	-	-	-	-	10.75	10.71
pSer-H_6	11.89	10.91	-	-	-	-	10.92	10.02	-	-	-	-	10.93	10.92
pSer-H_7	11.89	11.05	-	-	-	-	11.42	10.67	-	-	-	-	11.02	10.99
pSer-H_8	13.76	13.26	-	-	-	-	13.57	13.03	-	-	-	-	13.73	13.71

[a] Single points over the optimized BLYP/6-31+G(d) geometries. [b] B3LYP and BLYP calculations refer to all electron geometry optimizations performed with the Gaussian03 package. CPMD refers to DFT/BLYP/pseudopotential single point calculations performed with a plane-wave basis set of 90 or 110 Ry with the CPMD package. See text for more details. ZPE stands for Zero Point Energy.

better describe exchange and correlation potentials,[35,36] and first results on vibrational spectroscopy are promising.[37,38] Rothlisberger's group has also contributed to including dispersion terms in the DFT formalism in the context of the Car−Parrinello methodology.[39,40] When dealing with charged species, as is the case here, this latter contribution is less important since the dominant interactions correspond to electrostatic forces, which are reasonably well estimated at the DFT level.

Deprotonated phosphorylated serine, noted hereafter $[pSer\text{-}H]^-$, is the first building block of a series of deprotonated phosphorylated peptides of increasing size and complexity that we characterize by combined vibrational experiments and MD simulations. DFT-based MD simulations are applied in the present theoretical investigation in order to assess the dynamics approach to the calculation of the IR spectrum of such a "model" system and its subsequent application to bigger phosphorylated peptides. Our main contribution is to quantify temperature and nonharmonic effects of $[pSer\text{-}H]^-$ in the mid-IR domain, in a direct way.

The remainder of the paper is organized as follows: part 2 is dedicated to the description of the methods used. Results are divided into three parts, with (1) the characterization of the potential energy surface of $[pSer\text{-}H]^-$ and the dynamics at room temperature, (2) the infrared spectroscopy at room temperature from Car−Parrinello dynamics, and (3) the characterization of anharmonic effects on the vibrational bands.

2. Computational Methods

2a. Static Calculations. The potential energy surface (PES) of $[pSer\text{-}H]^-$ was initially explored at the B3LYP/6-31+G(d) and BLYP/6-31+G(d) levels. The choice of this basis set follows our previous calibration work on protonated and deprotonated phosphorylated amino acids.[24−26] Harmonic frequencies have been calculated in order to check that optimized structures are minima on the PES. The structures of lowest energies, *pSer-H_1*, *pSer-H_2*, *pSer-H_3*, and *pSer-H_4* (see below) have been subsequently reoptimized at the BLYP/6-311++G(d,p) and B3LYP/6-311++G(d,p) levels, followed by harmonic frequency calculations. Moreover, the two most stable conformers (*pSer-H_1* and *pSer-H_2*) have been further optimized at the BLYP/aug-cc-pVTZ and B3LYP/aug-cc-pVTZ levels, and we also have performed single point calculations for *pSer-H_3* and *pSer-H_4* at the BLYP/aug-cc-pVTZ using the BLYP/6-31+G(d) optimized geometries. This calibration study was carried out to check that the energetic order of the low energy structures was unchanged. Energies are reported in Table 1. All calculations were performed using the Gaussian03 package.[41]

2b. DFT-Based Molecular Dynamics Simulations. The DFT-based Car−Parrinello molecular dynamics (CPMD) simulations performed in this work follow the general setup of our previous simulations.[12,14,28,31,32] All simulations were carried out with the CPMD package.[27] We used the Becke, Lee, Yang, and Parr (BLYP) gradient-corrected functional[42,43] for the exchange and correlation terms. The one electron orbitals are expanded in a plane-wave basis set with a kinetic energy cutoff of 90 Ry restricted to the Γ point of the Brillouin zone. Medium soft norm-conserving pseudopotentials of the Martins-Trouillier type[44] are used. The core−valence interaction in C, N, and O is treated by *s* and *p* potentials with pseudization radii of 1.23, 1.12, and 1.05 au, respectively (taking the same radius for *s* and *p*), while H atoms are treated as an *s* potential with a 0.5 au radius. The core−valence interaction in (neutral) P is treated by *s*, *p*, and *d* potentials with pseudization radii of 1.5 (*s*), 1.5 (*p*), and 1.09 (*d*) (in au). We have added *d*-states taken from the $1s^2 2s^2 2p^6 3s^1 3p^{1.75} 3d^{0.25}$ configuration of the ion. Energy expectation values are calculated in reciprocal space using the Kleinman-Bylander transformation.[45]

The value of 90 Ry for the energy cutoff of the plane wave expansion of the wave function has been chosen with the following scheme. We have checked that the difference of energy between the different structures identified in the all electron BLYP/6-31+G(d) geometry optimizations is correctly reproduced with single point energy calculations using the plane wave basis set and the pseudopotentials of the Car−Parrinello setup. Two energy cutoff values for the plane-wave expansion have been checked, 90 and 110 Ry. Both give the same energy order and nearly identical energy differences with respect to the most stable conformer, as reported in Table 1. As there is no difference between the two plane wave basis sets, we chose the lower energy cutoff of 90 Ry. The energy differences given by the plane-wave

Table 2. Average Ionic Temperatures (Kelvin) Obtained in the Microcanonical Car-Parrinello Molecular Dynamics Simulations Performed in This Work

CPMD	<T>	<T_C>	<T_H>	<T_O>	<T_N>	<T_P>
17.1 ps	13 ± 2	18 ± 4	12 ± 2	12 ± 1	13 ± 0	11 ± 2
17.9 ps	245 ± 30	220 ± 25	232 ± 41	267 ± 22	281 ± 36	237 ± 8
13.9 ps	298 ± 23	297 ± 1	309 ± 37	290 ± 20	292 ± 5	268 ± 30

calculation are very similar to those obtained at the all electron BLYP/aug-cc-pVTZ level. Note that the BLYP and B3LYP energy differences are nearly identical, apart for isomers *pSer-H_3* and *pSer-H_4* for which the energy difference with respect to the conformation of lowest energy (*pSer-H_1*) is underestimated by the BLYP calculations (∼1.5 kcal/mol). This has however no consequence in the present work, as it turns out that these salt bridge conformations are not explored during the dynamics.

Car−Parrinello dynamics (CPMD) were performed in the microcanonical ensemble (at constant volume and internal energy) using a fictitious electron mass of 400 au and a time step of 4 au (0.096 fs), see our previous papers for further details.[12,14,28,31,32] Gas-phase simulations were carried out with the decoupling technique of Martyna and Tuckerman[46] in order to eliminate the effect of the periodic images of the charge density. A cubic box length of 20 Å was selected after performing a series of wave function optimizations of *pSer-H_1* and *pSer-H_2* isomers in boxes of increasing length. We found that from 20 Å on, the electronic energy of the isolated molecule is converged within 10^{-5} au, which ensures that the wave function of the isolated anion is entirely contained in the box cell. CPMD simulations reported here consist of two steps: an equilibration phase of 1−2 ps partially performed with a control of temperature through velocity rescaling, followed by data collection over trajectories of ∼14−18 ps where molecular dynamics are strictly microcanonical. We have performed three dynamics, one at ∼20 K (17.1 ps) and two at ∼300 K (17.9 and 13.9 ps). Room temperature has been chosen for consistency with the average temperature of the IR-MPD experiments. Initial velocities were chosen in a Boltzmann distribution centered at the desired temperature. The average molecular temperature and average temperatures of each atomic type obtained for the simulations are shown in Table 2. Equipartition of energy over all degrees of freedom is globally respected in each simulation (keeping in mind that definition of temperature for such a small system is always questionable). It is especially difficult to achieve a proper equipartition of energy for low temperature dynamics within the short time-scale that can be afforded by CPMD. One can thus note that, on average, only the carbon atoms are slightly too warm during the low temperature dynamics; we will come back to this point when discussing infrared intensities. The final "room temperature" infrared spectrum presented here has been averaged over the two trajectories.

The calculation of the infrared absorption coefficient, $\alpha(\omega)$, makes use of the relation involving the Fourier transform of the time correlation of the total dipole moment of the molecule $\mathbf{M}(t)$, according to ref 29

$$\alpha(\omega)n(\omega) = \frac{2\pi\beta\omega^2}{3cV}\int_{-\infty}^{+\infty}\langle \mathrm{M}(t)\mathrm{M}(0)\rangle e^{i\omega t}dt \qquad (1)$$

where $\beta = 1/k_BT$, $n(\omega)$ is the refractive index, c is the speed of light in vacuum, and V is the volume of the simulation box. The angular brackets in formula (1) indicate a statistical average. Note that in this formula we have taken into account a quantum correction factor (multiplying the classical line shape) of the form $\beta\hbar\omega/(1\text{-}e^{-\beta\hbar\omega})$, which was shown to give the most accurate results for IR intensities.[31,32] For a complete discussion on quantum corrections, we refer the reader to refs 47 and 48. The IR spectrum is defined as the product $\alpha(\omega)n(\omega)$, with ω in cm^{-1}. $\mathbf{M}(t)$ is the dipole moment of the molecule at time t, which is the sum of the nuclear and electronic contributions. The dipole moment of the box cell is calculated with the Berry phase representation, as implemented in the Car−Parrinello framework and described in details previously (see for instance ref 31). The final spectra were smoothed with a window filtering applied in the time domain, which corresponds roughly to the convolution of the bare spectrum by a 10−20 cm^{-1} width Gaussian function. This convolution has the only purpose to remove the numerical noise arising from the finite length of the Fourier transform of eq 1. We have checked that the durations of the dynamics performed here are sufficient to obtain converged IR intensities (i.e., the latter are not modified upon increasing the dynamics duration).

Comparison of IR absorption intensities calculated within either the static or the dynamics formalisms to the ones obtained in IR-MPD experiments is certainly not well understood. Equation 1 for IR signal relies on linear response theory and is strictly valid for one-photon linear IR absorption spectroscopy. IR-MPD on the other hand is a multiphoton IR absorption process leading to the fragmentation of the molecule: the recorded signal is the fragmentation yield as a function of the IR excitation wavelength. It is thus an indirect measurement of IR absorption, in contrast to the usual linear IR spectroscopy. Calculations and experiments are therefore not directly comparable for band intensities, giving rise to possible discrepancies. The direct simulation of IR-MPD spectra, with a clear theoretical expression of signal intensity in terms of dynamical quantities, remains an open question.

In all of our previous applications of CPMD to IR spectroscopy (gas and liquid phase calculations in the 800−2000 cm^{-1} range)[12,14,28,31,32] we have systematically found that our calculated infrared spectra have to be blue-shifted by 100 cm^{-1} so that all of the calculated bands are aligned with their experimental counterparts. As a consequence, although our CPMD calculations do not give the proper absolute values of band positions, they do yield accurate band-gaps between the active bands. We stress again that a *global translation* is applied to the spectrum, not a scaling factor. This empirical finding is in contrast to static *ab initio* calculations where a scaling factor is used to correct the theoretical predictions with respect to the observed frequencies (in order to compensate for both the level of theory and anharmonicities). The origin of this is at the moment unclear to us. Effects of the fictitious mass, which leads to instantaneous Car−Parrinello forces being different

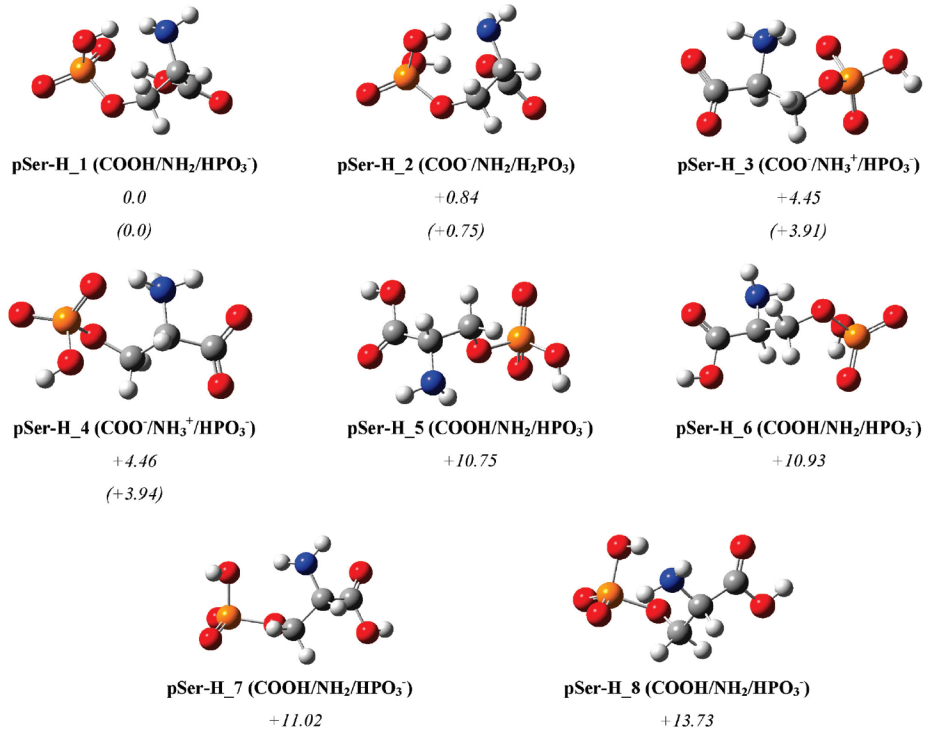


Figure 1. Schematic representation of the eight isomers of the deprotonated phosphorylated Serine $[pSer\text{-}H]^-$ identified in this work. Relative energies between the isomers have been reported using the CPMD setup (plane-wave basis set of 90 Ry and pseudopotentials) and in parentheses using the all electron BLYP/6-311++G** calculations (either optimized geometries or single point energies on the BLYP/6-31+G(d) geometry optimizations, see Table 1).

from Born−Oppenheimer (BO) ones whatever the fictitious mass value[49,50] is certainly important, and indeed the blue-shift of 100 cm^{-1} can be reduced when performing CPMD with smaller fictitious masses for the propagation of the electronic wave function. This is though at the cost of more expensive simulations. However, it has been demonstrated[49,50] that the CP forces can be brought into good agreement with the BO forces by simply rescaling the nuclear masses (thus leading to "dressed atoms"). Applying the mass scaling procedure[49] and averaging the results over all atoms of $[pSer\text{-}H]^-$, we find that blue shifts of 25 cm^{-1} and 100 cm^{-1} have to be applied, for the harmonic (20 K) and nonharmonic (300 K) dynamics performed in the present work, respectively. We have hence recovered the 100 cm^{-1} translation empirically found in our previous dynamics at 300 K. The 25 cm^{-1} translation nicely leads to the alignment of the harmonic CPMD spectrum (13 K) with the 0 K all electron spectrum computed at the BLYP/aug-cc-pVTZ level of theory (see Figure 1 in the Supporting Information and further details in part 3 below). In the rest of the paper, the above-mentioned translations have been applied to the calculated spectra. Note that performing a Born−Oppenheimer dynamics (free from the fictitious mass) remains about 10 times slower than a Car−Parrinello dynamics (within the CPMD code). This explains why vibrational spectroscopy calculations are performed with the CP algorithm despite the consequences of the use of the fictitious mass on absolute vibrational frequencies.

Assignment of the vibrational bands extracted from MD simulations has been achieved with the localization and decomposition procedure developed previously[51,52] with the associated Potential Energy Distribution (PED) quantification. This procedure goes beyond the Vibrational Density Of States (VDOS) analysis usually performed in the literature. Assignments have been done in terms of nonredundant Pulay internal coordinates (see ref 51 for more details). They are given for the two trajectories performed in this work, i.e. harmonic low-temperature dynamics and nonharmonic room-temperature dynamics. As presented in refs 51 and 52, our assignment method turns out to provide "effective normal modes". The ones extracted from the low-temperature harmonic dynamics are identical to the harmonic normal modes that are calculated by diagonalizing a Hessian matrix. The "effective normal modes" extracted from the 300 K nonharmonic dynamics take into account temperature and anharmonicities of the dynamics. We have compared the vibrational assignments for the harmonic and nonharmonic dynamics in order to assess how the modes can be modified by anharmonicities and temperature. Furthermore, we take the opportunity to quantify how much the room-temperature modes resemble or differ from the pure harmonic normal modes by projecting the room-temperature modes onto the normal modes extracted from the harmonic dynamics. In that way, we are able to assess how the harmonic normal modes can be relevant for the interpretation of the vibrational bands that are recorded at finite temperature in the experiment. Results will be presented in the following section.

3. Results

3.1. Geometry Optimizations. All electron geometry optimizations led to the eight isomers/conformers depicted in Figure 1. Their relative energies are gathered in Table 1. Three families can be identified, depending on the proton-

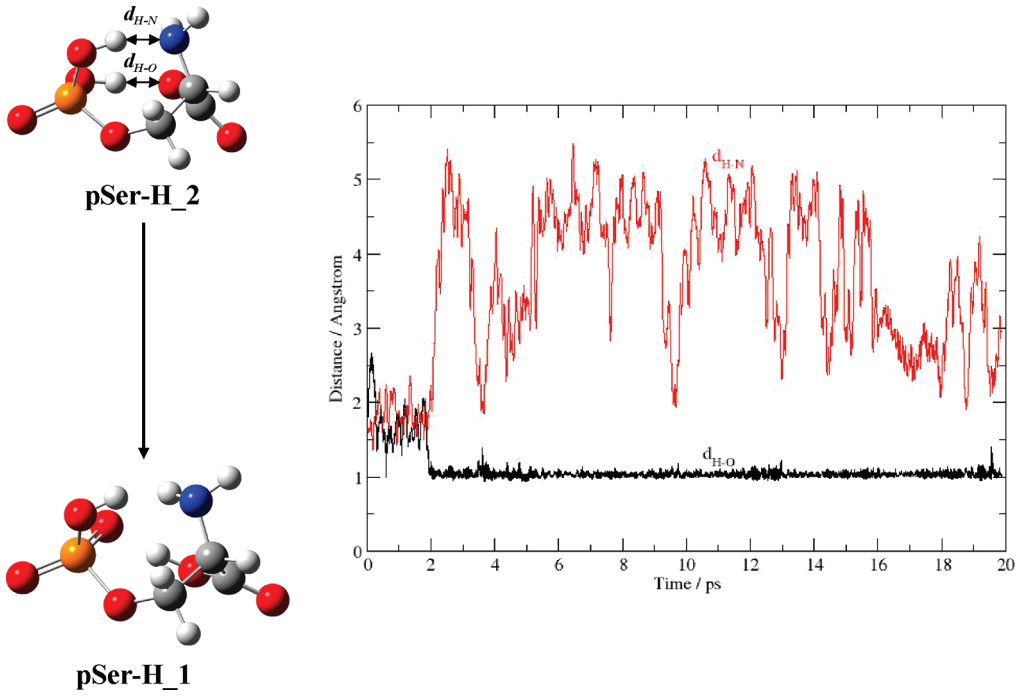


Figure 2. Car−Parrinello molecular dynamics simulation at 300 K. Evolution with time of (PO)H.. .N(H_2) and (PO)H.. .OCO distances, in order to illustrate the conformational dynamics between "pSer-H_1" and "pSer-H_2" types of geometries during the dynamics: proton transfer from POH to the acid.

ation state of the phosphate group and the N- and C-termini of the amino acid (OPO_3H_2/OPO_3H^-, NH_3^+/NH_2, COOH/COO^-). They are respectively identified as (1) deprotonated on the phosphate group (HPO_3), see structures *pSer-H_1*, *pSer-H_5*, *pSer-H_6*, *pSer-H_7*, and *pSer-H_8* in Figure 1, (2) deprotonated on the acid (COO^-), see structure *pSer-H_2*, and (3) salt-bridge structures with deprotonated phosphate and acid, and protonated amine, see *pSer-H_3* and *pSer-H_4*. Structures *pSer-H_1* and *pSer-H_2* are found to be the lowest in energy, with an energy difference smaller than ∼1.5 kcal/mol whatever the density functional and basis set employed. The energy difference is decreased to less than ∼1.0 kcal/mol when including the zero point vibrational energy, see Table 1. The BLYP and B3LYP relative energies are in excellent agreement with each other, with less than 0.5 kcal/mol difference in most cases (pSer-H_3, pSer-H_4, and pSer-H_5 have slightly larger differences of 1−1.5 kcal/mol).

pSer-H_1 and *pSer-H_2* can be seen as folded geometries displaying strong hydrogen bonds between the phosphate and amino groups. They differ by a simple proton transfer between the phosphate and the carboxylate. The conformations of higher energy mostly display less folded geometries that therefore give less opportunity for the phosphate group to form hydrogen bonds with the amine or the acid.

The geometry of *pSer-H_1* displays two strong and one weak H-bonds. The two strong H-bonds are formed between the COOH and one PO (COOH...OP H-bond distance and angle are respectively 1.529 Å and 172° at the BLYP/aug-cc-pVTZ level) and between the NH_2 and the same PO (2.009 Å and 140° at the BLYP/aug-cc-pVTZ level). A weak H-bond is formed between the amine and the POH (H_2N...HOP = 2.281 Å and 140° at the BLYP/aug-cc-pVTZ level). In the geometry of *pSer-H_2* the two POH groups are involved in strong hydrogen bonds, one with the carboxylate COO^-...HOP (1.597 Å and 178° at the BLYP/aug-cc-pVTZ level) and one with the amine H_2N...HOP (1.758 Å and 160° at the BLYP/aug-cc-pVTZ level).

3.2. Dynamics at 300 K. We have performed two dynamics at room temperature. The average temperatures obtained are reported in Table 2. One dynamics was begun from the optimized geometry *pSer-H_2*. As illustrated in Figure 2, this geometry changes within 2 ps of dynamics, when a proton transfer occurs from the POH group initially H-bonded to COO^-. This leads to an isomer bearing a deprotonated phosphate and a protonated acid. We never observed any proton transfer back from the COOH to the phosphate group during the dynamics. Figure 2 also shows that the remaining POH group is highly fluxional, with very little probability for a POH...NH_2 hydrogen bond. Note that the proton transfer occurs during the period of thermalization process of the dynamics, but we have checked that this event is not related to the velocity rescaling procedure, as the latter only takes place during the first 500 fs of thermalization, well before the proton transfer event. In order to shed some light as to why no proton transfer back to the phosphate group is observed during the length of our dynamics, we have extracted the free energy profile of the proton transfer from the dynamics. This is calculated as -kT lnP(H) where P(H) is the probability histogram related to the sampling of the reaction coordinate for the proton transfer between the phosphate and the acid along the dynamics. We found that the free energy barrier from pSer-H_2 to the transition state is ∼0.8 kcal/mol and that the free energy barrier from pSer-H_1 to the transition state is ∼3.1 kcal/mol. These values explain why the energy barrier is easily overcome during a room temperature dynamics taking place in the basin of pSer-H_2 but is not so easily overcome once trapped in the basin

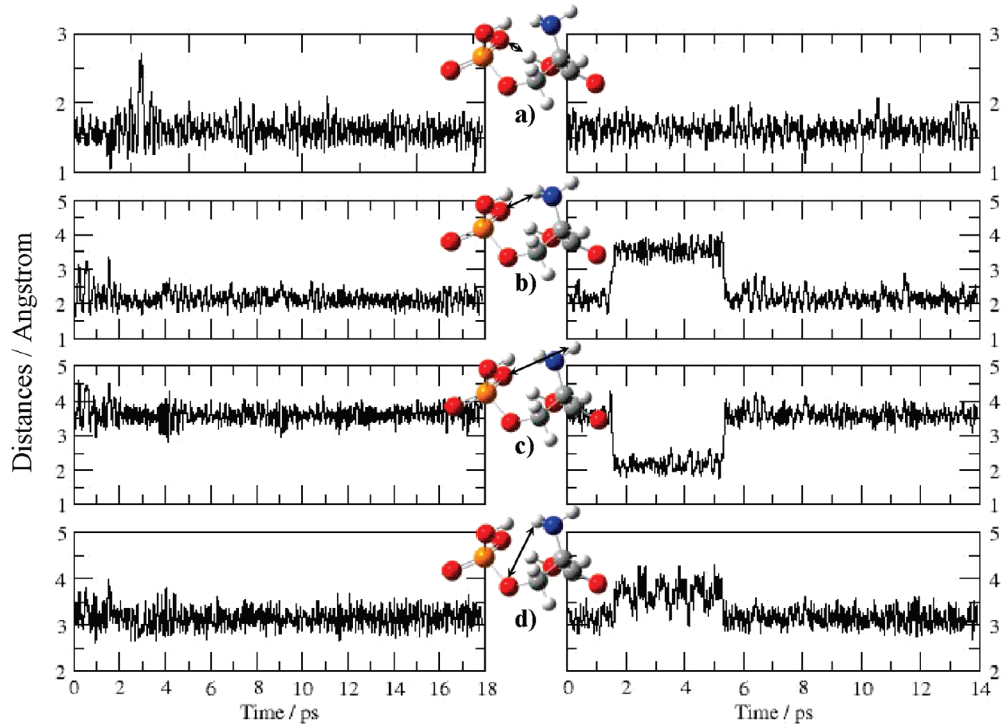


Figure 3. Car−Parrinello molecular dynamics simulations at 300 K. Evolution with time of selected possible hydrogen bond distances reported for the two room temperature dynamics performed in this work (average temperatures of the dynamics - see Table 2: 245 K (left) and 298 K (right)). Selected distances are illustrated with arrows.

of pSer-H_1. Note that the addition of nuclear quantum effects is expected to decrease these values. On the other hand, the free energies calculated from the optimized geometries of *pSer-H_1*, *pSer-H_2* and the saddle-point connecting them, within the harmonic approximation, give a slightly different view. Depending on the basis set used, the free energy barrier to be overcome from pSer-H_2 varies in the range 0.2/−0.2 kcal/mol and the one to be overcome from pSer-H_1 is 0.3/0.7 kcal/mol. These values suffer from the harmonic approximation though, which is not the case for the free energies extracted from the dynamics. Yet short propagation times may lead to an overestimation of the barriers. Low energy barriers would lead to the dynamical sampling of regions of the PES mostly outside the minima, while higher energy barriers would do the opposite. The latter is confirmed by the present dynamics and thence very good reproduction of the IR-MPD signatures, *vide infra*, therefore supporting the free energy values extracted from the dynamics.

The last configuration (positions and velocities) of this dynamics was used as a starting point for the second dynamics performed here, with an initial rescaling of velocities and thermalization procedure in order to randomize the velocities.

The two dynamics were propagated in the microcanonical ensemble for 17.9 and 13.9 ps, respectively, and we find that they both take place within the basin of "pSer-H_1", with small distortions of the skeleton from the optimized geometry: the skeleton dihedrals differ by no more than 20° from those of the optimized configuration. As reported in Table 2, the average temperatures of the dynamics are 245 and 298 K.

As reported in Figures 3 and 4, we can see that a PO...HOC=O H-bond is present at room temperature at a very short distance of ∼1.61 Å on average and is almost never broken (Figure 3a). PO is, on average, simultaneously involved in a supplementary weak H-bond with one of the hydrogen atoms of the amine. The average PO...HNH distance is ∼2.16 Å, and we can observe an exchange of the amine hydrogen involved in the H-bond over time (Figure 3b-c). There is thus enough internal energy at room temperature to overcome the barrier to the rotation of the amine group and the associated breaking/reforming of the H-bond with the neighboring PO group. Note that Figure 3d indeed confirms the rotation of the NH_2 group. There is no POH...NH_2 H-bond formed, on average (Figure 4a). This distance indeed evolves between 2.0 and 5.5 Å along the two dynamics, displaying transient very short periods of time during which a H-bond can actually be seen. At room temperature, there is an easy rotation around the P−OH bond, as nicely illustrated by the time evolution of the two POH...OP distances (Figure 4b-c): the H atom is seen to alternate short intermolecular distances with the two PO oxygens along the time. This rotation explains why no POH...NH_2 H-bond can be formed during the dynamics. On average, the phosphate group is therefore composed of one free PO bond, while the second PO is involved in two simultaneous hydrogen bonds with the amine N−H (weak H-bond) and the acid O−H (strong H-bond). POH is free of any hydrogen bonding, on average.

3.3. Infrared Spectroscopy at 300 K. The infrared spectrum extracted from the room temperature dynamics is shown in Figure 5, together with the experimental IR-MPD spectrum. We find that the calculated spectrum of $[pSer-H]^-$ at 300 K displays a good agreement of band widths with the IR-MPD experiment. The positions of the bands obtained in the present calculation differ by 10−20 cm^{-1} from their

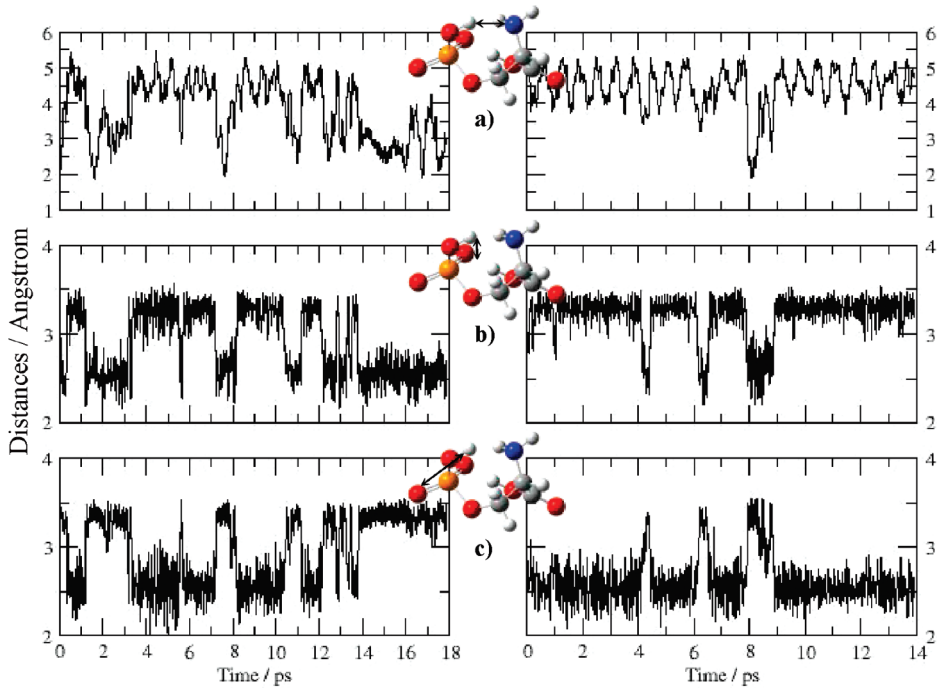


Figure 4. Car−Parrinello molecular dynamics simulations at 300 K. Evolution with time of selected possible hydrogen bond distances reported for the two room temperature dynamics performed in this work (average temperatures of the dynamics - see Table 2: 245 K (left) and 298 K (right)). Selected distances are illustrated with arrows.

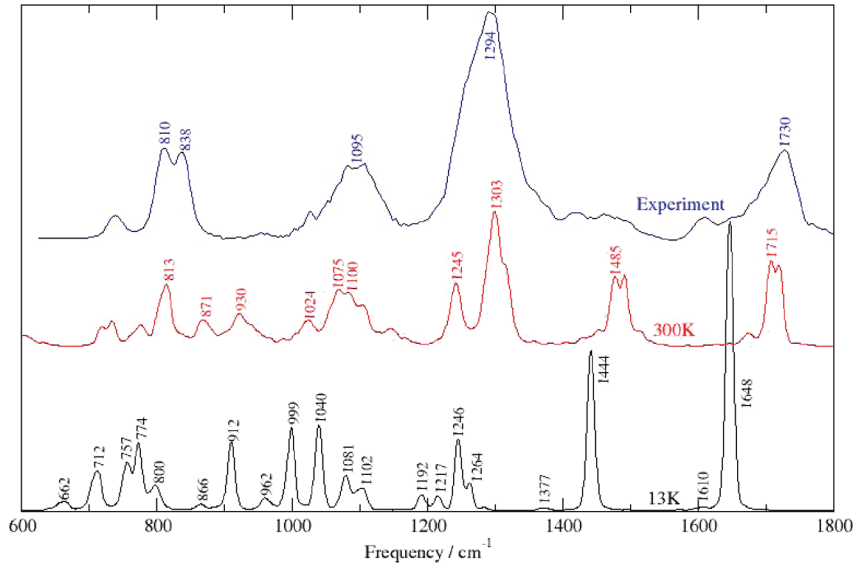


Figure 5. Infrared spectrum of $[pSer-H]^-$ extracted from the harmonic Car−Parrinello dynamics (13 K on average, bottom of the figure) and the room temperature nonharmonic Car−Parrinello dynamics (271 K as the average from the two CPMD performed in the present work, medium of the figure). The experimental IRMPD spectrum is reported at the top of the figure. Wavenumbers displayed on top of the bands are reported in cm^{-1}. Assignment of the bands can be found in the text. See Table 3 for the summary of the vibrational interpretation of the IRMPD spectrum as found from the room temperature dynamics performed in the present work.

IR-MPD values, while the experimental band gaps are very well reproduced by our calculation.

The following assignments of the vibrational bands are summarized in Table 3. The two bands located in the 1600−1800 cm^{-1} region of the spectrum are respectively assigned to the C═O stretching (∼1715 cm^{-1}) and to the N−H scissoring (∼1675 cm^{-1}) motions. The shape and breadth of the experimental feature at 1660−1740 cm^{-1} suggests that it is composed of two bands, as indeed our calculations show. Its higher frequency part is consistent with a carboxylic group as opposed to a carboxylate.[26,53] Note that the low intensity experimental band at ∼1600 cm^{-1} has no counterpart in our calculation.

The bands in the 1500−700 cm^{-1} calculated spectrum are related to vibrations predominantly arising from the phosphate and to a lesser extent from the acid. The band located between 1520 and 1420 cm^{-1} with a sharp feature at 1500−1470 cm^{-1} is due to C−OH stretching and bending,

Table 3. Summary of the Vibrational Interpretation of the IRMPD Spectrum As Found from the Room Temperature Dynamics Performed in the Present Work[a]

IRMPD band position (cm^{-1})	assignments from 300 K CPMD
1730	C═O stretch (acid)
1294	free PO stretch (phosphate)
1095	H-bonded PO stretch (phosphate)
810−838	P−OH stretch (phosphate)

[a] The interpretation reported here corresponds to the main assignments from the dynamics. See Text-Part 3, for the details.

while the ∼1430 cm^{-1} shoulder is due to the adjacent CH_2 rocking motion. The higher frequency part of the ∼1320−1220 cm^{-1} band arises from the free PO stretching (∼1303 cm^{-1}), also with a very low participation of the H-bonded PO stretching, and the lower frequency part is due to the stretching-bending of the C−OH (∼1245 cm^{-1}). The COH motions are thus split into two bands separated by ∼240 cm^{-1}. Within the large feature between 1000 and 1120 cm^{-1}, the main peak at ∼1075 cm^{-1} comes from the stretching motion of the hydrogen bonded PO. The two shoulders on both sides of this broadband are related to N−C stretching and NH_2 wagging around the C−N bond (also noted C-NH_2 wagging) (∼1145 cm^{-1}) and to combined skeleton OCC scissoring and C-NH_2 wagging (∼1024 cm^{-1}). Globally, the 1000−700 cm^{-1} calculated domain appears to suffer the largest band displacements with respect to experiment. The low intensity band calculated at ∼930 cm^{-1}, due to a combination of N−C stretching and POC bending, has no experimental counterpart. The ∼871 cm^{-1} band arises from the skeleton OCC scissoring motion, while stretching of the P−OH gives rise to the main ∼813 cm^{-1} peak. The 700−775 cm^{-1} bands come from more delocalized skeleton torsions. At least 2 peaks can be seen in the 800 cm^{-1} IR-MPD band. It is conceivable that the 813 cm^{-1} peak is well reproduced by our calculation, while the 871 cm^{-1} calculated band is blue-shifted by 32 cm^{-1} relative to experiment.

The IRMPD experimental active bands in the 1400−1800 cm^{-1} domain are therefore signatures of the acid and amine groups of the amino acid. In fact, the signatures of the acid (C═O at higher frequencies and COH at lower frequencies) are mainly observed, while the signature of the amine can only be seen in the tail of the C═O band. The three remaining experimental bands in the intermediate 800−1400 cm^{-1} range come solely from the phosphate group. There is a band gap of 228 cm^{-1} between the signatures of the free (1303 cm^{-1}) and H-bonded (1075 cm^{-1}) PO groups, while the band at lower frequency reflects the vibrations of the phosphate POH group. The signature of the H-bonded PO is red-shifted from the free PO, as expected upon formation of a hydrogen bond, and the large 228 cm^{-1} red-shift obtained here reflects the energetically strong H-bond that is formed between PO and COH.

All of these bands reproduce the IR-MPD bands in both positions and shapes. The present dynamical treatment thus appears to capture adequately the anharmonicities and mode couplings in such a way that most band positions and shapes are rather accurately reproduced.

3.4. Harmonic Spectra. In order to obtain a more detailed understanding of the vibrational signatures of $[pSer-H]^-$ at room temperature, it is useful to compare them to those of the harmonic spectrum. The latter has been calculated in two different ways: low temperature CPMD and static calculations (all-electron Gaussian calculations[41] and Hessian-based calculation through the CPMD package). At very low temperatures (an average temperature of 13 ± 2 K was used here, see Table 2) the dynamics becomes effectively harmonic to a very good approximation. If the exact same procedure were used to calculate the energy, this calculation would be nearly identical to the more common Hessian-based calculation at the optimized geometry, as shown in Table 3 of the Supporting Information. Note that in principle, even at the low temperature used here for the dynamics, very small barriers on the potential energy surface (e.g., 10 cm^{-1}) could be overcome, leading to deviations from the harmonic approximation. In order to evaluate whether this occurs in the present case, root-mean-square deviations on bond-lengths and bond-angles along the 13 K dynamics are presented: bond-lengths fluctuate in the range 0.004−0.009 Å (apart H-bond distances that fluctuate by up to 0.02−0.03 Å) and bond-angles vary within 1.9−3.0°. These very small variations of the molecular geometry confirm that the harmonic approximation of the potential energy surface is correct in the 13 K dynamics. Such small fluctuations cannot therefore have any significant influence on the calculated frequencies.

Although the same density functional is used (BLYP), the CPMD calculations (dynamics and Hessian-based) involve a plane wave valence basis set together with core pseudopotentials, instead of the all-electron Gaussian basis set used in the Hessian-based static calculation. Thus the low temperature CMPD dynamics calculation has a 2-fold goal: (1) calculate the harmonic CPMD IR spectrum and compare it to the static calculations in order to validate our CPMD setup and (2) discuss vibrational anharmonicities of $[pSer-H]^-$ at room temperature by comparing the low and room temperature spectra extracted at exactly the same level of theory. Furthermore, comparison of the CPMD Hessian-based harmonic spectrum to the 13 K CPMD dynamics harmonic spectrum is presented in order to demonstrate that the average 25 cm^{-1} translation that we apply to the 13 K dynamics harmonic spectrum (due to the use of the fictitious mass in the dynamics) is correct.

The static harmonic normal-mode frequencies of the conformation of lowest energy (*pSer-H_1*) calculated with the BLYP and B3LYP functionals and three different basis sets are reported in Table 1 in the Supporting Information (no scaling factor applied). For each functional investigated here, the frequencies vary on average by ±10 cm^{-1} when changing the basis set, and slight variations in some peak intensities can also be observed. The corresponding IR spectra are plotted in Figure 2 of the Supporting Information for the specific case of the aug-cc-pVTZ basis set. In this figure scaling factors of 0.9669 and 0.9962 have been applied to the B3LYP and BLYP frequencies, respectively.[54] These

values are taken from 6-311G(d,p) basis set calculations, as this is the largest basis set investigated in ref 54. BLYP and B3LYP spectra display exactly the same vibrational features, with the B3LYP frequencies being on average 10−20 cm^{-1} higher than the BLYP frequencies. The two functionals therefore perform similarly.

We performed a 17.1 ps CPMD dynamics at 13 K starting from *pSer-H_1*. As expected at such a low temperature, the average geometry obtained during the dynamics remains very close to the optimized one, as shown in Table 2 in the Supporting Information where structural parameters are reported. In particular, the three hydrogen bonds in the optimized geometry discussed above are maintained at 13 K. The only significant difference comes from the POH...N H-bond distance that is found to be 0.13 Å longer in the dynamics. This is mainly a basis set rather than a temperature effect, as this distance is optimized at 2.28 Å with the largest aug-cc-pVTZ basis set and increases with the basis set growth, to be compared to its average value of 2.33 Å from the dynamics.

The frequencies extracted from the 13 K dynamics harmonic spectrum are compared to the frequencies obtained from the Hessian-based calculation in Table 3 of the Supporting Information, all obtained within the CPMD setup. Two schemes have been applied for the Hessian-based calculation, Linear Response (LR) and Finite Difference (FD), that give identical results within 0−2 cm^{-1}. The frequencies extracted from the 13 K harmonic spectrum differ from the Hessian-based frequencies by an average 7 cm^{-1}. This is the result of the mean 25 cm^{-1} translation we apply to the dynamical spectrum to take care of the fictitious mass and not from deviations from harmonicities in the 13 K dynamics as already demonstrated above. It is worth noting that the frequencies of the movements implying heavy atoms are identically found in the spectra extracted from the dynamics and the static Hessian-based. In the rest of the paper we only make use of the harmonic spectrum extracted from the 13 K CPMD dynamics.

The harmonic CPMD IR spectrum is presented in Figure 5 together with the 300 K CPMD and the experimental IRMPD spectra. It is also shown in Figure 1 of the Supporting Information along with the all electron BLYP/aug-cc-pVTZ IR spectrum. Frequencies in the 600−1800 cm^{-1} region obtained with the 6-31+G(d), 6-311++G(d,p), and aug-cc-pVTZ basis sets using B3LYP and BLYP functionals are reported in Table 1 of the Supporting Information, for the sake of comparison. We immediately observe that the spectrum extracted from the dynamics displays the same features as the all electron BLYP spectrum, apart from slight shifts in the band positions. On average, CPMD frequencies differ by 10−20 cm^{-1} from the all electron frequencies, the largest discrepancies being observed below 800 cm^{-1}. Differences observed in the intensities can be traced back to the equipartition of energy, which is difficult to achieve at low temperature and within such a short dynamics time scale. As a consequence, the participation of cold atoms involved in active bands is underestimated (thus usually underestimating the corresponding IR intensities), while the participation of warm atoms involved in active bands is overestimated (thus usually overestimating IR intensities). As seen in Table 2, carbon atoms are, on average, the warmest atoms, which will be one main reason for the high intensities of the δ(COH) (1444 cm^{-1}) and ν(C═O) bands (1648 cm^{-1}) as well as for the 912 and 999 cm^{-1} bands whose signatures can be traced back to carbon atoms.

As for the 300 K spectrum, the 13 K harmonic spectrum in Figure 5 may be interpreted using our localization and decomposition procedure.[51,52] The bands in the 1800−1400 cm^{-1} domain are due to the stretching motion of C═O (1648 cm^{-1}) and combined stretch−bend of COH (1444 cm^{-1}). The shoulder at 1610 cm^{-1} arises from the amine scissoring motion. The main signature of the ∼1200−1300 cm^{-1} band is due to the free PO stretching (1246 cm^{-1}). The satellite bands are respectively arising from CH rocking and C−OH stretch (1264 cm^{-1}), NH_2 rocking (1217 cm^{-1}), and C−OH stretching (1192 cm^{-1}). The large 1100−800 cm^{-1} band is composed of POH bending and free PO stretch (1102 cm^{-1}), C−OH wagging (1081 cm^{-1}), CH_2 rocking (1040 cm^{-1}), H-bonded PO stretch (999 and 962 cm^{-1}), C-NH_2 wagging (912 cm^{-1}), N−C stretch (866 cm^{-1}), and COOH wagging around the C−C bond (800 cm^{-1}). The 757−774 cm^{-1} bands are assigned to the OH stretch of POH and the lower frequencies to more delocalized modes (NH_2/COOH part at 712 cm^{-1} and phosphate part at 662 cm^{-1}).

It is worth noting that the harmonic COH bending mode observed here at 1444 cm^{-1} is blue-shifted by about 200 cm^{-1} from the same bending motion of the COH group when it is not involved in hydrogen bonding as is the case for conformers pSer-H_5 to pSer-H_8, see Figure 1. Blue shifts on bending motions are expected from the formation of hydrogen bonds, and such a strong displacement reflects the energetically strong H-bond that is formed between the PO and the acid group of $[pSer\text{-}H]^-$.

3.5. Vibrational Anharmonicities and Mode Couplings at 300 K. Comparison of the spectra extracted from the harmonic (low temperature) and nonharmonic (room temperature) dynamics allows us to quantify vibrational shifts due to the combination of temperature driven conformational dynamics, vibrational anharmonicities (anharmonic oscillators and mode couplings), and dipole anharmonicities (beyond the electrical harmonic approximation), since band shifts are the result of these combined effects. The anharmonic spectrum is globally shifted toward higher frequencies with respect to the harmonic spectrum, with an average blue-shift of ∼50−60 cm^{-1}. Though blue-shifts upon anharmonicities and mode couplings have already been observed by Gerber et al. on other peptide models in the same frequency range,[55−58] the systematic blue-shift obtained here is puzzling, and might be entirely fortuitous and due to the DFT PES. However, it remains that upon this blue-shift the anharmonic spectrum is perfectly aligned with the IR-MPD spectrum and provides a very good account of the experimental signatures.

Following ref 55, we define the percentage of anharmonicity of a mode by $(\nu_{anharm} - \nu_{harm})*100/\nu_{anharm}$ where ν_{harm} and ν_{anharm} are respectively the harmonic and anharmonic frequencies of a given mode. In the frequency region investigated here, we find that this percentage is 4% on

average, apart for the stretching movements of the H-bonded PO (1075 cm^{-1} in the 300 K anharmonic spectrum) and free P−OH (813 cm^{-1} in the 300 K anharmonic spectrum), where the anharmonicities are respectively 10.5% and 7%. Interestingly, the C−OH stretching movement (1245 cm^{-1} in the 300 K anharmonic spectrum) is only 4% anharmonic, even though this group is hydrogen bonded to PO.

As described in the method section above, we are able to describe and quantify the anharmonic vibrational modes in terms of the harmonic modes. We find that the C═O stretching, NH_2 scissoring, and COH bending modes in the 1400−1800 cm^{-1} domain can be fully described by the corresponding harmonic modes, implying that these modes undergo no anharmonic coupling to other modes. As a consequence, the blue-shifts of these bands at 300 K relative to 13 K and the average 4% anharmonicity of these modes appear to be due to the intrinsic anharmonicities arising from the potential energy surface and transition dipole moment of the related vibrational motions.

On the contrary, the three phosphate bands located at 1303, 1075, and 813 cm^{-1} display couplings to several harmonic modes. Hence, the 1303 cm^{-1} free PO stretch anharmonic mode has two major components, 60% from the 1246 cm^{-1} PO stretch harmonic normal mode and 25% from the POH bending harmonic normal mode at 1102 cm^{-1}. The 1075 cm^{-1} H-bonded PO stretch is a stiffer vibrational mode as it displays less mode coupling, though two major components are found: 76% of this mode is represented by the 999 cm^{-1} PO stretch harmonic normal mode, together with 7% of the C−OH wagging 1081 cm^{-1} harmonic mode. The 813 cm^{-1} mode displays an intermediate state in mode couplings, with 69% and 18% from the 774 and 757 cm^{-1} harmonic modes, respectively, although both harmonic modes correspond to a splitting of the harmonic P−OH stretch. These couplings certainly participate to the blue shifts of the bands obtained at room temperature.

4. Conclusions

The IR spectrum of deprotonated phosphorylated serine $[pSer-H]^-$ extracted from the present DFT-based room temperature molecular dynamics simulations gives a good account of the experimental IRMPD spectrum. Band positions and shapes are in very good agreement with the experiment, and, as already mentioned, intensities obtained from the one-photon absorption calculations performed here should not be directly compared to the fragmentation yield recorded in the experiment. The comparison of the room temperature calculated spectrum to the low temperature harmonic one calculated at the same level allows for detailed insight into temperature effects. It is found that temperature induces large changes in the IR spectrum, as illustrated by the blue shifts of a number of the main bands observed from 13 to 300 K. These shifts, the subsequent merging of certain bands and changes in their assignments, the broadening of the bands, are direct results of anharmonicities (from both the potential energy surface and the dipole moment) and mode couplings that are naturally included in the room temperature dynamics.

It is clear that vibrational anharmonic effects probed in molecular dynamics depend on the temperature of the simulation, as recently shown in ref 59. The present investigation shows that room-temperature dynamics of $[pSer-H]^-$ (expected temperature of the reference IR-MPD experiment) provides a theoretical spectrum that convincingly agrees with the experiment. The same agreement cannot be achieved with harmonic calculations, implying that anharmonicities and mode-couplings are appropriately probed in our room-temperature simulations.

The room temperature dynamics of $[pSer-H]^-$ shows a rather geometrically stiff molecule, with small distortions of the skeleton observed from the optimized geometry. Only the POH and NH_2 groups appear to be significantly fluxional. The vibrational bands recorded in the mid-IR 1800−700 cm^{-1} region predominantly arise from the deprotonated phosphate and to a lesser extent from the acid. No significant participation from the amine is observed in this region, although it is contained in the 1730 cm^{-1} IR-MPD broadband. A large blue-shift and a rather spectacular change of band shape is observed for the C═O stretching band at 1720 cm^{-1}. The bump observed in the 1400−1500 cm^{-1} region in the IRMPD spectrum is a direct probe of the protonation of the acid, as it is due to the COH bending. The three main bands in the 1400−800 cm^{-1} experimental spectrum are the signatures of the deprotonated phosphate with two bands demonstrating that there is a free PO (1294 cm^{-1} in the IR-MPD) and a strongly hydrogen bonded PO (1095 cm^{-1} in the IR-MPD). The low frequency P−OH stretching (around 830 cm^{-1} in the IR-MPD) reflects the fluxional rotating POH group, not involved in hydrogen bonding.

The analyses of the harmonic and nonharmonic modes performed here have shown that the percentage of anharmonicity in the vibrational modes is 4% on average, but that they are larger for the stretching movements of the H-bonded PO and free P−OH, where the anharmonicities are respectively 10.5% and 7%. Moreover, the assignment of the high frequency 1800−1400 cm^{-1} bands arising from the protonated acid has been shown to be identical to the harmonic modes. This is not true anymore in the 1400−700 cm^{-1} region, where the nonharmonic assignments show mode couplings arising from two predominant harmonic modes. These couplings are expected to participate to the blue-shifts of the bands that allow for a good agreement with the IR-MPD experiment.

The importance of taking temperature effects into account has been demonstrated here for the model phosphorylated serine amino acid. It is expected that the relevance of these effects will be even larger for peptides of increasing size and complexity. Going further down to lower frequency regions of the vibrational spectrum is also likely to strengthen mode couplings and anharmonicities. Calculation of IR spectra through molecular dynamics simulations will enable the proper modeling of these features. This is where our combined experiments and calculations are currently heading.

Acknowledgment. The authors thank ANR-Probio for financial support (PostDoc position of A.C) and IDRIS (Orsay, France) for a generous allowance of computer time. This work was performed using HPC resources from GENCI-IDRIS

(Grant 2009-i2009082073). The IRMPD spectrum was recorded at the infrared laser center CLIO in Orsay, France. This was made possible by the support of the European Union through the EPITOPES project (NEST 15637).

Supporting Information Available: Car−Parrinello molecular dynamics and all electron geometry optimizations performed. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cohen, P. *Trends Biochem. Sci.* **2000**, *25*, 596–601.

(2) Sefton, B. M.; Hunter, T. In *Protein phosphorylation*; Academic Press: 1998.

(3) Johnson, L. N.; Lewis, R. J. *Chem. Rev.* **2001**, *101*, 2209–2242.

(4) Espinoza-Fonseca, L. M.; Kast, D. T. D. D. *J. Am. Chem. Soc.* **2008**, *130*, 12208–12209.

(5) Mandell, D. J.; Chorny, I.; Groban, E. S.; Wong, S. E.; Levine, E.; Rapp, C. S.; Jacobson, M. P. *J. Am. Chem. Soc.* **2007**, *129*, 820–827.

(6) Nakamoto, K. *Infrared and Raman Spectra of Inorganic and Coordination Compounds. Parts A and B*; New York, 1997.

(7) Lemaire, J.; Boissel, P.; Heninger, M.; Mauclaire, G.; Bellec, G.; Metsdagh, H.; Simon, A.; Caer, S. L.; Ortega, J. M.; Glotin, F.; Maître, P. *Phys. Rev. Lett.* **2002**, *89*, 273002.

(8) Maître, P.; Caer, S. L.; Simon, A.; Jones, W. D.; Lemaire, J.; Metsdagh, H.; Heninger, M.; Mauclaire, G.; Moissel, P.; Prazeres, R.; Glotin, F.; Ortega, J. M. *Nucl. Instrum. Methods Phys. Res. A* **2003**, *507*, 541–546.

(9) Oomens, J.; Roij, A. J. A. V.; Meijer, G.; Helden, G. V. *Astrophys. J.* **2000**, *542*, 404–410.

(10) Oomens, J.; Sartakov, B. G.; Meijer, G.; Helden, G. V. *Int. J. Mass Spectrom.* **2006**, *254*, 1–19.

(11) Balaj, O. P.; Kapota, C.; Lemaire, J.; Ohanessian, G. *Int. J. Mass Spectrom.* **2008**, *269*, 196–209.

(12) Gregoire, G.; Gaigeot, M. P.; Marinica, D. C.; Lemaire, J.; Schermann, J. P.; Desfrançois, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3082–3097.

(13) Lucas, B.; Gregoire, G.; Lemaire, J.; Maître, P.; Glotin, F.; Schermann, J. P.; Desfraçois, C. *Int. J. Mass Spectrom.* **2005**, *243*, 105–113.

(14) Marinica, D. C.; Gregoire, G.; Desfrançois, C.; Schermann, J. P.; Borgis, D.; Gaigeot, M. P. *J. Phys. Chem. A* **2006**, *110*, 8802–8810.

(15) Oomens, J.; Polfer, N. C.; Moore, D. T.; Meer, L. v. d.; Marshall, A. G.; Eyler, J. R.; Helden, G. v. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1345–1348.

(16) Polfer, N. C.; Oomens, J.; Dunbar, R. C. *ChemPhysChem* **2008**, *9*, 579–589.

(17) Polfer, N. C.; Oomens, J.; Suhai, S.; Paizs, B. *J. Am. Chem. Soc.* **2007**, *129*, 5887–5897.

(18) Polfer, N. C.; Paizs, B.; Snoek, L. C.; Compagnon, I.; Suhai, S.; Helden, G. v.; Oomens, J. *J. Am. Chem. Soc.* **2005**, *127*, 8571.

(19) Prell, J. S.; Demireva, M.; Oomens, J.; Williams, E. R. *J. Am. Chem. Soc.* **2009**, *131*, 1232.

(20) Vaden, T. D.; Boer, T. S. J. A. d.; Simons, J. P.; Snoek, L. C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1443–1447.

(21) Vaden, T. D.; Boer, T. S. J. d.; Simons, J. P.; Snoek, L. C.; Suhai, S.; Paizs, B. *J. Phys. Chem. A* **2008**, *112*, 4608–4616.

(22) Vaden, T. D.; Gowers, S. A. N.; Boer, T. S. J. A. d.; Steill, J. D.; Oomens, J.; Snoek, L. C. *J. Am. Chem. Soc.* **2008**, *130*, 14640–14650.

(23) Wu, R. H.; McMahon, T. B. *J. Am. Chem. Soc.* **2007**, *129*, 11312–11313.

(24) Correia, C. F.; Balaj, O. P.; Scuderi, D.; Maître, P.; Ohanessian, G. *J. Am. Chem. Soc.* **2008**, *130*, 3359–3370.

(25) Correia, C. F.; Clavaguera, C.; Erlekam, U.; Scuderi, D.; Ohanessian, G. *ChemPhysChem* **2008**, *9*, 2564.

(26) Scuderi, D.; Correia, C. F.; Balaj, O. P.; Ohanessian, G.; Lemaire, J.; Maître, P. *ChemPhysChem* **2009**, *10*, 1630–41.

(27) CPMD, copyright International Business Machines Corporation (1990−2008) and Max Planck Institute fuer Festkoerperforschung Stuttgart (1995−2001).

(28) Gaigeot, M. P. *J. Phys. Chem. A* **2008**, *112*, 13507.

(29) McQuarrie, D. A., *Statistical Mechanics*; Harper-Collins Publishers: New York, 1976.

(30) Cimas, A.; Vaden, T. D.; Boer, T. S. J. A. d.; Snoek, L. C.; Gaigeot, M. P. *J. Chem. Theor. Comput.* **2009**, *5*, 1068–1078.

(31) Gaigeot, M. P.; Sprik, M. *J. Phys. Chem. B* **2003**, *107*, 10344.

(32) Gaigeot, M. P.; Vuilleumier, R.; Sprik, M.; Borgis, D. *J. Chem. Theory Comput.* **2005**, *1*, 772.

(33) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. *J. Phys. Chem. B* **2006**, *110*, 3685.

(34) Guidon, M.; Schiffmann, F.; Hutter, J.; VandeVondele, J. *J. Chem. Phys.* **2008**, *128*, 214104.

(35) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(36) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(37) Jimenez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6621–6629.

(38) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(39) Lin, I. C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. *J. Phys. Chem. B* **2009**, *113*, 1127–1131.

(40) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 19.

(41) 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.; 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. *Gaussian03*; Gaussian: Pittsburgh, 2003.

(42) Becke, A. *Phys. Rev. A* **1988**, *38*, 3098.

(43) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(44) Trouillier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993.

(45) Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425.

(46) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 2810.

(47) Borysow, J.; Moraldi, M.; Frommhold, L. *Mol. Phys.* **1985**, *56*, 913.

(48) Ramirez, R.; Lopez-Ciudad, T.; Kumar, P.; Marx, D. *J. Chem. Phys.* **2004**, *121*, 3973.

(49) Tangney, P. *J. Chem. Phys.* **2006**, *124*, 044111.

(50) Tangney, P.; Scandolo, S. *J. Chem. Phys.* **2002**, *116*, 14.

(51) Gaigeot, M. P.; Martinez, M.; Vuilleumier, R. *Mol. Phys.* **2007**, *105*, 2857.

(52) Martinez, M.; Gaigeot, M. P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2006**, *125*, 144106.

(53) Oomens, J.; Steill, J. D. *J. Phys. Chem. A* **2008**, *112*, 3281–3283.

(54) Irikura, K. K.; Johnson, R. D.; Kacker, R. N. *J. Phys. Chem. A* **2005**, *109* (37), 8430.

(55) Brauer, B.; Chaban, G. M.; Gerber, R. B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2543.

(56) Brauer, B.; Dubnikova, F.; Zeiri, Y.; Kosloff, R.; Gerber, R. B. *Spectrochim. Acta, Part A* **2008**, *71*, 1438.

(57) Pele, L.; Gerber, R. B. *J. Chem. Phys.* **2008**, *128*, 165105.

(58) Chaban, G. M.; Gerber, R. B. *Theor. Chem. Acc.* **2008**, *120*, 273.

(59) Carbonniere, Ph.; Dargelos, A.; Ciofini, I.; Adamo, C.; Pouchan, C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4375−84.

CT900179D

JCTC Journal of Chemical Theory and Computation

Excited State Potential Energy Surfaces of Polyenes and Protonated Schiff Bases

Robert Send*

Institut für Physikalische Chemie, Universität Karlsruhe, Kaiserstrasse 12, D-76128 Karlsruhe, Germany

Dage Sundholm

Department of Chemistry, P.O. Box 55 (A.I. Virtanens plats 1), University of Helsinki, FI-00014 Helsinki, Finland

Mikael P. Johansson

Lundbeck Foundation Centre for Theoretical Chemistry, Aarhus University, Langelandsgade 140, DK-8000 Århus C, Denmark

Filip Pawłowski

Physics Institute, Kazimierz Wielki University, Plac Weyssenhoffa 11, PL-85-072 Bydgoszcz, Poland

Received May 14, 2009

Abstract: The potential energy surface of the ^1B_u and $^1A'$ states of all-*trans*-polyenes and the corresponding protonated Schiff bases have been studied at density functional theory and coupled cluster levels. Linear polyenes and protonated Schiff bases with 4 to 12 heavy atoms have been investigated. The calculations show remarkable differences in the excited state potential energy surfaces of the polyenes and the protonated Schiff bases. The excited states of the polyenes exhibit high torsion barriers for single-bond twists and low torsion barriers for double-bond twists. The protonated Schiff bases, on the other hand, are very flexible molecules in the first excited state with low or vanishing torsion barriers for both single and double bonds. Calculations at density functional theory and coupled cluster levels yield qualitatively similar potential energy surfaces. However, significant differences are found for some single-bond torsions in longer protonated Schiff bases, which indicate a flaw of the employed time-dependent density functional theory methods. The close agreement between the approximate second and third order coupled cluster levels indicates that for these systems calculations at second order coupled cluster level are useful in the validation of results based on time-dependent density functional theory.

1. Introduction

The 11-*cis*-retinal photoreceptor in rhodopsin is responsible for the 11-*cis* to all-*trans* isomerization reaction which triggers the human visual process. The absorption of a photon initiates a complex and fast photoreaction involving several intermediates; the first species are formed within 200 fs.[1−9] The product of the photoisomerization reaction after about 1 ps is bathorhodopsin which has a distorted all-*trans* structure.[8−10] Experimental information about the photoreaction mechanism can be obtained by, e.g., femtosecond Raman spectroscopy (FSRS) measurements,[9] but computational approaches are indispensable to complement the experimental results.

The size of the full retinal still poses something of a challenge for accurate computational approaches. Insight

* Corresponding author e-mail: robert.send@kit.edu.

10.1021/ct900240s CCC: $40.75
Published on Web 07/27/2009

gained from the study of smaller model systems, the focus of the present study, alleviates to some extent this problem. All-*trans* polyenes and the corresponding protonated Schiff bases (PSBs), where one $=CH_2$ end group is replaced by the isoelectronic $=NH_2^+$, are often adopted as models for retinal PSB chromophores[11−16] It has to be kept in mind that the polyenes and the PSBs have different properties,[13] and molecular orbital theory models based on polyenes cannot be used to understand the excited state behavior of the PSBs.

Computationally demanding approaches such as complete-active-space self-consistent-field (CASSCF)[17] calculations in combination with second order perturbation theory (CASPT2) corrections[18] have been considered the most reliable approach for computational studies of the retinal isomerization reaction. However, the CASSCF/CASPT2 approach suffers from some undesired limitations. At the CASSCF level, merely the static valence correlation effects are considered, because only a few valence orbitals and electrons can be included in the active space. Dynamic electron correlation effects can be taken into account in a subsequent CASPT2 calculation. Larger active spaces could be used in multireference calculations by employing the restricted active space SCF method (RASSCF)[19] which also can be augmented with the second-order perturbation theory (RASPT2) approach[20] for considering parts of the remaining dynamical correlation effects. Due to the high computational cost, the molecular structure is usually not optimized at the CASPT2 level. Using CASSCF structures can lead to very large errors in the computed excitation energies, as recently confirmed by Altun et al.[21] Further, the employed basis sets are generally only of double-ζ quality, which is somewhat inadequate for reliable energetics at correlated *ab initio* levels.[22−26] Basis set studies on acrolein, *cis*-butadiene, and diazomethane show that also the CASPT2 geometries are strongly basis set dependent, with large differences between double and triple-ζ results.[27]

Alternative computational approaches for retinal studies are necessary to allow excited state optimizations using analytical gradients and larger basis sets. The time-dependent density functional theory (TDDFT) approach[28−32] has proven to be a successful method for studies of excited states. Optimizations can be routinely performed as effective algorithms for molecular gradients have been developed and implemented.[33] We also note that recent developments within the "dressed TDDFT" approach[34−37] appear to be promising for treating double-excitation dominated transitions, which are troublesome for standard TDDFT approaches. However, extensive TDDFT calculations on retinal protonated Schiff-bases (PSBs) using large basis sets have yielded results in disagreement with previously published CASSCF/CASPT2 studies.[22,23,38−41] The molecular structure of the first excited state of retinal PSBs optimized at the TDDFT level has a long single bond connecting the β-ionone ring with the retinyl chain, leading to a $-90°$ orientation of the ring relative to the retinyl chain. For the ground state structure there is only a $-39°$ twist angle between the ring and the chain.[38] Such a structure of the first excited state is not supported at any other level of theory. The most popular explanation for these discrepancies is the inability of the TDDFT method to accurately describe long-range charge transfer effects,[42,43] a conception that is challenged by this work.

Coupled cluster methods offer an alternative to the above-mentioned computational procedures. The approximate second order coupled cluster approach (CC2) is still feasible for molecules as large as retinal and accounts for a significant amount of electron correlation effects.[44,45] It has an appealing black-box character and allows excited state optimizations with large basis sets using analytical gradients. The CC2 method is able to accurately describe excited states with dominant single excitation character. CC2 calculations on retinal PSBs indicated that this necessary condition holds for at least the two lowest excited states;[39] the CC2 excitation energies deviate less than 0.07 eV from available experimental data.[46] The CC2 method does not suffer from long-range charge transfer problems since the exact-exchange operator is employed. The accuracy of CC2 calculations can be further assessed at the computationally more demanding approximate third order coupled cluster (CC3) level[47] on smaller PSB model compounds.

Recent CC2 and TDDFT studies on the 11-*cis* retinal PSB yielded qualitatively the same results for excitation energies, excited state molecular structures, dipole moment changes on excitation, and for twists along two major reaction coordinates.[38−40,48] The CC2 and TDDFT studies proposed a new isomerization reaction mechanism supported by femtosecond spectroscopic studies.[9] The reaction involves twisted retinyl structures and a stable intermediate.[41] The main disagreement between the results obtained at the two levels concerns the torsion angle of the single bond connecting the retinyl chain with the β-ionone ring moiety. Zaari and Wong also noted discrepancies between the TDDFT and CC2 description of some excited states of retinal.[49] The objective of the present study is to perform a thorough benchmark comparison of the TDDFT and CC2 results.

Although not central to this study, and therefore addressed only briefly, another discrepancy between the CC2 and TDDFT results on the one side, and the CAS results on the other side, is found in the excited state bond length alternation. CASSCF calculations on PSBs by Page and Olivucci yielded excited state structures with an inverted bond length alternation as compared to the ground state;[27] the double bonds become longer and the single bonds shorter upon excitation. The inverted bond length alternation is less severe at the CASPT2 level. It was also found that the potential energy surface (PES) of the excited state of PSBs is flat. At the CASSCF level, the authors obtained two energetically almost degenerate geometries for the 2,4-pentadieniminium cation with an essentially reversed bond conjugation, whereas at the CASPT2 level only one minimum is obtained.[27] In contrast, at the CC2 and TDDFT levels, the bond length alternation in the excited state of the 11-*cis* retinal PSB is not inverted but enhanced; the double bonds become shorter and the single bonds longer.[14,40] Before studying the bond length discrepancy in detail, it is necessary to analyze the differences between the CC2 and TDDFT results, and we therefore

concentrate on the PESs of bond torsion twists at the CC2 and TDDFT level.

Here, we present systematic TDDFT and CC2 studies of the PESs of the first excited state for both single and double bonds of the polyenes and the PSBs. The CC3 method is used to assess the accuracy and the reliability of the CC2 calculations. If the MR character is important, the PESs obtained at the CC2 and CC3 levels significantly differ. This occurs especially for large torsion angles of the double bond twists. The energies in these regions are not well-described by the single-reference methods employed in this study but are at the same time of little importance for the discussion and conclusions, which only consider small-angle twists. Previous work has shown that as long as the structural region of interest is located far from a conical intersection, the performance of, e.g., CC2 is very satisfactory.[50,51]

The PESs are examined by performing single point calculations for torsion angles of the carbon−carbon (carbon−nitrogen) bonds yielding upper bounds for the torsion barriers. Relaxation of the remaining structural degrees of freedom would introduce unpredictable uncertainties. For some molecules, a remote part of the molecule might twist, whereas in other cases a small barrier can prevent such relaxations. Thus, significantly different results can be obtained for two molecules even though the differences in the PESs are small.

The present article is structured as follows. After an overview of the computational methods employed, Section 3 discusses basis set effects on the vertical excitation energies and the PESs as well as basis set effects on the bond length alternation of the ground and the first excited state. In Section 4, the performance of different density functionals is reported. The PESs of the polyenes and the PSBs calculated at the TDDFT, CC2, and CC3 levels are compared in Section 5.

2. Computational Details

2.1. Basis Sets. The basis set convergence of the excitation energies is investigated at the CC2 level by performing single point calculations using a systematic sequence of Dunning's correlation-consistent (cc) basis sets up to aug-cc-pV6Z quality.[52−54] In the corresponding B3LYP[55,56] TDDFT calculations, Dunning's cc basis sets up to aug-cc-pVQZ are employed. We also used the Karlsruhe basis sets of split valence quality with polarization functions on all atoms (SVP) and on all atoms except hydrogens (SV(P)),[57] the triple-ζ valence quality basis set with one (TZVP) and two (TZVPP) sets of polarization functions,[58,59] and the quadruple-ζ basis sets augmented with two sets of polarization functions (QZVPP).[60] The basis sets denoted aug-SV(P) are the SV(P) basis sets augmented with diffuse functions from Dunning's cc double-ζ basis sets (aug-cc-pVDZ). The aug-TZVP basis sets are the TZVP basis sets augmented with the diffuse functions from Dunning's cc triple-ζ basis sets (aug-cc-pVTZ).

2.2. Functionals. The performance of different density functionals is assessed by single point TDDFT calculations using functionals of the generalized gradient approximation (GGA) as well as hybrid functionals. We have used the BP86[61−63] and PBE[64] GGA functionals as well as the B3LYP[55,56] and PBE0[65] hybrid GGA functionals. The Coulomb-attenuated B3LYP (CAM-B3LYP) functional[66] was used to check the long-range charge transfer effects on the excitation energies. In the benchmark calculations of the functionals, the TZVPP basis set was used; in the CAM-B3LYP calculations, the cc-pVTZ basis set was employed.

2.3. Ground State Structures. In the density functional theory (DFT) studies, the ground state structures were optimized using the B3LYP functional and the TZVPP basis sets. For the excited state coupled cluster studies, the ground state structures were optimized at the second order Møller−Plesset (MP2) perturbation theory level using the resolution-of-identity (RI) approximation[67−69] and the TZVPP basis sets. The optimized Cartesian coordinates are given in Sections I and II of the Supporting Information (SI).

2.4. Excited State Potential Energy Surfaces. The B3LYP ground state structures were starting geometries for calculations at the TDDFT level. The MP2 structures were starting geometries for calculations at the CC2 and CC3 levels. At the CC2 level,[44] the RI approximation is employed to speed up the calculations.[45] CC3 is an iterative approximation to the coupled cluster singles, doubles, and triples (CCSDT) model.[47] The triples equation is approximated according to two criteria: (i) the triples equation is restricted to contain only terms that enter to second order in the fluctuation potential; (ii) the single excitations are treated as zeroth order parameters in the fluctuation potential. Keeping all terms that enter to second order in the fluctuation potential leads to an energy that is correct through fourth order.[47] An error analysis of the CC3 excitation energies has shown that the excitation energies of single-replacement dominated states are correct through third order.[44,70,71] This places CC3 between the CCSDT model and the coupled cluster singles and doubles (CCSD) and CC2 models. The excitation energies of double-replacement dominated states are correct in CC3 through second order,[70,71] which is the same as in the case of CCSDT.

The PESs were examined by twisting the ground state structure around all carbon−carbon (carbon−nitrogen) bond torsion coordinates. Energies of the ground and first excited state are obtained in single point calculations at intervals of 15° for each torsion angle.

2.5. Programs and Nomenclature. The CC3 and CAM-B3LYP calculations were carried out using a development version of the Dalton program package.[72] All other calculations were done with the Turbomole package, versions V5−9 and V5−10.[73] Difference density plots were produced with the gOpenMol package.[74,75]

In the following, we denote the polyenes by POLx and the protonated Schiff bases by PSBx, where x is the number of heavy atoms. The systematic chemical names and a figure with the corresponding structures are given in Section III of the SI. All PESs discussed in this work refer to the first excited states of either ^1B_u or $^1A'$ symmetry. We denote bond torsion curves repulsive, when the energy of the planar structure is lower than

Table 1. Basis Set Convergence of the 1^1B_u and 2^1A_g Excitation Energies (EE in eV) for POL4 (*trans*-1,3-Butadiene) Calculated at the CC2 Level using Dunning's Correlation-Consistent Basis Sets and the Karlsruhe Basis Sets[b]

basis set	EE(1^1B_u)	ΔEE(1^1B_u)	EE($2\,^1A_g$)	ΔEE($2\,^1A_g$)
cc-pVDZ	6.664		7.952	
cc-pVTZ	6.439		7.685	
cc-pVQZ	6.349		7.536	
cc-pV5Z	6.265		7.372	
cc-pV6Z	6.222		7.272	
aug-cc-pVDZ	6.165	−0.499	7.062	−0.890
aug-cc-pVTZ	6.156	−0.283	7.077	−0.608
aug-cc-pVQZ[a]	6.158	−0.191	7.081	−0.455
aug-cc-pV5Z	6.158	−0.108	7.074	−0.297
aug-cc-pV6Z	6.156	−0.065	7.071	−0.201
SV(P)	6.780		8.041	
SVP	6.743		7.996	
TZVP	6.508		7.662	
TZVPP	6.369		7.570	
QZVPP	6.273		7.389	
aug-TZVP	6.171	−0.337	7.054	
aug-TZVPP	6.157	−0.212	7.081	

[a] The excitation energies obtained with all electrons correlated are 6.162 and 7.091 eV, for $1\,^1B_u$ and $2\,^1A_g$, respectively. [b] The $1s_C$ orbitals are uncorrelated. The contributions from the diffuse basis functions (ΔEE in eV) are also given.

the energy of the perpendicular structure, and attractive, when the energy of the planar structure is higher than the energy of the perpendicular structure. The curvature of the PES at the planar orientation can either be convex yielding higher energies for slightly twisted structures or concave yielding lower energies for slightly twisted structures.

3. Basis Set Effects

3.1. Excitation Energies. The basis set requirements are tested by performing CC2 and B3LYP TDDFT calculations on the 1^1B_u and 2^1A_g states of POL4 and on the $2^1A'$ and $1^1A''$ states of PSB4. The CC2 excitation energies are given in Tables 1 and 2, and the B3LYP excitation energies are given in Tables 3 and 4.

For POL4, diffuse basis functions are necessary to reach the basis set limit at the CC2 and B3LYP level; they also speed up the basis set convergence. The CC2/aug-cc-pVTZ excitation energies deviate only a few meV from the CC2 limit obtained in the CC2/aug-cc-pV6Z calculation. The contributions from diffuse functions to the CC2/aug-cc-pV6Z excitation energies are 0.06 and 0.2 eV.

For PSB4, diffuse basis functions are less important. This is expected as the molecular orbitals in the cationic PSBs are more strongly bound than in the neutral polyenes. The use of diffuse functions gives almost the same excitation energies as basis sets of the next Cardinal number. The cc-pVTZ excitation energies agree within 0.1 eV with the excitation energies calculated at the CC2/cc-pV6Z levels.

The cc-pVTZ basis set is the most cost efficient cc basis set for the PSBs, whereas already the aug-cc-pVDZ basis set seems to be appropriate for the polyenes. Calculations using the augmented Karlsruhe basis sets yield excitation energies of comparable accuracy as the corresponding Dunning basis sets.

Table 2. Basis Set Convergence of the $2^1A'$ and $1^1A''$ Excitation Energies (EE in eV) for PSB4 (the *trans*-2-Propeniminium Cation) Calculated at the CC2 Level using Dunning's Correlation-Consistent Basis Sets and Karlsruhe Basis Sets[b]

basis set	EE($2^1A'$)	ΔEE($2^1A'$)	EE($1^1A''$)	ΔEE($1^1A''$)
cc-pVDZ	5.916		7.388	
cc-pVTZ	5.780		7.241	
cc-pVQZ	5.736		7.201	
cc-pV5Z	5.712		7.185	
cc-pV6Z	5.705		7.179	
aug-cc-pVDZ	5.760	−0.155	7.278	−0.110
aug-cc-pVTZ	5.714	−0.066	7.194	−0.046
aug-cc-pVQZ[a]	5.706	−0.030	7.182	−0.019
aug-cc-pV5Z	5.703	−0.009	7.178	−0.007
SV(P)	5.950		7.370	
SVP	5.911		7.396	
TZVP	5.787		7.355	
TZVPP	5.756		7.225	
QZVPP	5.718		7.193	
aug-TZVP	5.729	−0.058	7.228	
aug-TZVPP	5.714	−0.042	7.192	

[a] The excitation energies obtained with all electrons correlated are 5.698 and 7.165 eV for $2\,^1A'$ and $1\,^1A''$, respectively. [b] The $1s_C$ and $1s_N$ orbitals are uncorrelated. The contributions from the diffuse basis functions (ΔEE in eV) are also given.

Table 3. Basis Set Convergence of the 1^1B_u and 2^1A_g Excitation Energies (EE in eV) for POL4 (*trans*-1,3-Butadiene) Calculated at the B3LYP TDDFT Level using Dunning's Correlation-Consistent Basis Sets[a]

basis set	EE(1^1B_u)	ΔEE(1^1B_u)	EE(2^1A_g)	ΔEE(2^1A_g)
cc-pVDZ	5.996		7.162	
cc-pVTZ	5.857		7.002	
cc-pVQZ	5.788		6.891	
aug-cc-pVDZ	5.626	−0.370	6.561	−0.601
aug-cc-pVTZ	5.613	−0.244	6.543	−0.459
aug-cc-pVQZ	5.607	−0.181	6.528	−0.363

[a] The contributions from the diffuse basis functions (ΔEE in eV) are also given.

Table 4. Basis Set Convergence of the $2^1A'$ and $1^1A''$ Excitation Energies (EE in eV) for PSB4 (the *trans*-2-Propeniminium Cation) Calculated at the B3LYP TDDFT Level using Dunning's Correlation-Consistent Basis Sets[a]

basis set	EE($2^1A'$)	ΔEE($2^1A'$)	EE($1^1A''$)	ΔEE($1^1A''$)
cc-pVDZ	5.796		6.385	
cc-pVTZ	5.726		6.355	
cc-pVQZ	5.699		6.344	
aug-cc-pVDZ	5.677	−0.119	6.329	−0.056
aug-cc-pVTZ	5.673	−0.053	6.334	−0.021
aug-cc-pVQZ	5.673	−0.027	6.334	−0.010

[a] The contribution from the diffuse basis functions (ΔEE in eV) are also given.

The TDDFT excitation energies converge somewhat faster toward the basis set limit compared to the CC2 ones, though the difference is small. With diffuse basis functions, the CC2 and TDDFT calculations exhibit a similar basis set convergence. Other uncertainties such as the errors of the numerical integration might become significant, as the differences between the excitation energies using basis sets of different Cardinal numbers are very small. A systematic study of the basis set requirements for calculation of the electronic

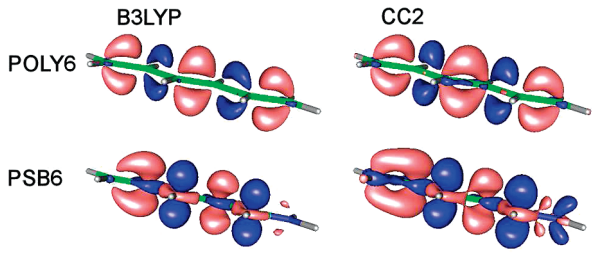


Figure 1. Density differences between the ground states of POL6 and PSB6, compared to the first excited states of ^1B_u and $^1A'$ symmetry, calculated for the planar conformations. The calculations were performed at both B3LYP and CC2 levels, using the aug-cc-pVTZ basis set. Light red color represents regions of electron loss upon excitation; dark blue represents regions of electron gain. An isocontour value of 0.003 e has been used throughout.

excitation spectra of POL4 and PSB4 will be published separately.[76]

Molecular structure effects on the excitation energies can be as large as 0.1 eV. For example, the CC2/aug-cc-pVTZ calculation on PSB6 using a BP86/TZVP structure yields the two lowest excitation energies of 5.613 and 7.052 eV, with the MP2/TZVPP structure one gets excitation energies of 5.714 and 7.194 eV. Using the more accurate MP2/TZVPP structure leads to slightly larger excitation energies than at the BP86/TZVP level. At the BP86 level, the double bonds are about 1 pm longer than the MP2 distances, whereas the single bonds obtained at the two levels are practically equal. Thus, the use of accurate molecular structures is important when aiming at very accurate vertical excitation energies.

3.2. Potential Energy Surfaces. In contrast to the excitation energies, the PESs of the first excited state of POL6 and PSB6 show rather similar basis set dependencies. Cancellation of errors seems to make diffuse basis functions less important for the PESs than for the excitation energies. The basis set convergence is faster at both the CC2 and B3LYP levels. Graphs of the basis set studies at the CC2 and B3LYP levels are given in Section IV of the SI.

Torsion barriers are practically independent of the basis set size at both computational levels. Twisted structures have a somewhat larger basis set dependence than the planar ones. The basis set dependence at the TDDFT level is somewhat smaller than at the CC2 level.

CC3 calculations on POL4 and PSB4 using Dunning's cc-pVDZ, aug-cc-pVDZ, and cc-pVTZ basis sets show a very similar basis set dependence as obtained at the CC2 level. Thus, the PESs calculated at the CC3/cc-pVDZ level can be used for benchmarking results obtained with computational levels that consider electron correlation less accurately. Graphs displaying the basis set dependence of the PESs at the CC3 level can be found in the Section V of the SI.

The methods employed in this study have difficulties in describing molecules with practically perpendicular bond orientations, where the single-configuration dominance breaks down. It should therefore be noted that the energies of the strongly distorted structures are unreliable and included for completeness. Tables with the minimum and maximum $\mathcal{T}_1$ diagnostic values[77] for the bond twists of all polyenes and PSBs can be found in Section VI of the SI.

3.3. Ground and Excited State Optimization. The first excited state of polyenes and PSBs is optimized at the B3LYP and CC2 levels. To avoid bond twists, the excited state structures are assumed to belong to the C_s point group. The structure optimizations of the excited states have been performed using the Karlsruhe SV(P), SVP, TZVP, TZVPP, and aug-TZVPP basis sets; POL12 and PSB12 were not optimized at the aug-TZVPP level. A table containing the minimum and maximum carbon−carbon bond distances as well as the C═N bond lengths is given in Section VII of the SI.

At the B3LYP and MP2 levels, the bond lengths of the ground state structure change by less than 0.2 and 0.3 pm, respectively, when augmenting the TZVP basis set to TZVPP. A similar comparison of the structures obtained using the SVP and TZVP basis sets yields bond length differences that are less than 0.9 and 0.8 pm at the B3LYP and MP2 levels, respectively.

For the first excited state, the structure optimizations at the B3LYP and CC2 levels also yielded very small changes in the bond distances of 0.6 and 0.7 pm, respectively, when augmenting TZVP to TZVPP. The corresponding changes in the bond distances are 0.9 and 1.1 pm when increasing the basis set from SVP to TZVP. For all levels, the bond lengths changed by less than 1 pm when augmenting the TZVPP basis set to aug-TZVPP. Thus, the molecular structures obtained using the TZVPP basis sets are accurate enough for comparisons of the structures obtained with B3LYP or CC2.

The bond lengths obtained in the ground state optimizations at the B3LYP DFT and MP2 levels agree within 1 pm. For the excited state structures, the bond lengths obtained at the B3LYP and CC2 levels agree within 3 pm.

3.4. Bond Length Alternation. Since the pattern for the bond length alternation and the changes in bond lengths upon excitation are similar at the CC2 and DFT levels, the discussion below is valid for both levels. As noted previously for ground state structures of polyenes, the bond length alternation is somewhat sensitive to the amount of exact exchange in the functional.[78]

For the polyenes, the bond lengths change significantly upon excitation. For POL4, the bond length alternation is inverted in the excited state. That is, single bonds become shorter and double bonds longer. For POL6, the bond lengths in the excited state are shorter at the ends of the polyene chain and increase toward the middle. For POL8, POL10, and POL12, the bonds at the ends of the polyenes are shortest, whereas all the other carbon−carbon bonds have almost equal lengths of 140 ± 1 pm. Similar polyene structures were obtained in a recent combined DFT/multi-reference configuration interaction study.[79]

PSBs behave differently. The shortest bond is the C═N bond, whose length is little affected upon excitation. For all PSBs of the present study, the bond length alternation of the ground state is preserved after excitation. The formal single bonds of the ground state remain longer than the

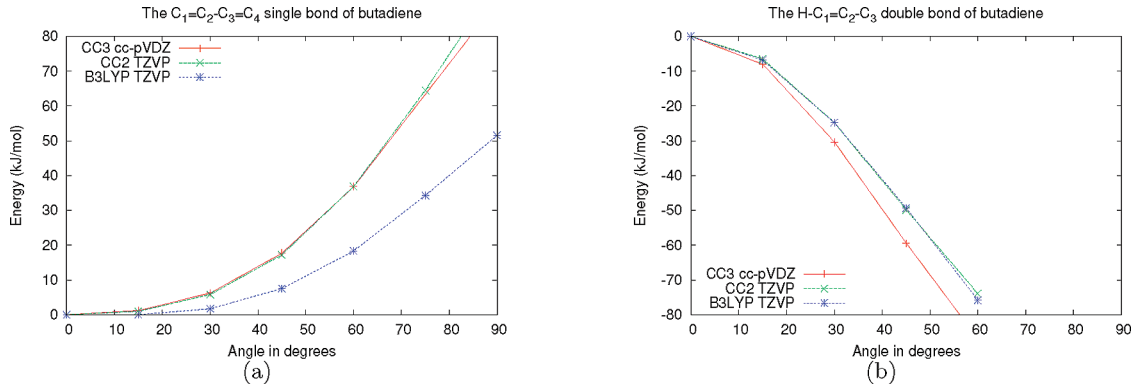


Figure 2. A comparison of potential energy surfaces for torsions of the first excited state of POL4 (*trans*-1,3-butadiene) calculated at different levels of theory. The zero angle corresponds to the planar orientation. (a) The C−C single bond. (b) The C═C double bonds.

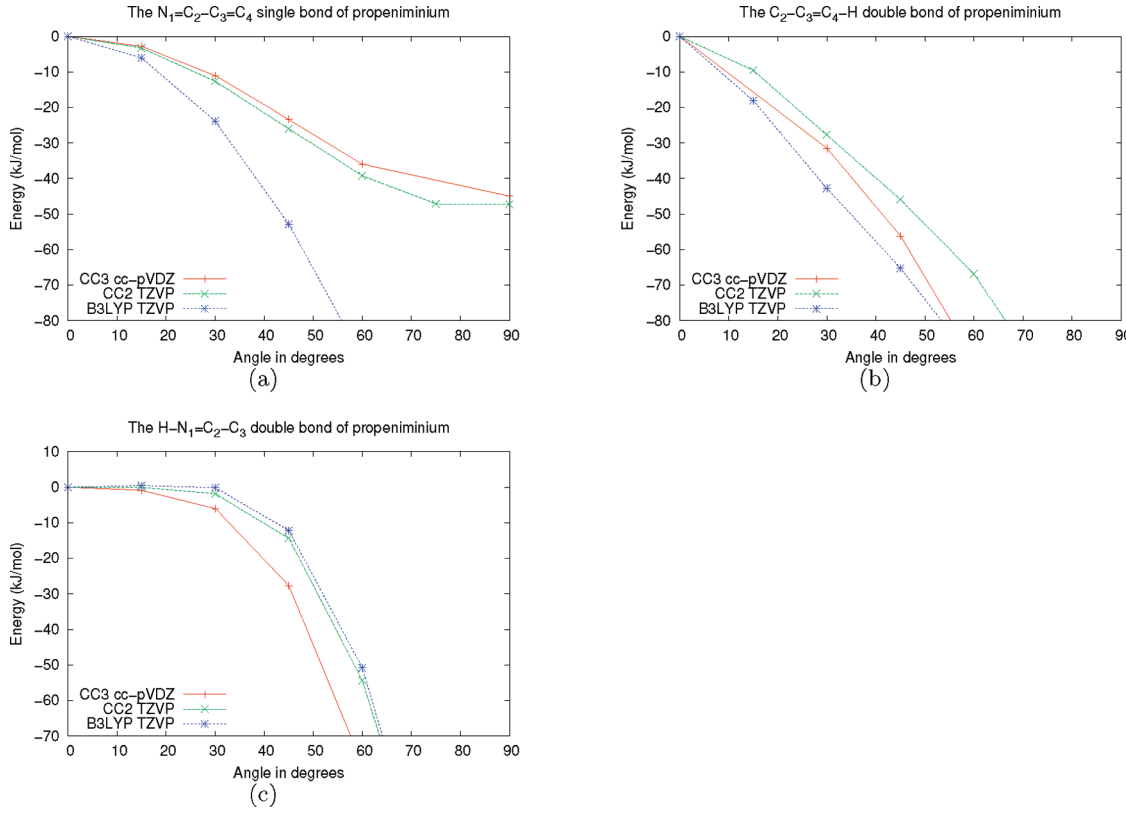


Figure 3. A comparison of the potential energy curves for torsion twists of the carbon−carbon (carbon−nitrogen) bonds for the first excited state of PSB4 (the all-*trans*-2-propeniminium cation) calculated at different levels of theory. The zero angle corresponds to the planar orientation. (a) The C−C single bond. (b) The C═C double bond. (c) The C═N double bond.

formal double bonds. All single bonds extend upon excitation; the C═C double bond at the end shrinks. The other C═C bond lengths remain constant or increase slightly. Compared to the ground state, the bond length alternation is generally enhanced, the effect is more pronounced at the B3LYP level than in the CC2 optimization (see tables in Section VII of the SI).

The different behavior of the polyenes and PSBs can also be seen in the density difference between the ground and excited states for the two classes. As an example, Figure 1 shows the density difference for the untwisted POL6 and PSB6 molecules, for the first excited states of ^1B_u and $^1A'$, respectively. For the polyene, the charge is mainly shifted from the space between atoms, the bonding region, consistent with bond length changes. For the PSB, on the other hand, the charge is mainly shifted from *p*-type orbitals surrounding one atom to another. The exception is the C═C bond most distant from the NH_2^+ group. Here, the charge density is shifted away from the bond region, similarly to what is observed for all C═C bonds in the polyene. With smaller density changes in the bonding region, smaller bond length variations are expected and observed. For PSB6, the electron density in the molecular plane also changes upon excitation, whereas for POL6 only the π density is affected.

For the *cis*-3-pentadieniminium cation, Page and Olivucci found two excited state structures at the CASSCF level, one with inverted and one with enhanced bond length alternation.

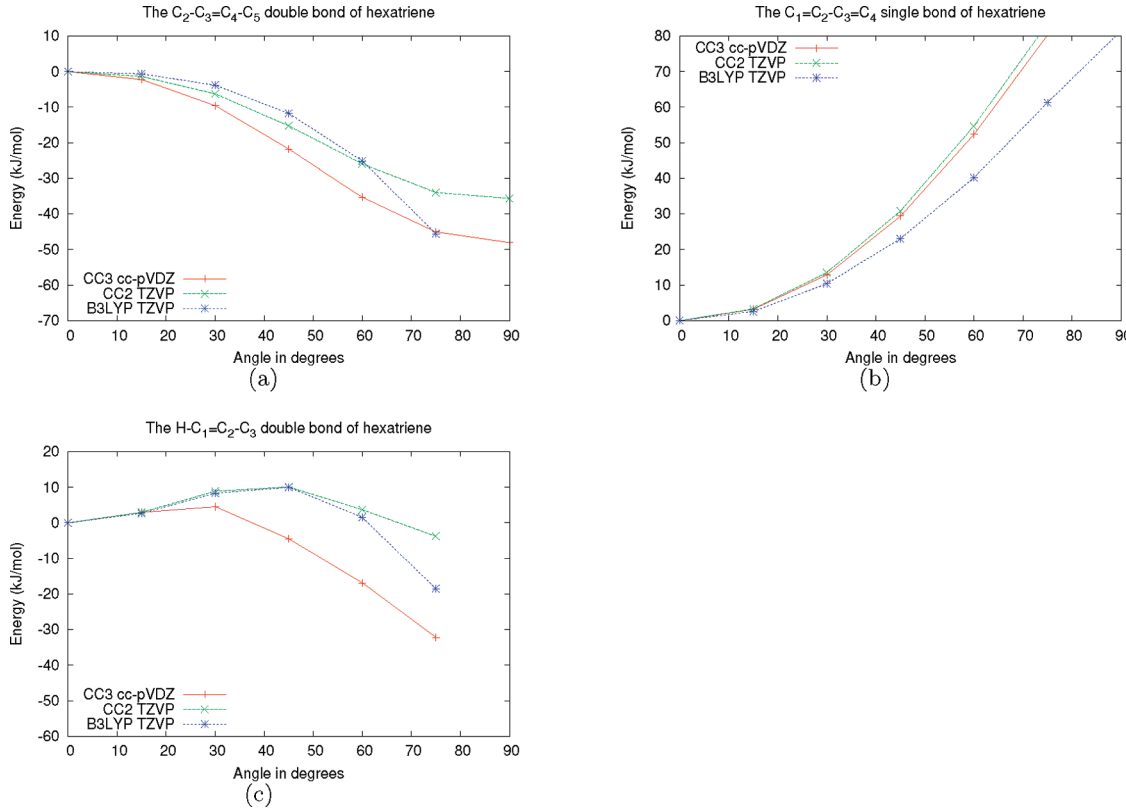


Figure 4. A comparison of the potential energy curves for torsion twists of the carbon−carbon bonds for the first excited state of POL6 (all-*trans*-1,3,5-hexatriene) calculated at different levels of theory. The zero angle corresponds to the planar orientation. (a) The C═C double bond in the middle. (b) The C−C single bonds. (c) The C═C double bonds at both ends.

The CASPT2 optimizations yielded only one minimum with inverted bond length alternation.[27] The present excited state optimizations at the CC2 and B3LYP levels have therefore been repeated using start structures with strongly inverted bond length alternations. These optimizations converged to the same structures as those obtained when starting from the optimized ground state structures. The bond length pattern obtained for excited states of PSBs has been discussed in previous articles and is not addressed in more detail here.[14,80]

4. Effects of Functionals

The PESs of POL6 and PSB6 calculated using different exchange-correlation functionals agree qualitatively. The graphs are given in Section VIII of the SI. The choice of the functional is relevant for torsion barriers when the PES is flat, that is, when the energy remains close to that of the planar structure even for large torsion angles. When the PESs obtained with the various functionals significantly differ, the PESs of the hybrid functionals lie between those of the CAM-B3LYP functional and those of the GGA functionals. Small torsion barriers of less than 5 kJ/mol at the CAM-B3LYP level disappear in the GGA calculations. The PES curve at the C-end of PSB6 calculated at the GGA level has a minimum at a twist of 30°. The potential well vanishes when hybrid functionals are used.

For large torsion angles, the PESs can differ by several tens of kJ/mol. However, strongly twisted structures usually have significant multireference character implying that they are not well described with single-reference methods such as contemporary DFT, which at most can treat mild multireference cases.[81−85] The energies calculated for twisted structures at the TDDFT level therefore become unreliable with increasing torsion angles. In conclusion, the choice of the functional is important when the height of torsion barriers is discussed or PESs remain rather flat.

5. Comparison of B3LYP with CC2 and CC3

5.1. General Trends. For the polyenes, the CC2, CC3, and B3LYP results agree well. The PESs for twists around single bonds are convex and repulsive. For the double-bond twists, they are attractive for POL4 and POL6 and repulsive for POL8, POL10, and POL12. We note that the 1^1B_u state, studied here, is not anymore the lowest excited state for the longer polyenes.

For the PSBs, the situation is much more complicated. No general trends for the PESs are obvious. The PESs have in some cases local minima at a twist of about 30°, and sometimes they have barriers at torsion angles of about 45°. The CC2, CC3, and TDDFT results are in qualitative agreement for most of the bonds. Exceptions are single-bond twists of PSB6, PSB8, PSB10, and PSB12, mainly where the PES is essentially flat at the CC2 level. The agreement between the TDDFT level and the two coupled cluster (CC) levels is better for double-bond twists than for single-bond twists. For small deviations from the planar structure, the PESs of the B3LYP calculations are in very close agreement with those obtained at the CC levels. The comparison of CC2 and CC3 results shows that triple excitations are more

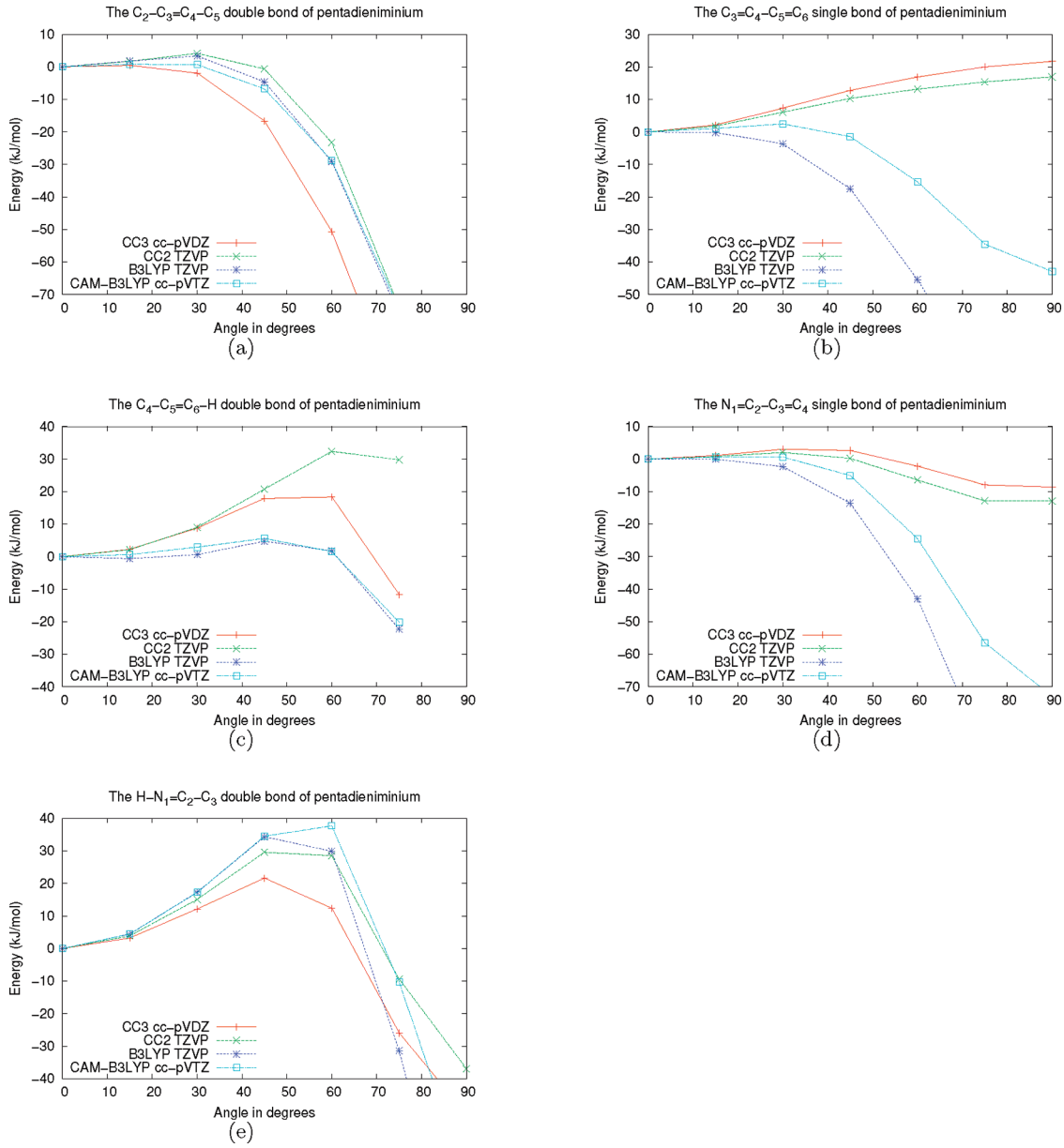


Figure 5. A comparison of the potential energy curves for torsion twists of the carbon−carbon (carbon−nitrogen) bonds for the first excited state of PSB6 (the all-*trans*-2,4-pentadieniminium cation) calculated at different levels of theory. The zero angle corresponds to the planar orientation. (a) The C=C double bond in the middle. (b) The C−C bond closer to the carbon end. (c) The C=C bond at the carbon end. (d) The C−C bond closer to the nitrogen end. (e) The C=N double bond.

important for double-bond twists than for single-bond twists, as expected.

To investigate the effect of the Franck−Condon relaxation on the torsion barriers, the molecular structures of the excited states are optimized, assuming that the molecules belong to the C_s point group. The calculations of the PESs for the bond twists yield similar curves as obtained when the ground state structures are used as starting geometry. The studied molecules have large torsion barriers for all bonds in the ground state. The ground state PESs and the excited state PESs for POL6 and PSB6 are given in Section VIII of the SI.

In the following, we discuss only the molecules of chain length 4, 6, and 8 in detail. The data for molecules of chain length 10 and 12 are given in Section X of the SI.

5.2. POL4 and PSB4. The PESs for the single- and double-bond twists of the first excited state of POL4 and PSB4 are shown in Figures 2 and 3. The PES for the single-bond twist of POL4 is convex and repulsive, whereas for the double-bond twist it is concave and attractive. The PESs for the torsion twists of all the PSB4 bonds are attractive and concave. PSB4 has no torsion barriers for bond twists in the first excited state. The calculations at the B3LYP, CC2, and CC3 levels yield qualitatively similar PESs. The largest differences are obtained for the single bonds. The torsion barrier for the single-bond twist of POL4 is lowest at the B3LYP level. For PSB4, the PES of the single-bond twist falls steeper at the B3LYP level than at the CC level.

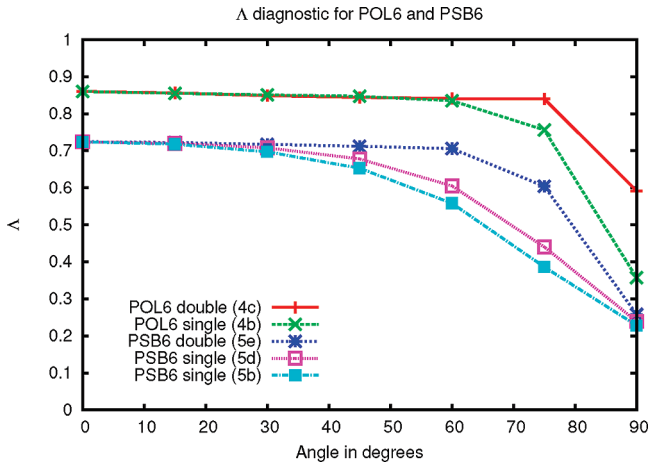


Figure 6. The Λ-diagnostic values for twists around selected single and double bonds of POL6 and PSB6, calculated at the B3LYP/TZVP level. The corresponding figures for the potential energy surfaces are indicated within parentheses.

5.3. POL6 and PSB6. The PESs for the single- and double-bond twists of POL6 and PSB6 are shown in Figures 4 and 5. For POL6, the double-bond twists are attractive. The PES of the double bond at the polyene end is convex yielding a small barrier of 5−10 kJ/mol depending on the computational level.

For the PSB6 double-bond twists, the PESs are attractive and convex except for the double bond in the middle of the molecule at the CC3 level. The small barrier of less than 5 kJ/mol for the mid C═C bond, obtained at the B3LYP and CC2 levels, is probably due to missing higher order correlation effects. The two other double bonds have barrier maxima at around 60°. The torsion barrier for the C═C bond at the end is larger at the CC levels than at the B3LYP level. The lowest C═N torsion barrier of about 20 kJ/mol is obtained at the CC3 level. For the double bonds, the B3LYP and CAM-B3LYP calculations yielded very similar PESs, indicating that the charge transfer problem does not significantly affect these PESs. The charge density differences upon excitation confirm this view, see Figure 1.

For the PSB6 single-bond twists, the PESs obtained at the CC and B3LYP levels do not completely agree. For the single bond at the N-end, all methods give attractive PESs. The CC and CAM-B3LYP calculations yield convex PESs with a tiny barrier of 0.7−3.0 kJ/mol, whereas the B3LYP calculations yield a concave PES. The coupled cluster PES is flat with the energy of the perpendicular structure less than 15 kJ/mol below that of the planar one. The PESs calculated at the TDDFT level are attractive, barrierless, and steep. For the single bond at the C-end, the CC curves are convex, essentially flat, but repulsive. The energy for the perpendicular structure is about 20 kJ/mol above that for the planar one. Opposed to this, the TDDFT curves are steep and attractive. The PES is barrierless at the B3LYP level but has a 2.5 kJ/mol barrier at the CAM-B3LYP level. For the longer PSBs, similar discrepancies are obtained between the PESs at the CC2 and B3LYP levels.

The CAM-B3LYP functional has been introduced to improve the performance of the TDDFT approach for charge transfer excitations. For the single-bond twists where the PESs calculated at the B3LYP level do not reproduce the CC ones, the CAM-B3LYP calculations yield PESs lying between the B3LYP and the CC ones. However, the CAM-B3LYP curves agree better with the B3LYP ones than with the CC curves. This indicates that nonlocal charge transfer might contribute but does not explain the difficulties of the TDDFT method to provide accurate PESs for single-bond twists of the PSBs. This is also seen in the comparison of the PESs calculated with various functionals given in Section VIII of the SI.

The PESs show that the TDDFT problem to describe the energy surfaces of the single-bond twists can in this case not be reduced only to the charge transfer issue. This is further confirmed by the Λ-diagnostic recently proposed by Peach et al.[86] Λ is a measure of the orbital overlap between the occupied and virtual orbitals involved in the excitation and is between zero and one; a higher value corresponds to a more local excitation. Excitations with a small Λ value were proposed to be badly described by standard GGAs and hybrid functionals.[86]

Figure 6 shows how Λ varies with torsion angle for selected double and single bonds for POL6 and PSB6. For the double bonds as well as the POL6 single bond Λ is quite high, even for angles approaching 90 degrees. For these twists, the TDDFT methods also compare well with the coupled cluster results. The troublesome single bond twists of PSB6 behave differently. With increasing twist angle, Λ falls off sooner than for the other bonds. A correlation between increasing error and a decreasing Λ value can be noted. The deviation between DFT and CC kicks in much before the diagnostic falls below the suggested critical value of $\Lambda < 0.4$,[86] however. Thus, the Λ diagnostic indicates no severe charge-transfer issue.

5.4. POL8 and PSB8. The PESs for POL8 and PSB8 calculated at the B3LYP and CC2 levels are shown in Figure 7. For POL8, the PESs for all double-bond twists are convex with the barrier maximum at 45−60°. For large angles, significant differences between the CC2 and B3LYP curves appear. However, the multiconfiguration character of the wave function is large for structures with perpendicular bond orientation, again implying that the PESs are unreliable at large torsion angles. The PESs for the double-bond twists of PSB8 calculated at the CC2 and B3LYP levels agree better than for POL8. The smallest double-bond torsion barrier for PSB8 of 20 kJ/mol is obtained for the C═C double bond closer to the C═N end. The corresponding C═C double bond for POL8 has a torsion barrier of 10 kJ/mol. The torsion barriers for the other double bonds are also higher for PSB8 than for POL8.

For the single-bond twists of POL8, the PESs calculated at the CC2 and B3LYP levels agree well, whereas significant differences are obtained for PSB8. For the single bond at the N-end and in the middle, the CC2 curves are convex and repulsive but flat. For these two single bonds, the B3LYP calculations yield convex and attractive PESs. The torsion barrier for the single bond at the N-end

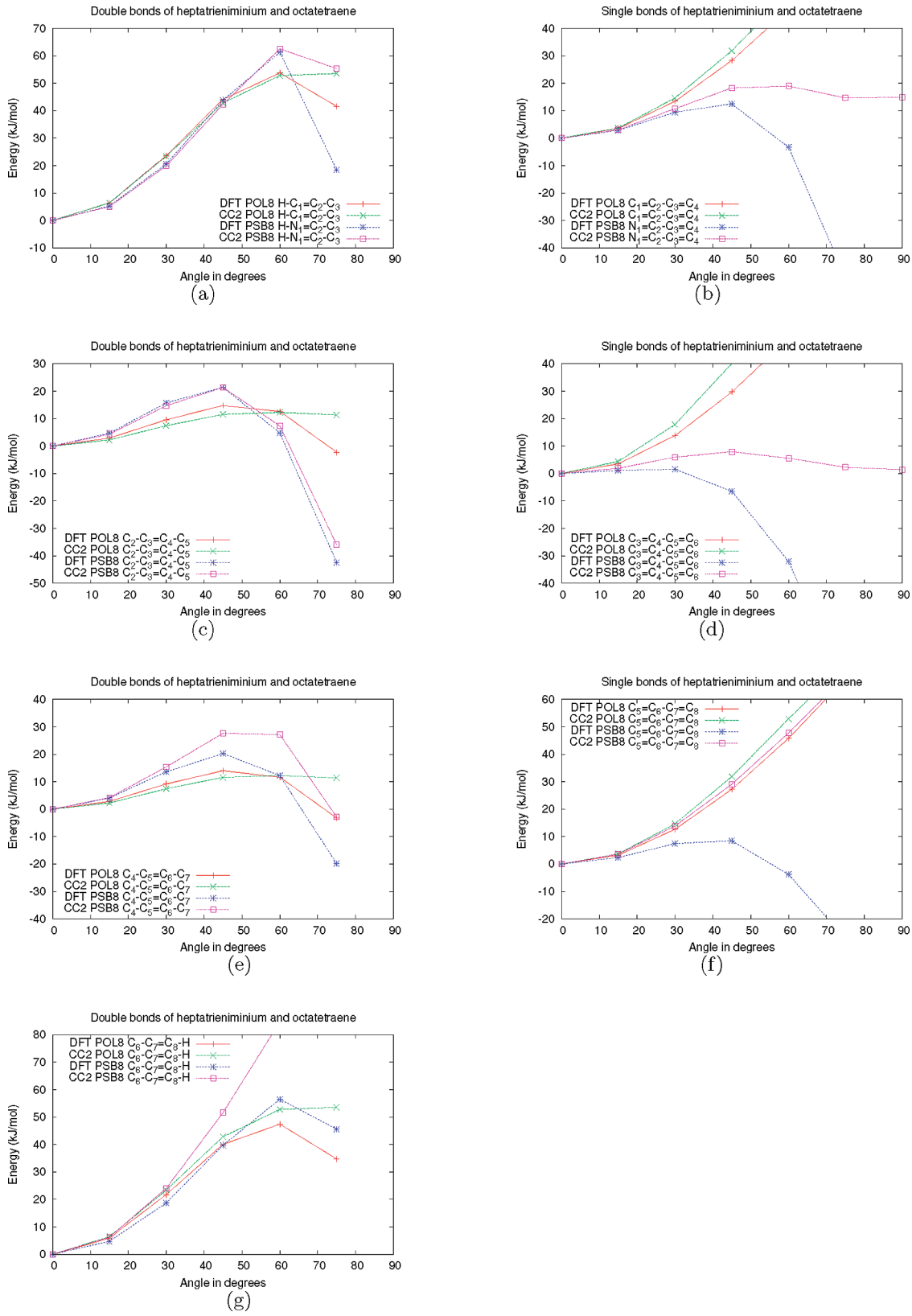


Figure 7. A comparison of the torsion barriers of the single and double bonds of POL8 (all-*trans*-1,3,5,7-octatetraene) and PSB8 (the all-*trans*-2,4,6-heptatrieniminium cation) calculated at the B3LYP TDDFT and CC2 levels. (a) The $C_1{=}C_2$ ($N_1{=}C_2$) bond, (b) $C_2{-}C_3$, (c) $C_3{=}C_4$, (d) $C_4{-}C_5$, (e) $C_5{=}C_6$, (f) $C_6{-}C_7$, and (g) the $C_7{=}C_8$ bond.

is less than 15 kJ/mol and almost vanishes for the single bond in the middle. The CC2 curve for the single bond at the C-end of PSB8 is convex and strongly repulsive, whereas the B3LYP curve is convex and strongly attractive leading to a torsion barrier of about 10 kJ/mol. In this case, the B3LYP calculations fail to reproduce the main trends of the PESs calculated at the CC2 level, in contrast to the two other single bonds.

The PESs of POL10, PSB10, POL12, and PSB12 are similar to those of their shorter homologues, POL8 and PSB8. The agreement of the CC2 and B3LYP levels is generally better than for POL8 and PSB8. The PESs for all bond torsions can be found in Section X of the SI.

6. Conclusion and Summary

Systematic studies of the potential energy surfaces (PESs) for the torsion twists of the first excited state of polyenes and PSBs, with chain lengths of 4−12 heavy atoms, revealed a remarkably different behavior of the two compound groups. While the excited states of the polyenes exhibit high torsion barriers for single-bond twists, the PSBs have low or vanishing torsion barriers for both single- and double-bond twists. PSBs are very flexible molecules in the first excited state, and most of the bonds are likely to contribute to the excited state dynamics, whereas in polyenes the dynamics are limited to motions around double bonds. The flexibility of PSBs in the excited state and the thermal ground state flexibility of 11-*cis*-retinal[87] are typical for the complexity of the retinal photoisomerisation.

Benchmark calculations of the excitation energies show that diffuse functions must be included in the basis set to reach the basis set limit of the polyenes, whereas diffuse functions are less important for the lowest states of the cationic PSBs. For an accurate description of the potential energy surfaces of the polyenes and PSBs, diffuse basis functions are not mandatory, due to cancellation of errors in the excitation energies.

Exchange-correlation functional studies at the TDDFT level employing GGA, hybrid, and Coulomb-attenuated functionals yield similar PESs. The PESs obtained with hybrid functionals lie in between those calculated at the GGA values on the one side and the CAM-B3LYP values on the other side. The influence of the functional is important when the potential barriers or wells are very flat.

Comparisons of the calculated excitation energies and PESs for the polyenes and the corresponding PSBs indicate that CC2 calculations are useful for retinal studies. The obtained PESs are in close agreement with those calculated at the CC3 level, where triple excitations are taken into account. The single-bond torsion barriers obtained at the CC2 and CC3 levels are in excellent agreement, whereas for the double-bond twists, CC2 gives somewhat higher barriers than CC3. The differences between the PESs obtained at the two CC levels are, however, small.

Comparison of the DFT and CC calculations shows that the PESs agree well for double-bond twists, whereas significant differences are found for the single-bond torsions of the longer PSBs. The low or vanishing barriers for torsions around the single bond farthest away from the C=N moiety are apparently a flaw of the TDDFT method. The TDDFT results are only slightly improved by using the CAM-B3LYP functional in those cases where calculations at the B3LYP level fail to reproduce the CC2 results. The TDDFT problems are not merely due to the inability to describe long-range charge transfer effects. TDDFT performs well for the polyenes and for PSB double-bond twists. In TDDFT studies on the 11-*cis* retinal PSB, the largest difference between the PESs obtained at the TDDFT and CC2 level appears at the bond connecting the retinyl chain with the β-ionone ring. In the first excited state, TDDFT optimizations yield a perpendicular orientation of the β-ionone-ring plane with respect to the retinyl-chain plain.[38] Single-bond twists of longer PSBs show that the TDDFT results are satisfactory for small torsion angles. We conclude that even though TDDFT might suffer from long-range charge transfer problems and that it apparently provides incorrect PESs at large torsion angles, one should not exclude it from the toolbox of retinal studies. However, it is necessary to confirm excited state TDDFT studies on retinal PSBs by performing calculations at *ab initio* correlation levels such as CC2.

Acknowledgment. This research has been supported by the Academy of Finland through its Centers of Excellence Programme 2006−2011, the OPNA research project (118195), the Centre for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project C3.9, the Lundbeck Foundation, and the Foundation for Polish Science (FNP) (Homing program grant no. HOM/2008/10B) within the EEA Financial Mechanism. The research collaboration is supported by the Nordic Centre of Excellence in Computational Chemistry (NCoECC) project funded by NordForsk (070253). We also thank CSC - the Finnish IT Center for Science and the Danish Center for Scientific Computing (DCSC) for computer time. We are grateful to Prof. Reinhart Ahlrichs for helpful discussions.

Supporting Information Available: Optimized Cartesian coordinates for ground and excited states (in C_s symmetry) including C−C and C−N distances; structures and systematic chemical names of the studied molecules; basis set studies of the potential energy surfaces at the CC2, CC3, and B3LYP levels; potential energy surfaces studied using different exchange-correlation functionals; potential energy curves for the double-bond twists of PSB10, POL10, PSB12, and POL12; $\mathcal{T}_1$ diagnostic values. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Shichida, Y.; Kobayashi, T.; Ohtani, H.; Yoshizawa, T.; Nagakura, S. Picosecond Laser Photolysis of Squid Rhodopsin at Room and Low Temperatures. *Photochem. Photobiol.* **1978**, *27*, 335–341.

(2) Yoshizawa, T.; Shichida, Y.; Matuoka, S. Primary intermediates of rhodopsin studied by low temperature spectrophotometry and laser photolysis: Bathorhodopsin, hypsorhodopsin and photorhodopsin. *Vision Res.* **1984**, *24*, 1455–1463.

(3) Shichida, Y.; Matuoka, S.; Yoshizawa, T. Formation of photorhodopsin, a precursor of bathorhodopsin, detected by a picosecond laser photolysis at room temperature. *Photobiochem. Photobiophys.* **1984**, *7*, 221–228.

(4) Schoenlein, R. W.; Peteanu, L. A.; Mathies, R. A.; Shank, C. V. The 1st Step in Vision - Femtosecond Isomerization of Rhodopsin. *Science* **1991**, *254*, 412–415.

(5) Kandori, H.; Shichida, Y.; Yoshizawa, T. Photoisomerization in Rhodopsin. *Biochemistry (Moscow)* **2001**, *66*, 1197–1209.

(6) Okada, T.; Ernst, O. P.; Palczewski, K.; Hoffmann, K. P. Activation of rhodopsin: new insights from structural and biochemical studies. *Trends Biochem. Sci.* **2001**, *26*, 318–324.

(7) Schreiber, M.; Sugihara, M.; Okada, T.; Buss, V. Quantum Mechanical Studies on the Crystallographic Model of Bathorhodopsin. *Angew. Chem., Int. Ed.* **2006**, *45*, 4274–4277.

(8) Nakamichi, H.; Okada, T. Crystallographic Analysis of Primary Visual Photochemistry. *Angew. Chem., Int. Ed.* **2006**, *45*, 4270–4273.

(9) Kukura, P.; McCamant, D. W.; Yoon, S.; Wandschneider, D. B.; Mathies, R. A. Structural Observation of the Primary Isomerization in Vision with Femtosecond-Stimulated Raman. *Science* **2005**, *310*, 1006–1009.

(10) Yoshizawa, T.; Wald, G. Pre-Lumirhodopsin and the Bleaching of Visual Pigments. *Nature* **1963**, *197*, 1279–1286.

(11) Vreven, T.; Bernardi, F.; Garavelli, M.; Olivucci, M.; Robb, M. A.; Schlegel, H. B. Ab Initio Photoisomerization Dynamics of a Simple Retinal Chromophore Model. *J. Am. Chem. Soc.* **1997**, *119*, 12687–12688.

(12) Garavelli, M.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. The $C_5H_6NH_2^+$ Protonated Shiff Base: An ab Initio Minimal Model for Retinal Photoisomerization. *J. Am. Chem. Soc.* **1997**, *119*, 6891–6901.

(13) Garavelli, M.; Vreven, T.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. Photoisomerization Path for a Realistic Retinal Chromophore Model: Nonatetraeniminimum Cation. *J. Am. Chem. Soc.* **1998**, *120*, 1285–1288.

(14) Aquino, A. J. A.; Barbatti, M.; Lischka, H. Excited-State Properties and Environmental Effects for Protonated Schiff Bases: A Theoretical Study. *Chem. Phys. Chem.* **2006**, *7*, 2089–2096.

(15) Barbatti, M.; Granucci, G.; Persico, M.; Ruckenbauer, M.; Vazdar, M.; Eckert-Maksić, M.; Lischka, H. The on-the-fly surface-hopping program system Newton-X: Application to ab initio simulation of the nonadiabatic photodynamics of benchmark systems. *J. Photochem. Photobiol. A: Chem.* **2007**, *190*, 228–240.

(16) Szymczak, J. J.; Barbatti, M.; Lischka, H. Mechanism of Ultrafast Photodecay in Restricted Motions in Protonated Schiff Bases: The Pentadieniminium Cation. *J. Chem. Theory Comput.* **2008**, *4*, 1189–1199.

(17) Roos, B. O.; Taylor, P. R. A Complete Active Space SCF Method (CASSCF) using a Density-Matrix Formulated Super-CI Approach. *Chem. Phys.* **1980**, *48*, 157–173.

(18) Andersson, K.; Malmqvist, P. Å.; Roos, B. O. 2nd-order perturbation-theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(19) Olsen, J.; Roos, B. O.; Jørgensen, P.; Jensen, H. J. A. Determinant Based Configuration-Interaction Algorithms for Complete and Restricted Configuration-Interaction Spaces. *J. Chem. Phys.* **1988**, *89*, 2185–2192.

(20) Malmqvist, P. Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. The restricted active space followed by second-order perturbation theory method: Theory and application to the study of CuO_2 and Cu_2O_2 systems. *J. Chem. Phys.* **2008**, *128*, 204109.

(21) Altun, A.; Yokoyama, S.; Morokuma, K. Mechanism of Spectral Tuning Going from Retinal in Vacuo to Bovine Rhodopsin and its Mutants: Multireference ab Initio Quantum Mechanics/Molecular Mechanics Studies. *J. Phys. Chem. B* **2008**, *112*, 16883–16890.

(22) Cembran, A.; González-Luque, R.; Altoè, P.; Merchán, M.; Bernardi, F.; Olivucci, M.; Garavelli, M. Structure, Spectroscopy, and Spectral Tuning of the Gas-Phase Retinal Chromophore: The β-Ionone "Handle" and the Alkyl Group Effect. *J. Phys. Chem. A* **2005**, *109*, 6597–6605.

(23) Cembran, A.; Bernardi, F.; Olivucci, M.; Garavelli, M. The retinal chromophore/chloride ion pair: Structure of the photoisomerization path and interplay of charge transfer and covalent states. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6255–6260.

(24) González-Luque, R.; Garavelli, M.; Bernardi, F.; Merchán, M.; Robb, M. A.; Olivucci, M. Computational evidence in favor of a two-state, two-mode model of the retinal chromophore potoisomerization. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *92*, 9379–9384.

(25) Sekharan, S.; Weingart, O.; Buss, V. Ground and Excited States of Retinal Schiff Base Chromophores by Multiconfigurational Perturbation Theory. *Biophys. J.* **2006**, *91*, L07–L09.

(26) Johansson, M. P.; Olsen, J. Torsional Barriers and Equilibrium Angle of Biphenyl: Reconciling Theory with Experiment. *J. Chem. Theory Comput.* **2008**, *4*, 1460–1471.

(27) Page, C. S.; Olivucci, M. Ground and Excited State CASPT2 Geometry Optimizations of Small Organic Molecules. *J. Comput. Chem.* **2003**, *24*, 298–309.

(28) Gross, E. K. U.; Kohn, W. Time-dependent density functional theory. *Adv. Quantum Chem.* **1990**, *21*, 255–291.

(29) Bauernschmitt, R.; Ahlrichs, R. Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. *Chem. Phys. Lett.* **1996**, *256*, 454–464.

(30) van Leeuwen, R. Key concepts in time-dependent density-functional theory. *Int. J. Mod. Phys. B* **2001**, *15*, 1969–2023.

(31) Furche, F.; Burke, K. Time-dependent density functional theory in quantum chemistry. In *Annual Reports in Computational Chemistry, 1*; Spellmeyer, D., Ed.; Elsevier: Amsterdam, 2005; pp 19−30.

(32) Burke, K.; Werschnick, J.; Gross, E. K. U. Time-dependent density functional theory: Past, present, and future. *J. Chem. Phys.* **2005**, *123*, 062206.

(33) Furche, F.; Ahlrichs, R. Adiabatic time-dependent density functional methods for excited state properties. *J. Chem. Phys.* **2002**, *117*, 7433–7447.

(34) Maitra, N. T.; Zhang, F.; Cave, R. J.; Burke, K. Double excitations within time-dependent density functional theory linear response. *J. Chem. Phys.* **2004**, *120*, 5932–5937.

(35) Cave, R. J.; Zhang, F.; Maitra, N. T.; Burke, K. A dressed TDDFT treatment of the 2^1A_g states of butadiene and hexatriene. *Chem. Phys. Lett.* **2004**, *389*, 39–42.

(36) Mikhailov, I. A.; Tafur, S.; Masunov, A. E. Double excitations and state-to-state transition dipoles in π-π^* excited singlet states of linear polyenes: Time-dependent density-functional theory versus multiconfigurational methods. *Phys. Rev. A* **2008**, *77*, 012510.

(37) Mazur, G.; Włodarczyk, R. Application of the Dressed Time-Dependent Density Functional Theory for the Excited States of Linear Polyenes. *J. Comput. Chem.* **2009**, *30*, 811–817.

(38) Send, R.; Sundholm, D. The role of the β-ionone ring in the photochemical reaction of rhodopsin. *J. Phys. Chem. A* **2007**, *111*, 27–33.

(39) Send, R.; Sundholm, D. Coupled-Cluster Studies of the Lowest Excited States of the 11-cis-Retinal Chromophore. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2862–2867.

(40) Send, R.; Sundholm, D. Stairway to the Conical Intersection: A Computational Study of the Retinal Isomerization. *J. Phys. Chem. A* **2007**, *111*, 8766–8773.

(41) Send, R.; Sundholm, D. The molecular structure of a curl-shaped retinal isomer. *J. Mol. Model.* **2008**, *14*, 717–726.

(42) Wanko, M.; Garavelli, M.; Bernardi, F.; Miehaus, T. A.; Frauenheim, T.; Elstner, M. A global investigation of excited state surfaces within time-dependent density-functional response theory. *J. Chem. Phys.* **2004**, *120*, 1674–1692.

(43) Dreuw, A.; Head-Gordon, M. Single-Reference ab initio Methods for the Calculation of Excited States of Large Molecules. *Chem. Rev.* **2005**, *105*, 4009–4037.

(44) Christiansen, O.; Koch, H.; Jørgensen, P. The 2nd-order Approximate Coupled-Cluster Singles and Doubles Model CC2. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(45) Hättig, C.; Weigend, F. CC2 excitation energy calculations on large molecules using the resolution of the identity approximation. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(46) Nielsen, I. B.; Lammich, L.; Andersen, L. H. S_1 and S_2 excited states of gas-phase Schiff-base retinal chromophores. *Phys. Rev. Lett.* **2006**, *96*, 018304.

(47) Koch, H.; Christiansen, O.; Jørgensen, P.; Sanchez de Merás, A. M.; Helgaker, T. The CC3 model: An iterative coupled cluster approach including connected triples. *J. Chem. Phys.* **1997**, *106*, 1808–1818.

(48) Pescitelli, G.; Sreerama, N.; Salvadori, P.; Nakanishi, K.; Berova, N.; Woody, R. W. Inherent Chirality Dominates the Visible/Near-Ultraviolet CD Spectrum of Rhodopsin. *J. Am. Chem. Soc.* **2008**, *130*, 6170–6181.

(49) Zaari, R. R.; Wong, S. Y. Y. Photoexcitation of 11-Z-cis-7,8-dihydro retinal and 11-Z-cis retinal: A comparative computational study. *Chem. Phys. Lett.* **2009**, *469*, 224–228.

(50) Chmura, B.; Rode, M. F.; Sobolewski, A. L.; Lapinski, L.; Nowak, M. J. A Computational Study on the Mechanism of Intramolecular Oxo-Hydroxy Phototautomeris Driven by Repulsive $\pi\sigma^*$ State. *J. Phys. Chem. A* **2008**, *112*, 13655–13661.

(51) Rode, M. F.; Sobolewski, A. L.; Dedonder, C.; Jouvet, C.; Dopfer, O. Computational Study on the Photophysics of Protonated Benzene. *J. Phys. Chem. A* **2009**, *113*, 5865–5873.

(52) Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(53) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the $H+H_2 \rightarrow H_2+H$ reaction. *J. Chem. Phys.* **1994**, *100*, 7410–7415.

(54) Wilson, A.; van Mourik, T.; Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon. *J. Mol. Struct. (Theochem)* **1997**, *388*, 339–349.

(55) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(56) 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.

(57) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(58) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(59) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(60) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(61) 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.

(62) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(63) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(64) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(65) Perdew, J. P.; Burke, K.; Ernzerhof, M. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(66) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(67) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

(68) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(69) Weigend, F.; Häser, M. RI-MP2: first derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97*, 331–340.

(70) Christiansen, O.; Koch, H.; Jørgensen, P. Perturbative triple excitation corrections to coupled cluster singles and doubles excitation energies. *J. Chem. Phys.* **1996**, *105*, 1451–1459.

(71) Hald, K.; Jorgensen, P.; Olsen, J.; Jaszuñski, M. An analysis and implementation of a general coupled cluster approach to excitation energies with application to the B_2 molecule. *J. Chem. Phys.* **2001**, *115*, 671–679.

(72) Dalton, an ab initio electronic structure program, Release 2.0; 2005. See http://www.kjemi.uio.no/software/dalton/dalton.html (accessed month day, year).

(73) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic Structure Calculations on Workstation Computers: The Program System TURBOMOLE. *Chem. Phys. Lett.* **1989**, *162*, 165–169. Current version. See http://www.turbomole.com (accessed month day, year).

(74) Laaksonen, L. A graphics program for the analysis and display of molecular dynamics trajectories. *J. Mol. Graph.* **1992**, *10*, 33–34.

(75) Bergman, D. L.; Laaksonen, L.; Laaksonen, A. Visualization of solvation structures in liquid mixtures. *J. Mol. Graphics Modell.* **1997**, *15*, 301–306.

(76) Lehtonen, O.; Sundholm, D.; Send, R.; Johansson, M. P. Density Functional Theory and Coupled-Cluster Studies of the Electronic Excitation Spectra of trans-1,3-butadiene and trans-2-propeniminium. *J. Chem. Phys.* **2009**, *131*, 024301.

(77) Lee, T. J.; Taylor, P. R. A Diagnostic for Determining the Quality of Single-Reference Electron Correlation Methods. *Int. J. Quant. Chem. Symp.* **1989**, *23*, 199–207.

(78) Peach, M. J. G.; Tellgren, E. I.; Sałek, P.; Helgaker, T.; Tozer, D. J. Structural and Electronic Properties of Polyacetylene and Polyyne from Hybrid and Coulomb-Attenuated Density Functionals. *J. Phys. Chem. A* **2007**, *111*, 11930–11935.

(79) Marian, C. M.; Gilka, N. Performance of the Density Functional Theory/Multireference Configuration Interaction Method on Electronic Excitation of Extended π-Systems. *J. Chem. Theory Comput.* **2008**, *4*, 1501–1515.

(80) Tavernelli, I.; Röhrig, U. F.; Rothlisberger, U. Molecular dynamics in electronically excited states using time-dependent density functional theory. *Mol. Phys.* **2005**, *103*, 963–981.

(81) Handy, N. C.; Cohen, A. J. Left-right correlation energy. *Mol. Phys.* **2001**, *99*, 403–412.

(82) Polo, V.; Kraka, E.; Cremer, D. Electron correlation and the self-interaction error of density functional theory. *Mol. Phys.* **2002**, *100*, 1771–1790.

(83) Polo, V.; Kraka, E.; Cremer, D. Some thoughts about the stability and reliability of commonly used exchange-correlation functionals - coverage of dynamic and nondynamic correlation effects. *Theor. Chem. Acc.* **2002**, *107*, 291–303.

(84) Cremer, D.; Filatov, M.; Polo, V.; Kraka, E.; Shaik, S. Implicit and Explicit Coverage of Multi-reference Effects by Density Functional Theory. *Int. J. Mol. Sci.* **2002**, *3*, 604–638.

(85) Johansson, M. P.; Sundholm, D. Spin and charge distribution in iron porphyrin models: A coupled cluster and density-functional study. *J. Chem. Phys.* **2004**, *120*, 3229–3236.

(86) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. Excitation energies in density functional theory: An evaluation and a diagnostic test. *J. Chem. Phys.* **2008**, *128*, 044118.

(87) Silva López, C.; Álvarez, R.; Domínguez, M.; Nieto Faza, O.; de Lera, Á. R. Complex Thermal Behavior of 11-cis-Retinal, the Ligand of Visual Pigments. *J. Org. Chem.* **2009**, *74*, 1007–1013.

CT900240S

JCTC Journal of Chemical Theory and Computation

Singlet−Triplet Transitions in Real-Time Time-Dependent Hartree−Fock/Density Functional Theory

Christine M. Isborn and Xiaosong Li*

Department of Chemistry, University of Washington, Seattle, Washington 98195-1700

Received May 23, 2009

Abstract: Real-time time-dependent Hartree−Fock (TDHF)/density functional theory (TDDFT) has been gaining in popularity because of its ability to treat phenomena beyond the linear response and because it has the potential to be more computationally powerful than frequency domain TDHF/TDDFT. Within real-time TDHF/TDDFT, we present a method that gives the excited state triplet energies starting from a singlet ground state. Using a spin-dependent field, we break the spin-symmetry of the α and β density matrices, which incorporates a triplet contribution into the superposition state. The α electron density follows the applied field, and the β electron density responds to the perturbation from the changing α electron density. We examine the individual α/β responses during the electron density propagation. Singlet−triplet transitions appear as 'dark' states: they are present in the α/β responses but are absent from the total electron density response.

Introduction

Investigation of electronic excitations is of fundamental importance in spectroscopy and photochemistry. Not only is accurate understanding of the photoallowed transitions necessary, but because much photochemistry also occurs on electronic surfaces that are not directly photoaccessible, such as triplet states, it is also of great importance to accurately model these 'dark' states. Because of their affordability and generally good accuracy, single configuration excited state methods such as time-dependent Hartree−Fock (TDHF)[1−4] and density functional theory (TDDFT)[5,6] have become the modern-day workhorses for modeling electronic excitations. While the majority of TDHF and TDDFT calculations are performed within the linear response matrix formalism in the frequency domain,[1,2,7,8] real-time methods are gaining in popularity because of their ability to treat phenomena beyond the linear response and because they have the potential to be more computationally efficient for large-scale systems when a large matrix equation must be solved in the frequency domain.[9−18]

While the matrix formalism of linear response TDHF/TDDFT can easily include transitions from a reference ground state singlet to triplets, singlet−triplet transitions are generally not accessible within real-time TDHF/TDDFT approaches because the applied perturbation is often spin-independent, and the α and β electron density responses are identical. With a spin-independent perturbation, such as the ubiquitous electric field within the dipole approximation, both the α and β electron densities are simultaneously excited, maintaining a closed-shell singlet state. The resulting perturbed wave function is then a superposition of the singlet reference ground state with other excited singlet states.[19] However, no contribution from any triplet state is included in the superposition.

In the present investigation, we show that a spin-dependent perturbation within real-time TDHF/TDDFT incorporates triplets into the superposition state wave function. This can be done via a spin-dependent field,[14] in which only the α electron density directly experiences the applied perturbation. Because the α and β electron densities are coupled through the Coulomb and exchange (for HF)/exchange-correlation (for DFT) terms, the β electron density is then perturbed by the change in the α electron density, leading to breaking of the singlet spin-symmetry and incorporation of the triplet wave function. We monitor the coupled α and β electron density response and the time-dependent expectation value of S^2.

Method

In our real-time electron dynamics, we use the method described in refs 12, and 19−22 in which the electron density

* Corresponding author e-mail: li@chem.washington.edu.

10.1021/ct900264b CCC: $40.75
Published on Web 08/12/2009

is propagated in the time domain and the time-dependent Hamiltonian takes into account the evolving electron distribution. We use atomic units consistently throughout this paper, where $m_e = \hbar = 1$. The TDHF/TDDFT equation for the density matrix is given by

$$i\frac{d\mathbf{P}(t)}{dt} = [\mathbf{F}(t), \mathbf{P}(t)] \tag{1}$$

where $\mathbf{P}$ and $\mathbf{F}$ are density and Fock/Kohn−Sham (KS) matrices. The MOs are represented as a linear combination of atomic orbital basis functions χ_μ as $\varphi_i(t) = \Sigma_\mu c_{\mu,i}(t)\chi_\mu$. The density matrix $\mathbf{P}(t)$ has elements given by the product of the time-dependent coefficients $P_{\mu\nu}(t) = \Sigma_i c^\dagger_{\mu,i}(t)c_{\nu,i}(t)$. In the unrestricted TDHF/TDDFT electron dynamics, the α and β electron densities are propagated with coupled α and β Fock/KS operators. With a spin-dependent perturbation, the α and β Fock matrices are

$$\mathbf{F}^\alpha(t) = \mathbf{h} + \mathbf{J}[\boldsymbol{\rho}^\alpha(t) + \boldsymbol{\rho}^\beta(t)] - \mathbf{K}[\boldsymbol{\rho}^\alpha(t)] + V_{ext} \tag{2}$$

$$\mathbf{F}^\beta(t) = \mathbf{h} + \mathbf{J}[\boldsymbol{\rho}^\alpha(t) + \boldsymbol{\rho}^\beta(t)] - \mathbf{K}[\boldsymbol{\rho}^\beta(t)] \tag{3}$$

where $\mathbf{h}$, $\mathbf{J}$, and $\mathbf{K}$ are the core Hamiltonian, two-electron Coulomb, and exchange matrices. For the KS matrices, $\mathbf{K}$ is replaced by the exchange-correlation potential $\mathbf{V}_{xc}$. V_{ext} is an external perturbation applied to α electrons only. In this work we include a spin-dependent perturbation within the Hamiltonian by adding the field term $\mathbf{E}(t) = \mathbf{E}_{max}\mathbf{d}\sin(\omega t)$ to the α Fock/KS matrix $\mathbf{F}^\alpha(t)$, where $\mathbf{d}$ is the electric dipole integral $d_{\mu\nu} = \langle\chi_\mu|\hat{r}|\chi_\nu\rangle$.

The electron density is propagated with a midpoint unitary transformation written in terms of the eigenvectors $\mathbf{C}$ and eigenvalues ε of the time-dependent Fock/KS matrix at time t_k

$$\mathbf{P}(t_{k+1}) = \mathbf{U}(t_k)\mathbf{P}(t_{k-1})\mathbf{U}^\dagger(t_k) \tag{4}$$

$$\begin{aligned}\mathbf{U}(t_k) &= \exp(i\mathbf{F}(t_k)2\Delta t)\\ &= \mathbf{C}(t_k)\exp(i\varepsilon(t_k)2\Delta t)\mathbf{C}^\dagger(t_k)\end{aligned} \tag{5}$$

This unitary transformation propagation naturally retains the idempotency of the density matrix ($\mathbf{P}\cdot\mathbf{P} = \mathbf{P}$).

In real-time TDHF/TDDFT electron dynamics, a perturbation to the ground state orbitals, φ, mixes occupied and virtual molecular orbitals (MOs) to give rise to a set of perturbed orbitals, φ', and a superposition state, ψ'

$$\phi' = \sum_i a_i\phi_i \tag{6}$$

$$\psi' = |\phi'_1\phi'_2....\phi'_N| \tag{7}$$

We have previously shown that this superposition state includes not only singly excited states but also doubly excited states within a closed-shell configuration, when using a spin-independent perturbation.[19] With a spin-dependent perturbation (eqs 2 and 3), the α and β spatial components of the wave function are no longer identical. We herein use a two-electron, two-orbital system with bonding MO φ_σ and virtual MO φ_{σ^*}, beginning in S_0, with both electrons in φ_σ, to show how the triplet component is introduced into the superposition wave function. The single-configuration superposition state ψ' in eq 7 can be written as (where $\mathbf{r}$ and τ are spatial and spin variables, respectively)

$$\begin{aligned}\psi' &= \frac{1}{\sqrt{(1+c_\alpha^2)(1+c_\beta^2)}}|\phi'(\mathbf{r}_1)\alpha(\tau_1)\phi'(\mathbf{r}_2)\beta(\tau_2)|\\ &= \frac{1}{\sqrt{(1+c_\alpha^2)(1+c_\beta^2)}}\times\\ &\quad |[\phi_\sigma + c_\alpha\phi_{\sigma^*}](\mathbf{r}_1)\alpha(\tau_1)[\phi_\sigma + c_\beta\phi_{\sigma^*}](\mathbf{r}_2)\beta(\tau_2)|\\ &= \frac{1}{\sqrt{(1+c_\alpha^2)(1+c_\beta^2)}}\left(\psi_{S_0} + \left(\frac{c_\alpha + c_\beta}{\sqrt{2}}\right)\psi_{S_1} + \left(\frac{c_\beta - c_\alpha}{\sqrt{2}}\right)\psi_{T_0} + c_\alpha c_\beta\psi_{S_2}\right)\end{aligned} \tag{8}$$

Here ψ_{S_1} and ψ_{T_0} are the spin-adapted singlet and triplet configurations, and ψ_{S_2} is the doubly excited configuration. The coefficients c_α and c_β determine the degree of individual α and β perturbation from the ground state and are governed by both the strength of the spin-dependent perturbation and the system-dependent response. Eq 8 indicates that a spin-dependent perturbation gives rise to a mixing of triplet component in the superposition state.[23]

Because the field is applied only to the α electron density, the spin-symmetry is broken from a pure singlet state, and we have an unrestricted wave function, whose spin-symmetry time evolution can be monitored by the time-dependent expectation value of the squared total spin angular momentum operator $\hat{S}^2$. For a single determinant open-shell wave function in an orthonormal basis, $\langle\hat{S}^2(t)\rangle$ can be written in terms of the time-dependent density matrices as[24,25]

$$\langle\hat{S}^2(t)\rangle = \left(\frac{N_\alpha - N_\beta}{2}\right)^2 + \frac{N_\alpha + N_\beta}{2} - \mathrm{Tr}[\mathbf{P}_\alpha(t)\cdot\mathbf{P}_\beta(t)] \tag{9}$$

Because $\hat{S}^2 = 0$ for singlets, such as S_0, S_1, and S_2, and $\hat{S}^2 = 2$ for triplets, such as T_0, a nonzero $\hat{S}^2$ value in eq 9 becomes an indicator of the singlet−triplet mixing during the time-evolution of the superposition wave function.

Results and Discussion

The development version of the Gaussian code[26] is used to obtain the initial wave function and to calculate the one- and two-electron integrals. We use a minimal STO-3G basis set for simplicity of analysis and also a 6-31G(d,p) basis for comparison. The simulations begin with the molecule initially in its field-free ground state. The electric field vector is applied along the molecular (z) axis for three cycles with a frequency of $\omega = 0.06$ au. We use an integration time step of 0.002 fs and propagate the electron density for 50 fs. As we and other groups have shown,[19,27] shifting of transition energies can also occur when contributions from excited states in the superposition state become large. For the spin-dependent perturbations discussed herein, a stronger field or a field frequency closer to resonance not only shifts the peak energies but also can cause extreme broken spin symmetry, with $\langle\hat{S}^2(t)\rangle$ values approaching 1. The intensity of peaks in

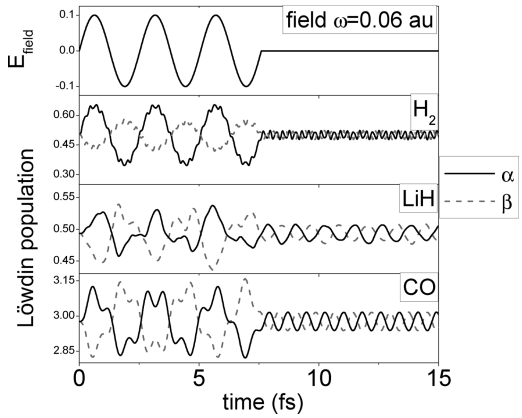


Figure 1. Electric field and Löwdin populations. Top panel: applied electric field with $\omega = 0.06$ au. Other panels: α and β Löwdin populations on one atom of the diatomics H_2, LiH, and CO. The α population (black lines) follows the field, and the β population (gray dash lines) responds by shifting onto the other atom. This breaks the spin-symmetry of the system, creating a superposition state that includes triplet contributions. After the field is off, the α and β populations oscillate at the same frequencies but out of phase with each other.

the spectrum is directly related to population of various states in the superposition wave function, making the strength of the spin-dependent perturbation directly correlated to these intensities. In order to compare peak energies with those obtained from linear response theory, we have kept the perturbation weak and away from resonant energies in order to not significantly alter the population away from that of the reference ground state.

A spin-dependent perturbation in the form of an off-resonant field was applied to the ground state of three different diatomic molecules: H_2, LiH, and CO. Figure 1 shows the applied electric field for $\omega = 0.06$ au, and the resulting Löwdin populations on one atom, obtained using the TDHF Hamiltonian with the STO-3G basis. A field strength of $E_{max} = 0.1$ au was used for H_2, and $E_{max} = 0.01$ au was used for LiH and CO. Initially, the diatomics are closed-shell with identical α and β populations on each atom. Once the spin-dependent perturbation is applied, the α electron density (black lines) follows the field, building up electron density on one atom in each of the diatomics. The β electron density (gray dash lines) responds to the changing α electron density by decreasing on that atom, and building up on the other atom, breaking the spin-symmetry. Once the field is removed the magnitude of the α and β population oscillations decrease significantly, as the field is no longer directly driving the α electron density, and the β electron density no longer is responding to the large change in the α electron density. While the change in magnitude of the α and β populations is much smaller after the field is off, they continue to oscillate out of phase. It is this out-of-phase oscillation that indicates the triplet contribution within the superposition state, as suggested by eq 8.

As discussed in the previous section, the $\langle \hat{S}^2 \rangle$ value is an indicator of the singlet−triplet mixing in the superposition state. The corresponding $\langle \hat{S}^2 \rangle$ values during the dynamics are shown in Figure 2. A nonzero value clearly indicates that the superposition state is no longer made up of pure singlets,

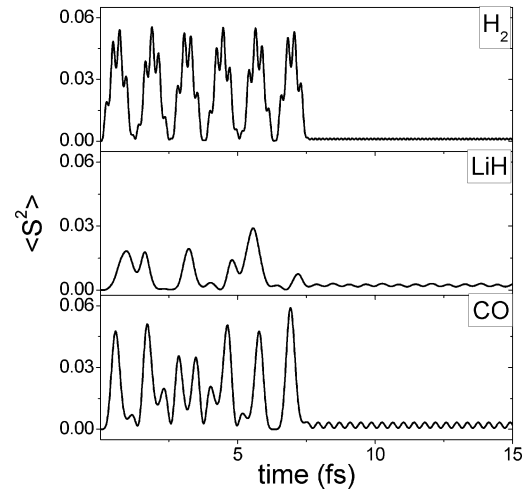


Figure 2. $\langle \hat{S}^2 \rangle$ value for H_2, LiH, and CO, using the same field conditions as given in Figure 1. The magnitude of $\langle \hat{S}^2 \rangle$ represents the triplet contribution to the superposition state.

and the magnitude of $\langle \hat{S}^2 \rangle$ indicates the degree of triplet contribution to the superposition state wave function. Because $\langle \hat{S}^2 \rangle$ is calculated as the trace of the product of the α and β electron densities, see eq 9, the value takes into account the phase of the wave function from the imaginary part of the density matrix. With the same field conditions as for the populations given in Figure 1, the value of $\langle \hat{S}^2 \rangle$ remains small throughout the dynamics. For H_2, using a maximum field strength of $E_{max} = 0.1$ au, the maximum value of $\langle \hat{S}^2 \rangle$ during the field application was 0.056 and was 0.0014 once the field was removed. For LiH and CO, with $E_{max} = 0.01$ au, the maximum of $\langle \hat{S}^2 \rangle$ with the field on was 0.029 and 0.059, respectively, and after field removal it was 0.0032 and 0.0033. A weaker field was used because the LiH and CO transition energies are closer to the $\omega = 0.06$ au field frequency, leading to a larger response.

Despite the α and β Löwdin populations oscillating out-of-phase with each other, eq 8 indicating a triplet contribution to the wave function, and $\langle \hat{S}^2 \rangle$ no longer being zero, the Fourier transformation (FT) of the total residual dipole moment $\mathrm{Tr}[\mathbf{d} \cdot (\mathbf{P}^{\alpha}(t) + \mathbf{P}^{\beta}(t))]$ gives an absorption spectrum that agrees with the closed-shell singlet results, see Figure 3 (black lines). In the absorption spectrum of H_2 there is a single peak, which is at the $S_0 \rightarrow S_1$ transition energy of 0.94 au (Table 1). However, examinations of the individual α and β contributions to the total dipole moment show that the out-of-phase oscillation provides additional information. In addition to the FT of the total dipole moment, Figure 3 also shows the FT of the contribution of the α electron density to the total residual dipole moment (gray dash lines), calculated via $\mathrm{Tr}[\mathbf{d} \cdot \mathbf{P}^{\alpha}(t)]$. The FT of the β contribution is identical. For H_2, this 'absorption spectrum' shows a peak at the same $S_0 \rightarrow S_1$ transition energy but also an additional peak at 0.57 au, which corresponds to the linear response TDHF excitation energy from the S_0 state into the triplet T_0 state. This result shows that broken-spin real-time TDHF/TDDFT simulations yield the 'dark' transitions between singlet and triplet states in the individual α or β contributions to the total dipole moment. These 'dark' transitions do not show up in the total dipole allowed absorption spectrum, as

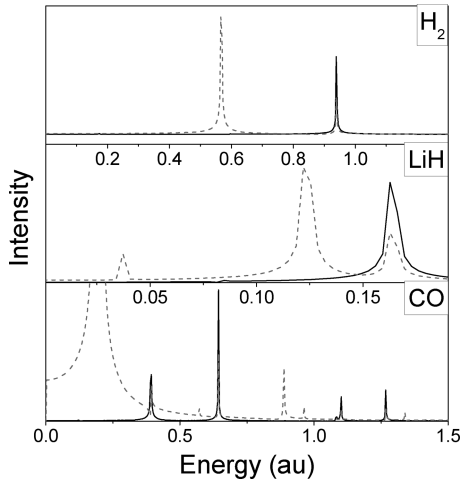


Figure 3. TDHF/STO-3G absorption spectra for H_2, LiH, and CO. Fourier transformation of the total dipole moment (black) shows dipole allowed singlet-to-singlet transitions. These peak energies agree with the closed-shell linear-response results (see Table 1). Fourier transformation of the α or β contribution to the dipole moment (gray dash) shows additional peaks corresponding to triplet transitions.

Table 1. Linear Response Excitation Energies (au)

	z-allowed singlet	triplet
	TDHF/STO-3G	
H_2	0.94	0.57
LiH	0.17	0.13
CO	0.64	0.20
	1.10	0.89
	1.27	1.13
	TDHF/6-31G(d,p)	
H_2	0.55	0.37
LiH	0.15	0.11
CO	0.54	0.22
	0.67	0.58
	0.95	0.77
	1.06	0.97
	TDPBE/STO-3G	
H_2	0.95	0.62
LiH	0.13	0.11
CO	0.64	0.31
	0.98	0.79
	1.14	1.00

the out of phase α/β oscillations cancel out in the total dipole response. This spin-dependent perturbation will only yield transitions in which the transition dipole matrix element $\langle\chi_\mu|\hat{r}|\chi_\nu\rangle$ is nonzero. While singlet−triplet transitions are not formally spin-symmetry allowed, the individual spin-allowed transitions for either the α or β electron can yield the singlet−triplet transitions as observed in the FT of the time evolution of the α or β electron density dipole moment. Just as with singlet−singlet transitions, the strength of the singlet−triplet transitions is governed by the magnitude of the dipole matrix element, and additional singlet−triplet transitions can be achieved by changing the field direction or using field terms beyond the dipole approximation.

The spectra for LiH and CO in Figure 3 are consistent with those found for the two-electron system H_2. For LiH, the FT of the α electron density dipole moment shows both the $S_0 \rightarrow S_1$ peak, and the $S_0 \rightarrow T_0$ peak, both of which are

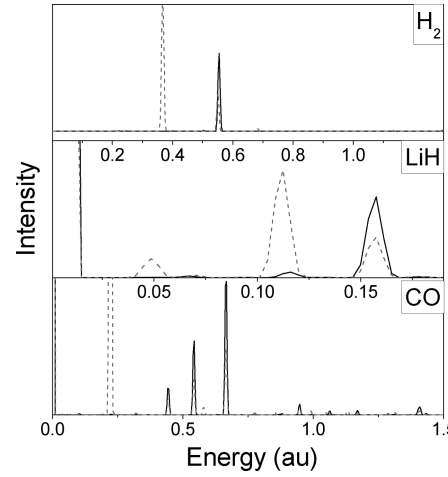


Figure 4. TDHF/6-31G(d,p) absorption spectra for H_2, LiH, and CO. Fourier transformation of the total dipole moment (black) shows dipole allowed singlet-to-singlet transitions. Fourier transformation of the α or β contribution to the dipole moment (gray dash) shows additional peaks corresponding to triplet transitions.

dominant HOMO−LUMO transitions. Because of a small population transfer into the excited state, the peak energies are slightly shifted from the linear response results (see Table 1). The dynamics give the $S_0 \rightarrow S_1$ transition energy at 0.16 au, and the $S_0 \rightarrow T_0$ transition energy at 0.12 au. There is also an additional small peak at 0.04 au which corresponds to transitions between the two excited states. This low-energy peak disappears at weaker field strengths. For CO the overall pattern is the same, with the FT of the α electron dipole moment showing both the singlet−singlet and the singlet−triplet transitions. The low energy transitions are $\pi \rightarrow \pi^*$, while the two higher energy transitions are $\sigma \rightarrow \sigma^*$. There is no shifting of peak transition energies compared to linear response theory, indicating very little population transfer. Because the $S_0 \rightarrow T_0$ transition at 0.20 au is near the field frequency of $\omega = 0.06$ au, the $S_0 \rightarrow T_0$ peak is very intense due to a larger T_0 contribution in the superposition state.

We also present two comparisons, one using the larger 6-31G(d,p) basis with the HF Hamiltonian (Figure 4) and the other using the same simple STO-3G basis with the PBE density functional Hamiltonian (Figure 5). The larger basis results show peaks shifting to lower energies as expected from linear response theory (Table 1). For the larger basis, we decreased the maximum field strength for H_2 and LiH to 0.03 au and 0.005 au, respectively. Because of the relatively large field strength of 0.01 au for CO, there is more population of the lower energy states, increasing the intensity of the corresponding lower energy peaks. The TDDFT results are comparable to those obtained with TDHF, with results agreeing with those from linear response theory (Table 1). The real-time TDDFT simulations used field conditions identical to the TDHF/STO-3G simulations. Some of the TDPBE excitations energies are different from the TDHF energies, changing the peak locations. Also, in addition to changes in the excitation energies, there are changes in the energies of peaks that correspond to transitions between excited states. For example, because the TDPBE LiH $S_0 \rightarrow T_0$ and $S_0 \rightarrow S_1$ transitions are closer in energy than the TDHF

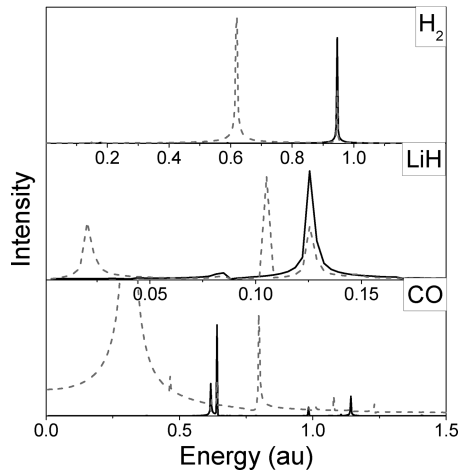


Figure 5. TDPBE/STO-3G absorption spectra for H_2, LiH, and CO. Fourier transformation of the total dipole moment (black) shows dipole allowed singlet-to-singlet transitions. Fourier transformation of the α or β contribution to the dipole moment (gray dash) shows additional peaks corresponding to triplet transitions.

transition energies, the peak due to transition between the excited states is red-shifted by ∼0.02 au compared to TDHF.

Conclusion

In this work we present a method that models singlet−triplet transitions within real-time TDHF and TDDFT methods. With the introduction of a spin-dependent perturbation, we show the incorporation of the triplet wave function into the superposition state achieved through real-time electron density propagation via monitoring of the α and β populations as well as the value of $\langle \hat{S}^2 \rangle$. With this mixed-spin superposition state, we observe both $S_0 \rightarrow$ singlet transitions and $S_0 \rightarrow$ triplet transitions as well as transitions between excited states.

Acknowledgment. Support is provided by NSF grants PHY-CDI 0835546 and CHE-CAREER 0844999 and the ACS Petroleum Research Fund (46487-G6). Enlightening discussions with Ernest Davidson are gratefully acknowledged.

References

(1) Simons, J. *J. Chem. Phys.* **1971**, *55*, 1218–1230.

(2) Jørgensen, P.; Simons, J. *Second quantization-based methods in quantum chemistry*; Academic Press: New York, 1981.

(3) Kulander, K. C.; Devi, K. R. S.; Koonin, S. E. *Phys. Rev. A* **1982**, *25*, 2968.

(4) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.

(5) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.

(6) Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. *Phys. Rev. Lett.* **1996**, *76*, 1212.

(7) Casida, M. E. *Recent Advances in Density-Functional Methods*; World Scientific: Singapore, 1995; pp 155−193.

(8) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.

(9) Yabana, K.; Bertsch, G. F. *Phys. Rev. B* **1996**, *54*, 4484.

(10) Tsolakidis, A.; Sánchez-Portal, D.; Martin, R. M. *Phys. Rev. B* **2002**, *66*, 235416.

(11) Marques, M. A. L.; López, X.; Varsano, D.; Castro, A.; Rubio, A. *Phys. Rev. Lett.* **2003**, *90*, 258101.

(12) Li, X.; Smith, S. M.; Markevitch, A. N.; Romanov, D. A.; Levis, R. J.; Schlegel, H. B. *Phys. Chem. Chem. Phys.* **2005**, *7*, 233.

(13) Smith, S.; Markevitch, A.; Romanov, D.; Li, X.; Levis, R.; Schlegel, H. *J. Phys. Chem. A* **2004**, *108*, 11063.

(14) Wang, F.; Yam, C. Y.; Chen, G.; Fan, K. *J. Chem. Phys.* **2007**, *126*, 134104.

(15) Wang, F.; Yam, C. Y.; Chen, G. *J. Chem. Phys.* **2007**, *126*, 244102.

(16) Andrade, X.; Botti, S.; Marques, M. A. L.; Rubio, A. *J. Chem. Phys.* **2007**, *126*, 184106.

(17) Takimoto, Y.; Vila, F. D.; Rehr, J. J. *J. Chem. Phys.* **2007**, *127*, 154114.

(18) Eshuis, H.; Balint-Kurti, G. G.; Manby, F. R. *J. Chem. Phys.* **2008**, *128*, 114113.

(19) Isborn, C. M.; Li, X. *J. Chem. Phys.* **2008**, *129*, 204107.

(20) Schlegel, H. B.; Smith, S. M.; Li, X. *J. Chem. Phys.* **2007**, *126*, 244110.

(21) Isborn, C. M.; Li, X.; Tully, J. C. *J. Chem. Phys.* **2007**, *126*, 134307.

(22) Li, X.; Tully, J. C. *Chem. Phys. Lett.* **2007**, *439*, 199.

(23) Schlegel, H. B. *Encyclopedia of Computational Chemistry*; Wiley: Chichester, 1998; pp 2665−71.

(24) Chen, W.; Schlegel, H. B. *J. Chem. Phys.* **1994**, *101*, 5957.

(25) Staroverov, V. N.; Davidson, E. R. *Chem. Phys. Lett.* **2000**, *330*, 161.

(26) Frisch, M. J. *Gaussian Development Version, Revision D.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(27) Hu, C.; Sugino, O.; Miyamoto, Y. *Phys. Rev. A* **2006**, *74*, 032508.

CT900264B

JCTC Journal of Chemical Theory and Computation

Extensive TD-DFT Benchmark: Singlet-Excited States of Organic Molecules

Denis Jacquemin,*,† Valérie Wathelet,† Eric A. Perpète,† and Carlo Adamo*,‡

Groupe de Chimie-Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium, and Ecole Nationale Supérieure de Chimie de Paris, Laboratoire Electrochimie et Chimie Analytique, UMR CNRS-ENSCP no. 7575, 11, rue Pierre et Marie Curie, F-75321 Paris Cedex 05, France

Received June 10, 2009

Abstract: Extensive Time-Dependent Density Functional Theory (TD-DFT) calculations have been carried out in order to obtain a statistically meaningful analysis of the merits of a large number of functionals. To reach this goal, a very extended set of molecules (∼500 compounds, >700 excited states) covering a broad range of (bio)organic molecules and dyes have been investigated. Likewise, 29 functionals including LDA, GGA, *meta*-GGA, global hybrids, and long-range-corrected hybrids have been considered. Comparisons with both theoretical references and experimental measurements have been carried out. On average, the functionals providing the best match with reference data are, one the one hand, global hybrids containing between 22% and 25% of exact exchange (X3LYP, B98, PBE0, and mPW1PW91) and, on the other hand, a long-range-corrected hybrid with a less-rapidly increasing HF ratio, namely LC-ωPBE(20). Pure functionals tend to be less consistent, whereas functionals incorporating a larger fraction of exact exchange tend to underestimate significantly the transition energies. For most treated cases, the M05 and CAM-B3LYP schemes deliver fairly small deviations but do not outperform standard hybrids such as X3LYP or PBE0, at least within the vertical approximation. With the optimal functionals, one obtains mean absolute deviations smaller than 0.25 eV, though the errors significantly depend on the subset of molecules or states considered. As an illustration, PBE0 and LC-ωPBE(20) provide a mean absolute error of only 0.14 eV for the 228 states related to neutral organic dyes but are completely off target for cyanine-like derivatives. On the basis of comparisons with theoretical estimates, it also turned out that CC2 and TD-DFT errors are of the same order of magnitude, once the above-mentioned hybrids are selected.

1. Introduction

Developing methodological approaches able to accurately deliver the transition energies corresponding to electronically excited-states remains a major challenge for theoretical chemists. Historically, the first computational schemes developed relied on semiempirical theories.[1] The most successful model, namely ZINDO,[2] was purposed-designed to allow quick estimates of the main features of UV/visible spectra and remains popular today. However, the quantitative aspect of the obtained results (absorption wavelengths and transition probabilities) was found to be highly system-dependent, a problematic feature.[3−5] More recently, calculations carried out for organic dyes have indicated that PM5[6] could be a promising approach,[7] but such a claim remains to be tested on a broader set of transitions and molecules. At the other extreme of the theoretical palette, one finds highly correlated ab initio approaches such as SAC-CI, EOM-CC, MR-CI, or CAS-PT2 that allow very accurate

* Corresponding author e-mail: denis.jacquemin@fundp.ac.be (D.J.); carlo-adamo@enscp.fr (C.A.).

† FUNDP, Namur.

‡ ENSCP, Paris.

10.1021/ct900298e CCC: $40.75
Published on Web 08/11/2009

estimates but are limited to rather small systems due to their extreme computational cost. Two of the most extensive investigations performed with such high-accuracy approaches (CC3 and CCSDR) have been published very recently by Thiel and its co-workers.[8,9] Although these contributions certainly represent a very large computational effort, they have been "limited" to molecules of about 15 atoms (naphthalene was the largest compound treated with CC3) and relied on a diffuse-free polarized triple-ζ basis set. While it is true that CAS-PT2 and CC2, the two "lighter" wave function approaches, could possibly be applied to molecules containing about 40 atoms,[10−18] the current implementations of these *ab intio* theories often do not permit a systematic inclusion of medium effects. This is a problematic drawback, as it is well-known that excited-state properties tend to be more solvent-sensitive than their ground-state counterparts.[19] Clearly, the difficulties to apply the highly correlated approaches to a broad set of molecules in a real-life environment have yet not been completely solved. In terms of computational cost, one finds an intermediate between semiempirical theories and wave function approaches, namely time-dependent density functional theory (TD-DFT).[20−23] TD-DFT is the most widely applied ab initio tool for modeling the electronic spectra of organic and inorganic molecules[24,25] and can be extended to incorporate environmental effects either through a modeling of the bulk environment[19,26−29] or through a variety of QM/MM approaches.[30−35] Despite its successes and versatility, TD-DFT is limited and suffers an important drawback: the quality of the obtained results is profoundly functional-dependent. Indeed, the appropriate selection of the exchange-correlation form is often crucial to grasp chemically sound conclusions. For most excited-states, hybrid functionals that incorporate a fraction of exact exchange (EE) tend to provide more accurate estimates than pure functionals. Anyway, transition wavelengths to excited-states presenting a doubly excited character or a significant charge-transfer nature are traditionally poorly estimated, as are the electronic spectrum of molecules having a strong multideterminantal nature. For sure these deficiencies are related to the approximate nature of today's implementation, as illustrated by the recently developed long-range-corrected hybrids (LCH),[36−44] that appear to correctly appraise the charge-transfer properties. Contrary to the global hybrids (GH), LCH presents an EE percentage depending on the interelectronic distance, allowing a physically correct asymptotic behavior when the two electrons are far apart.

It is quite astonishing that only a limited number of contributions collated the *pros* and *cons* of functionals in the TD-DFT framework, for a significant set of molecules. In Table 1, we summarize the selected methodologies and training sets for twelve investigations tackling this question. One can certainly find many other TD-DFT studies using GH or LCH but often specific to a specific class of molecules.[5,45−53] As can be seen, not only the training set but also the details of the methodologies selected for benchmarking (including the size of the basis set and the possible modeling of solvent effects) differ significantly from one work to the other. We believe it is especially striking that most studies include only a very small number of functionals (typically three) and that only four works used more than 100 excitations to obtain statistically meaningful conclusions. Considering the different training sets and procedures, it is to be expected that the conclusions of these investigations are not perfectly uniform. While the obtained mean absolute deviation (MAE) for the "best" functional is typically close to 0.25 eV, the actual findings are in fact partly antagonistic, making it difficult to appreciate the "general" functional performance in the TD-DFT framework:

1. Tozer and co-workers concluded that CAM-B3LYP[40] leads to much smaller deviations than B3LYP[54] for a variety of transitions of medium-size chromogens.[55] The average B3LYP error being completely unacceptable (>1.0 eV) for both Rydberg and charge-transfer states.[55,56] For valence transitions, all tested functionals (PBE, B3LYP, and CAM-B3LYP) provided similar MAE (0.27, 0.26, and 0.27 eV, respectively).[55]

2. Rohrdanz and Herbert found that an accurate description of both the ground-state and excited-state properties of large molecules was uneasy with common LCH functionals[57] and subsequently design a LCH functional working for both ground- and excited-states.[58] This new LCH functional provided a MAE of about 0.3 eV.[58]

3. For the λ_{max} related to $\pi \rightarrow \pi^{\star}$ transitions in 100 organic dyes, we found, within the vertical approximation, that PBE0 outperforms LCH and provides a MAE close to 0.15 eV,[59] the errors being of the same order of magnitude for $n \rightarrow \pi^{\star}$ transitions.[60,61] For the same set of $\pi \rightarrow \pi^{\star}$ transitions, CAM-B3LYP provided significantly larger deviations (0.26 eV).[59]

4. Thiel's group used BP86, B3LYP, and BHHLYP and they obtained MAE of 0.52 eV, 0.27, and 0.50 eV, respectively, for more than 100 transitions in small molecules,[62] using their own "best theoretical estimates"[8] as reference values.

5. Dierksen and Grimme concluded from an extensive vibronic investigation of (mainly) hydrocarbons that the optimal global hybrid should contain between 30% and 40% of EE.[63] Comparing their vertical (0−0) TD-DFT data to their solvent-corrected experimental references, we calculated MAE of 0.43 (0.57) eV, 0.21 (0.34) eV, and 0.31 (0.18) eV for BP86, B3LYP, and BHHLYP, respectively.

6. Very recently,[64] Goerigk et al. used the CAS-PT2 result of ref 8 to benchmark double-hybrid functionals[65] and found a MAE of 0.22 eV for B2PLYP and B2GPPLYP, significantly smaller than with B3LYP (0.30 eV) and confirmed this finding on a set of five large chromophores.

Consequently, given an arbitrary molecule, it remains difficult to know without testing what is (are) "reasonably" the most adequate functional(s) to evaluate the electronic spectra. Should one choose a GH or a LCH? Would the error be much larger with a GGA than with a GH? What is the "expected" accuracy with today's computational procedure? Are *ab intio* functionals outperforming (or not) parametrized functionals? Should the chosen functional vary for molecules of different size? Of course, all these questions have been tackled in part in the above-mentioned works, but with no generic answer embraced by a large community. Here, we

Table 1. Training Set and Method Used in Recent Benchmark TD-DFT Calculations[n]

citation		training set			method				
group(s)	ref.	states	molecules	functionals	basis set	solvent	transitions	geometry	comparison
Boeij	14	25 mixed	19 molecules	SVWN, SVWN-VK	ET-pVQZ	none	vertical	BP/TZ2P	mixed[a]
Fabian	3	163 mixed	76 sulfur-bearing dyes	B3LYP	6-31+G(d)	none	vertical	B3LYP/6-31+G(d)	experiment
Fabian	12	54 mixed	21 sulfur-free dyes	B3LYP, B3LYP(TB)	6-31+G(d)	none	vertical	B3LYP/6-31+G(d)	experiment
Grimme	13	22 mixed	14 molecules	B3LYP	TZVP	none	vertical	B3LYP/TZVP	mixed
Grimme	63	42 $\pi-\pi^{\star}$ [b]	40 large molecules[c]	BP86, B3LYP, BHHLYP	TZVP	none[d]	0−0	DFT/TZVP	experiment
Grimme	64	142/5 mixed	28/5 small/large molecules	6/17 functionals[e]	TZVP	none/PCM	vertical/0−0	MP2/DFT	theory[f]/experiment
Herbert	57	29 mixed	9 molecules	LC-PBE, LC-BLYP, LC-PBE0	aug-cc-PVDZ	none	vertical	B3LYP/6-31G(d)	experiment
Matsumura	7	50 $\pi-\pi^{\star}$	50 organic dyes	B3LYP	6-31G(d), cc-pVDZ	none	vertical	B3LYP	experiment
Tozer	55	59 mixed	18 model molecules	PBE, B3LYP, CAM-B3LYP	cc-pVTZ[g]	none	vertical	Mixed[h]	mixed[i]
Thiel	62	102 mixed	28 small molecules	BP86, B3LYP, BHHLYP, MR-DFT	TZVP	none	vertical	MP2/6-31G(d)	theory[j]
Us	60	34 $n-\pi^{\star}$	34 small dyes	12 pure and hybrid functionals[k]	Large w. diffuse[l]	PCM	vertical	PCM-DFT	experiment
Us	59	118 $\pi-\pi^{\star}$	115 organic dyes	6 pure and hybrid functionals[m]	6-311+G(2d,p)	PCM	vertical	PCM-PBE0/6-311G(d,p)	experiment

[a] CAS-PT2, gas-phase measurements, or solvent measurements empirically corrected for solvatochromism. [b] Mainly $\pi-\pi^{\star}$ transitions corresponding to singlet and doublet excited-states but a few other states. [c] Mainly aromatic and aliphatic hydrocarbons or oligomeric structures with a few heteroatoms. [d] No solvent model in the theory, but the experimental values have been shifted by a constant 0.15 eV (for all solvents) to include solvatochromism. [e] BP86, B3LYP, B2LYP, B2GPLYP, B2PLYP, and B2GPPLYP for the small molecules with PBE, OPBE, BLYP, mPWLYP, TPSS, VSXC, O3LYP, B98, PBE0, BMK, and BHHLYP for the large ones. [f] CAS-PT2 from ref 9, using the same basis set and geometry. [g] d-aug-cc-pVTZ for Rydberg states. [h] Experimental, B3LYP/TZVP, CAM-B3LYP/6-31G(d), or MP2/6-31G(d) geometries, depending on the molecule. [i] CAS-PT2, CC2, or gas-phase experiment, depending on the molecule. [j] Best estimates (generally CC3 or CAS-PT2) from their own ref 8, using the same basis set and geometry. [k] HF, BLYP, PBE, TPSS, B3LYP, PBE0, BMK, LC-BLYP, LC-PBE, LC-TPSS, LC-ωPBE, and CAM-B3LYP. [l] 6-311++G(3d,3p) for nitroso dyes and 6-311+G(2df,p) for thiocarbonyl chromophores. [m] HF, PBE, PBE0, LC-PBE, LC-ωPBE, and CAM-B3LYP. [n] For each contribution, we list the nature of the selected excited-stated, molecules, and functionals as well as a summary of the methodological scheme. In this table, i) "basis set" refers to the basis set used for TD-DFT calculations; ii) "solvent" indicates the consideration or/not of environmental effects; iii) "transitions" indicates if full vibronic calculations have been computed or if vertical values have been used; iv) "geometry" gives the method used to obtain the molecular ground-state structures; and v) "comparison" indicates the origin of the values used as reference data during the statistical analysis.

have performed benchmarks that are more complete than any previously published data, both from the point of view of the number of molecules considered and of the set of pure and hybrid functionals incorporated.

2. Methodology

2.1. Strategy. As can be seen in Table 1, two philosophies can be used to benchmark TD-DFT functionals: versus experiment (VE) or versus theory (VT). Both approaches have advantages and disadvantages. Trying to closely match experiment (VE) is generally desired in most practical applications and allows to include in the training set a wide range of molecules and compounds. On the other hand, one would normally need to compute the full vibronic spectra (and not "simply" vertical transitions) and to perfectly model the experimental setup (pressure, temperature, full environmental effects, ...), both tasks being impossible for a large set of solvated molecules. Additionally, it is not always straightforward to pinpoint the theoretical transition actually corresponding to the experimental measures, especially for highly excited states. Comparisons with accurate wave function estimates (VT) allows straightforward and physically meaningful comparisons (same conditions, same transitions) but is obviously limited by the availability of theoretical data, i.e. only small molecules can be included. In many cases, CC2 results have been used as reference values for medium size molecules, a strategy that we think unsatisfying. Indeed, we computed a MAE of 0.27 eV (0.30 eV) between the CC2/TZVP and the CAS-PT2/TZVP ("best estimates") values for the 103 singlet-excited excited-states of ref 8.[66] Even for low-lying excited-states, CC2 is often off the theoretical limit by 0.1 eV,[8] a value equal to one-half or one-third of the typical TD-DFT error.

In the following, we will use both philosophies so to be as general as possible. In what concerns the versus theory scheme, we have selected Thiel's set (VT set in the following) and mimic exactly the computational procedure (basis set and geometry). For the VE set, we have used a computational strategy that is at the limit of today's possibilities for such a set of molecules, trying to circumvent the possible limitations of our computational procedure. For the sake of consistency, we have chosen to use a uniform methodology (basis set, solvent effects, ...) for all VE molecules.

2.2. General Computational Procedure. All calculations have been performed with the Gaussian suite of programs, using both the commercial and development versions[67,68] with a tight self-consistent field convergence threshold (10^{-8} to 10^{-10} au). For the VE set, we have followed a well-established three-step approach:[25] i) the ground-state geometry of each compound has been optimized until the residual mean force is smaller than 1.0×10^{-5} au (so-called *tight* threshold in Gaussian); ii) the vibrational spectrum is analytically determined to confirm that the structure is a true minimum; and iii) the vertical transition energies to the valence excited states are computed with TD-DFT. For the VT set, the geometries have been taken from ref 8 and step iii) directly performed.

As the majority of experimental data are obtained in condensed phase, we have included bulk solvent effects in our VE model (all VT calculations are in gas-phase). This was performed at each stage, including geometry optimizations and Hessian calculations, using the well-known Polarizable Continuum Model (PCM),[19] that is able to obtain a valid approximation of solvent effects as long as no specific interactions link the solute and the solvent molecules. Typically solvent−solute hydrogen bonds tend to influence more significantly the $n \rightarrow \pi^\star$ transitions than their $\pi \rightarrow \pi^\star$ counterparts, and we have tried to select aprotic solvent for the former, at least when different experimental values are available. The list of solvent selected is given in the Supporting Information. The default PCM Gaussian parameters have generally been used, though for a few calculations if was necessary to change the atomic raddi (UAKS instead of UA0) or to switch off the presence of smoothing sphere (NoAddSph) to converge the force minimizations. For the records, note that some default PCM parameters might differ between the two versions of the program used. All TD-DFT calculations have been performed within the nonequilibrium approximation, valid for absorption spectra.[19]

2.3. Functionals and Basis Sets. As we want to assess the *pros* and *cons* of a series of DFT approaches, a very extended set of functionals has been used. Apart from the Time-Dependent Hartree−Fock approach (TD-HF, refereed to as HF in the following), the selected functionals can be classified in five major categories: LDA, GGA, *meta*-GGA, GH, and LCH. In the first category, that is expected to be the less efficient we have selected only one functional, SVWN5.[69,70] We have chosen four GGAs, namely BP86,[71,72] BLYP,[71,73] OLYP[73,74] and PBE,[75] whereas we have picked up three popular *meta*-GGA: VSXC,[76] τ-HCTH[77] and TPSS.[78] Twelve global hybrids have been used: TPSSh (10%),[79] O3LYP (11.61%),[80] τ-HTCHh (15%),[77] B3LYP (20%),[54,81] X3LYP (21%),[82] B98 (21.98%),[83] mPW1PW91 (25%),[84] PBE0 (25%),[85,86] M05 (28%),[87] BMK (42%),[88] BHHLYP (50%),[89] and M05−2X (56%).[90] The LCH constitute the last category and use a growing fraction of EE when the interelectronic distance increases. This is formally performed by partitioning the two-electron operator as[36,40,91]

$$\frac{1}{r_{12}} = \frac{1 - [\alpha + \beta \operatorname{erf}(\omega r_{12})]}{r_{12}} + \frac{\alpha + \beta \operatorname{erf}(\omega r_{12})}{r_{12}} \quad (1)$$

The first term of the rhs of this equation describes the so-called short-range effect and is modeled through DFT exchange, whereas the second term corresponds to the long-range contribution calculated with the HF exchange formula. In eq 1 ω is the range separation parameter, while α and $\alpha + \beta$ define the EE percentage at $r_{12} = 0$ and $r_{12} = \infty$, respectively. The LC model uses $\alpha = 0.00$, $\beta = 1.00$, and $\omega = 0.33$ au^{-1} in eq 1[37,38] and has been applied to both GGA and *meta*-GGA to give LC-BLYP, LC-OLYP, LC-PBE, LC-τ-HCTH, and LC-TPSS. The approach designed by Vydrov and Scuseria,[42,43] namely LC-ωPBE, with $\omega = 0.40$ au^{-1} and $\alpha = 0$, $\beta = 1$, has been used as well. Note that in LC-ωPBE, the short-range exchange functional can be rigorously derived[41,92] by integration of the model

exchange hole.[42,43] We have also used a variation of the LC-ωPBE functional using $\omega = 0.20$ au^{-1} (all other parameters are the same as the original model), here denoted LC-ωPBE(20). Indeed, such smaller ω has been recently found promising for TD-DFT calculations on large molecules.[57,58] Note that the functional designed in ref 58 differs from LC-ωPBE(20) by the use of 20% of short-range exchange. Additionally, the well-known CAM-B3LYP model ($\alpha = 0.19$, $\beta = 0.46$ and $\omega = 0.33$ au^{-1}) has been included in our set.[40] As the sum of α and β is not strictly equal to 1.00 in CAM-B3LYP, the exact asymptote of the exchange potential is lost, whereas a larger percentage of HF exchange is included at short-range. Eventually, we note that all LCH functionals selected in this work use full-range semilocal correlation.

In the VE set, steps i) and ii) of section 2.2 have been performed a split-valence triple-ζ 6-311G(d,p) basis set that delivers fully converged geometrical parameters for most molecules.[93] As our main focus is the TD-DFT part, all optimizations have been achieved with the PBE0 functional, that is suitable for most organic molecules, so to avoid that the quality of the geometry interferes with the evaluation of the performances of the functional for transition energies. In particular, it has been shown that several LCH lead to relatively poor geometries, so that performing geometry optimization and transition energy calculations with the same LCH may yield unsatisfactory results.[57,61] It has been tested that choosing another GH such as B3LYP indeed delivers very similar structural parameters for most molecular families in the VE set. The electronic excitations (step iii) were evaluated with the 6-311+G(2d,p) basis set, as such a basis set often bestows converged transitions wavelengths for medium and large chromophores,[25] as long as no Rydberg state is considered. The second *d* polarization function has been shown compulsory for indigoids,[94] coumarins,[95] and diarylethenes:[5] this second function is therefore necessary to get closer to converged results. The accuracy of 6-311+G(2d,p) for low-lying excited-states of medium and large molecules can be illustrated by numerous examples: 1) for the λ_{max} of two typical diarylethenes, the differences between the 6-311+G(2d,p) and 6-311++Gp(2df,2pd) results are limited to +0.007 and −0.004 eV;[5] 2) the differences are also negligible (±0.007 eV at most) when adding additional diffuse or polarization functions on the selenoindigo,[96] thioindigo,[45] and indigo[94] structures; 3) for four diphenylamine dyes one notes no variation when using 6-311++G(3d,3p) instead of 6-311+G(2d,p);[97] 4) the first $n \rightarrow \pi^{\star}$ transition of thioacetone computed with 6-311+G(2d,p) is within 0.005 eV of the 6-311++G(3df,3pd) results, for a constant geometry;[48] 5) among the two strong transitions of five 1,4-naphthoquinones, the largest discrepancy noted between 6-311++G(3df,3pd) and 6-311+G(2d,p) is 0.012 eV;[98] 6) the largest deviation between the results calculated with these two basis sets is limited to 0.011 eV both neutral and anionic dinitrophenylhydrazones;[99] 7) the three major transitions of a large tetrakis hydrocarbon undergo no change when shifting from 6 to 311+G(2d,p) to 6-311++G(3df,3pd).[100] Of course, for the small molecule subset and, more specifically, the tiny systems listed in Table XXVII of the Supporting Information, the errors induced by the 6-311+G(2d,p) choice are certainly non-negligible, and we are far from convergence. In the VT set, we have used the TZVP basis set during the TD-DFT step to be consistent with ref 62. Note that the basis set effects can be large for some cases of the VT set, as discussed previously.[8] In other words, basis sets including diffuse functions would certainly modify the TD-DFT estimates of the VT set, but we have conserved this basis set for the sake of consistency (see below).

2.4. Building the VE Training Set. As we have discussed in the Introduction, building a meaningful training set of molecules is certainly important. In this first work, we focus on the singlet-excited states of (bio)organic molecules that are the focus of most TD-DFT investigations. We have tried to obtain a set of molecules as large as possible and as inclusive as possible. Indeed, our VE training set includes all molecules of refs 1, 7, 10, 12, 59, 60, and 61 as well as the majority of the compounds of refs 55 and 63. We want to highlight that absolutely no structure was simplified with respect to the actual experimental structure, e.g. no *t*-Bu side chain was replaced by a methyl group or a simple hydrogen atom, as common in many theoretical works. Additionally, no system was discarded because of the probable inefficiency of TD-DFT to model them correctly. For instance, we have included many cyanine dyes that, due to their multideterminantal nature,[101] are not satisfactorily modeled even with the most refined functionals.[65,102] The VE set (see the Supporting Information) contains 483 molecules, for a total of 614 excited-states. We have divided our full set of molecules in various subgroups for neutral dyes, charged dyes, hydrocarbons, biomolecules, oligomers, ... In what concerns the dyes, azobenzene, anthraquinone, and triphenylmethane derivatives are strongly represented as they constitute the three most important classes of "absorption" dyes.[103,104] The design of subgroups has been achieved not only to ease the reader's work but also to be able to test the consistency of the TD-DFT estimates within specific structural families. For the record, we note that consistently with ref 8, the VT set contains 28 molecules and 103 excited-states.

2.5. Reference Values. In the VT set, we have selected the reference values which are either the "best theoretical" estimates or the CAS-PT2/TZVP values,[8,62] the latter ensuring perfectly meaningful comparisons from the basis set point of view.[105] In the VE set, choosing appropriate references is more difficult. In most cases, the experimental works only report the longest wavelength of maximal absorption (λ_{max}) with possibly the related molar absorption coefficient. For most dyes or large conjugated molecules, this corresponds to the first low-lying transition with a large oscillator strength, and the comparison between theory and experiment is straightforward. In other cases (typically small chromogens or unconjugated molecules), comparisons could be more difficult, and we used the relative oscillator strength and symmetry of the excited-states to pinpoint the correct transition. For sure, a few specific assignments could be discussed although, on the one hand, their statistical weight is indeed limited, and, on the other hand, a reasonable

chemical assignment (as the one used here) is often performed in practice, the explicit experimental information for the nature of the excited-state being frequently missing. When the vibronic structure is clearly defined experimentally, as for most hydrocarbons, we have listed in the Supporting Information tables both the transition with the largest molar absorption coefficient (selected in most comparisons for the sake of consistency with other experimental data) and the likely 0−0 peak. Our experience is that both GH and LCH generally provide theoretical spectra that are more easily comparable to experiment, as only a few transitions do present large oscillator strength. With pure functionals, the interpretation tends to be less immediate. Finally, when several experimental values in the same conditions are available, the average value was used for comparisons.

2.6. Limitations. For the VE comparisons, several limitations can be pointed out. The first is certainly the lack of vibronic couplings,[63,106] in our model, meaning that we incorrectly compare purely vertical transitions to experimental transitions. Unfortunately, computations of the Franck−Condon factors require the determination of the Hessian of the relevant excited-state, a task that is very far from today's possibilities with a large basis set, for a large number of functionals and molecules in a solvated environment. The differences between vertical and 0−0 transitions could be sizable:[13,63] in a recent work Grimme and co-workers obtained variations in the 0.24 eV−0.41 eV range for five large chromophores.[64] This clearly indicates that computation of the 0−0 spectra for all molecules would certainly lead to different conclusions regarding the merits of each functionals. On the bright side, using vertical transitions has a practical advantage: such calculations are much faster and are used in the large majority of TD-DFT works. The second possible origin of theory/experiment deviation is the modeling of solvent effect that is limited to a bulk linear-response approach. It is obvious that protic solvents might interact specifically with many chromogens, tuning the computed spectra,[26,107,108] while for molecules undergoing a large change of dipole moment between the ground- and excited-state, like coumarins, a state-specific PCM modeling would be more suited.[109] On the contrary, we believe that the selected basis sets are large enough to induce a negligible error, whereas the choice of PBE0 for optimizing the ground-state geometries should not cause a significant bias but for long oligomeric chains.[102,110,111]

All these limitations do not exist for the VT set, where only the reliability of the theoretical reference could be criticized. While CAS-PT2, CC3, MR-CI, MRMP, ... values used to determine the "best estimates" in ref 8 are certainly not perfect, they are probably close enough to the theoretical limit, so that we assume in the following that the main error originates from the selected functionals, not from the theoretical reference. Of course, when comparing with these "best estimates" (obtained with different basis sets), the lack of diffuse functions in TD-DFT/TZVP might also induce an error. Using CAS-PT2/TZVP results for reference allows to lift this problem, although the errors associated with CAS benchmarks still remains non-negligible for a few specific states.[112]

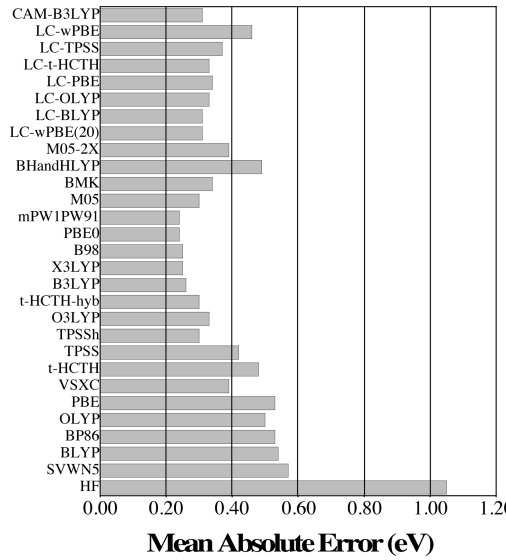


Figure 1. Mean Absolute Error for the full VT set (eV). The "best estimates" of ref 8 are used as references.

3. Results

For such extensive benchmark calculations, it is useless to analyze the computed spectra molecule-per-molecule, as one can always find a case for which a specific functional does provide the most accurate results. Only statistical analysis, allowing to unravel general trends does matter. The transition energies obtained for each molecules/functionals as well statistical results are catalogued in the Supporting Information.

3.1. Analysis of the VT Set. Let us start our comparisons with the VT set. The computed MSE, MAE, and the R^2 obtained through linear fitting for BP86, B3LYP, and BHHLYP are within 0.01 eV of the one reported by Silva-Junior and co-workers[62] for the same methods, confirming that the computational details (DFT integration grid, exact implementation of each functional, ...), that may differ from code to code, have a totally negligible impact. From the Supporting Information, it is clear that benchmarking the TD-DFT/TZVP values wrt the "best estimates" or wrt the CAS-PT2/TZVP references yield similar average deviations and error patterns. Therefore, in the following, we discuss the comparison with the former set of references, except when noted. The amplitude of the MAE for the 29 functionals are depicted in Figure 1. It turns out that HF provides very large errors (|MSE|, MAE, and RMS > 1.00 eV) and overestimates the transition energies in nearly 90% of the cases. It is worth pointing out that the HF errors are even larger for $n \rightarrow \pi^\star$ states with a MAE (RMS) of 1.36 eV (1.47 eV). The especially small HF R^2 (0.73) also indicates inconsistent predictions, and one can definitively rule out the uncorrelated approximation. In fact, using any DFT scheme does reduce the errors by a factor ranging from two to four. As expected, the pure functionals tend to provide too small transition energies, with MSE > 0.2 eV, although two of the *meta*-GGA (VSXC and TPSS) yield significantly more accurate spectral data than SVWN5, BLYP, BP86, OLYP, and PBE, in agreement with the ladder of functionals proposed by Perdew.[22] For instance, VSXC delivers a MAE of 0.39 eV

and a RMS of 0.47 eV, both 0.15 eV smaller than their BP86 counterpart: if only pure functionals are available, using the most refined one is indeed useful. All pure functionals are characterized by R^2 of 0.91 or 0.92, indicating that the chemical ordering is only reasonably restored by these approximations. Global hybrids are more accurate than the GGA and *meta*-GGA, and adding more and more EE tends to shift the transition energies to larger values. Consequently, the MSE evolves quite smoothly with the EE percentage (at the notable exception of M05), remaining slightly positive, as in pure functionals, for TPSSh and becoming largely negative for BHHLYP, that overestimates the transition energies by an average 0.42 eV (0.38 eV when CAS-PT2 values are selected as reference). Pinpointing the four GH presenting the MSE the closest to zero (X3LYP, B98, mPW1PW91, and PBE0) allows to state that functionals containing between 22% and 25% of EE are on the spot. As the exact same four GH provide the smallest MAE (see Figure 1), the smallest RMS and the largest R^2, and as these findings hold for both sets of reference values ("best estimates" and CAS-PT2/TZVP), these methods can be considered as best choice, at least for small molecules. The MAE of all LCH with an $\omega = 0.33$ au is close to 0.3 eV, due to an overestimation of the transition energies. We note that LC-ωPBE, characterized by a larger range-separation parameter, has a larger MAE (0.46 eV), while LC-ωPBE(20) relying on a less rapidly increasing fraction of EE presents a small and positive MSE. While CAM-B3LYP allows more accurate results than LC-BLYP, it remains slightly less efficient than B3LYP. This contradicts the results of Peach and co-workers.[55] A possible explanation of this discrepancy is the small size of the VT molecules: charge-transfer states are not significantly represented, which penalizes the CAM model.

For a more thorough discussion of the accuracy within each chemical family of the VT set, we refer the reader to ref 62 in which BP86, B3LYP, and BHHLYP performances are discussed in detail. The conclusions can be extended to other pure functionals and global hybrids. Nevertheless, as LCH have not been used before for the same set of molecules, it is probably worth discussing further this category of functionals. For the acene series, the B_u states are accurately estimated by LCH, including the correct evolution with oligomer length. For instance, the vertical transition energy to the first B_u state decreases by 1.52 eV (1.77 eV with CAS-PT2) from butadiene to octatetraene, a value nicely reproduced by CAM-B3LYP (1.65 eV). This conclusion is in agreement with the investigation of Peach et al.,[55] though they selected different ground-state geometries and reference values. On the contrary, for the A_g states, that are already poorly described by GH, the LCH are of no help and yield very large errors, e.g. LC-TPSS is 1.97 eV off for hexatriene. Therefore, it appears that all monodeterminantal DFT approaches fail to recover the correct ordering of the A_g and B_u states, at least for medium-sized polyacetylene oligomers.[62,64] For unsaturated cyclic hydrocarbons, LCH tend to outperform the commonly used GH. This is especially striking for norbornadiene for which CAM-B3LYP provides A_2/B_2 states at 5.13/6.03 eV in good agreement with the reference data (5.34/6.11 eV), whereas PBE0 strongly undervalues these energies (4.91/5.67 eV). For benzene and naphthalene, we found that CAM-B3LYP transitions energies are larger than their B3LYP counterpart for the low-lying states, exactly as in ref 55. This difference between LCH and GH pertains for higher transitions, leading to large deviations for the higher-excited states that are strongly overestimated with all LCH [but LC-ωPBE(20)] but reasonably reproduced with GH like B3LYP or PBE0. This latter finding follows our previous work on Rydberg states.[113] For the heterocyclic structures, the LCH do not cure the most significant GH deficiencies. For instance the first B_{2u} states of pyrazine is located at 5.44 eV with PBE0 and 5.40 eV with LC-PBE but at 4.64 eV with the best wave function scheme (4.85 eV with CAS-PT2). Likewise, the first $\pi \rightarrow \pi^\star$ state of *s*-triazine should be close to 5.79 eV but is overvalued by all hybrids (PBE0: 6.24 eV and LC-PBE: 6.21 eV). In addition, we note that the $n \rightarrow \pi^\star$ states are extremely sensitive to the EE percentage. Indeed, for the 20 $n \rightarrow \pi^\star$ transitions of the heterocycle subset, the MSE (MAE) is 0.66 eV (0.66 eV) with BLYP, 0.13 eV (0.16 eV) with B3LYP, −0.56 eV (0.56 eV) with BHHLYP, −0.09 eV (0.15 eV) with LC-BLYP, and −0.18 eV (0.21 eV) with CAM-B3LYP. For these 20 states, PBE0 and mPW1PW91 give the smallest MAE (0.13 eV), while, on the contrary, LC-ωPBE(20) that was efficient on average is not appropriate (MAE of 0.36 eV). For aldehydes and ketones, LC-GGA, LC-*meta*GGA, and CAM-B3LYP give accurate estimates of the transition energies. The MAE of CAM-B3LYP for this subset is limited to 0.15 eV with only one of the high-energy states of benzoquinone being poorly evaluated. For the amides, the $n \rightarrow \pi^\star$ are correctly evaluated by all LCH, but the $\pi \rightarrow \pi^\star$ transition energies are significantly overestimated. Eventually, for the four nucleobases, the largest systems of the VT set, we found that the MAE of LCH are significantly larger than these of 25%-GH (LC-OLYP: 0.31 eV, CAM-B3LYP: 0.25 eV versus PBE0: 0.09 eV), the LCH's errors being largest for $n \rightarrow \pi^\star$ transitions. For the low-lying states of uracil, a more complete investigation of the merits of different (PCM-)TD-DFT approaches has recently been published by Improta and Barone.[114] Overall, it is worth pointing out that, for the small molecules of the VT set, LCH tend to behave like GH containing a sizable fraction of EE. This is quite obvious by comparing the CAM-B3LYP and BMK columns in the Supporting Information. The similarity is large enough so that the mean absolute difference between the two sets of data is limited to 0.11 eV, whereas the R^2 relating the results of the two functionals attains 0.99.

Previous works on the same set of molecules also used different approaches.[62,64] The DFT MR/CI scheme of ref 62 is more accurate than the four best GH listed above (X3LYP, B98, PBE0, and mPW1PW91), though the differences remains trifling, the multireference MAE being 0.22 eV (instead of 0.24 eV) and the related R^2 reaching 0.96 (instead of 0.95). Determining if these improvements justify the computational effort related to the DFT MR/CI approach probably depends of the nature of the case under scrutiny. CC2 yields larger MAE (0.30 eV) but better correlation ($R^2 = 0.97$)[8] than the 22−25% functionals, clearly hinting that

CC2 cannot be viewed as "systematically more accurate" than TD-DFT, nor can it be used to benchmark DFT functionals without adequate testing. Of course, CC2 values are expected to be more basis set dependent than the TD-DFT ones, and our conclusion holds only for medium-size basis sets. However, using the CAS-PT2/TZVP as reference, we obtain a MAE of 0.27 eV for CC2/TZVP and 0.26 eV for PBE0/TZVP, meaning that their performances are extremely similar. At the very least, these results indicate that, although CC2 remains a method of choice to tackle specific problems, its blind and straightforward application does not always guarantee outperforming TD-DFT. Starting with the B2LYP and B2PLYP of ref 64 we have performed a statistical analysis for the same set of reference, that is the "best estimates" not the CAS-PT2 values. We obtain rather poor results with B2LYP (MSE=0.45 eV, MAE=0.52 eV, rms=0.62 eV and R^2=0.90) but very accurate ones for B2PLYP (MSE=0.01 eV, MAE=0.18 eV, rms=0.25 eV and R^2=0.97). These findings are in perfect agreement with the conclusions of Grimme and co-workers:[64] it indeed appears that B2PLYP surpasses significantly all other hybrids (including the DFT MR/CI scheme) as well as CC2. This indicates that such double-hybrid functional might allow the taking of the inner track for accurate spectroscopic estimates.

3.2. The Complete VE Set. A statistical analysis performed for the 29 functionals applied on the 614 excited-states of the VE set can be found in Table 2. We remind that the molecules of this set tend to be (much) larger than in the VT set. The mean signed error (MSE) indicates that pure functionals tend to underestimate the transition energies by approximatively 0.3 eV, though, as in the VT set, the errors are significantly smaller with two of the three *meta*-GGA: VSXC (0.15 eV) and TPSS (0.20 eV). Including a small fraction of exact exchange is sufficient to be much closer to the spot, as illustrated by TPSSh (MSE = 0.05 eV). In fact, the MSE varies quite steadily with the amount of EE included in the GH, being close to zero for about 21% of EE, slightly negative with 25%, and strongly negative for BMK, BHHLYP, and M05-2X. All GH containing between 15% and 27% of EE deliver |MSE| below the 0.10 eV mark and could therefore be considered as satisfactory for this criterion, especially B3LYP and X3LYP. For most LCH, the MSE are similar to those of pure functionals, though with the opposed sign: standard LCH tend to overestimate the transition energies. The error is only acceptable for LC-ωPBE(20) that uses significantly less EE, confirming the conclusions of the Herbert's group.[57,58] On the other hand, we found that CAM-B3LYP, that is characterized by a smaller fraction of EE at a large interelectronic distance (65%), appears more accurate than Hirao's LC-GGA methods, though it still yields a sizable MSE (−0.25 eV). Let us now turn toward the MAE and RMS errors. If HF is clearly an extremely poor approximation for transition energies (MAE of 0.85 eV), the errors remain large for all pure functionals, ranging from 0.32 eV (VSXC) to 0.41 eV (SVWN5). Not surprisingly, they are significantly reduced by adding EE, the minimal MAE for GH being obtained for functionals containing between 22% and 25% of EE (X3LYP, B98, PBE0 and mPW1PW91). The exact same four approaches have been found most efficient for the VT set. This percentage can therefore be viewed as optimal for computing transition energies of organic derivatives. This is good news: the same amount of EE yields accurate ground-state geometries and spectroscopic properties for the same kind of compounds. LC-ωPBE(20) delivers the same MAE as the best GH, in agreement with refs 57 and 58, whereas CAM-B3LYP remains acceptable (MAE of 0.30 eV). On the contrary, all other LCH yield too large MAE and RMS and could probably be discarded. These trends can be further rationalized by considering the selected error profiles depicted in Figure 2. One clearly notes that the error profile of BP86 is quite loose and moved to the right (too small transition energies). VSXC improves the pattern, confirming that it is one of the most satisfactory pure functional. TPSSh and PBE0 profiles are centered close to zero and are much more tight, the most probable PBE0 error being close to zero. The LC-ωPBE(20) profile is even tighter explaining the small rms, while the CAM-B3LYP sketch remains tight but is clearly unbalanced to the left (too large transition energies).

In Table 2, one notes that the R^2 obtained through linear regression are quite large as could be expected for a broad set of transitions ranging experimentally from 1.56 to 10.27 eV! Obviously, pure functionals are systematically character-

Table 2. Statistical Analysis for the Full VE Set (614 Excited-States)[a]

	before fitting			after linear regression		
functional	MSE	MAE	RMS	R^2	MAE	RMS
HF	−0.82	0.85	0.96	0.88	0.37	0.48
SVWN5	0.32	0.41	0.48	0.94	0.28	0.35
BLYP	0.32	0.40	0.47	0.94	0.28	0.35
BP86	0.29	0.38	0.46	0.94	0.28	0.34
OLYP	0.29	0.38	0.45	0.94	0.28	0.35
PBE	0.29	0.39	0.46	0.94	0.28	0.34
VSXC	0.15	0.32	0.39	0.94	0.27	0.34
τ-HCTH	0.27	0.37	0.44	0.94	0.28	0.34
TPSS	0.20	0.34	0.41	0.94	0.27	0.34
TPSSh	0.05	0.26	0.32	0.95	0.24	0.30
O3LYP	0.11	0.26	0.32	0.95	0.24	0.30
τ-HCTH-hyb	0.06	0.24	0.31	0.96	0.23	0.29
B3LYP	0.01	0.23	0.29	0.96	0.22	0.28
X3LYP	−0.01	0.22	0.28	0.96	0.22	0.28
B98	−0.04	0.22	0.29	0.96	0.22	0.28
PBE0	−0.08	0.22	0.29	0.96	0.21	0.27
mPW1PW91	−0.08	0.22	0.29	0.96	0.21	0.27
M05	−0.02	0.25	0.31	0.95	0.25	0.30
BMK	−0.26	0.32	0.39	0.96	0.22	0.27
BHHLYP	−0.36	0.40	0.47	0.95	0.23	0.29
M05−2X	−0.29	0.38	0.45	0.95	0.25	0.32
LC-ωPBE(20)	−0.08	0.22	0.27	0.96	0.20	0.26
LC-BLYP	−0.31	0.35	0.41	0.96	0.22	0.27
LC-OLYP	−0.34	0.37	0.43	0.96	0.21	0.27
LC-PBE	−0.34	0.38	0.44	0.96	0.21	0.26
LC-τ-HCTH	−0.32	0.36	0.42	0.96	0.22	0.28
LC-TPSS	−0.38	0.40	0.46	0.96	0.21	0.26
LC-ωPBE	−0.46	0.48	0.54	0.96	0.23	0.29
CAM-B3LYP	−0.25	0.30	0.36	0.96	0.21	0.26
MLR, eq 2				0.98	0.16	0.20
MLR-P, eq 3				0.97	0.20	0.25
MLR-B, eq 4				0.97	0.20	0.26

[a] MSE stands for the mean signed error (experiment-theory), MAE stands for the mean absolute error, and RMS is the residual mean-squared error. At the bottom of the table, the results obtained through MLR are detailed (see the text for more details). All values are in eV.

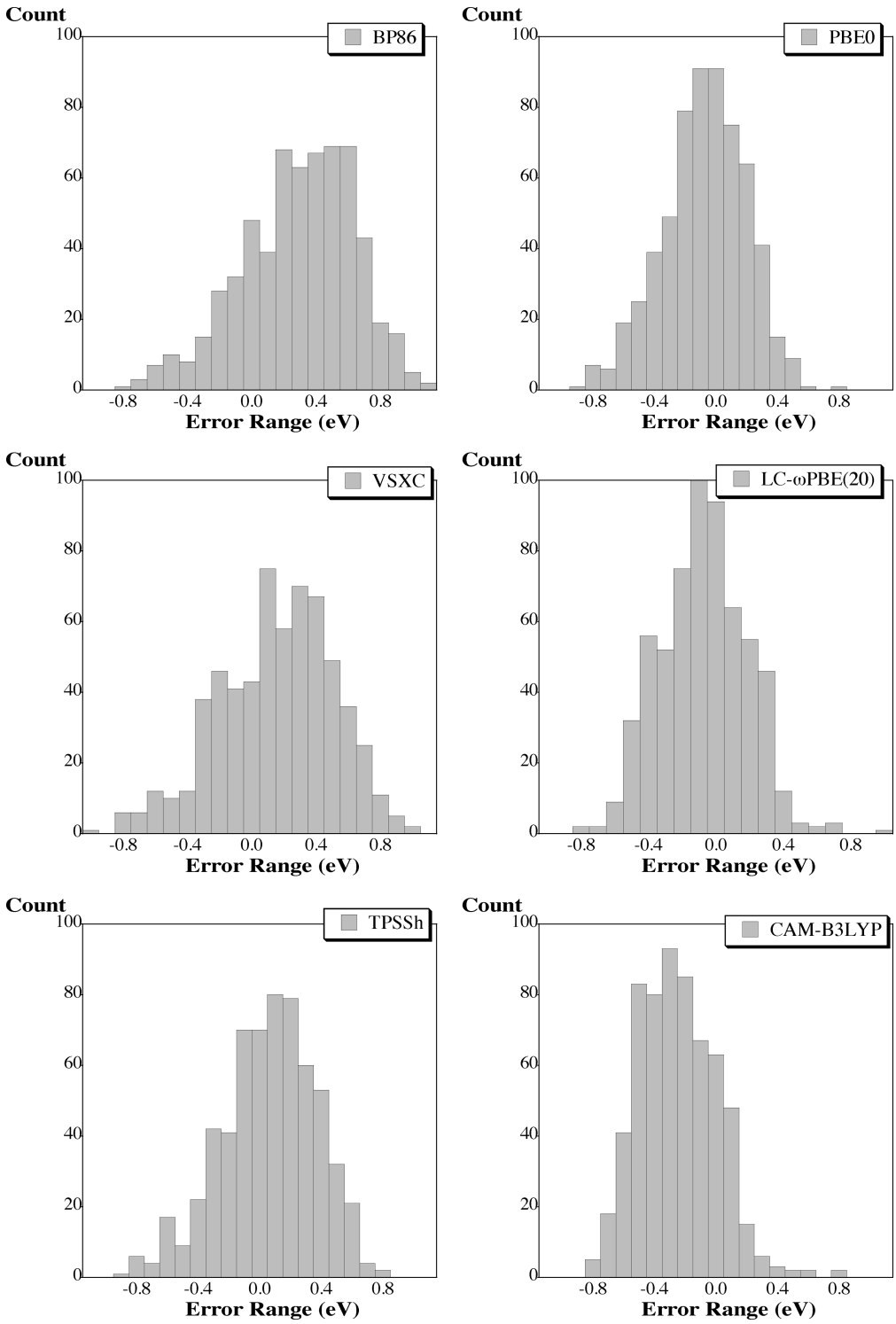


Figure 2. Histogram of the errors (eV) computed on the VE set for six representative functionals.

ized by smaller R^2. Subsequently, linearly correcting the raw GGA estimates decreases the MAE but not to a level competitive with the one obtained with the best unfitted hybrids: when hybrids are at hand, they have to be preferred. All GH and LCH grant a R^2 of 0.95 or 0.96 and a MAE equal or smaller than 0.25 eV when a simple linear correction is applied. Actually, after fitting, all commonly used hybrids deliver similar results, the most accurate results being achieved, on the one hand, with LC-ωPBE(20) for LCH and, on the other hand, with PBE0 and mPWPW91 for the GH. Let us point out that performing such simple linear regression has an insignificant effect for these "best" functionals, the MAE and rms decreasing by only 0.01 or 0.02 eV. In an attempt to improve the predictive accuracy of TD-DFT, we have associated the results of several functionals through multiple linear regression (MLR).[115−117] Indeed, combining the results obtained through several functionals has been found extremely efficient for anthraquinone derivatives.[25,118,119] In addition, the present set is large enough so to allow meaningful statistical analysis. By removing step-by-step the less significant functionals, we obtained a MLR equation relying on seven functionals

$$E^{\mathrm{MLR}} = -0.17 - 1.49E^{\mathrm{PBE}} + 1.55E^{\tau-\mathrm{HCTH}} + 1.62E^{\mathrm{mPW1PW91}} - 1.18E^{\mathrm{M05}} - 0.56E^{\mathrm{BMK}} + 1.99E^{\mathrm{LC-PBE}} - 0.99E^{\mathrm{LC}-\tau-\mathrm{HCTH}} \quad (2)$$

that outperforms any of the simple linear regression by providing a R^2 of 0.98, a MAE of 0.16 eV, and a rms of 0.20 eV. For the record, we note that the *P*-value in the ANOVA table indicates that there is a statistically significant relationship between theory and experiment at the 99% confidence level, while all seven functional coefficients are also significant at the 99% confidence level. Of course using eq 2 requires to roughly multiply the computational effort by a factor of 7 for a gain of only ∼20% of MAE and ∼23% of rms, compared to the best SLR. Such performance is unlikely to be perceived as efficient for practical applications: there often exists more powerful approaches for such computational cost. For this reason, we have built two other MLR combining one global hybrid and one long-range-corrected hybrid of the Perdew and Becke's families

$$E^{\mathrm{MLR-P}} = -0.10 + 0.44E^{\mathrm{PBE0}} + 0.53E^{\mathrm{LC-PBE}} \quad (3)$$

$$E^{\mathrm{MLR-B}} = -0.13 + 0.24E^{\mathrm{B3LYP}} + 0.74E^{\mathrm{CAM-B3LYP}} \quad (4)$$

For both equations, the sum of the coefficients is close to one, confirming that the raw evolution (slope) provided by TD-DFT is reasonable. These equations improve the R^2 to 0.97 (compares to 0.96 for SLR) but only slightly tune the computed MAE and rms (see Table 2). Indeed, eq 3 provides a MAE and RMS of 0.20 and 0.25 eV, respectively, only 0.02 eV smaller than the one of the raw LC-ωPBE(20) values. Subsequently, using one of the above equations for correcting the transition energies computed on a new molecular structure is unnecessary, though statistical procedures can be very useful if one investigates a single family of compounds.

As an intermediate conclusion of the statistical analysis of the VT and VE sets, we can already state that the "expected TD-DFT" error for an unknown singlet excited-state should be close to 0.22 eV, in quite good agreement with the previous works mentioned in the Introduction. Such accuracy could be obtained by using, either a GH containing between 22% and 25% of EE (X3LYP, B98, PBE0, mPW1PW91) or with a LCH with a relatively small damping parameter ($\omega \simeq 20$), such as LC-ωPBE(20). More specifically, we have to point out that X3LYP yields the smallest MSE, MAE, and rms of all GH for the VE set. Performing a linear correction on the raw computed values levels out the results obtained with all hybrids but does not top the "best" hybrids mentioned above.

3.3. Analysis of VE Subsets. Statistical analysis for various VE subsets can be found in the Supporting Information. As $\pi \rightarrow \pi^\star$ transitions constitute the major portion of the excited-states in the full VE set, it is not surprising that the errors found for these 510 states are very similar to these of Table 2. For the 79 $n \rightarrow \pi^\star$ transitions considered, one finds the smallest deviations for functionals containing between 20% and 25% of exact exchange (B3LYP, X3LYP, B98, PBE0, mPW1PW91) that yield nearly zero MSE as well as MAE and rms close to 0.15 eV. LCH also deliver small MAE, though slightly larger MSE and RMS than GH. The smaller errors for $n \rightarrow \pi^\star$ than for $\pi \rightarrow \pi^\star$ transition energies is probably related to the more local character of the former in our set. The correlation coefficient being also large for all these schemes, these findings support the conclusions of our previous work:[60] $n \rightarrow \pi^\star$ states could be accurately described by both GH and LCH.

For the subset of charged molecules, significantly larger errors appear, e.g. MAE of 0.34 eV with PBE0 and 0.37 eV with LC-ωPBE(20), due to the large overestimation of the transition energies with otherwise-successful 20−25% GH. Consequently, the inaccuracies are minimal with pure functionals or hybrids containing a very small fraction of EE. This phenomenon is explained by the dominance of cyanine structures in the subset of charged molecules: cyanine, triphenylmethane, and acridine derivatives represent more than half of the transitions investigated. These compounds present a strong multideterminantal nature, at least for systems with more than two or three double bonds,[101] and all DFT functionals are inadequate.[65,102] Therefore, the large MAE listed in the Supporting Information are related to the nature of the molecule, rather than to the presence of a charge. Indeed, by removing the cyanine-like structures from the set, one obtains a MAE of 0.26 eV and a RMS of 0.33 eV for PBE0, similar to the one listed in Table 2. This is also illustrated by anionic hydrazones, for which GH obviously work more accurately than pure functionals. For neutral molecules (502 states), the conclusion follows the one obtained for the full VE set, with minimal discrepancies and maximal correlation coefficients for X3LYP, B98, PBE0, and mPW1PW91 for GH and LC-ωPBE(20) for LCH. This latter functionals delivers a MAE as small as 0.18 eV, illustrating, on the one hand, its efficiency for monodeterminantal structures and, on the other hand, the possibility of going significantly below a ∼0.3 eV average error with TD-DFT, even when absolutely no statistical treatment of the results is performed.

For the 178 states belonging to the family of $\pi \rightarrow \pi^\star$ chromophores, the errors are completely similar to these obtained in our previous work relying on a less extensive set of chromogens,[59] with a MAE of 0.46 eV for PBE (0.45 eV in ref 59), 0.14 eV for PBE0 (0.14 eV in ref 59), and 0.25 eV for CAM-B3LYP (0.26 eV in ref 59). From the tables in the Supporting Information, it is striking that, within our functional list, PBE0, mPW1PW91, and LC-ωPBE(20) deliver the smallest MSE, MAE, and rms; the latter functional additionally providing the largest R^2. For $n \rightarrow \pi^\star$ chromophores (nitrosos, thiocarbonyls, and azobenzenes), the errors are small with all functionals, especially with LC-BLYP, LC-OLYP, and CAM-B3LYP (MAE of 0.07 eV) that also lead to excellent corrections (R^2 of 0.99). The best GH presenting between 20% and 25% of EE are also on the spot (MSE smaller than 0.03 eV and MAE between 0.10 and 0.13 eV), though the correlation with experiment is slightly less impressive (R^2 of 0.99). On the contrary, we have to point out the comparatively large errors of M05 and M05-2X. If one searches for the smallest errors for the neutral dye set (228 states, both types of transitions being incorporated), one

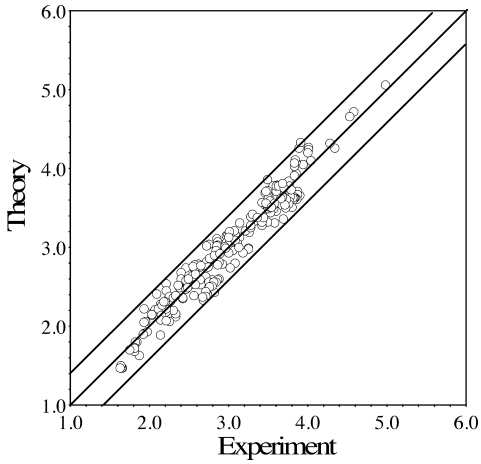


Figure 3. Comparison between the LC-ωPBE(20) and experimental transition energies (eV) for neutral dyes. The central lines indicate a perfect theory/experiment match, whereas the side lines are borders for ±0.4 eV deviations.

finds MAE of 0.14 eV for PBE0, mPW1PW91, and LC-ωPBE(20). Figure 3 compares the LC-ωPBE(20) and experimental transition energies for these 228 states, and the nice match is obvious: only one case presents a deviation larger than 0.4 eV. For the record, B3LYP (CAM-B3LYP) gives MAE of 0.16 eV (0.21 eV) for the same set of derivatives. In that sense, PBE0 remains perfectly adequate for investigating neutral dyes, clearly supporting the conclusions of our previous investigations on specific families of organic dyes or photochroms.[5,61,95,108,120−126] For sure, there is a partial error compensation between the lack of vibronic contribution in our model and the form of the functional (see the next section), but our calculations confirm the practical ability of the PCM-TD-PBE0 scheme, even when no fitting procedure is activated. For charged dyes, the same conclusion as above pertains: the errors are large due to the multideterminantal nature of both the ground- and excited-states in many cyanine compounds. Clearly TD-DFT can only be used for such systems to obtain qualitative knowledge or to compare systems with the same delocalization length. In that case, fitting the results might allow much more accurate estimates. For instance for arylcarbonium (**AC-a** in the Supporting Information), the B3LYP (CAM-B3LYP) MAE is 0.29 eV (0.54 eV) prior to linear correction but only 0.10 eV (0.04 eV) afterward, although these systems are characterized by two strong absorptions with different physical nature. Note that the impressive performance of CAM-B3LYP after fitting is related to the better consistency of LCH (compared to GH) when only one family of molecules is considered. This effect has already been pointed our previously.[59,102]

For hydrocabons (the set contains both aromatic and aliphatic conjugated compounds), the two 25%-GH and especially LC-ωPBE(20) remain the most efficient (MAE of 0.18 eV and R^2 of 0.96 for the latter), but with errors significantly larger than for neutral dyes. Another striking evolution wrt neutral dyes is that functionals containing a large fraction of EE (BMK, BHHLYP, CAM-B3LYP, ...) are fairly accurate for this subset with relatively small MAE, e.g. 0.23 eV for BMK and 0.22 eV for CAM-B3LYP. This is in good agreement with the work of Dierksen and Grimme,[63] that concluded that a EE percentage between 30%−40% should be optimal (on average) for spectral calculations on similar compounds. For the 126 excited-states measured on small structures (five or six member rings as well as molecules with less than 14 atoms), the MSE are very close to zero with VSXC, τ-HCTH-hyb, and LC-ωPBE(20). All GH containing between 10 and 25% of EE and LC-ωPBE(20) yielding MAE smaller than 0.30 eV, but not below 0.26 eV, whereas the difference between GH and LCH is more limited for small molecules than for dyes or hydrocarbons. In fact, the vertical TD-DFT approach appears significantly less potent for predicting the λ_{max} of these small structures than for other sets, partly explaining the nonidentical conclusions obtained by different groups previously (see the Introduction). In any case, for small molecules, LC-ωPBE(20), clearly emerges as the most accurate approach with the smallest deviations and the largest correlation coefficient. The biomolecules treated here (33 states) are characterized by a small chromophoric unit, though they can be large molecules. Consequently, they behave analogously to the small structures with minimal theory/experiment deviations for hybrids presenting a small fraction of EE.

In Figure 4, we report, for five functionals of the Becke's family (BLYP, B3LYP, BHHLYP, LC-BLYP, and CAM-B3LYP), the evolution of the MSE, MAE, and RMS for small, medium, and large chromogens. In the Supporting Information, these subsets respectively correspond to the 33 states computed on small molecules (**SC**, less than 14 atoms), 51 states calculated for medium chemicals (**MC**, between 14 and 29 atoms), and 32 states obtained for large compounds (**LC**, more than 30 atoms). Due to the difficulty to build generic and consistent sets for molecules of various sizes, the analysis of Figure 4 allows only qualitative conclusions. Nevertheless, it is clear that, for a given functional, the MSE tend to become more negative (or less positive) as the size of the molecules increases, e.g. for B3LYP the MSE equals 0.09 eV, −0.01, and −0.07 eV for small, medium, and large chromophores, respectively. In other words, as the size of the molecule increases, functionals with a larger EE fraction tend to produce larger errors, which is consistent with our above analysis for the different subsets. Indeed, the MAE and rms of B3LYP, LC-BLYP, and CAM-B3LYP are similar for **SC** but decrease for the former when going to **MC** and **LC**, whereas the average errors of the two LCH tend to increase for larger compounds. Therefore, the excited-states of very extended molecules are not necessarily better described by long-range-approaches, at least in the vertical approximation (see the next section). For these three subsets, B3LYP clearly outperforms the other functionals, as it (almost) systematically yields the smallest deviations. However, it is striking that no functional has a flat error profile for different size of molecules, as one would fancy.

3.4. Importance of Vibronic Effects. As we pointed out in the methodological section, the main weakness of our approach is the lack of vibronic modeling in the VE set: vertical transitions do not physically correspond to λ_{max}. Despite the practical computational advantage of vertical calculations, it is worth estimating the impact of this

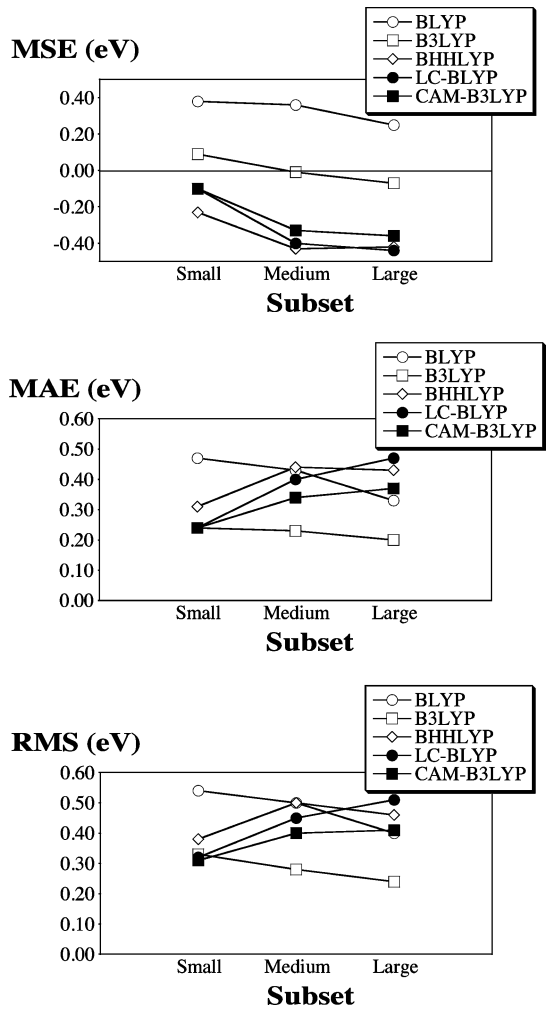


Figure 4. Evolution of the MSE (top panel), MAE (central panel), and RMS (bottom panel) calculated for the transition energies of small (**SC**), medium (**MC**), and large (**LC**) chromogens.

approximation. A first procedure to tackle this problem is to compare the errors obtained for the VT set (no vibronic effects, purely vertical reference values) and for similar molecules (i.e., small structures) of the VE set (vibronic effects neglected). Such an inspection not only highlights similarities, as the most accurate functionals in terms of MAE and rms are the same in both sets, but also sheds light on the discrepancies, the zero of the MSE being obtained for an EE percentage of ~23% in the VT set but ~14% in the VE subset. Nevertheless, both sets appear to present alike error pattern wrt the chosen functional. Such comparisons are however limited because the transitions considered are different: in the VE case, only bright low-lying states are included.

An alternative scheme is to define a subset of VE molecules with clearly measured 0−0 transition (see tables in the Supporting Information). This leads to 75 states measured for hydrocarbons and oligomers as well as medium and large chromogens (**FS**, **HC**, **OL**, **MC**, and **LC**), that is molecular structures alike the one of ref 63.[127] For this subset, when performing a statistical analysis using the experimental λ_{max} as reference, one obtains MSE and MAE similar to those obtained for the full set of neutral molecules (see Table 3).

Table 3. Analysis of the Importance of Vibronic Effects for Eight Selected Functionals[a]

	vertical vs λ_{max}		vertical vs 0−0		corrected vertical vs 0−0	
functional	MSE	MAE	MSE	MAE	MSE	MAE
BLYP	0.44	0.48	0.35	0.40	0.57	0.58
VSXC	0.29	0.39	0.20	0.33	0.42	0.47
B3LYP	0.13	0.28	0.04	0.24	0.32	0.37
PBE0	0.03	0.24	−0.06	0.22	0.24	0.32
BHHLYP	−0.23	0.30	−0.32	0.34	0.05	0.26
LC-ωPBE(20)	−0.05	0.20	−0.14	0.20	0.16	0.24
LC-TPSS	−0.36	0.36	−0.45	0.45	−0.09	0.18
CAM-B3LYP	−0.19	0.26	−0.28	0.30	0.04	0.21

[a] All values are in eV and are calculated on the 75 states for which experimental 0-0 transitions could be defined. See the text of section 3.4 for more details.

The two most accurate functionals remain PBE0 and LC-ωPBE(20) with MAE of 0.24 and 0.20 eV, respectively. For the same set of excited-states, one can use the measured 0−0 transitions as references for benchmarking functionals. These 0−0 peaks either correspond to the λ_{max} or appear at smaller transition energies than the maximal absorption, the average experimental difference between these two peaks attaining 0.09 eV for the present set. Therefore functionals incorporating less EE, that statistically reduce the gap between ground and excited-states, tend to produce smaller deviations when the 0−0 absorption is used as reference. This is illustrated by the difference between the "vertical vs λ_{max}" and the "vertical vs 0−0" columns of Table 3. Indeed, the MAE of VSXC is reduced from 0.39 to 0.33 eV, whereas the MSE of PBE0 changes from positive (0.03 eV) to negative (−0.06 eV). Nevertheless, the minimal deviations for the selected functionals are once more reached with PBE0 and LC-ωPBE(20). Of course, if a proper Franck−Condon calculation was performed, the calculated transition energies would be smaller than our vertical values. To estimate this effect, we have analyzed the theoretical data reported in ref 63, and we have found that the TD-DFT vertical transition energies of closed-shell molecules are larger than their computed 0−0 counterpart by an average of 0.22 eV for BLYP, 0.28 eV for B3LYP, and 0.37 eV for BHHLYP.[128] There is therefore a smooth relationship between this difference and the EE fraction: ~+0.03 eV per 10% of EE. This allowed us to very roughly estimate the average difference for the other functionals.[129] By shifting our vertical transition energies by this average values, we obtain a qualitative information about the impact of the vertical model ("corrected vertical vs 0−0" columns in Table 3). We are well aware that this represents a extremely crude approximation, as we incorrectly apply a constant shift to all molecules (they are significantly different in practice), but general trends may still emerge. The smallest MAE are now obtained with global hybrids including a large fraction of EE and LCH that become closer to the spot. On the contrary, B3LYP and PBE0 present much larger MSE and MAE, BHHLYP outperforming them significantly. For the record, we note that LC-ωPBE(20) apparently remains satisfactory with a MSE of 0.16 eV and a MAE of 0.24 eV. Of course, as LCH tend to deliver poorer geometries than global hybrids, their use for vibronic calculation certainly remains an open question.

4. Conclusions

Benchmark calculations aiming at identifying the most efficient functionals for TD-DFT calculations have been performed. Large panels of pure functionals, global hybrids, and long-range-corrected hybrids have been tested for more than 700 excited-states. Assessments have been performed using both highly correlated wave function results and experimental wavelengths as references. It appears that the most accurate estimates are obtained, by using a GH containing between 22% and 25% of EE (X3LYP, B98, PBE0, mPW1PW91) or a LCH with a small damping parameter (LC-ωPBE(20), with $\omega = 0.20$). The four GH provide a mean absolute error smaller than 0.25 eV for both types of benchmarks, although the training set includes compounds known to be difficultly described by TD-DFT. GH containing less (more) exact exchange tend to underestimate (overestimate) the transition energies, LCH with large damping parameter suffering the same problem as global hybrids with 40−50% of exact exchange. For almost all cases, the errors obtained by LDA and GGA are about 50% larger than with the 25%-GH. *Meta*-GGA, especially VSXC and TPSS, yield transition energies in better agreement with reference data than other pure functionals, though they cannot outperform the hybrids. CAM-B3LYP appears to be one of the most satisfying LCH, although the deviations with respect to experiment are larger than with B3LYP for most compounds. The accuracy significantly depends on the set of molecules considered, the errors being very large for cyanine-like derivatives but smaller than average for neutral molecules, dyes, and $n \rightarrow \pi^\star$ excited-states. Indeed, for neutral molecules, the best choice, namely the LC-ωPBE(20) LCH, provides a MAE as small as 0.18 eV, whereas the errors are even smaller for organic dyes: 0.14 eV with PBE0, mPW1PW91, and LC-ωPBE(20). It also turned out that functionals are sensitive to the size of the chromogens investigated (Figure 4), whereas a crude estimation of vibronic effects revealed that hybrids including a large fraction of EE (e.g., BHHLYP) may be more accurate in the framework of Franck−Condon applications than for vertical estimates.

Acknowledgment. D.J. and E.A.P. thank the Belgian National Fund for Scientific Research for their research associate and senior research associate positions, respectively. The authors thank Prof. G. E. Scuseria for the use of LCH and are deeply indebted to Dr. I. Ciofini for many fruitful discussions. They also acknowledge the constructive comments of the two anonymous referees. Several calculations have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the "Loterie Nationale" for the convention number 2.4578.02 and of the FUNDP. The collaboration between the Belgian and French group is supported by the *Wallonie-Bruxelles International*, the *Fonds de la Recherche Scientfique*, the *Ministère Français des Affaires étrangères et européennes*, the *Ministère de l'Enseignement supérieur et de la Recherche* in the framework of Hubert Curien Partnership.

Supporting Information Available: Statistical analysis for various sets of molecules, representation of all investigated chemicals, full tables with all transition energies, bibliographic information for the experimental references. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Adachi, M.; Nakamura, S. *Dyes Pigm.* **1991**, *17*, 287–296.

(2) Bacon, A. D.; Zerner, M. C. *Theor. Chim. Acta* **1979**, *53*, 21–54.

(3) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199–217.

(4) Caricato, M.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2004**, *108*, 6248–6256.

(5) Jacquemin, D.; Perpète, E. A. *Chem. Phys. Lett.* **2006**, *429*, 147–152.

(6) Stewart, J. J. P. *MOPAC2002*; Fujitsu Ltd.: Tokyo, Japan, 2001.

(7) Matsuura, M.; Sato, H.; Sotoyama, W.; Takahashi, A.; Sakurai, M. *J. Mol. Struct. (THEOCHEM)* **2008**, *860*, 119–127.

(8) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(9) Sauer, S. P. A.; Schreiber, M.; Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 555–564.

(10) Guillaumont, D.; Nakamura, S. *Dyes Pigm.* **2000**, *46*, 85–92.

(11) Serrano-Andrès, L.; Roos, B. O. *Chem.—Eur. J.* **1997**, *3*, 717–725.

(12) Fabian, J.; Diaz, L. A.; Seifert, G.; Niehaus, T. *J. Mol. Struct. (THEOCHEM)* **2002**, *594*, 41–53.

(13) Parac, M.; Grimme, S. *J. Phys. Chem. A* **2002**, *106*, 6844–6850.

(14) van Faasen, M.; Boeij, P. L. *J. Chem. Phys.* **2004**, *120*, 8353–8363.

(15) Blancafort, L.; Robb, M. A. *J. Phys. Chem. A* **2004**, *108*, 10609–10614.

(16) Blancafort, L.; Voityuk, A. A. *J. Phys. Chem. A* **2007**, *111*, 4714–4719.

(17) Sobolewski, A. L.; Shemesh, D.; Domcke, W. *J. Phys. Chem. A* **2009**, *113*, 542–550.

(18) Shemesh, D.; Sobolewski, A. L.; Domcke, W. *J. Am. Chem. Soc.* **2009**, *131*, 1374–1375.

(19) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

(20) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.

(21) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(22) Perdew, J. P.; Ruzsinsky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062001.

(23) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.

(24) Barone, V.; Polimeno, A. *Chem. Soc. Rev.* **2007**, *36*, 1724–1731.

(25) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326–334.

(26) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708–4717.

(27) Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 094107.

(28) Caricato, M.; Mennucci, J.; Tomasi, B.; Ingrosso, F.; Cammi, R.; Corni, S.; Scalmani, G. *J. Chem. Phys.* **2006**, *124*, 124520.

(29) Preat, J.; Loos, P. F.; Assfeld, X.; Jacquemin, D.; Perpète, E. A. *J. Mol. Struct. (THEOCHEM)* **2007**, *808*, 85–91.

(30) Bondar, A.-N.; Fischer, S.; Smith, J.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004**, *126*, 14668–14677.

(31) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.

(32) Curutchet, C.; Scholes, G. D.; Mennucci, B.; Cammi, R. *J. Phys. Chem. B* **2007**, *111*, 13253–13265.

(33) Loos, P.-F.; Preat, J.; Laurent, A. D.; Michaux, C.; Jacquemin, D.; Perpète, E. A.; Assfeld, X. *J. Chem. Theory Comput.* **2008**, *4*, 637–645.

(34) Wanko, M.; Hoffmann, M.; Frähmcke, J.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2008**, *112*, 11468–11478.

(35) Jacquemin, D.; Perpète, E. A.; Laurent, A. D.; Assfeld, X.; Adamo, C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1258–1262.

(36) Savin, A. In *Recent Developments and Applications of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996; Chapter *9*, pp 327−354.

(37) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(38) Tawada, T.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.

(39) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.

(40) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–56.

(41) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274–7280.

(42) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(43) Vydrov, O. A.; Heyd, J.; Krukau, V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(44) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.

(45) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpète, E. A. *J. Mol. Struct. (THEOCHEM)* **2005**, *731*, 67–72.

(46) Petit, L.; Quartarolo, A.; Adamo, C.; Russo, N. *J. Phys. Chem. B* **2006**, *110*, 2398–2404.

(47) Pezzella, A.; Panzella, L.; Crescenzi, O.; Napolitano, A.; Navaratman, S.; Edge, R.; Land, E.; Barone, V.; d'Ischia, M. *J. Am. Chem. Soc.* **2006**, *128*, 15490–15498.

(48) Jacquemin, D.; Wathelet, V.; Perpète, E. A. *J. Phys. Chem. A* **2006**, *110*, 9145–9152.

(49) Marian, C. M.; Gilka, N. *J. Chem. Theory Comput.* **2008**, *4*, 1501–1515.

(50) Wong, B. M.; Cordaro, J. G. *J. Chem. Phys.* **2008**, *129*, 214703.

(51) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.

(52) Mennucci, B.; Cappelli, C.; Guido, C. A.; Cammi, R.; Tomasi, J. *J. Phys. Chem. A* **2009**, *113*, 3009–3020.

(53) Tsuji, T.; Onoda, M.; Otani, Y.; Ohwada, T.; Nakajima, T.; Hirao, K. *Chem. Phys. Lett.* **2009**, *473*, 196–200.

(54) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(55) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

(56) Peach, M. J. G.; Cohen, A. J.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4543–4549.

(57) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.

(58) Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M. *J. Chem. Phys.* **2009**, *130*, 054112.

(59) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.

(60) Jacquemin, D.; Perpète, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 094102.

(61) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *Chem. Phys. Lett.* **2008**, *465*, 226–229.

(62) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(63) Dierksen, M.; Grimme, S. *J. Chem. Phys.* **2004**, *120*, 3544–3554.

(64) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.

(65) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

(66) Note that both values have been computed for the same set of states, that is excited-states for which a theoretical best estimate has been provided. This set is used in the following.

(67) 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 D.02 and E.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(68) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Kudin, K. N.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; 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.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Brothers, E.; Staroverov, V.; Kobayashi, R.; Normand, J.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; 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.; Farkasv, 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.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian DV, Revision H.01*; Gaussian, Inc.: Wallingford, CT, 2008.

(69) Slater, J. C. *Quantum Theory of Molecular ans Solids*; McGraw-Hill: New York, 1974; Vol. 4.

(70) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(71) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(72) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(73) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(74) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(75) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(76) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400–410.

(77) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559–9569.

(78) Tao, J.; Perdew, J.; Staroverov, V.; Scuseria, G. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(79) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(80) Baker, J.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 1441–1449.

(81) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(82) Xu, X.; Goddard III, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.

(83) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624–9631.

(84) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242–250.

(85) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(86) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(87) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(88) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405–3416.

(89) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(90) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

(91) Leininger, T.; Stoll, H.; Werner, H. J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.

(92) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

(93) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.

(94) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpète, E. A. *J. Chem. Phys.* **2006**, *124*, 074104.

(95) Preat, J.; Jacquemin, D.; Perpète, E. A. *Chem. Phys. Lett.* **2005**, *415*, 20–24.

(96) Jacquemin, D.; Preat, J.; Wathelet, V.; Fontaine, M.; Perpète, E. A. *J. Am. Chem. Soc.* **2006**, *128*, 2072–2083.

(97) Jacquemin, D.; Bouhy, M.; Perpète, E. A. *J. Chem. Phys.* **2006**, *124*, 204321.

(98) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpète, E. A. *Chem. Phys.* **2006**, *328*, 324–332.

(99) Jacquemin, D.; Perpète, E. A. *J. Mol. Struct. (THEOCHEM)* **2007**, *804*, 31–34.

(100) Jacquemin, D.; Perpète, E. A.; Adamo, C. *J. Mol. Struct. (THEOCHEM)* **2008**, *863*, 123–127.

(101) Schreiber, M.; Bub, V.; Fülscher, M. P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3906–3912.

(102) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(103) Christie, R. M. *Colour Chemistry*; The Royal Society of Chemistry: Cambridge, U.K., 1991; p 228.

(104) Zollinger, H. *Color Chemistry, Syntheses, Properties and Applications of Organic Dyes and Pigments*, 3rd ed.; Wiley-VCH: Weinheim, 2003; p 647.

(105) Note that we have selected the nonrelativistic values of the most recent paper as reference. The difference wrt to the relativistic estimates of the former contribution is completely negligible.

(106) Santoro, F.; Improta, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 084509.

(107) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **2000**, *330*, 152–160.

(108) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Theor. Chem. Acc.* **2008**, *120*, 405–410.

(109) Improta, R.; Barone, V.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 054103.

(110) Peach, M. J. G.; Tellgren, E.; Salek, P.; Helgaker, T.; Tozer, D. J. *J. Phys. Chem. A* **2007**, *111*, 11930–11935.

(111) Sancho, García.; Perez-Jimenez, A. J. *Phys. Chem. Chem. Phys.* **2007**, 5874–5879.

(112) The MAE computed between the "best estimates" and the CAS-PT2/TZVP value is limited to 0.09 eV, but it is difficult to judge if this error should mainly be ascribed to the diffuseless basis set or to the inherent limitations of CAS-PT2.

(113) Ciofini, I.; Adamo, C. *J. Phys. Chem. A* **2007**, *111*, 5549–5556.

(114) Improta, R.; Barone, V. *J. Mol. Struct. (THEOCHEM)* 2009. in press, doi: 10.1016/j.theochem.2009.02.021.

(115) *Statgraphics Plus 5.1.*; Manugistics Inc.: Herndon, Virginia, U.S.A., 2000.

(116) Dagnelie, P. *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique*; De Boeck and Larcier: Bruxelles and Paris, 1998; p 516.

(117) Dagnelie, P. *Statistique théorique et appliquée. Tome 2. Inférence statistique à une et deux dimensions*; De Boeck and Larcier: Bruxelles and Paris, 1998; p 664.

(118) Perpète, E. A.; Wathelet, V.; Preat, J.; Lambert, C.; Jacquemin, D. *J. Chem. Theory Comput.* **2006**, *2*, 434–440.

(119) Jacquemin, D.; Wathelet, V.; Preat, J.; Perpète, E. A. *Spectrochim. Acta, Part A* **2007**, *67*, 334–341.

(120) Briquet, L.; Vercauteren, D. P.; Perpète, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2006**, *417*, 190–195.

(121) Jacquemin, D.; Perpète, E. A. *Chem. Phys. Lett.* **2006**, *420*, 529–533.

(122) Preat, J.; Jacquemin, D.; Wathelet, V.; André, J. M.; Perpète, E. A. *J. Phys. Chem. A* **2006**, *110*, 8144–8150.

(123) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Assfeld, X.; Ciofini, I.; Adamo, C. *J. Chem. Phys.* **2006**, *125*, 164324.

(124) Briquet, L.; Vercauteren, D. P.; André, J. M.; Perpète, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2007**, *435*, 257–262.

(125) Perpète, E. A.; Maurel, F.; Jacquemin, D. *J. Phys. Chem. A* **2007**, *111*, 5528–5535.

(126) Preat, J.; Michaux, C.; Lewalle, A.; Perpète, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2008**, *451*, 37–42.

(127) In this reference, 30 states of neutral closed-shell molecules and 13 states for open-shell molecules, have been selected.

(128) For the full set, including open-shell structures, the vertical transition energies are larger than their 0−0 counterpart by an average 0.22 eV for BLYP, 0.25 eV for B3LYP, and 0.32 eV for BHHLYP.

(129) For VSXC, the same as for BLYP: 0.22 eV; for PBE0 and LC-ωPBE(20) that behave similarly: 0.30 eV; for LC-TPSS: 0.36 eV; and for CAM-B3LYP: 0.32 eV. These latter values have been estimated from the amount of exact exchange at intermediate interelectronic distance. For BLYP, B3LYP, and BHHLYP, we stick to Dierksen and Grimme's differences.

CT900298E

JCTC Journal of Chemical Theory and Computation

1-Octanol/Water Partition Coefficients of *n*-Alkanes from Molecular Simulations of Absolute Solvation Free Energies

Nuno M. Garrido,[†,‡] António J. Queimada,[†] Miguel Jorge,[†] Eugénia A. Macedo,[†] and Ioannis G. Economou*[,‡]

Laboratory of Separation and Reaction Engineering (LSRE), Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua do Dr. Roberto Frias, 4200-465 Porto, Portugal and Molecular Thermodynamics and Modeling of Materials Laboratory, Institute of Physical Chemistry, National Center for Scientific Research "Demokritos", GR-153 10, Aghia Paraskevi Attikis, Greece

Received April 30, 2009

Abstract: The 1-octanol/water partition coefficient is an important thermodynamic variable usually employed to understand and quantify the partitioning of solutes between aqueous and organic phases. It finds widespread use in many empirical correlations to evaluate the environmental fate of pollutants as well as in the design of pharmaceuticals. The experimental evaluation of 1-octanol/water partition coefficients is an expensive and time-consuming procedure, and thus, theoretical estimation methods are needed, particularly when a physical sample of the solute may not yet be available, such as in pharmaceutical screening. 1-Octanol/water partition coefficients can be obtained from Gibbs free energies of solvation of the solute in both the aqueous and the octanol phases. The accurate evaluation of free energy differences remains today a challenging problem in computational chemistry. In order to study the absolute solvation Gibbs free energies in 1-octanol, a solvent that can mimic many properties of important biological systems, free energy calculations for *n*-alkanes in the range C_1-C_8 were performed using molecular simulation techniques, following the thermodynamic integration approach. In the first part of this paper, we test different force fields by evaluating their performance in reproducing pure 1-octanol properties. It is concluded that all-atom force fields can provide good accuracy but at the cost of a higher computational time compared to that of the united-atom force fields. Recent versions of united-atom force fields, such as Gromos and TraPPE, provide satisfactory results and are, thus, useful alternatives to the more expensive all-atom models. In the second part of the paper, the Gibbs free energy of solvation in 1-octanol is calculated for several *n*-alkanes using three force fields to describe the solutes, namely Gromos, TraPPE, and OPLS-AA. Generally, the results obtained are in excellent agreement with the available experimental data and are of similar accuracy to commonly used QSPR models. Moreover, we have estimated the Gibbs free energy of hydration for the different compounds with the three force fields, reaching average deviations from experimental data of less than 0.2 kcal/mol for the case of the Gromos force field. Finally, we systematically compare different strategies to obtain the 1-octanol/water partition coefficient from the simulations. It is shown that a fully predictive method combining the Gromos force field in the aqueous phase and the OPLS-AA/TraPPE force field for the organic phase can give excellent predictions for *n*-alkanes up to C_8 with an absolute average deviation of 0.1 log *P* units to the experimental data.

10.1021/ct900214y CCC: $40.75
Published on Web 08/17/2009

1. Introduction

In several biochemical processes and for successful drug design strategies in the pharmaceutical industry, a correct understanding of the interactions of a given solute in both aqueous (hydrophilic) and biological (lipophilic) media is necessary.[1−4] Together with the Gibbs free energy of solute transfer, the corresponding partition coefficient between the 1-octanol and the water phases is probably the most important input parameter used in quantitative structure−property relationships (QSPR) to correlate and predict many solute properties.[5] Especially in the pharmaceutical industry, the prediction of drug partitioning, hydrophobicity, and even pharmacokinetic characteristics in biological systems can be quantified by expressions based on the 1-octanol/water partition coefficient[2,3,6] (commonly known as P or even log P). Furthermore, log P is also used as a measure of the activity of agrochemicals, the degree of purity in metallurgy, and the hydrophobicity in environmental problems. Partition coefficient data are also useful to estimate the solubility of a solute in a solvent.[7,8]

The partition coefficient of a solute between 1-octanol and water was first introduced in 1964 by Hansch and Fujita,[9] and since then, many different approaches have been developed in an attempt to estimate this property. In the beginning, mostly semiempirical approaches based on the sum of fragment contributions or atom-derived group equivalents were proposed.[1−3,10] Nowadays, fragment additive schemes remain a standard method to estimate solvation free energies and partition coefficients,[11] but the most common methods to estimate solvation properties are procedures based on QSPR that (cor)relate partition coefficients or solvation properties with other calculated or available molecular properties.[12−14] Although these methods are considerably fast and applicable to large databases of molecular structures, they require large multiparameter tables having the disadvantage that whenever new molecules/compounds are under study, these need to be similar to the ones contained in the training set. This is evidenced by the lack of existing parameters to calculate log P for new chemical groups.[15−17] In short, we can conclude that QSPR methods are statistically rather than physically based. Simulations based on linear response theory and molecular descriptors to derive empirical relationships for estimating log P values have been carried out by Duffy and Jorgensen.[18] Finally, approaches based on continuum models have also been investigated.[15,16,19]

Besides the above-mentioned estimation methods, the partition coefficient can also be obtained from experiments, by applying e.g., the shake-flask method[20−22] for generating the saturated liquid phases, followed by sampling and quantitative solute analysis (e.g., high-performance liquid chromatography[23]). Still, this can be a very expensive and time-consuming procedure, and thus, has limited practical use for product design, such as in pharmaceutical screening.

* Corresponding author. E-mail: economou@chem.demokritos.gr.
† Laboratory of Separation and Reaction Engineering.
‡ Molecular Thermodynamics and Modeling of Materials Laboratory.

A different approach to all of the above is to use information of the free energy of solvation in water and in octanol to estimate the partition coefficient. From Gibbs free energies of solvation in two different phases at temperature T, one can calculate the corresponding partition coefficient, according to the following expression:

$$\log P^{1-\text{octanol/water}} = \frac{\Delta_{\text{hyd}}G - \Delta_{\text{solv}}G}{2.303RT} \quad (1)$$

where $\Delta_{\text{hyd}}G$ is the hydration free energy, and $\Delta_{\text{solv}}G$ is the Gibbs free energy of solvation in 1-octanol. The first computational approaches involving this relationship go back to the 1980s.[24−26] Recent developments in simulation methods and increased computing power allow today the calculation of the absolute solvation free energies of complex molecules, such as amino acid analogues, directly from molecular simulations.[27−30] Thus, we propose here an innovative approach to predict the 1-octanol/water partition coefficient without (or at least with a minimum) experimental information, based on the estimation of absolute solvation energies in water and 1-octanol, obtained from molecular simulation.

Regarding solvation, the majority of previously published studies focused on aqueous media (e.g., see a review paper by Tomasi and Persico[31]), but nowadays, computer simulations can be used to model and understand molecular-level interactions of biological membranes, proteins, and lipids. It is now possible to simulate the interactions of small solutes with complex biological membranes by explicit simulation of the lipid-bilayers,[32] an approach that has the disadvantage of being very computationally expensive.[33] Therefore, alternatives are sought to mimic the fundamental characteristics of biological systems using simpler molecules. Numerous solvents, such as oils,[1] chloroform,[5−9] or alkanes,[34] have been tested to study and reproduce the hydrophobic properties of organic systems, but 1-octanol remains today the most important reference solvent for this kind of study. The amphiphilic nature of the 1-octanol molecule (a polar headgroup attached to a flexible nonpolar tail) gives this molecule similar characteristics to the main constituents of lipid biomembranes. 1-octanol molecules can also mimic the complex behavior of the soil and, thus, play an important role in the prediction of solute partitioning in environmental fate and toxicological processes.[35] Although 1-octanol cannot form long, stable complex structures such as bilayers,[36] which are typical of lipid solutions, it can form liquid aggregates[33,35] and mimic successfully many of the properties of biologically relevant systems. Consequently, it has been widely used for this purpose.

Several simulation studies related to 1-octanol systems have been reported in the literature. In the work of Debolt and Kollman,[33] pure 1-octanol and water-saturated 1-octanol physical properties were studied in detail. More recently, MacCallum and Tieleman[36] investigated 1-octanol mixtures at different hydration levels, including the calculation of pure 1-octanol physical properties using various force fields (FF). In that study, formation of hydrogen-bonded chains in 1-octanol/water systems were observed, which interestingly become more spherical with increasing water concentration.

On the contrary, in pure 1-octanol these clusters are long and thin. Chen and Siepmann[35] identified these microscopic structural differences in the aggregate/micelle formation between dry and water-saturated 1-octanol using configurational-bias Metropolis Monte Carlo simulations in the Gibbs ensemble. Regarding free energy calculations, most studies in 1-octanol have reported only relative free energy changes, i.e., the free energy associated with a mutation from one solute into another solute of similar structure, which is a different approach than the one followed here. Studied systems include: benzene to phenol,[33] ethylbenzene to phenol, pyridine to benzene, cyclopentane to tetrahydrofuran, methanol to methylamine, iso-propanol to iso-propane, acetamide to acetone, and phenol to benzene.[15] Finally, Gibbs free energies of transfer of *n*-alkanes and primary alcohols between water and (dry or wet) 1-octanol were obtained by Chen et al.[37]

Our starting point in the present study is to evaluate/predict the Gibbs free energy of solvation of *n*-alkanes up to C_8 in 1-octanol. The availability of free energy data can be used to understand the behavior of complex systems and has the potential to revolutionize several scientific and technological fields,[38,39] particularly in the pharmaceutical industry.[40] Solvation free energy can also be an important input parameter in order to predict solubility.[17,41] Several investigations regarding free energy calculations in aqueous systems have been reported in the literature,[27,28,42−47] and it is now well established that accurate results can be obtained directly from molecular simulation methods. However, simulations of solvation in nonaqueous solvents are less common. In particular, for 1-octanol, apart from the works of Chen and Siepmann,[35,37] there is clearly a lack of a systematic study, particularly for solvation of longer alkanes. We aim here to fill this gap by presenting calculations of absolute solvation free energies of eight alkanes in 1-octanol. Initially, a comparison is made between several FF's, including all-atom (AA) and united-atom (UA) descriptions, in reproducing pure 1-octanol physical properties. Afterward, we present a comparison of three popular FF's, namely TraPPE, Gromos, and OPLS-AA, to represent solute molecules by analyzing their performance in predicting the 1-octanol absolute Gibbs free energy of solvation for *n*-alkanes up to C_8. Finally, calculation of the hydration free energies and 1-octanol/water partition coefficient by molecular simulation is discussed.

The remainder of this paper is organized as follows: in Section 2, we describe the computational methods used for the Gibbs free energy calculation, particularly the thermodynamic integration, the molecular dynamics (MD) simulation details, and the FF's tested; in Section 3.1, results for the pure 1-octanol physical properties predicted using different FF's are shown, while the capability of molecular simulation methods in predicting solvation free energies and 1-octanol/water partition coefficients are discussed in Sections 3.2−3.4. The main conclusions of this work are summarized in Section 4.

2. Computational Methods

2.1. Thermodynamic Integration. The solvation process consists of the transfer of a compound from a well-defined state (gas/vacuum) to another state (solution), and the solvation free energy may be defined as the free energy difference given by the total reversible work associated with changing the Hamiltonian of the system from the gas to the liquid state.[48] Solvation can be measured experimentally or calculated using an appropriate model and methodology. Experimental free energies are commonly estimated from solute concentration measurements in two-phase systems (vapor and liquid solution) in which, after reaching equilibrium, one evaluates the transfer of molecules between the two phases (see refs 49 and 50 for equations and details). From the theoretical point of view, in the ideal gas approximation, the interaction of a solute with its environment in the gas state is effectively zero, and only the interactions of the solute with a particular solvent environment need to be considered. Free energy is a state function and can, thus, be calculated by molecular simulation based on the construction of a thermodynamic cycle that may include nonphysical transformations necessary to make the calculation feasible. Thus, the 1-octanol solvation free energy at temperature T and pressure P, $\Delta_{solv}G(P,T)$, can be calculated using the following thermodynamic cycle:[51]

$$\begin{array}{ccc} \text{Solute (1-octanol)} & \xrightarrow{\Delta_{oct}G} & \text{Dummy (1-octanol)} \\ \Delta_{solv}G\uparrow & & \downarrow -\Delta_{dummy}G \\ \text{Solute (Vacuum)} & \xrightarrow{\Delta_{vac}G} & \text{Dummy (Vacuum)} \end{array}$$

where $\Delta_{oct}G$ is the free energy associated with the mutation of the solute molecules into dummy molecules in a 1-octanol media, $\Delta_{vac}G$ is the free energy associated with the same process in a vacuum, and finally $\Delta_{dummy}G$ can be seen as the hypothetical solvation free energy of a dummy species. Dummy molecules do not interact with their environment. In practice, these molecules have no electrostatic or van der Waals interactions, but their intramolecular bonded interactions are the same as in the solute molecules. As a consequence, $\Delta_{dummy}G$ is equal to zero, and we can write the following equation for the thermodynamic cycle:

$$\Delta_{solv}G = \Delta_{vac}G - \Delta_{oct}G - \Delta_{dummy}G = \Delta_{vac}G - \Delta_{oct}G \quad (2)$$

The separate calculation in vacuum is necessary to compensate for changes in solute−solute intramolecular nonbonded interactions that take place when the intermolecular interactions are switched off. For each case (solvent and vacuum), the associated free energy (expressed in terms of ΔG for the NPT ensemble) is estimated here using the thermodynamic integration method,[48,52] whose algorithm is as follows: let us consider two generic well-defined states, an initial reference state (state 0) and a final target state (state 1) with Hamiltonians $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. A coupling parameter, λ, can be added to the Hamiltonian, $\mathcal{H}(\mathbf{p}, \mathbf{q}; \lambda)$, where $\mathbf{p}$ is the linear momentum and $\mathbf{q}$ the atomic position,

and used to describe the transition between the two states: $\mathcal{H}(\mathbf{p}, \mathbf{q}; 0) \rightarrow \mathcal{H}(\mathbf{p}, \mathbf{q}; 1)$. Considering several discrete and independent λ values between 0 and 1, equilibrium averages can be used to evaluate derivatives of the free energy with respect to λ. One then integrates the derivatives of the free energy along a continuous path connecting the initial and final states in order to obtain the energy difference between them:

$$\Delta G = \int_0^1 \left\langle \frac{\partial \mathcal{H}(\mathbf{p}, \mathbf{q}, \lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \quad (3)$$

In practice, the solvation free energy can be estimated as follows: i) simulate the system in 1-octanol at different λ values; ii) simulate the system in vacuum at different λ values; and iii) compute the solvation free energy from eq 4:

$$\Delta_{solv} G = \int_0^1 \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_\lambda^{vac} d\lambda - \int_0^1 \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_\lambda^{oct} d\lambda \quad (4)$$

Notice that because we are using thermodynamic integration, which involves equilibrium runs at independent λ values, the direction of the process is irrelevant, and the results for $\Delta_{solv}G$ are free of hysteresis. This is an important advantage relative to other methods (e.g., slow growth) where the results depend on the direction of the calculation.[27,47] As a final remark, one should notice that, since we are studying nonpolar solute molecules (*n*-alkanes), the only contribution to the free energy comes from the process of "turning off" the Lennard-Jones (LJ) interactions. There is no need to separately account for a Coulombic contribution (i.e., "turning off" the solute charges) to the free energy in eq 4, as is normally done for solvation of polar molecules.

2.2. Molecular Dynamics Simulations. Molecular dynamics (MD) simulations were performed with the GROMACS[53] simulation package. The integration of Newton's equations of motion was carried out using the leapfrog dynamic algorithm[54] with a time step of 2 fs. Langevin (stochastic) dynamics[55] were used to control the temperature with a frictional constant of 1 ps^{-1} and a reference temperature of 298 K. This approach eliminates several problems that may arise from the use of conventional thermostats in free energy calculations.[44] For constant pressure simulations, the Berendsen barostat[56] with a time constant of 0.5 ps and an isothermal compressibility of 4.5×10^{-5} bar^{-1} was used to enforce pressure coupling, where the box size was scaled at every time step. The reference pressure was always set to 1 bar. Each simulation box was cubic, with periodic boundary conditions in all directions, and contained 200 1-octanol molecules. Simulations of systems with different numbers of molecules revealed this to be the optimum system size: larger systems yielded statistically similar results but at a higher computational cost, while smaller systems exhibited finite-size effects.

The initial configuration for the pure 1-octanol simulations was generated by randomly placing 200 molecules in a large cubic box. We then run an energy minimization, followed by a constant volume equilibration of 100 ps, and finally a 5 ns long NPT production stage. Two minimization procedures were employed: first, minimization was performed using the limited-memory Broyden−Fletcher−Goldfarb−Shanno (L-BFGS) algorithm of Nocedal[57] for 5 000 steps, followed by a steepest descent minimization for 500 steps. Analysis of several observables ensured that the simulations were properly equilibrated during the NPT run. Average properties were computed by discarding the time steps pertaining to the equilibration period.

To calculate solvation free energies, it is necessary to carry out several independent simulations of each solute (from methane to *n*-octane) in each solvent (1-octanol and water) for different values of the coupling parameter, as described in Section 2.1. The starting configuration for each of these simulations was obtained by immersing a solute molecule into an equilibrated box of 200 1-octanol solvent molecules or 500 water solvent molecules. The equilibrated 1-octanol box was obtained from the NPT simulations for pure 1-octanol, described above, and a similar approach was used for water. In these simulations an energy minimization was initially performed using the same protocol as for the pure liquid simulations, followed by a constant volume equilibration of 100 ps, a constant pressure equilibration of 1 ns (enough to fully equilibrate the box volume and correctly reproduce solvent density), and finally a NVT production run of 5 ns. This procedure was repeated for each of the following 16 λ values:

$$\lambda \in \{0.0, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00\}$$

where $\lambda = 0$ refers to a fully interacting solute, and $\lambda = 1$ to a noninteracting solute. We have used such a large number of intermediate λ states because in the thermodynamic integration, the accuracy of the $\Delta_{solv}G$ value depends strongly on the smoothness of the $\partial\mathcal{H}\partial\lambda$ vs λ curve, where a smooth profile is necessary in order to minimize numerical integration errors. In the present work, the reported statistical uncertainties were obtained from block averaging,[54] and integrals were computed via the trapezoidal rule.[58] Finally, it should be noted that in the transformation process between states with different λ values, the λ dependence of the LJ potential was interpolated between the neighboring states via soft-core interactions. The soft-core expression of Beuler et al.[59] eliminates singularities in the calculation as the LJ interactions are turned off.[60] As suggested in the literature,[27,44] the soft-core parameter used was 0.5, which is the optimized value when the power for λ in the soft-core function is 1, and the soft-core σ value used was 0.3 nm.

2.3. Force Fields. MD simulations for pure 1-octanol were performed using six different FF's. The FF's examined included Gromos (versions 43A2,[61] 53A5,[29] and 53A6[29]), OPLS-UA,[62,63] OPLS-AA,[64] and TraPPE.[65−67] We have decided to test three different versions of the Gromos FF since they were parametrized for different purposes, all relevant to this work. Version 43A2 was parametrized in order to reproduce only pure solvent properties. More recently, the Gromos parameter set 53A5 was optimized to reproduce thermodynamic properties of pure liquids and the solvation Gibbs free energy of amino acid analogs in

Table 1. 1-Octanol Density and Heat of Vaporization at 1 Bar from MD Simulations and Experimental Measurements[a]

	T (K)								
	280		340		400		298		
FF	ρ (kg/m^3)	dev (%)	ρ (kg/m^3)	dev (%)	ρ (kg/m^3)	dev (%)	$\Delta_{vap}H$ (kJ/mol)	dev (%)	production times (hr/ns)
G43A2	864.4 ± 0.9	3.4	822.3 ± 0.3	3.9	779.7 ± 0.7	5.0	64.4	−10.5	1.09
G53A5	867.9 ± 0.8	3.8	827.0 ± 0.9	4.5	785.3 ± 0.6	5.7	59.5	−17.3	1.09
G53A6	868.0 ± 0.7	3.8	827.1 ± 1.4	4.5	785.3 ± 0.8	5.7	59.5	−17.3	1.09
OPLS-UA	859.5 ± 0.7	2.8	818.8 ± 0.6	3.5	773.5 ± 0.6	4.1	72.3	0.4	1.19
OPLS-AA	841.8 ± 0.9	0.7	781.2 ± 1.3	−1.3	719.5 ± 1.2	−3.1	70.7	−1.8	8.00
TraPPE	837.0 ± 0.9	0.08	793.4 ± 0.5	0.3	744.5 ± 0.8	0.2	67.0	−6.9	1.15
Experimental	836.26[b]		791.39[b]		742.75[b]		71.98[c]		–

[a] Computational production times per node (Intel Xeon at 3.0 GHz) for each FF is shown. [b] Data from refs 84−86. [c] Data from ref 72.

cyclohexane, while parameter set 53A6 was optimized to reproduce free energies in water.[29] The TraPPE FF was also chosen because it was optimized to provide accurate descriptions of pure liquids and vapor−liquid equilibria (VLE).[65−67] It should be noted that, contrary to the original version of TraPPE where all bonds were fixed, bond stretching was modeled in our studies by a harmonic potential with force constants taken from CHARMM,[68] except for bonds involving hydrogen atoms that were constrained using LINCS.[69] Finally, we have tested the popular OPLS FF, which are designed to be transferrable to a wide range of organic molecules in the liquid phase. We have compared the united-atom (UA) against the all-atom (AA) FF because the former are expected to be computationally much cheaper.

In this work, the modified extended simplified point charge (MSPC/E)[70] model was used for the simulation of water. MSPC/E is an accurate FF for pure water and water−hydrocarbon thermodynamic properties, and it was chosen over other popular FF's for water. This FF also includes a polarization correction expected to improve the hydration predictions.[28]

Several of the above FF's (in particular OPLS-AA, TraPPE, and Gromos 53A6) were also used to model the alkane molecules, solvated in either 1-octanol or water. Different combinations of solute−solvent FF's were tested in order to assess the influence of this choice on the free energy and the partition coefficient predictions. Dummy molecules were considered to be identical to real solute molecules in terms of mass, while their LJ interaction parameters were set to zero. In all cases, electrostatic interactions were calculated using the reaction field[71] method with $\varepsilon_{rf,oct} = 10.3$ (the dielectric constant for pure 1-octanol[72]) or $\varepsilon_{rf,wat} = 80$ (the dielectric constant for pure water[72]). Tests performed with the more computationally demanding particle mesh Ewald method yielded similar results. The cutoff radii used were 1 nm for the electrostatic interactions, 1 nm for the short-range neighbor list, and 0.8−0.9 nm switched cutoff for the LJ interactions. It was also observed that the use of higher cutoff radii induces minimal perturbations in the absolute energy values. Overall, the simulation parameters described here were chosen so that the computational cost was minimized without sacrificing the accuracy of the calculations. Long range corrections for energy and pressure were also employed as it was concluded that they significantly improve the accuracy of the predicted solvation energies.[27]

Detailed van der Waals parameters, point charges, bond stretching, bond angle bending and torsional force constants are provided in the Supporting Information for all compounds and FF's. Coordinate and topology files were built manually or with the help of the Molden[73] and PRODRG[74] software.

3. Results and Discussion

3.1. Pure 1-Octanol Physical Properties. The accuracy of different FF's for the prediction of pure 1-octanol properties was initially evaluated. The calculated 1-octanol densities over a wide temperature range from NPT MD and the heat of vaporization at 298 K are shown in Table 1. Densities were directly obtained from the GROMACS suite using the g_energy tool,[54] while heats of vaporization were estimated by taking the difference of enthalpy in the vapor and liquid phases:

$$\Delta_{vap}H = E_g - E_L + RT \quad (5)$$

where $E_g$is the total energy in the gas phase, and $E_L$is the total energy per mole in the liquid phase.

Based on the data reported in Table 1, one concludes that Gromos generally overestimates the 1-octanol densities, in line with previous studies of this FF,[47] and also significantly underestimates the vaporization enthalpy. As expected, version 43A2 of Gromos performed better that the other two versions because it was optimized to reproduce pure solvent liquid properties. The TraPPE FF provides excellent accuracy for the density over the temperature range but slightly underestimates the enthalpy of vaporization. Conversely, OPLS-UA overestimates the density at all temperatures but does an excellent job at predicting the enthalpy of vaporization. OPLS-AA improves significantly over OPLS-UA yielding good predictions of both density and vaporization enthalpy but at the cost of an increased computational time. In fact, computational production times, included in Table 1, show that this AA FF is more than 7 times more expensive compared to the UA approaches. One should also notice that for higher temperatures, OPLS-AA accuracy decreases. In general, an AA FF is preferable than a UA FF, provided that one can afford the additional computational cost. For simulations where minimization of computing cost is an important issue, such as in the highly demanding free energy calculations performed in this work, it is reasonable to use a UA approximation, at least for the solvent. In this case, TraPPE is probably the best option, since it performs well

Table 2. Comparison of $\Delta_{vac}G$, $\Delta_{oct}G$, and $\Delta_{solv}G$ (all in kcal/mol) Predictions for *n*-Alkanes in 1-Octanol Using TraPPE, Gromos and OPLS-AA/TraPPE FF's against Available Experimental Data at 298 K[16]

	TraPPE			Gromos			OPLS-AA/TraPPE			expt
solute	$\Delta_{vac}G$	$\Delta_{oct}G$	$\Delta_{solv}G$	$\Delta_{vac}G$	$\Delta_{oct}G$	$\Delta_{solv}G$	$\Delta_{vac}G$	$\Delta_{oct}G$	$\Delta_{solv}G$	$\Delta_{solv}G$
methane	0	-0.5 ± 0.2	**0.5 ± 0.1**	0	-0.4 ± 0.1	**0.4 ± 0.1**	0	-0.2 ± 0.1	**0.2 ± 0.1**	0.5
ethane	0	0.4 ± 0.2	**−0.4 ± 0.2**	0	0.9 ± 0.2	**−0.9 ± 0.2**	0	0.5 ± 0.2	**−0.5 ± 0.2**	−0.6
propane	0	1.0 ± 0.2	**−1.0 ± 0.2**	0	1.9 ± 0.2	**−1.9 ± 0.2**	-0.6 ± 0.1	0.6 ± 0.2	**−1.2 ± 0.2**	−1.2[a]
n-butane	0	1.4 ± 0.2	**−1.4 ± 0.2**	0.0 ± 0.1	2.9 ± 0.2	**−2.9 ± 0.2**	-1.3 ± 0.1	0.6 ± 0.2	**−1.9 ± 0.2**	−1.8[a]
n-pentane	0.1 ± 0.1	2.3 ± 0.2	**−2.2 ± 0.2**	-0.1 ± 0.1	3.3 ± 0.2	**−3.4 ± 0.2**	-2.1 ± 0.1	0.7 ± 0.2	**−2.8 ± 0.2**	−2.3[a]
n-hexane	0.2 ± 0.1	2.9 ± 0.2	**−2.7 ± 0.2**	-0.1 ± 0.1	4.4 ± 0.2	**−4.5 ± 0.2**	-2.9 ± 0.1	0.5 ± 0.2	**−3.4 ± 0.2**	−3.3
n-heptane	0.3 ± 0.1	3.5 ± 0.2	**−3.2 ± 0.2**	-0.1 ± 0.1	4.7 ± 0.2	**−4.8 ± 0.2**	-3.8 ± 0.1	0.2 ± 0.2	**−4.0 ± 0.2**	−4.1
n-octane	0.4 ± 0.1	3.7 ± 0.2	**−3.4 ± 0.2**	-0.1 ± 0.1	6.0 ± 0.2	**−6.1 ± 0.2**	-4.9 ± 0.2	-0.1 ± 0.3	**−4.7 ± 0.3**	−4.6

[a] Values estimated from eq 6.

for pure solvent liquid properties and is also able to accurately describe VLE.[65]

3.2. Free Energies of Solvation in 1-octanol. The Gibbs free energy of solvation of *n*-alkanes in 1-octanol at 298 K was calculated from MD, as described above. Simulations were performed using three different FF's for the representation of both 1-octanol and *n*-alkane molecules, namely Gromos 53A6,[29] TraPPE,[65−67,75−77] and OPLS-AA.[64] The 53A6 version of Gromos was preferred over the other two versions as it was parametrized to reproduce solvation properties in a polar solvent. Preliminary calculations using the OPLS-AA FF to model both alkanes and octanol showed that the computational time required for the accurate estimation of $\Delta_{solv}G$ was very high. As shown in Section 3.1, this is due to the high cost associated to an AA description of 1-octanol. Consequently, 1-octanol molecules were modeled with the TraPPE FF instead, and these calculations are referred to as OPLS-AA/TraPPE in the remainder of this paper. We have also tested a combination of OPLS-AA for the solutes with OPLS-UA for the solvent for consistency. Unfortunately, differences between experimental data and simulations from a preliminary test with propane were as high as 1 kcal/mol, and this combination of FF was not pursued further. It should be noted that deficiencies of the OPLS-UA FF in reproducing hydration free energies and hydrocarbon solubilities in water were also reported by MacCallum and Tieleman.[36]

Thermodynamic integration was performed using the three FF's for the solutes in both vacuum and solvent medias. Representative results for the integrand of eq 4 in the octanol phase are shown in Figure 1 based on the Gromos 53A6 FF, while similar results for all FF's are given in the Supporting Information. Furthermore, MD calculations for $\Delta_{vac}G$, $\Delta_{oct}G$ and $\Delta_{solv}G$ from the different FF's are shown in Table 2, with experimental data reported for comparison, while the different data sets for $\Delta_{solv}G$ are shown in Figure 2. It should be noted that the experimental values in Table 2 and Figure 2 represent solvation free energies of *n*-alkanes in water-saturated 1-octanol solutions, since there are no data available for anhydrous 1-octanol, which is used in the simulations. However, the difference between the free energy of solvation determined in pure and water-saturated 1-octanol is typically small, on the order of 0.2−0.4 kcal/mol[16] (and refs 78−80). Moreover, for the case of propane, *n*-butane, and *n*-pentane, there are no available experimental data. To allow for a more comprehensive comparison of our simulations with experimental results, experimental values presented in Table 2 marked with *a* were estimated from

$$\Delta_{solv}G^{octanol} = \Delta_{solv}G^{water} - \log P^{octanol/water} \times 2.303 \times RT \quad (6)$$

where $\Delta_{solv}G^{water}$ are experimental data from Michielan et al.,[81] and $\log P^{octanol/water}$ are the 1-octanol/water partition coefficients suggested by Sangster.[3]

In general, the calculated $\Delta_{solv}G$ decrease with increasing chain length is consistent with the experimental data. Calculations based on OPLS-AA/TraPPE FF provide the best agreement with experimental data, while Gromos predicts lower $\Delta_{solv}G$ values and TraPPE predicts higher $\Delta_{solv}G$ than the experiments. The average deviation between the experi-

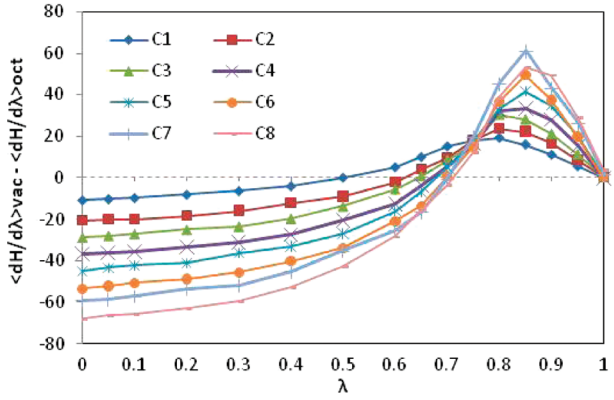


Figure 1. Derivative of the Hamiltonian with respect to λ as a function of λ for *n*-alkanes in 1-octanol using the Gromos FF.

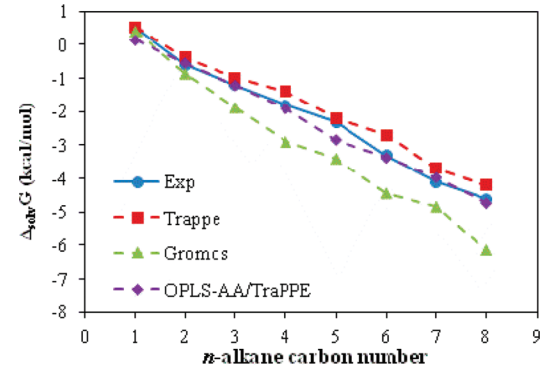


Figure 2. $\Delta_{solv}G$ for *n*-alkanes in 1-octanol at 298 K as a function of carbon number: Experimental data and MD simulations.

Table 3. $\Delta_{vac}G$, $\Delta_{wat}G$, and $\Delta_{hyd}G$ (all in kcal/mol) Predictions for *n*-alkanes in MSPC/E Water Using TraPPE, Gromos, and OPLS-AA/TraPPE FF's against Available Experimental Data[a] at 298 K[49,87]

	TraPPE			Gromos			OPLS-AA/TraPPE			expt	simulation
solute	$\Delta_{vac}G$	$\Delta_{wat}G$	$\Delta_{hyd}G$	$\Delta_{vac}G$	$\Delta_{wat}G$	$\Delta_{hyd}G$	$\Delta_{vac}G$	$\Delta_{wat}G$	$\Delta_{hyd}G$	$\Delta_{hyd}G$	$\Delta_{hyd}G$
methane	0	-2.3 ± 0.1	**2.3 ± 0.1**	0	-2.0 ± 0.1	**2.0 ± 0.1**	0	-2.4 ± 0.1	**2.4 ± 0.1**	1.98	2.0−2.6[27,28,42−47]
ethane	0	-2.1 ± 0.1	**2.1 ± 0.1**	0	-1.8 ± 0.1	**1.8 ± 0.1**	0	-2.6 ± 0.1	**2.6 ± 0.1**	1.81	1.7−2.6[42,43,46,47]
propane	0	-2.4 ± 0.1	**2.4 ± 0.1**	0	-1.9 ± 0.1	**1.9 ± 0.1**	-0.6 ± 0.1	-3.7 ± 0.1	**3.1 ± 0.1**	2.02	1.9−2.7[27,28,42−47]
n-butane	0	-2.8 ± 0.1	**2.8 ± 0.1**	-0.0 ± 0.1	-1.7 ± 0.2	**1.7 ± 0.2**	-1.3 ± 0.1	-4.7 ± 0.2	**3.4 ± 0.2**	2.18	1.9−3.5[27,28,42−47]
n-pentane	0.1 ± 0.1	-3.0 ± 0.1	**3.1 ± 0.1**	-0.1 ± 0.1	-2.1 ± 0.2	**2.0 ± 0.2**	-2.1 ± 0.1	-5.6 ± 0.2	**3.5 ± 0.2**	2.36	2.7−3.7[42,47]
n-hexane	0.2 ± 0.1	-3.2 ± 0.1	**3.4 ± 0.1**	-0.1 ± 0.1	-2.3 ± 0.2	**2.2 ± 0.2**	-2.9 ± 0.1	-7.1 ± 0.2	**4.2 ± 0.2**	2.58	n.a.
n-heptane	0.3 ± 0.1	-3.5 ± 0.1	**3.7 ± 0.1**	-0.1 ± 0.1	-2.4 ± 0.2	**2.3 ± 0.2**	-3.8 ± 0.1	-7.9 ± 0.2	**4.2 ± 0.2**	2.65	n.a.
n-octane	0.4 ± 0.1	-3.8 ± 0.2	**4.2 ± 0.2**	-0.1 ± 0.1	-2.4 ± 0.2	**2.3 ± 0.2**	-4.9 ± 0.2	-9.7 ± 0.2	**4.8 ± 0.2**	2.93	n.a.

[a] For comparison, literature values based on molecular simulation are included.

mental data and the simulations is 0.1 kcal/mol for OPLS-AA/TraPPE, 0.8 kcal/mol for Gromos, and 0.4 kcal/mol for TraPPE. In the pharmaceutical industry, accuracies of 0.5−1.0 kcal/mol are required for predicting affinities in drug binding.[43] In this respect, the polarizable continuum model MST, originally developed by Miertus et al.,[82] was recently reparameterized[16] for reproducing solvation free energies in 1-octanol, and differences of 0.4−0.6 kcal/mol were observed for *n*-alkanes from C_6−C_8. Even more, this (re)parametrization required the knowledge of the solvation experimental data, which for complex molecules is a clear disadvantage. Indeed, the methodology used in this work can provide molecular level details and insights that cannot be obtained using continuous models, since solvent molecules are modeled explicitly. The average absolute deviation (AAD) observed in this work for the organic phase are considerably smaller than the typical AAD published in the literature for aqueous systems (see Section 3.3).

In short, the accuracy of the OPLS-AA/TraPPE combination of FF's for solute/solvent to describe the Gibbs energy of solvation in 1-octanol is clearly better in comparison with other published studies. These calculations also verify that an AA description of the solute molecules improves the accuracy in the prediction of solvation energies.

3.3. Free Energies of Hydration of *n*-Alkanes. Contrary to the case of 1-octanol, there are many experimental data and simulation studies available in the literature concerning $\Delta_{hyd}G$ of *n*-alkanes. In Table 3, a compilation of such data is presented (last two columns). In our simulations, the same molecular models as above were used for *n*-alkanes. Simulation results for $\Delta_{vac}G$, $\Delta_{wat}G$, and $\Delta_{hyd}G$ from the various FF's are presented in Table 3. A graphical comparison of simulation results with experimental data for $\Delta_{hyd}G$ is shown in Figure 3. We can observe that, while in 1-octanol solvation free energies are negative and decrease with the chain length so that the solubility in octanol increases, the opposite is found in water, and the solubility decreases with the chain length. These facts are supported both by experiments and simulation.

For the hydration calculations, the deviation between experimental data and our MD results is larger than in the case of 1-octanol, although in the same accuracy range of previously published studies for these systems.[27,28,44,45,47] Typical average absolute deviations for hydration Gibbs free energy calculations available in the literature range from 0.8 to 1.5 kcal/mol, as can be found in the study of Shirts et al.[27] for 15 amino acid side chain analogs: 1.2 kcal/mol for AMBER, 1.1 kcal/mol for CHARMM, and 0.8 kcal/mol for OPLS-AA. Furthermore, for the hydration of alkanes (up to C_5), average deviations of 0.5 kcal/mol were reported.[42,43,46,47]

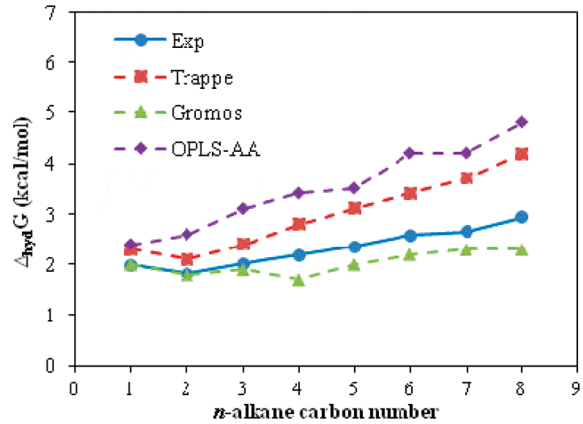


Figure 3. $\Delta_{hyd}G$ for *n*-alkanes at 298 K as a function of carbon number: Experimental data and MD simulations.

Gromos provides the best agreement with the experimental data with an AAD lower than 0.3 kcal/mol, while OPLS-AA/TraPPE predictions deviate by an average of 1.2 kcal/mol, and TraPPE by an average of 0.6 kcal/mol from the experimental data. This good performance of the Gromos FF is to be expected a priori since this FF was parametrized to reproduce free energies of hydration. Interestingly, the use of an AA description of the solute in hydration free energy calculations seems to be less important than the optimization of the interaction parameters. This is in marked contrast to the case of solvation free energies in 1-octanol, as described above. Thus, it appears that it is important to take hydration free energies into consideration during the parametrization of a FF, if accurate predictions of this property are desired. Previous simulation studies have also revealed the importance of the FF used for water in the description of the hydration free energy.[28,45]

3.4. 1-Octanol/Water Partition Coefficients. The 1-octanol/water partition coefficient at 298 K for the various *n*-alkanes can be readily estimated from eq 1 using the Gibbs free energies of solvation calculated from our MD simulations. In Table 4, simulation predictions are shown for the different FF's employed together with literature experimental data for comparison.

The overall AAD between the experimental data and the simulation results for log *P* is equal to 0.4 (in log *P* units)

Table 4. Experimental data[2,3,88] and Simulation Predictions for the Logarithm of 1-Octanol/Water Partition Coefficient (log *P*) using Different FF Combinations[a]

	log *P*				
solute	Gromos	TraPPE	OPLS-AA/TraPPE	Gromos + OPLS-AA/TraPPE	expt
methane	1.2	1.3	1.6	1.3	1.1
ethane	2.0	1.8	2.3	1.7	1.8
propane	2.8	2.5	3.2	2.3	2.4
n-butane	3.4	3.1	3.8	2.6	2.9
n-pentane	4.0	3.9	4.6	3.5	3.4
n-hexane	4.9	4.3	5.6	4.1	3.9
n-heptane	5.2	5.1	6.0	4.6	4.7
n-octane	6.2	5.6	7.0	5.1	5.2
AAD	0.4	0.3	0.9	0.1	–

[a] The ADD's between experiment and simulation are also included.

for Gromos, 0.3 for TraPPE and 0.9 for OPLS-AA/TraPPE. Interestingly, the TraPPE FF provides accurate log *P* predictions, while the corresponding solvation energies are not so accurately estimated; this can be attributed to cancellation of errors between the two phases – the high overestimation of the hydration free energy (Figure 3) is partially compensated by an overestimation of the octanol solvation free energy (Figure 2). A similar effect occurs in the Gromos predictions but from the opposite direction – underestimation of both water and octanol free energies. On the other hand, the OPLS-AA/TraPPE FF combination is much more accurate in the organic phase than in the aqueous phase, leading to larger deviations in log *P*.

However, if one calculates log *P* using the most accurate simulation predictions for both $\Delta_{hyd}G$ (from Gromos) and $\Delta_{solv}G$ (from OPLS-AA/TraPPE), then an AAD of 0.14 is obtained. Clearly, this approach provides a very accurate prediction within the experimental uncertainty. Comparing accuracies of different methods can be merely qualitative, since the method performance is highly dependent on the validation set used, which may vary on size, complexity, or the overlap of information used in the training set/model correlation. Even so, similar calculations using a continuous model resulted in an AAD of 0.75 log *P* units,[16] verifying that our predictions should be considered very satisfactory. Another published work[83] reports deviations of 0.6 log *P* units using a continuum method based on a continuous electrostatic model using atomic point charges combined with a nonelectrostatic term function of surface tension for a set of 2 116 molecules.

A final remark should be made regarding the accuracy of the available experimental data. As previously explained, log *P* and Gibbs free energy of solvation data are estimated following different experimental methodologies. At the same time, eq 1 provides a means to check the consistency between different data. A compilation of different data results in deviations of up to 0.8 log *P* units with an AAD of 0.24 log *P* units.

4. Conclusions

In order to predict the partition coefficient of a solute between 1-octanol and water, absolute free energy calculations were performed in 1-octanol and water solvents for different *n*-alkanes up to *n*-octane using MD and thermodynamic integration. The absolute free energies of solvation were estimated by fully decoupling the solute from the solvent, which must be distinguished from previous studies where the relative free energies were calculated from mutations between two solutes. The method we used here is more flexible and not limited to mutations between similar structures. However, this complete decoupling requires large changes in the Hamiltonian, and potentially higher errors are introduced in the calculations as more intermediate states are required. It is also worthwhile to notice that, contrary to many other methodologies presented in the literature, we do not need the knowledge of the experimental solvation data in advance, which is a clear advantage.

Our method is capable of predicting solvation free energies of nonpolar solutes such as *n*-alkanes in 1-octanol with good accuracy. A comparison between different FF's permitted to conclude that the OPLS-AA FF for the solute in combination with the TraPPE FF for 1-octanol produces the most accurate results with differences to experimental data of 0.1 kcal/mol, which is approximately the precision of the experimental methods. The results are much improved by using an AA model for the *n*-alkanes, relative to UA models, with very little increase in computational cost. Arguably, the predictions could be further improved by adopting an AA description of the 1-octanol solvent as well, since this yielded a better representation of pure liquid properties. However, the associated high computational cost currently precludes this approach.

Moreover, we reproduced experimental hydration free energies of the same *n*-alkanes with average deviations of 0.3 kcal/mol, using the Gromos FF. For hydration free energies, a correct parametrization of the interaction potentials seems to be more important than using an AA description of the solute. For this reason, Gromos, which included hydration free energies in its parametrization, performed better than OPLS-AA.

Combining the simulated values of solvation free energy of the *n*-alkanes in water and 1-octanol, we were able to predict the corresponding partition coefficients with an accuracy that is within the experimental uncertainty. All FF combinations that were tested here performed well, in some cases due to the cancellation of errors in both solvation free energies. The most accurate log *P* predictions are afforded by the combination of the Gromos FF in the water phase with the OPLS-AA/TraPPE FF in the organic phase, reaching

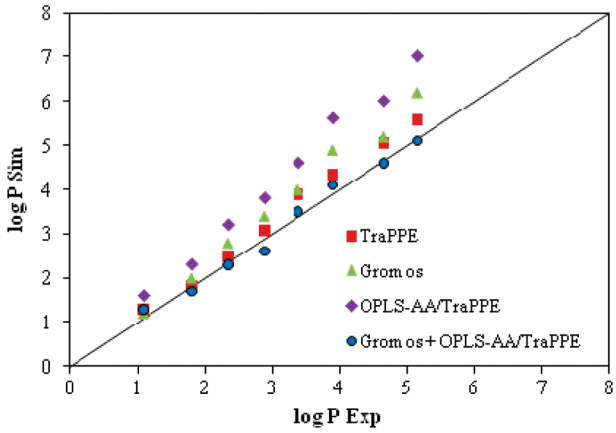


Figure 4. Comparison of log *P* predictions using different FF's against experimental data.

absolute deviations to experimental data of 0.1 log *P* units, which can be comparable to the widely used QSPR statistical methods.

Acknowledgment. The authors are grateful for the support provided by Fundação para a Ciência e a Tecnologia (FCT) Portugal, through projects FEDER/POCI/2010 and REEQ/1164/EQU/2005. N.M.G. acknowledges his FCT Ph.D. scholarship SFRH/BD/47822/2007 and financial support for his visit to NCSR "Demokritos". Financial support from the Greek Secretariat of Research and Technology was provided for N.M.G. to stay in NCSR "Demokritos". A.J.Q. acknowledges financial support from POCI/N010/2006, and M.J. acknowledges financial support from Ciência 2008.

Supporting Information Available: Detailed van der Waals parameters, point charges, bond stretching, bond angle bending and torsional force constants are provided for all compounds and FF. Detailed bonded and nonbonded potential parameters for all the compounds and for the different FF under study as well as plots of the derivatives of the Hamiltonian with respect to the coupling parameter for all the case studies are available. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71* (6), 525–616.

(2) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Hydrophobic, Electronic and Steric Constants*. American Chemical Society: Washington, DC, 1995.

(3) Sangster, J. *Octanol-Water Partitioning Coefficients: Fundamentals and Physical Chemistry*. John Willey & Sons: Chichester, U.K., 1997.

(4) Perlovich, G. L.; Kurkov, S. V.; Kinchin, A. N.; Bauer-Brandl, A. Solvation and Hydration Characteristics of Ibuprofen and Acetylsalicylic Acid. *AAPS PharmSciTech* **2004**, *6* (1), 1–9.

(5) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. American Chemical Society: Washington DC, 1995.

(6) Betageri, G. V.; Rogers, J. A. Correlation of Partitioning of Nitroimidazoles in the Normal Octanol Saline and Liposome Systems with Pharmacokinetc Parameters and Quantitative Structure Activity Relationships (QSAR). *Pharmaceut. Res.* **1989**, *6* (5), 399–403.

(7) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*. Oxford University: Oxford, U.K., 1999.

(8) Pinho, S. P.; Macedo, E. A. Solubility in Food, Pharmaceuticals, and Cosmetic Industries. In *Developments and Applications in Solubility*, Letcher, T. M., Ed. Royal Society of Chemistry: Cambridge, U.K., 2003; pp 309−326.

(9) Hansch, C.; Fujita, T. ρ-π-σ analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616–1626.

(10) Leo, A. J. Calculating log P(Oct) from Structures. *Chem. Rev.* **1993**, *93* (4), 1281–1306.

(11) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of Solvation Free energies of Small organic Molecules: Additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (2), 405–412.

(12) Bodor, N.; Buchwald, P. Molecular Size Based Approach to Estimate Partition Properties for Organic Solutes. *J. Phys. Chem. B* **1997**, *101* (17), 3404–3412.

(13) Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Marcus, Y.; Taft, R. W. Linear Solvation Energy Relationships 46: An Improved Equation for Correlation and Prediction of Octanol Water Partition Coefficients of Organic Nonelectrolytes (Including Strong Hydrogen-Bond Donor Solutes). *J. Phys. Chem.* **1988**, *92* (18), 5244–5255.

(14) Karelson, M. *Molecular Descriptors in QSAR/QSPR*. Wiley Interscience: New York, 2000.

(15) Best, S. A.; Merz, K. M.; Reynolds, C. H. Free Energy Perturbation Study of Octanol/Water Partition Coefficients: Comparison with Continuum GB/SA Calculations. *J. Phys. Chem. B* **1999**, *103* (4), 714–726.

(16) Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22* (11), 1180–1193.

(17) Westergren, J.; Lindfors, L.; Hoglund, T.; Luder, K.; Nordholm, S.; Kjellander, R. In Silico Prediction of Drug Solubility: 1. Free Energy of Hydration. *J. Phys. Chem. B* **2007**, *111* (7), 1872–1882.

(18) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122* (12), 2878–2888.

(19) Orozco, M.; Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **2000**, *100* (11), 4187–4225.

(20) Bergstrom, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and Computational Screening Models for Prediction of Aqueous Drug Solubility. *Pharmaceut. Res.* **2002**, *19* (2), 182–188.

(21) Glomme, A.; Marz, J.; Dressman, J. B. Comparison of a Miniaturized Shake-Flask Solubility Method with Automated Potentiometric Acid/Base Titrations and Calculated Solubilities. *J. Pharm. Sci.* **2005**, *94* (1), 1–16.

(22) Loftsson, T.; Hreinsdottir, D. Determination of Aqueous Solubility by Heating and Equilibration: A technical Note. *AAPS PharmSciTech* **2006**, *7* (1), E1–E4.

(23) Chen, X. Q.; Venkatesh, S. Miniature Device for Aqueous and Non-aqueous Solubility Measurements during Drug Discovery. *Pharmaceut. Res.* **2004**, *21* (10), 1758–1761.

(24) Jorgensen, W.; Briggs, J.; Contreras, M. Relative Partition Coefficients for Organic Solutes from Fluid Simulations. *J. Phys. Chem.* **1990**, *94*, 1683–1686.

(25) Essex, J. W.; Reynolds, C. A.; Richards, W. G. Relative Partition Coefficients from Partition Functions: A Theoretical Approach to Drug Transport. *Chem. Commun.* **1989**, (16), 1152–1154.

(26) Jorgensen, W. Free Energy Calculations: A Breaktrough for Modeling Organic Chemistry in Solution. *Acc. Chem. Res.* **1989**, *22*, 184–189.

(27) 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* (11), 5740–5761.

(28) Hess, B.; van der Vegt, N. F. A. Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models. *J. Phys. Chem. B* **2006**, *110* (35), 17616–17626.

(29) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS ForceField Parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676.

(30) Shivakumar, D.; Deng, Y.; Roux, B. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. *J. Chem. Theory Comput.* **2009**, *5* (4), 919–930.

(31) Tomasi, J.; Persico, M. Molecular Interactions in Solution - an Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94* (7), 2027–2094.

(32) Tieleman, D. P.; Marrink, S. J.; Berendsen, H. J. C. A Computer Perspective of Membranes: Molecular Dynamics Studies of Lipid Bilayer Systems. *Biochim. Biophys. Acta, Rev. Biomembr.* **1997**, *1331* (3), 235–270.

(33) Debolt, S. E.; Kollman, P. A. Investigation of Structure, Dynamics and Solvation in 1-Octanol and its Water-Saturated Solution: Molecular Dynamics and Free-Energy Perturbation Studies. *J. Am. Chem. Soc.* **1995**, *117* (19), 5316–5340.

(34) Shih, P.; Pedersen, L. G.; Gibbs, P. R.; Wolfenden, R. Hydrophobicities of the Nucleic Acid Bases: Distribution Coefficients from Water to Cyclohexane. *J. Mol. Biol.* **1998**, *280* (3), 421–430.

(35) Chen, B.; Siepmann, J. I. Microscopic Structure and Solvation in Dry and Wet Octanol. *J. Phys. Chem. B* **2006**, *110* (8), 3555–3563.

(36) MacCallum, J. L.; Tieleman, D. P. Structures of Neat and Hydrated 1-Octanol from Computer Simulations. *J. Am. Chem. Soc.* **2002**, *124* (50), 15085–15093.

(37) Chen, B.; Siepmann, J. I. Partitioning of Alkane and Alcohol Solutes between Water and (Dry or Wet) 1-Octanol. *J. Am. Chem. Soc.* **2000**, *122* (27), 6464–6467.

(38) Kollman, P. Free Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93* (7), 2395–2417.

(39) Kollman, P. A. Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc. Chem. Res.* **1996**, *29* (10), 461–469.

(40) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818.

(41) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharmaceut.* **2008**, *5* (2), 266–279.

(42) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. Hydration and Conformational Equilibria of Simple Hydrophobic and Amphiphilic Solutes. *Biophys. J.* **1998**, *75* (2), 755–768.

(43) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.* **1994**, *98*, 13077–13082.

(44) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of Charge Models for Fixed-Charge Force Fields: Small-molecule Hydration Free Energies in Explicit Solvent. *J. Phys. Chem. B* **2007**, *111* (9), 2242–2254.

(45) Shirts, M. R.; Pande, V. S. Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. *J. Chem. Phys.* **2005**, *122* (13), 134508.

(46) Slusher, J. T. Accurate Estimates of Infinite-dilution Chemical Potentials of Small Hydrocarbons in Water via Molecular Dynamics Simulation. *J. Phys. Chem. B* **1999**, *103* (29), 6075–6079.

(47) Wescott, J. T.; Fisher, L. R.; Hanna, S. Use of Thermodynamic Integration to Calculate the Hydration Free Energies of n-alkanes. *J. Chem. Phys.* **2002**, *116* (6), 2361–2369.

(48) Christophe Chipot; Pohorille, A. *Free Energy Calculations - Theory and Applications in Chemistry and Biology*. Springer: Berlin, Germany, 2007.

(49) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistry* **1981**, *20* (4), 849–855.

(50) Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016–2027.

(51) Leach, A. , *Molecular Modeling: Principles and Applications*. Prentice-Hall: 2001.

(52) Kirkwood, J. G. Statistical Mechanics of Pure Fluids. *J. Chem. Phys.* **1935**, *3*, 300–313.

(53) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(54) Spoel, D. L., E.; Hess, B.; Buuren, A.; Apol, E.; Meuulenhof, P.; Tieleman, D.; Sijbers, A.; Feenstra, K.; Drunen, R.; Berendsen, H. *Gromacs User Manual - version 3.3*. The Netherlands, 2006.

(55) van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Molec Sim.* **1988**, *1* (3), 173–185.

(56) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(57) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45* (3), 503–528.

(58) Chapra, S.; Canale, R. *Numerical Methods for Engineers*, 5th ed.; McGraw-Hill: New York, USA, 2006.

(59) Beuler, T., M. R.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(60) Pitera, J. W.; Van Gunsteren, W. F. A Comparison of Non-bonded Scaling Approaches for Free Energy Calculations. *Mol. Simul.* **2002**, *28* (1−2), 45–65.

(61) 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: GROMOS96 Manual and User Guide;. Vdf Hochschulverlag AG an der ETH Zürich*: Zürich, Switzerland, 1996.

(62) Jorgensen, W. L. Optimized Intermolecular Potential Functions for Liquid Alcohols. *J. Phys. Chem.* **1986**, *90* (7), 1276–1284.

(63) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *J. Am. Chem. Soc.* **1984**, *106* (22), 6638–6646.

(64) 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* (45), 11225–11236.

(65) Chen, B.; Potoff, J. J.; Siepmann, J. I. Monte Carlo Calculations for Alcohols and their Mixtures with Alkanes. Transferable Potentials for Phase Equilibria. 5. United-atom description of Primary, Secondary, and Tertiary Alcohols. *J. Phys. Chem. B* **2001**, *105* (15), 3093–3104.

(66) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-atom description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102* (14), 2569–2577.

(67) Martin, M. G.; Siepmann, J. I. Novel Configurational-bias Monte Carlo Method for Branched Molecules. Transferable Potentials for Phase Equilibria. 2. United-atom description of Branched Alkanes. *J. Phys. Chem. B* **1999**, *103* (21), 4508–4517.

(68) Brooks, B. R.; Bruccoleri, R. E.; Olafson, D. J.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(69) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(70) Boulougouris, G. C.; Economou, I. G.; Theodorou, D. N. Engineering a Molecular Model for Water Phase Equilibrium over a Wide Temperature Range. *J. Phys. Chem. B* **1998**, *102* (6), 1029–1035.

(71) Lee, F. S.; Warshel, A. A Local Reaction Field Method for Fast Evaluation of Long-range Electrostatic Interactions in Molecular Simulations. *J. Chem. Phys.* **1992**, *97* (5), 3100–3107.

(72) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 85 ed.; CRC Press: Boca Raton, FL, 2005.

(73) Schaftenaar, G.; Noordik, J. H. Molden: a Pre- and Post-processing Program for Molecular and Electronic Structures. *J. Comput.-Aided Mol. Des.* **2000**, *14* (2), 123–134.

(74) Schuettelkopf, A. W.; Aalten, D. M. F. v., PRODRG - a Tool for High-throughput Crystallography of Protein-ligand Complexes. *Acta Cryst. D* **2004**, *60*, 1355–1363.

(75) Stubbs, J. M.; Potoff, J. J.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 6. United-atom description for Ethers, Glycols, Ketones, and Aldehydes. *J. Phys. Chem. B* **2004**, *108* (45), 17596–17605.

(76) Wick, C. D.; Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 4. United-atom description of Linear and Branched Alkenes and Alkylbenzenes. *J. Phys. Chem. B* **2000**, *104* (33), 8008–8016.

(77) Wick, C. D.; Stubbs, J. M.; Rai, N.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 7. Primary, Secondary, and Tertiary Amines, Nitroalkanes and Nitrobenzene, Nitriles, Amides, Pyridine, and Pyrimidine. *J. Phys. Chem. B* **2005**, *109* (40), 18974–18982.

(78) Bernazzani, L.; Cabani, S.; Conti, G.; Mollica, V. Thermodynamic Study of the Partitioning of Organic Compounds between Water and Octan-1-ol. *J. Chem. Soc., Faraday Trans.* **1995**, *91* (4), 649–655.

(79) Berti, P.; Cabani, S.; Conti, G.; Mollica, V. Thermodynamic Study of Organic Compunds in Octan-1-ol: Processes of Transfer from Gas to and from Dilute Aqueous Solution. *J. Chem. Soc., Faraday Trans. 1* **1986**, *82*, 2547–2556.

(80) Dallas, A. J.; Carr, P. W. A Thermodynamic and Solvatochromic Investigation of the Effect of Water on the Phase Transfer Properties of Ocatn-1-ol. *J. Chem. Soc., Perkin Trans. 2* **1992**, (12), 2155–2161.

(81) Michielan, L.; Bacilieri, M.; Kaseda, C.; Moro, S. Prediction of the Aqueous Solvation Free Energy of Organic Compounds by Using Autocorrelation of Molecular Electrostatic Potential Surface Properties Combined with Response Surface Analysis. *Bioorg. Med. Chem.* **2008**, *16* (10), 5733–5742.

(82) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum: a Direct Utilization of Ab-initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55* (1), 117–129.

(83) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. Accurate transferable model for water, n-octanol, and n-hexadecane solvation free energies. *J. Phys. Chem. B* **2002**, *106* (42), 11009–11015.

(84) Daubert, T.; Danner, R. *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation, version 4.1.1*; Hemisphere: New York, 2003.

(85) Hales, J. L.; Ellender, J. H. Liquid Densities from 293 to 490 K of 9 Aliphatics Alcohols. *J. Chem. Thermodyn.* **1976**, *8* (12), 1177–1184.

(86) Smith, B. D.; Srivastava, R. *Thermodynamic Data for Pure Compounds. Part B. Halogenated Hydrocarbons and Alcohols*. Elsevier: Amsterdam, The Netherlands, 1986.

(87) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solut. Chem.* **1981**, *10* (8), 563–595.

(88) Rytting, E.; Lentz, K. A.; Chen, X. Q.; Qian, F.; Venkatesh, S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS J.* **2005**, *7* (1), E78–E105.

CT900214Y

JCTC Journal of Chemical Theory and Computation

Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening

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 20, 2009

Abstract: We present a new self-consistent reaction field continuum solvation model based on the generalized Born (GB) approximation for the bulk electrostatic contribution to the free energy of solvation. The new model improves on the earlier SM8 model by using the asymmetric descreening algorithm of Grycuk to treat dielectric descreening effects rather than the Coulomb field approximation; it will be called Solvation Model 8 with asymmetric descreening (SM8AD). The SM8AD model is applicable to any charged or uncharged solute 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 a group; all parameters are unique and continuous functions of geometry. This model employs a single set of parameters (solvent acidity-dependent intrinsic Coulomb radii for the treatment of bulk electrostatics and solvent description-dependent atomic surface tensions coefficients for the treatment of nonelectrostatic and short-range electrostatic effects). The SM8AD model was optimized over 26 combinations of theoretical levels including various basis sets (MIDI!, 6-31G*, 6-31+G*, 6-31+G**, 6-31G**, cc-pVDZ, DZVP, 6-31B*) and electronic structure methods (M05-2X, M05, M06-2X, M06, M06-HF, M06-L, mPW1PW, mPWPW, B3LYP, HF). It may be used with confidence with any level of electronic structure theory as long as self-consistently polarized Charge Model 4 or other self-consistently polarized charges compatible with CM4 charges are used, for example, CM4M charges can be used. With M05-2X/6-31G*, the SM8AD model achieves a mean unsigned error of 0.6 kcal/mol on average over 2 560 solvation free energies of tested aqueous and nonaqueous neutral solutes and a mean unsigned error of 3.9 kcal/mol on average over 332 solvation free energies of aqueous and nonaqueous ions.

1. Introduction

The electrostatic contribution[1] to the free energy of solvation results from the interaction of a solute with its reaction field, which is the electric field produced by the polarized charge density that the solute induces in the solvent. In self-consistent reaction field theory, the solute is polarized self-consistently by the reaction field, and it is the interaction of the mutually polarized solute and solvent subsystems that is called the electrostatic contribution.[2−4] The electrostatic contribution can be evaluated by solving the nonhomogeneous Poisson equation (NPE, by which we mean the Poisson equation for a nonhomogeneous medium in which the dielectric constant is unity inside the solute cavity but has a nonunit value outside it) in terms of the continuous charge density or by using alternative approaches, for instance, the generalized Born (GB) approximation,[5−11] which does not start with the NPE but instead employs a starting point based on representation of the solute as a collection of point charges, located at the nuclear positions. We have previously introduced a series of successively improved self-consistent reaction field solvation models based on the GB approxima-

* Corresponding authors. E-mail: cramer@umn.edu (C.J.C.) and truhlar@umn.edu (D.G.T.).

10.1021/ct900312z CCC: $40.75
Published on Web 08/12/2009

tion (SM5.4,[12] SM5.42,[13−15] SM5.43,[16] SM6,[17] SM8[18]) or the nonhomogeneous Poisson equation (SM5C,[19] SMD[20]) for bulk electrostatics combined with empirical atomic surface tensions[21] that account for cavity formation, dispersion, and solvent structure effects and for shorter-range nonbulk electrostatic effects.

The GB models involve partial atomic charges, whose interaction with the solvent and with each other is dielectrically screened by the polarized solvent and descreened by other parts of the solute. Conventional GB models treat dielectric descreening effects in terms of the so-called Born radii of individual atoms in the solute molecule. The SM5.4, SM5.42, SM5.43, SM6, and SM8 models employ the Born radius based on the Coulomb field (CF) approximation of Still et al.[10] for the electric displacement induced by the partial atomic charge in a dielectric. In the case of the CF approximation, a charge-induced dipole interaction varies as r^{-4}, where r is the distance between the partial atomic charge and a volume element of the continuum solvent. Starting with the Kirkwood distributed monopole model[22] for biopolymer electrostatics, Grycuk has shown[23] that, when the individual partial atomic charges are asymmetrically situated in the molecule, i.e., located near the dielectric boundary rather than at the center of the molecular surface, one can apparently estimate the dielectric descreening more accurately by using a shorter-range function proportional to r^{-6} in evaluation of the polarization component of the free energy of solvation. The shorter-range bulk electrostatics can be approximated with a corrected formula for the Born radius suggested by Grycuk.[23] Tjong and Zhou have demonstrated[24,25] that, as measured against electrostatic energies calculated by solving the NPE for a set of 55 proteins in water as well as in low-dielectric media,[24,25] the GB method using the improved Born radius formula is more accurate than any of the GB/CF models tested in their study. We refer the reader to several other studies[26−37] dealing with Grycuk's method[23] or other approaches to improve the Coulomb field approximation of Still et al.[10]

The key element of the present article is the incorporation of the new descreening algorithm of Grycuk[23] into a self-consistent reaction field solvation model that should improve on conventional GB models with regard to predicting solvation free energies, liquid-phase molecular geometries, solute response properties, such as NMR chemical shifts in solution, and other molecular properties. The resulting model will be called Solvation Model 8 with Asymmetric Descreening (SM8AD) because the SM8AD model extends the earlier SM8 model based on the CF algorithm of Still et al.[10] to those solutes for which the conventional GB/CF approach might be particularly poor, for instance, in situations when one or more polar residues in the solute molecule lie near the dielectric boundary. Both SM8AD and SM8 use the universal cavity dispersion solvent structure (CDS) formalism[38] to account for the nonbulk electrostatic contributions to the solvation free energy, arising from interactions between the solute and the solvent molecules in the first solvation shell. The solute electronic relaxation and, therefore, the solvent-induced changes in the atomic charges result from all the solute−solvent bulk electrostatic interactions and — in the case of clustered ions — from solute−cluster molecule interactions, but the CDS terms are added post-SCF and do not affect the solute charge distribution. The CDS terms are parametrized to include all of the deviations of the electrostatics from the assumed bulk model, such as the inexactness of the solute charge model and the inexactness of the solvent permittivity model including uncertainties in the precise definition of a solute cavity. The SM8AD model is a universal continuum model where "universal" denotes its applicability to any charged or uncharged solute 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).

The SM8AD model has been tested against the earlier SM8 and SMD models for 2 892 solvation data. In contrast to SM8AD and SM8, both of which employ GB approximation for bulk electrostatics and represent the solute molecule as a collection of partial atomic charges in a cavity, the SMD model[20] is based on the polarized continuous quantum mechanical charge density of the solute (the "D" in the name stands for "density"). The SMD bulk electrostatic contribution to the free energy of solvation arises from a self-consistent reaction field treatment that involves solution of the nonhomogeneous Poisson equation by the Integral Equation Formalism Polarizable Continuum Model (IEFPCM) algorithm.[39−42]

2. Computational Details

The free energy of solvation is defined as the standard-state free energy of transfer from the gas phase to the condensed phase according to

$$\Delta G_S^o = \Delta E_E + \Delta E_N + G_P + G_{CDS} + \Delta G_{conc}^o \quad (1)$$

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, 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. Because all calculations reported here are based on gas-phase geometries, the ΔE_N component is assumed to be zero in this article, although not in the model in general. The final term in eq 1 accounts for the concentration change between the gas-phase standard state and the liquid-phase standard state. Since here the same concentration (1 mol/L) is used in both the gaseous and solution phases, ΔG_{conc}^o is zero.[43] It would be 1.89 kcal/mol if we instead used a gas-phase standard state of 1 atm.

Bulk Electrostatics Formalism. The electronic relaxation term ΔE_E and the polarization term G_P in eq 1 comprise the bulk electrostatic contribution ($\Delta G_{EP} = \Delta E_E + G_P$) to the solvation free energy. The bulk electrostatic contribution is calculated from a self-consistent molecular orbital calculation,[13] where the generalized Born approximation[5−11] is used to calculate the polarization component G_P according to

$$G_P = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_{k,k'}^{\text{atoms}} q_k \gamma_{kk'} q_{k'} \quad (2)$$

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'. The Coulomb integrals $\gamma_{kk'}$ are calculated according to ref 10:

$$\gamma_{kk'} = (R_{kk'}^2 + \alpha_k \alpha_{k'} \exp[-R_{kk'}^2/d\alpha_k\alpha_{k'}])^{-1/2} \quad (3)$$

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 set to the value of 3.7, which was found to be optimal in earlier work.[17] For atoms and monatomic ions, the GB result reduces to the original Born formula:[44]

$$G_P = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\frac{q^2}{\alpha} \quad (4)$$

which is the exact classical result for the case where the solute is a conducting sphere of radius α, the charge q being located in the center of the sphere.

The effective Born radius of atom k in eq 3 can be expressed using the assumption that the electric displacement field induced by the charge q_k is a Coulomb field. In this case the charge-induced dipole interaction (G_{Pk}) varies as r^{-4}, according to the formula:

$$G_{P_k} = -\frac{q_k^2}{2}\left(1 - \frac{1}{\varepsilon}\right)\int_{\rho_k}^{\infty} \frac{dV}{4\pi r^4} \quad (5)$$

where r is the distance between the partial atomic charge q_k and the volume element of the continuum solvent, ρ_k is the so-called Coulomb radius of atom k that defines the boundary between the solute cavity ($r < \rho_k$) and the bulk solvent ($r \geq \rho_k$). The Coulomb field approximation leads to the following formula for the Born radius:[45,46]

$$\alpha_k = \left(\frac{1}{R'} + \int_{\rho_{Z_k}}^{R'} \frac{A_k(r)}{4\pi r^4} dr\right)^{-1} \quad (6)$$

In eq 6, R' is the radius of the sphere centered on atom k that completely engulfs all the other spheres centered on the other atoms of the solute, ρ_{Z_k} is the intrinsic Coulomb radius of atom k, which in the present study depends only on the atomic number Z_k, and $A_k(r)$ is the exposed (solvent accessible) area of a sphere of radius r that is centered on atom k. This area calculated here using the ASA algorithm[45] depends on the geometry of the solute and the radii of the spheres centered on all the other atoms in the solute.[47] If the analytical gradient of G_P with respect to the position $\mathbf{R}_a$ of an arbitrary atom a is desirable, the derivative of α_k must be taken according to[46]

$$\frac{\partial \alpha_k}{\partial \mathbf{R}_a} = -\alpha_k^2 \int_{\rho_{Z_k}}^{R'} \frac{\partial A_k(r)}{\partial \mathbf{R}_a} \frac{dr}{4\pi r^4} \quad (7)$$

where the derivative of $A_k(r)$ is evaluated analytically using the ASA formulation.[45] The GB approximation that uses eq 6 for the Born radius will be called the generalized Born Coulomb field (GB/CF) approximation hereafter. The earlier SM8 model[18] uses the GB/CF approximation for bulk electrostatics.

Following Grycuk's formulation,[23] the performance of the GB approximation can be improved by replacement of eq 6 with an alternative functional form for the Born radius α_k, which is given as follows:

$$\alpha_k = \left(\frac{1}{R'^3} + \int_{\rho_{Z_k}}^{R'} \frac{3A_k(r)}{4\pi r^6} dr\right)^{-1/3} \quad (8)$$

The GB approximation that uses eq 8 for the Born radius will be called the GB approximation with asymmetric descreening (GB/AD). The SM8AD model proposed in the present study uses the new GB/AD approximation for bulk electrostatics. The derivative of α_k with respect to the position $\mathbf{R}_a$ of an arbitrary atom a is given as follows:

$$\frac{\partial \alpha_k}{\partial \mathbf{R}_a} = -\alpha_k^4 \int_{\rho_{Z_k}}^{R'} \frac{\partial A_k(r)}{\partial \mathbf{R}_a} \frac{dr}{4\pi r^6} \quad (9)$$

where the derivative of $A_k(r)$ is evaluated analytically using the ASA formulation.[45] The $\partial\alpha_k/\partial\mathbf{R}_a$ derivative in eq 9 can be used in analytical computation of the $\partial G_P/\partial \mathbf{R}_a$ gradient within the GB/AD formalism along with the other components of $\partial G_P/\partial \mathbf{R}_a$ derived in earlier work.[46]

According to eq 2, the GB models are based on partial atomic charges, and therefore, their accuracy for a particular level of electronic structure theory depends on whether meaningful partial charges can be computed for that theoretical level. Like the earlier SM8 model, the new SM8AD model is designed to employ class IV charge models, in particular, Charge Model 4 (CM4)[17] and Charge Model 4M (CM4M).[48] 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,[49] Löwdin,[50] or redistributed Löwdin[51] population analyses. This allows one to shift the focus of the modeling effort away from the description of the solute toward the various components of the solvation process.[18] In addition, CM4 and CM4M charges yield more accurate long-range electrostatic potentials than population analysis charges, and this makes the solvation models based on such charges more physical.

Cavity Dispersion Solvent Structure Formalism. The G_{CDS} contribution to the free energy of solvation in eq 1 is given by

$$G_{CDS} = \sum_k^{\text{atoms}} \sigma_k A_k(\mathbf{R}, \{R_{Z_k} + r_s\}) + \sigma^{[M]} \sum_k^{\text{atoms}} A_k(\mathbf{R}, \{R_{Z_k} + r_s\}) \quad (10)$$

where σ_k is the atomic surface tension of atom k, A_k is the solvent accessible surface area (SASA)[52,53] of atom k, and $\sigma^{[M]}$ is the molecular surface tension. The SASA depends on the geometry $\mathbf{R}$, the set $\{R_Z\}$ 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 according to the ASA algorithm.[45] The van der Waals radii used in eq 10 for the SASA calculation are not the same as the intrinsic Coulomb radii used in eqs 6−9 for solution of the bulk electrostatic problem. In fact, in eq 10 we use the values of R_Z fixed at Bondi's values[54] and the value of r_s fixed at the value[16] of 0.4 Å, whereas the intrinsic Coulomb radii used in eqs 6−9 have been optimized according to the algorithm that will be described later in the article.

The atomic surface tensions are given by

$$\sigma_k = \tilde{\sigma}_{Z_k} + \sum_{k'}^{\text{atoms}} \tilde{\sigma}_{Z_k Z_{k'}} T_k(\{Z_{k'}, R_{kk'}\}) \quad (11)$$

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 $T_k(\{Z_{k'}, R_{kk'}\})$ is a geometry-dependent switching function called a cutoff tanh, or COT; this function is described in the Supporting Information.

The atomic surface tensions in eq 11 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 as follows:

$$\tilde{\sigma}_i = \tilde{\sigma}_i^{[n]} n + \tilde{\sigma}_i^{[\alpha]} \alpha + \tilde{\sigma}_i^{[\beta]} \beta \quad (12)$$

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[55−58] hydrogen bond acidity parameter of the solvent (which Abraham denotes as $\sum\alpha_2$), β is Abraham's hydrogen bond basicity parameter of the solvent (which Abraham denotes as $\sum\beta_2$), and $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, and $\tilde{\sigma}_i^{[\beta]}$ are empirical parameters that depend on i.

The molecular surface tension in eq 10 is also a function of solvent descriptors, and it is given by

$$\sigma^{[M]} = \tilde{\sigma}^{[\gamma]}({}^{\gamma}/_{\gamma_o}) + \tilde{\sigma}^{[\phi^2]}\phi^2 + \tilde{\sigma}^{[\psi^2]}\psi^2 + \tilde{\sigma}^{[\beta^2]}\beta^2 \quad (13)$$

where γ is the macroscopic surface tension of the solvent at air/solvent interface at 298.15 K; we express surface tension in units of cal mol^{-1} Å^{-2} (note that 1 dyn/cm = 1.43932 cal mol^{-1} Å^{-2}), and γ_o = 1 cal mol^{-1} Å^{-2}, ϕ^2 is the square of the fraction of non-hydrogenic atoms in the solvent molecule that are aromatic carbon atoms (carbon aromaticity), ψ^2 is the square of the fraction of non-hydrogenic atoms in the solvent molecule 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. According to eq 10, the molecular surface tension is multiplied by the total SASA of the given solute. The latter is equal to the sum of the SASAs of each of the individual atoms in the solute.

The SM8AD model may be applied to any medium for which the relevant macroscopic descriptors such as dielectric constant, refractive index, bulk surface tension, and acidity and basicity parameters are either known or may be estimated. Because water is so important as a solvent, there are advantages to using a less general and more specifically parametrized solvation model for this solvent. Therefore, water is treated as a special solvent that is given its own set of surface tension coefficients, so that eqs 12 and 13 are not used for water, and the molecular surface tension $\sigma^{[M]}$ in eq 10 is set to equal zero. Thus, in the case when the model is employed to compute solvation free energies in aqueous solvent, the parameters $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ used in eq 11 to obtain σ_k are simply numbers that do not depend on solvent descriptors.

SM8AD Training Set. The SM8AD training set is part of the Minnesota Solvation Database[59] and it contains 2 892 experimental solvation data for 233 ionic and 482 neutral solutes composed of H, C, N, O, F, Si, P, S, Cl, or Br. All standard-state 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. The SM8AD primary testing set contains six data subsets:

(i) 274 aqueous free energies of solvation for 274 neutral compounds;[17,18,20]

(ii) 71 aqueous free energies of solvation for an additional 71 neutral compounds;[60,61]

(iii) 2 072 free energies of solvation in 90 nonaqueous solvents for 276 neutral solutes (232 of the 276 solutes are also included in the set of 274 aqueous solutes, and 44 solutes are additional);[18,20]

(iv) 143 transfer free energies between water and 15 organic solvents for an additional 93 neutral solutes;[18,20]

(v) 112 aqueous free energies of solvation for 112 selectively clustered singly charged ions (there are 81 unclustered ions and 31 clustered ions);[17,18,20,62]

(vi) 220 free energies of solvation in acetonitrile, dimethyl sulfoxide (DMSO), and methanol for 166 singly charged ions (45 ions out of 166 are also included in the set of 81 unclustered aqueous ions, and 121 ions are additional).[18,20,63]

Several technical points should be mentioned here. The SM8AD training set is similar to those used in the parametrization of the SM8[18] and SMD models,[20] except for 71 new solutes added to the previously used set of 274 aqueous neutrals.[18,20] The addition of the 71 compounds, many of which are compounds of complex functionality (for instance, agricultural pesticides), is essential in extending the applicability of the SM8AD model to the classes of compounds which are poorly represented in the training sets of many solvation models (for instance, compounds with oxidized sulfur and phosphorus functionalities). The subset of 71 aqueous solutes includes 13 compounds from the set of 17 compounds described in ref 60 and 58 compounds from the set of 63 compounds presented in the SAMPL1 challenge[61] organized by Openeye Software (February, 2008). The remaining four of the 17 solutes[60] (benzylbromide, benzylchloride, diethyl sulfide, and 1,4-dioxane) were already included in the set of 274 aqueous solutes. Two compounds out of the 63 SAMPL1 solutes,[61] in particular cup08042 and cup08062, were discarded from consideration here for the reason explained in ref 64, and the other three compounds

in SAMPL1 but not in subset ii (dichlobenil, fenuron, methyl parathion) were already among the 274 aqueous solutes of subset i. The SM8AD training set has five fewer data than the SM8AD primary testing set because subset ii includes five sulfonylureas (bensulfuron methyl, chlorimuron ethyl, metsulfuron methyl, sulfometuron methyl, thifensulfuron), which were used for testing the SM8AD model but not used in the SM8AD parametrization because we considered the experimental solvation free energy targets assigned for these compounds to be suspect, see ref 64 for more detail. Thus, for the SM8AD parametrization, we use only 66 out of 71 additional aqueous solutes.

The 143 transfer free energies associated with transferring the solute from aqueous solution to an organic solvent were determined directly from the corresponding experimental partition coefficients. The transfer free energy data are included in this training set 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.

The single-ion solvation free energies were evaluated in previous work[17,63] based on the corresponding thermochemical cycle that relates the solvation free energy of the cation BH^+ or the anion A^- to the gas-phase basicity of the base B or the gas-phase acidity of the acid AH using the reference solvation free energy of the proton.[18,20] For the 1:1 M standard-state free energies of solvation for the proton in acetonitrile, DMSO, methanol, and water, we use −260.2, −273.3, −263.5, and −265.9 kcal/mol, respectively.[62,63] For aqueous ions, 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 an unclustered form). The criteria for whether to cluster an ion are explained elsewhere.[17] In all cases, the clusters were represented by a single, lowest-energy conformation.

The estimated average uncertainty for experimental free energies of solvation and transfer free energies of neutral solutes in subsets of 274, 2072, and 143 data is 0.2 kcal/mol.[17,60] The uncertainty for free energies of solvation for a set of 71 aqueous data is about 1 kcal/mol on average in the range of 0.1−1.93 kcal/mol.[61] The estimated average uncertainty for solvation free energies of ionic solutes is 3 kcal/mol.[17]

Table 1 lists 92 solvents including water used in the SM8AD parametrization. The corresponding values of solvent descriptors such as dielectric constant, refractive index, bulk (macroscopic) surface tension, and acidity and basicity parameters were taken from the Minnesota Solvent Descriptor Database[65] (these values are given in the Supporting Information). Experimental values for the 2 892 solvation free energies are also given in the Supporting Information. All computed solvation free energies in this study are based on rigid, gas-phase geometries. The molecular geometries of all unclustered neutral and ionic solutes were optimized at the mPW1PW[66]/MIDI![67,68] level of electronic structure theory.[59] The molecular geometries of aqueous clustered ions were optimized at the B97−1[69]/MG3S[70] level of theory.[17]

Table 1. Solvents Used in the SM8AD Training Set[a]

acetic acid	*dibutyl ether*	methylene chloride
acetonitrile*	*o*-dichlorobenzene	*N*-methylformamide
acetophenone	*1,2-dichloroethane*	4-methyl-2-pentanone
aniline	*diethyl ether*	2-methylpyridine
anisole	diisopropyl ether	*nitrobenzene*
benzene	*N,N*-dimethylacetamide	nitroethane
benzonitrile	*N,N*-dimethylformamide	nitromethane
benzyl alcohol	2,6-dimethylpyridine	*o*-nitrotoluene
bromobenzene	dimethyl sulfoxide*	nonane
bromoethane	dodecane	nonanol
bromoform	ethanol	octane
bromooctane	ethoxybenzene	*octanol*
n-butanol	*ethyl acetate*	pentadecane
sec-butanol	ethylbenzene	pentane
butanone	fluorobenzene	pentanol
butyl acetate	1-fluoro-*n*-octane	perfluorobenzene
n-butylbenzene	*heptane*	phenyl ether
sec-butylbenzene	heptanol	propanol
t-butylbenzene	hexadecane	pyridine
carbon disulfide	hexadecyl iodide	tetrachloroethene
carbon tetrachloride	*hexane*	tetrahydrofuran
chlorobenzene	hexanol	tetrahydrothiophene dioxide
chloroform	iodobenzene	tetralin
chlorohexane	isobutanol	toluene
m-cresol	isooctane	tributylphosphate
cyclohexane	isopropanol	triethylamine
cyclohexanone	isopropylbenzene	1,2,4-trimethylbenzene
decalin (mixture)	*p*-isopropyltoluene	undecane
decane	mesitylene	water*
decanol	methanol*	xylene (mixture)
1,2-dibromoethane	methoxyethanol	

[a] All solvents except methanol have data for free energies of solvation for neutral solutes. The asterisk denotes the solvents which have data for free energies of solvation for ionic solutes. The names of 15 solvents for which we used solvent−water transfer free energies of neutral solutes are italicized.

SM8AD Parametrization. As for the earlier SM8 model,[18] the SM8AD parametrization effort is focused on two types of parameters: (i) the intrinsic Coulomb radii used for construction of the cavities for the bulk electrostatic calculation; and (ii) the atomic surface tensions $\tilde{\sigma}_Z$ and $\tilde{\sigma}_{ZZ'}$ in eq 11 and the parameters $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, $\tilde{\sigma}^{[\beta^2]}$ in eq 13 used for the nonbulk electrostatic calculation within the CDS formalism.

We recall here that there is no thermodynamically unique way to separate the electrostatic contribution to the free energy of solvation from the nonelectrostatic one because only their sum is a state function and physical observable.[71−73] Therefore, continuum solvation models differ from one another in the way in which the electrostatic and nonelectrostatic components are defined. It is also widely recognized that the electrostatic terms may depend strongly on the model radii.[12,74−77] Keeping in mind that the magnitudes of solvation free energies of ions are much larger than those of neutral solutes and are dominated by large electrostatic contributions, an optimization of intrinsic atomic Coulomb radii to provide accurate solvation free energies of ions is a reasonable way to determine these parameters. If one assumes that the solute cavity is charge independent, then by using the same radii one might also achieve a reasonable estimation of the bulk electrostatics for cases where electrostatic and nonelectrostatic terms are comparable. The nonelectrostatic terms can then be defined as the difference between the experimentally available and path-independent total free

Table 2. Intrinsic Coulomb Radii (Å) of Various Models and Bondi's van der Waals Radii (Å)

atom	Z	SM8AD(aq)[a]	SM8AD (DMSO)[a]	SM8(aq)[b]	SM8 (DMSO)[b]	SMD(aq)[c]	SMD (DMSO)[c]	Bondi[d]
H	1	1.02	0.80	1.02	0.80	1.20	1.20	1.20
C	6	1.75	1.75	1.57	1.57	1.85	1.85	1.70
N	7	1.94	1.94	1.61	1.61	1.89	1.89	1.55
O	8	1.52	2.29	1.52	2.18	1.52	2.29	1.52
F	9	1.68	1.68	1.47	2.63	1.73	1.73	1.47
Si	14	2.47	2.47	2.10	2.10	2.47	2.47	2.10
P	15	2.12	2.12	1.80	2.13	2.12	2.12	1.80
S	16	2.16	2.16	2.12	2.45	2.49	2.49	1.80
Cl	17	2.40	2.40	2.02	2.63	2.38	2.38	1.75
Br	35	2.62	2.62	2.60	2.85	3.06	3.06	1.85

[a] The SM8AD radii for H and O are defined as a function of Abraham's hydrogen bond acidity parameter (α) for a given solvent according to eqs 14 and 15. We adopted the SMD values[20] for the SM8AD radii for Si and P. [b] The intrinsic Coulomb radii used by the SM8 model for any solute in water and DMSO.[18] [c] The intrinsic Coulomb radii used by the SMD model for any solute in water and DMSO.[20] [d] Bondi's values of van der Waals radii.[54]

energy of solvation and the modeled electrostatic contribution. We use this approach in the present study.

As in earlier work,[18] we optimize the SM8AD radii in calculations only on ions, then fix these parameters and optimize the nonbulk electrostatic term on data for neutrals. The optimization of SM8AD radii was done by minimizing the sum of mean squared errors calculated over 332 data points corresponding to 220 ions in acetonitrile, dimethyl sulfoxide, methanol and 112 selectively clustered ions in water. The ΔG_{EP} values used in the SM8AD optimization of radii were calculated with a locally modified version of Gaussian 03[78] called the Minnesota Gaussian Solvation Module (MN-GSM),[79] the M05-2X density functional,[80] and the 6-31G* basis set.[81] For simplicity, we did not use basis sets with diffuse functions in the optimization of radii. Although diffuse functions are important for calculations on small gas-phase anions, especially for atoms and diatomic molecules, they are less important for large anions and in solution where diffuse charge clouds of gas-phase anions are contracted (although perhaps only a little) by solvation effects.

After a testing of various optimization schemes we have found that the hydrogen and oxygen radii strongly depend on the solvent's value of Abraham's hydrogen bond acidity parameter α (the parameter that Abraham calls $\sum\alpha_2$),[55−58] whereas the radii for other elements do not demonstrate such a dependency. To approximate the hydrogen radius we adopted the scheme previously elaborated for the SM8 model:[18]

$$\rho = \begin{cases} 1.02 & \alpha \geq 0.43 \\ 1.02 - 0.52\,(0.43 - \alpha) & \alpha < 0.43 \end{cases} \quad (14)$$

where α is the solvent's value of Abraham's hydrogen bond acidity parameter. For the oxygen radius we adopted the SMD dependence of the radius on the α parameter:[20]

$$\rho = \begin{cases} 1.52 & \alpha \geq 0.43 \\ 1.52 + 1.8\,(0.43 - \alpha) & \alpha < 0.43 \end{cases} \quad (15)$$

Using eqs 14 and 15 to define the hydrogen and oxygen radii, we optimized the radii for the remaining elements (C, N, F, S, Cl, Br) by taking them to be independent of the solvent. Since the training set of 332 ions used in the SM8AD optimization does not include any solute containing silicon or phosphorus we have opted to fix the SM8AD radii for Si and P at their SMD values.[20] Table 2 lists the optimized values of SM8AD intrinsic Coulomb radii given for water (α = 0.82) and DMSO (α = 0) compared to the values of the radii used by our most recent solvation models SM8[18] and SMD[20] and to the values of Bondi.[54] This set includes the radii for H, C, N, O, F, Si, P, S, Cl, and Br. For any other atom, the SM8AD model can use the van der Waals radii of Bondi[54] and Mantina et al.[82] for those atoms for which they defined radii; in cases where the atomic radius is not given in those papers, a radius of 2.0 Å is used.

The cavity dispersion solvent structure term (eq 1) associated with nonbulk electrostatic effects is parametrized by means of the atomic surface tension coefficients $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$ (eq 12) and the molecular surface tension coefficients $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$ (eq 13) for nonaqueous solvents and the atomic surface tension coefficients $\tilde{\sigma}_i$ (eq 11) for water. The optimization of the sigma parameters involves a minimization of the following error function:

$$\chi = \sum_{J=1}^{2\,555} \left[\Delta G_S^o(\text{expt.}, J) - \frac{1}{26} \sum_{j=1}^{26} \Delta G_{EP}(j, J) - G_{CDS}(J) \right]^2 \quad (16)$$

where the *J*-summation runs over all data points in the neutral training set, including 2 412 solvation free energies and 143 transfer free energies, the *j*-summation runs over all levels of electronic structure theory used in the parametrization, ΔG_s^o(expt., *J*) is the experimental standard-state solvation or transfer free energy, $\Delta G_{EP}(j, J)$ is the bulk electrostatic energy computed for a given theoretical level (or bulk electrostatic contribution to a transfer free energy), and $\Delta G_{CDS}(J)$ is the nonbulk electrostatic energy defined by eqs 10−13 (or the corresponding contribution to a transfer free energy). The 2 412 data points do not include the five sulfonylureas in water mentioned before, which were used in the SM8AD testing but not in the SM8AD parametrization. The CDS term was parametrized by averaging over 26 combinations of charge models, electronic structure methods, and basis sets as listed in Table 3. Namely, we used two charge models (CM4[17] or CM4M[48]), nine density functionals (M05-2X,[80] M05,[80,83] M06-2X,[84,85] M06,[84,85] M06-HF,[85,86] M06-L,[85,87] mPW-PW,[66] mPW1PW,[66] and B3LYP[88−91]), the Hartree−Fock (HF) method, and nine basis sets for which the charge models

Table 3. Twenty Six Combinations of Charge Models, Electronic Structure Levels, and Basis Sets Used in the SM8AD Parametrization

charge model	functional	basis set
CM4	M05-2X	MIDI!
CM4	M05-2X	MIDI!6D
CM4	M05-2X	6-31G*
CM4	M05-2X	6-31+G*
CM4	M05-2X	6-31+G**
CM4	M05-2X	6-31G**
CM4	M05-2X	cc-pVDZ
CM4	M05-2X	DZVP
CM4	M05-2X	6-31B*
CM4M	M06-2X	MIDI!
CM4M	M06-2X	MIDI!6D
CM4M	M06-2X	6-31G*
CM4M	M06-2X	6-31+G*
CM4M	M06-2X	6-31+G**
CM4M	M06-2X	6-31G**
CM4M	M06-2X	cc-pVDZ
CM4M	M06-2X	DZVP
CM4M	M06-2X	6-31B*
CM4	M05	6-31G*
CM4M	M06	6-31G*
CM4M	M06-HF	6-31G*
CM4M	M06-L	6-31G*
CM4	mPWPW	6-31G*
CM4	mPW1PW	6-31G*
CM4	B3LYP	6-31G*
CM4	HF	6-31G*

Table 4. Nine Additional Combinations of Charge Models, Electronic Structure Levels, and Basis Sets Used in the SM8AD Testing but not Used in the SM8AD Parametrization[a]

charge model	electronic structure level	basis set
CM4	M05-2X	6-31B**
CM4M	M06-2X	6-31B**
CM4	M05-2X	cc-pVTZ
LPA	M05-2X	6-31G*
RLPA	M05-2X	6-31+G**
CM2	AM1	n.a.
CM3	AM1	n.a.
CM2	PM3	n.a.
CM3	PM3	n.a.

[a] The n.a. denotes not applicable.

are available (MIDI!,[67,68] MIDI!6D,[67,68] 6-31G*,[81] 6-31+G*,[81] 6-31+G**,[81] 6-31G**,[81] cc-pVDZ,[92] DZVP,[93] or 6-31B*[94]). We used the CM4M charge model only with the M06 suite[85] of density functionals for which this model was designed,[48] and we use the CM4 model with any other density functional or with the Hartree−Fock method.

Table 4 lists nine additional levels of theory that were not used in parametrization, but that are tested in the present study. In particular, we tested the SM8AD model using partial atomic charges obtained from Löwdin population analysis (LPA)[50] or redistributed LPA (RLPA),[51] the 6-31B**[94] and cc-pVTZ[92] basis sets, and the semiempirical models Austin Model 1 (AM1)[95] and Parametrized Model 3 (PM3)[96] combined with Charge Model 2 (CM2)[97] and Charge Model 3 (CM3).[98]

In the present study we have employed essentially the same strategy for optimizing the sigma coefficients $\tilde{\sigma}_i^{[n]}$, $\tilde{\sigma}_i^{[\alpha]}$, $\tilde{\sigma}_i^{[\beta]}$, $\tilde{\sigma}_i^{[\text{water}]}$, $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\phi^2]}$, $\tilde{\sigma}^{[\psi^2]}$, and $\tilde{\sigma}^{[\beta^2]}$, as we did in our earlier work,[18,20] with one exception. In the previous studies,[18,20] we optimized these parameters separately for atoms involving at most H, C, N, and O, atoms involving F, S, Cl, and Br, and atoms involving Si and P. In the present study, we have abandoned this scheme, and we optimized all the sigma coefficients simultaneously. The final set of SM8AD surface tension coefficients is listed in Tables 5 and 6. The SM8AD model uses 46 nonzero surface tension parameters compared to the 54 parameters that are used by SM8. Functional forms for atomic surface tensions used by SM8AD are given in the Supporting Information.

Table 5. Atomic Surface Tension Parameters (cal mol^{-1} Å^{-2}) for SM8AD that Depend on Atomic Numbers[a]

i	$\tilde{\sigma}_i^{[\text{water}]}$	$\tilde{\sigma}_i^{[n]}$	$\tilde{\sigma}_i^{[\alpha]}$	$\tilde{\sigma}_i^{[\beta]}$
H	32.74	25.35		
C	65.00	32.05	146.39	
H, C	−41.80	−69.24		
C, C	−50.56	−44.81	−89.04	
O	−79.20	−15.09		−41.39
H, O	−54.88	−84.48		−160.75
O, C	183.85		276.59	
O, O	76.58			
N		44.96	−100.55	
H, N	−111.09	−93.29		
C, N	41.77	−73.49	174.33	
N, C	−57.40		−82.30	
O, N	176.32			136.00
F	27.28			
Cl	−3.36	−23.09		
Br	−13.16	−31.68		
S	−20.49	−34.96		
O, P	151.00		283.85	
O, S	277.16			
S, P	76.43			
Si		−78.22		

[a] Any possible atomic number-dependent surface tension parameter that is not in this table is set equal to zero in SM8AD.

Table 6. Molecular Surface Tension Parameters (cal mol^{-1} Å^{-2}) for SM8AD that Do Not Depend on Atomic Numbers

$\tilde{\sigma}^{[\gamma]}$	0.19
$\tilde{\sigma}^{[\phi^2]}$	−2.71
$\tilde{\sigma}^{[\psi^2]}$	−8.25
$\tilde{\sigma}^{[\beta^2]}$	2.10

3. Results

Tables 7 and 8 show the mean signed errors (MSE) and the mean unsigned errors (MUE) in 2 892 solvation energies calculated by SM8AD, SM8, and SMD used in combination with five selected electronic structure methods (M05-2X, M06-2X, mPW1PW, B3LYP, HF) and four basis sets (MIDI!6D, 6-31G*, 6-31+G**, 6-31G**). The mean signed and unsigned errors as well as the root mean squared errors over all theoretical levels used in the present study are given in the Supporting Information. Tables 9−12 show in more detail the errors in SM8AD solvation energies calculated using M05-2X/6-31G*. Table 9 gives a breakdown of the errors in calculated aqueous solvation free energies for neutrals by solute class. In Tables 10 and 11, the errors are broken down by solute class for calculated solvation free energies of neutral solutes in nonaqueous solvents and for calculated transfer free energies, respectively. Table 12 gives

Table 7. Mean Signed Errors (kcal/mol) in the Free Energies of Solvation Calculated using SM8AD, SM8, and SMD[a]

				2 560 neutral data					332 ionic data	
salvation model	charge model	functional	basis set	aqueous data			nonaqueous data	transfer energies	aqueous data	nonaqueous data
				274	66[b]	5[c]	2 072	143	112	220
SM8AD	CM4	M05-2X	MIDI!6D	0.09	0.78	−2.00	0.04	−0.09	0.62	−1.69
SM8	CM4	M05-2X	MIDI!6D	0.28	0.97	−8.26	0.14	−0.14	1.55	−1.59
SMD		M05-2X	MIDI!6D	0.56	1.60	−4.35	0.34	−0.48	4.59	0.95
SM8AD	CM4	M05-2X	6-31G*	−0.02	0.40	−3.02	−0.02	0.04	−0.18	−2.36
SM8	CM4	M05-2X	6-31G*	0.20	0.69	−8.96	0.08	−0.08	0.71	−2.31
SMD		M05-2X	6-31G*	−0.06	0.40	−8.64	−0.06	−0.02	3.74	0.42
SM8AD	CM4	M05-2X	6-31+G**	0.00	0.03	−6.05	−0.01	−0.22	0.22	−2.62
SM8	CM4	M05-2X	6-31+G**	0.20	0.29	−12.10	0.07	−0.21	1.01	−2.61
SMD		M05-2X	6-31+G**	−0.73	−0.80	−10.75	−0.50	0.24	4.32	0.60
SM8AD	CM4	M05-2X	6-31G**	0.03	0.45	−2.19	−0.01	−0.05	−0.16	−2.30
SM8	CM4	M05-2X	6-31G**	0.24	0.70	−8.38	0.09	−0.13	0.66	−2.30
SMD		M05-2X	6-31G**	−0.12	0.24	−8.93	−0.09	0.00	3.79	0.38
SM8AD	CM4M	M06-2X	6-31G*	−0.33	−0.30	−5.50	−0.16	0.36	−0.32	−2.44
SM8	CM4M	M06-2X	6-31G*	−0.11	0.12	−10.77	−0.08	0.19	0.62	−2.37
SMD		M06-2X	6-31G*	0.22	0.87	−7.43	0.10	−0.21	4.08	0.78
SM8AD	CM4	mPW1PW	6-31G*	−0.30	−0.34	−5.43	−0.14	0.29	−0.23	−2.24
SM8	CM4	mPW1PW	6-31G*	−0.08	0.10	−10.73	−0.06	0.13	0.71	−2.18
SMD		mPW1PW	6-31G*	0.23	0.94	−6.99	0.10	−0.24	4.25	0.92
SM8AD	CM4	B3LYP	6-31G*	−0.12	−0.08	−3.68	−0.05	0.26	0.04	−1.94
SM8	CM4	B3LYP	6-31G*	0.12	0.34	−9.14	0.07	0.10	0.98	−1.85
SMD		B3LYP	6-31G*	0.59	1.38	−6.13	0.31	−0.38	4.72	1.42
SM8AD	CM4	HF	6-31G*	0.03	0.18	−3.49	−0.01	0.05	−0.57	−2.90
SM8	CM4	HF	6-31G*	0.30	0.51	−9.29	0.14	−0.05	0.36	−2.79
SMD		HF	6-31G*	−0.71	−1.54	−14.14	−0.33	0.71	2.67	−0.70

[a] Description of the SM8AD data set is given in Section 2. [b] Part of a subset of 71 aqueous data used in the SM8AD parametrization but not used in the SM8 and SMD parametrization. [c] Five sulfonylureas as part of a subset of 71 aqueous data not used in the parametrization of either model.

Table 8. Mean Unsigned Errors (kcal/mol) in the Free Energies of Solvation Calculated using SM8AD, SM8, and SMD[a]

				2 560 neutral data					332 ionic data	
salvation model	charge model	functional	basis set	aqueous data			nonaqueous data	transfer energies	aqueous data	nonaqueous data
				274	66[b]	5[c]	2 072	143	112	220
SM8AD	CM4	M05-2X	MIDI!6D	0.59	1.62	2.85	0.57	0.63	3.33	3.98
SM8	CM4	M05-2X	MIDI!6D	0.65	1.93	8.26	0.59	0.72	3.36	4.57
SMD		M05-2X	MIDI!6D	0.88	2.34	4.62	0.70	0.82	4.83	4.19
SM8AD	CM4	M05-2X	6-31G*	0.53	1.72	3.23	0.56	0.63	2.96	4.38
SM8	CM4	M05-2X	6-31G*	0.59	2.01	8.96	0.57	0.73	3.00	4.93
SMD		M05-2X	6-31G*	0.60	1.85	8.64	0.63	0.64	4.13	4.15
SM8AD	CM4	M05-2X	6-31+G**	0.67	2.41	6.05	0.61	0.80	2.50	4.90
SM8	CM4	M05-2X	6-31+G**	0.72	2.34	12.10	0.61	0.71	3.11	5.70
SMD		M05-2X	6-31+G**	0.96	1.97	10.75	0.79	0.68	4.64	4.08
SM8AD	CM4	M05-2X	6-31G**	0.53	1.74	2.71	0.56	0.59	2.90	4.38
SM8	CM4	M05-2X	6-31G**	0.59	2.01	8.38	0.57	0.68	2.94	4.95
SMD		M05-2X	6-31G**	0.62	1.84	8.93	0.64	0.64	4.15	4.13
SM8AD	CM4M	M06-2X	6-31G*	0.62	1.69	5.50	0.59	0.72	2.79	4.38
SM8	CM4M	M06-2X	6-31G*	0.57	1.91	10.77	0.58	0.76	2.92	4.94
SMD		M06-2X	6-31G*	0.62	1.94	7.43	0.63	0.69	4.41	4.13
SM8AD	CM4	mPW1PW	6-31G*	0.60	1.69	5.43	0.59	0.71	2.86	4.28
SM8	CM4	mPW1PW	6-31G*	0.56	1.96	10.73	0.58	0.77	2.96	4.84
SMD		mPW1PW	6-31G*	0.63	1.97	6.99	0.63	0.69	4.54	4.20
SM8AD	CM4	B3LYP	6-31G*	0.59	1.60	3.68	0.57	0.72	3.00	4.14
SM8	CM4	B3LYP	6-31G*	0.58	1.87	9.14	0.58	0.79	3.07	4.68
SMD		B3LYP	6-31G*	0.81	2.11	6.13	0.67	0.74	4.94	4.24
SM8AD	CM4	HF	6-31G*	0.57	1.59	3.52	0.56	0.65	3.01	4.82
SM8	CM4	HF	6-31G*	0.65	1.93	9.29	0.58	0.74	2.91	5.26
SMD		HF	6-31G*	0.92	2.29	14.14	0.73	0.97	3.39	4.50

[a] See footnote a of Table 7. [b] See footnote b of Table 7. [c] See footnote c of Table 7.

a breakdown of MSEs and MUEs in 332 solvation free energies for ions by solute class. The errors in the M05-2X/6-31G* solvation energies for neutrals broken down by solvent name are given in the Supporting Information.

The SM8AD and SM8 solvation energies based on DFT or the Hartree−Fock method were calculated with a locally modified version of Gaussian 03[78] called MN-GSM.[79] The SMD solvation energies were calculated with the GESOL

Table 9. Mean Signed and Mean Unsigned Errors in Aqueous Solvation Free Energies Calculated using SM8AD and SM8 with CM4/M05-2X/6-31G* by Solute Class[a]

		SM8AD		SM8	
solute class	*N*	MSE[b]	MUE[b]	MSE[b]	MUE[b]
274 Data					
H_2, NH_3, H_2O, $(H_2O)_2$	4	−1.74	1.74	−2.06	2.06
unbranched alkanes	8	−0.81	0.81	−0.85	0.85
branched alkanes	5	−0.75	0.75	−0.77	0.77
cycloalkanes	5	−0.47	0.47	−0.70	0.70
alkenes	9	−0.37	0.41	−0.30	0.40
alkynes	5	0.20	0.24	0.43	0.43
arenes	8	0.06	0.23	−0.09	0.24
alcohols	12	0.52	0.53	0.59	0.59
phenols	4	0.46	0.52	0.60	0.60
ethers	12	0.13	0.43	0.54	0.61
aldehydes	6	0.19	0.29	0.16	0.27
ketones	12	−0.12	0.30	0.30	0.36
carboxylic acids	5	0.42	0.42	0.74	0.74
esters	13	−0.43	0.47	−0.03	0.16
peroxides	3	−0.16	0.28	0.14	0.14
bifunctional H, C, O compounds	5	0.49	0.59	0.79	0.79
aliphatic amines	15	0.13	0.63	0.17	0.61
anilines	7	0.37	0.37	0.54	0.54
aromatic N-heterocycles (1 N)	10	0.41	0.41	0.46	0.46
aromatic N-heterocycles (2 Ns)	3	−0.70	0.70	−0.11	0.77
nitriles	4	−0.43	0.43	0.89	0.89
hydrazines	3	0.47	0.95	0.19	0.88
bifunctional H, C, N compounds	3	0.32	0.34	0.44	0.61
amides	4	0.45	0.62	0.85	0.97
ureas	2	0.27	0.59	0.12	0.12
nitrohydrocarbons	7	−0.11	0.30	0.30	0.40
bifunctional H, C, N, O compounds	3	−0.35	0.35	0.18	0.18
fluoroalkanes	5	−0.31	0.42	−0.07	0.30
fluoroarene	1	−0.09	0.09	−0.05	0.05
chloroalkanes	13	−0.17	0.32	0.10	0.26
chloroalkenes	6	0.17	0.27	0.68	0.68
chloroarenes	8	−0.43	0.43	−0.28	0.28
bromoalkanes	9	−0.11	0.12	−0.21	0.21
bromoalkene	1	−0.08	0.08	0.04	0.04
bromoarenes	4	−0.23	0.23	−0.41	0.41
multihalogenhydrocarbons	12	−0.34	0.37	0.03	0.25
halogenated bifunctional compounds	9	0.30	0.95	1.33	1.43
thiols	4	0.73	0.73	0.63	0.63
sulfides	5	0.87	1.01	0.77	0.86
disulfides	2	0.44	0.44	0.12	0.12
sulfur heterocycle	1	0.50	0.50	0.30	0.30
halogenated sulfur compounds	2	−0.98	2.63	−0.13	1.78
phosphorus compounds	14	0.20	1.08	0.78	1.62
silicon compound	1	−0.49	0.49	−0.12	0.12
all data	274	−0.02	0.53	0.20	0.59
66 Data					
compounds containing H, C, O	9	−0.70	1.06	0.30	0.70
compound containing H, C, N	1	1.76	1.76	0.84	0.84
compounds containing H, C, N, O	17	0.18	2.57	0.80	2.37
compound containing H, C, F	1	−0.34	0.34	0.39	0.39
compounds containing H, C, Cl	3	0.48	0.48	1.58	1.58
halogen compounds containing H, C, N, and/or O	15	1.24	1.85	2.15	2.43
sulfur compounds not containing P	9	0.04	1.33	−1.11	2.55
phosphorus compounds	11	0.74	1.57	0.09	1.89
all data	66	0.41	1.72	0.69	2.01
5 Data (Sulfonylurea Subset)					
compounds containing H, C, N, O, S	4	−3.16	3.43	−9.22	9.22
compound containing H, C, N, O, S, Cl	1	−2.44	2.44	−7.93	7.93
all data	5	−3.02	3.23	−8.96	8.96

[a] *N* is the number of data in a given solute class. [b] MSE and MUE are in kcal/mol.

Table 10. Mean Signed and Mean Unsigned Errors in Nonaqueous Solvation Free Energies Calculated using SM8AD and SM8 with CM4/M05-2X/6-31G* by Solute Class[a]

		SM8AD		SM8	
solute class	*N*	MSE[c]	MUE[c]	MSE[c]	MUE[c]
H_2, NH_3, H_2O	29	−1.72	1.73	−1.70	1.72
unbranched alkanes	85	0.19	0.43	0.23	0.45
branched alkanes	7	0.16	0.33	0.33	0.41
cycloalkanes	13	−0.12	0.35	−0.32	0.45
alkenes	18	−0.22	0.50	0.11	0.41
alkynes	9	−0.13	0.60	0.42	0.51
arenes	134	−0.43	0.54	−0.29	0.46
alcohols	272	−0.04	0.47	0.01	0.38
phenols	109	0.24	0.45	0.49	0.62
ethers	87	0.42	0.66	0.29	0.69
aldehydes	32	−0.01	0.73	−0.10	0.60
ketones	195	−0.51	0.57	−0.51	0.58
carboxylic acids	120	0.32	0.55	0.48	0.64
esters, including lactones [b]	243	0.12	0.35	0.27	0.45
peroxides	17	0.45	1.29	−0.07	0.58
bifunctional H, C, O compounds	24	0.81	1.38	1.01	1.33
aliphatic amines	154	0.18	0.38	0.19	0.43
anilines	61	−0.28	0.39	0.10	0.38
aromatic N-heterocycles (1 N)	52	0.00	0.58	−0.06	0.62
aromatic N-heterocycles (2 Ns)	9	0.63	0.82	0.68	0.82
nitriles	20	−0.33	0.59	0.18	0.46
hydrazines	5	0.83	1.19	0.80	1.27
bifunctional H, C, N compounds	2	−0.91	0.94	−0.74	0.79
amides	26	0.19	0.65	0.45	0.71
ureas	7	0.75	0.75	0.85	0.96
lactams	4	0.91	0.99	0.75	0.90
nitrohydrocarbons	86	−0.05	0.68	0.03	0.51
bifunctional H, C, N, O compounds	3	0.27	1.00	0.26	0.75
fluoroalkanes	5	−0.71	0.71	−0.12	0.62
fluoroarenes	11	−0.43	0.81	0.15	0.59
chloroalkanes	26	−0.61	0.61	−0.43	0.44
chloroalkenes	15	0.59	0.64	0.76	0.76
chloroarenes	31	−0.23	0.38	−0.06	0.33
bromoalkanes	21	−0.46	0.46	−0.67	0.68
bromoalkenes	2	−0.38	0.38	−0.16	0.16
bromoarenes	16	−0.40	0.54	−0.44	0.52
multihalogenhydrocarbons	14	−0.15	0.31	−0.26	0.36
halogenated bifunctional compounds	37	0.42	0.87	0.85	0.99
thiols	10	0.53	0.55	0.15	0.25
sulfides	13	−0.22	0.82	−0.30	0.91
disulfides	4	0.57	0.57	0.23	0.43
sulfur heterocycles	4	0.69	0.69	0.67	0.67
sulfoxide	1	−0.19	0.19	−0.41	0.41
phosphorus compounds	37	0.15	1.43	0.94	1.67
silicon compounds	2	0.97	0.97	1.61	1.61
all data	2 072	−0.02	0.56	0.08	0.57

[a] *N* is the number of data in a given solute class. [b] Five lactones and 238 other esters. [c] MSE and MUE are in kcal/mol.

program[99] that employs the external option of Gaussian 03. All solvation calculations using AM1 and PM3 were carried out with a locally modified version of GAMESS[100] called GAMESSPLUS.[101]

4. Discussion of Model Performance

First, we discuss the performance of SM8AD for predicting the solvation free energies of neutral solutes in aqueous solution. For the 274 neutral data in water that constitute subset i defined in Section 2, the MSE in the SM8AD free energies of solvation calculated using 26 theoretical levels listed in Table 2 is between −0.37 (M06-HF/6-31G*) and 0.58 kcal/mol (M05-2X/6-31B*). The corresponding MUE varies from 0.53 (M05-2X/6-31G*, M05-2X/6-31G**) to 0.87 kcal/mol (M05-2X/6-31B*). The MUE in the SM8 solvation free energies for the same data set varies from 0.55 (M06/6-31G*) to 1.04 kcal/mol (M05-2X/6-31B*), and the corresponding MUE in the SMD solvation free energies is between 0.60 (M05-2X/6-31G*) and 1.15 kcal/mol (M05-2X/6-31B*). For the 66 neutral data in water that constitute the training part of ii defined in Section 2, the MSE averaged over 26 theoretical levels used in these calculations is 0.1, 0.4, and 0.5 kcal/mol for SM8AD, SM8, and SMD, respec-

Table 11. Mean Signed and Mean Unsigned Errors in Transfer Free Energies between Water and Organic Solvents Calculated using SM8AD and SM8 with CM4/M05-2X/6-31G* by Solute Class[a]

		SM8AD		SM8	
solute class	*N*	MSE[b]	MUE[b]	MSE[b]	MUE[b]
lactones	10	0.40	0.84	0.06	0.89
aromatic N-heterocycles	6	−0.14	0.53	0.02	0.37
bifunctional H, C, N compounds	2	0.94	0.94	0.82	0.82
amides	13	−0.57	0.61	−1.05	1.05
ureas	11	−0.07	0.35	0.02	0.27
lactams	4	−1.35	1.35	−1.71	1.71
thymines and uracils	12	0.37	0.63	0.31	0.67
bifunctional H, C, N, O compounds	5	−0.15	0.81	−0.14	0.35
halogenated bifunctional compounds	39	0.32	0.64	0.38	0.67
sulfur compounds	19	0.10	0.37	−0.06	0.49
phosphorus compounds	9	−0.77	0.91	−1.36	1.38
silicon compounds	13	0.20	0.61	0.16	0.82
all data	143	0.04	0.63	−0.08	0.73

[a] *N* is the number of data in a given solute class. [b] MSE and MUE are in kcal/mol.

Table 12. Mean Signed and Mean Unsigned Errors in Ionic Solvation Free Energies Calculated using SM8AD and SM8 with CM4/M05-2X/6-31G*[a]

		SM8AD		SM8	
solute class	*N*	MSE[e]	MUE[e]	MSE[e]	MUE[e]
Acetonitrile					
H, C, N, O cations[b]	36	2.6	5.3	4.0	6.5
S cations[c]	3	9.1	9.1	16.1	16.1
All cations	39	3.1	5.6	4.9	7.2
H, C, N, O anions[b]	19	−3.2	3.2	−4.5	4.5
F, Cl, Br, S anions[c]	11	−4.2	4.3	−3.4	3.4
All anions	30	−3.6	3.6	−4.1	4.1
All ions	69	0.2	4.7	1.0	5.9
DMSO					
H, C, N, O cations[b]	4	0.1	0.7	−1.5	1.8
All cations	4	0.1	0.7	−1.5	1.8
H, C, N, O anions[b]	52	−4.2	5.0	−8.0	8.3
F, Cl, Br, S anions[c]	15	−5.8	5.8	−3.7	4.1
All anions	67	−4.6	5.2	−7.0	7.3
All ions	71	−4.3	4.9	−6.7	7.0
Methanol					
H, C, N, O cations[b]	26	−3.8	4.1	−1.3	2.1
Cl, Br cations[c]	3	−3.8	3.8	−1.0	1.0
All cations	29	−3.8	4.1	−1.2	2.0
H, C, N, O anions[b]	36	−2.5	3.4	−1.0	2.4
F, Cl, Br anions[c]	15	−1.7	2.9	−1.8	2.5
All anions	51	−2.3	3.3	−1.2	2.4
All ions	80	−2.8	3.6	−1.2	2.3
Water[d]					
H, C, N, O cations[b]	48	−2.1	3.0	0.4	2.4
Cl, S cations[c]	4	−0.7	2.7	1.5	2.5
All cations	52	−2.0	3.0	0.5	2.4
H, C, N, O anions[b]	43	1.5	3.0	1.3	3.7
F, Cl, Br, S anions[c]	17	1.3	2.8	0.0	3.0
All anions	60	1.4	3.0	0.9	3.5
All ions	112	−0.1	3.0	0.7	3.0

[a] *N* is the number of data in a given solute class. [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] 112 selectively clustered ions from the SM6 model training set as defined in ref 17. [e] MSE and MUE are in kcal/mol.

tively; the corresponding MUE is 1.9, 2.1, and 2.1 kcal/mol. Thus, SM8AD slightly outperforms SM8 and SMD on these data sets.

For five sulfonylureas (namely, bensulfuron methyl, chlorimuron ethyl, metsulfuron methyl, sulfometuron methyl, thifensulfuron methyl) in water, the MSE averaged over 26 theoretical levels is equal to −4.2, −10.1, and −8.0 kcal/mol, respectively for SM8AD, SM8, and SMD, and the corresponding MUE is 4.6, 10.1, and 8.1 kcal/mol. These errors are rather large. However, one should keep in mind that Guthrie[61] assigned these five compounds the largest possible uncertainty (1.9 kcal/mol), meaning that the corresponding solubility and vapor pressure could not be found in the open source literature or that the primary vapor pressure temperature data were not available.[61] As the compounds are highly nonvolatile, their vapor pressures were measured by assessing slow mass loss from solid samples heated to rather high temperatures and then by extrapolating back to room temperature.[102] In addition to uncertainties associated with the extrapolation procedure, there is the possibility of thermal degradation of the sulfonylureas.[102] In view of the uncertainty in the experimental targets for these compounds, we did not include them in the parametrization of any of the solvation models tested in our study. Nevertheless, the SM8AD model is able to predict the free energies of solvation for such difficult compounds more accurately than any other tested model apparently due to more robust parametrization of its nonbulk electrostatic term with the use of a more diverse training set in spite of the fact that no sulfonylureas were used in the SM8AD parametrization. In this regard, we recall that the SM8AD training set has been extended by addition of 66 aqueous solutes (see Section 2) many of which are compounds of complex functionality, including compounds with oxidized sulfur that are not present in the SM8 and SMD training sets of aqueous neutrals.

Now we discuss the performance of SM8AD as compared to that of SM8 and SMD for predicting the free energies of solvation for neutral solutes in nonaqueous solvents. For 2 072 neutral data in 90 organic solvents, the MSE in the SM8AD free energies of solvation calculated using 26 theoretical levels listed in Table 2 is between −0.20 (M06-HF/6-31G*) and 0.31 kcal/mol (M05-2X/6-31B*). The corresponding MUE varies from 0.56 to 0.61 kcal/mol, only slightly depending on density functional and basis set. The MUE in the SM8 solvation free energies for the same data set varies from 0.57 to 0.69 kcal/mol, and the corresponding MUE in the SMD solvation free energies is between 0.63 (M05-2X/6-31G*) and 0.84 kcal/mol (M05-2X/6-31B*). For 143 transfer energies, the MSE averaged over 26 theoretical levels used in these calculations is 0.03, −0.04, and −0.13 kcal/mol for SM8AD, SM8, and SMD, respectively; the corresponding MUE is 0.71, 0.74, and 0.74 kcal/mol.

For 112 aqueous ions, the MSE in the SM8AD free energies of solvation calculated using 26 theoretical levels is between −0.59 (M06-HF/6-31G*) and 0.62 kcal/mol (M05-2X/MIDI!6D), and the corresponding MUE varies from 2.50 (M05-2X/6-31+G**) to 3.37 kcal/mol (M05-2X/MIDI!). The MSE in the SM8 free energies of solvation for the same data set is between 0.12 (M05-2X/DZVP) and 1.55 kcal/mol (M05-2X/MIDI!6D), and the corresponding MUE ranges from 2.75 (M06-HF/6-31G*) to 3.36 kcal/mol (M05-2X/MIDI!). The MSE in the SMD free energies of solvation for the same data set is between 2.67 (HF/6-31G*) and 5.17

kcal/mol (mPWPW/6-31G*), and the corresponding MUE ranges from 3.39 (HF/6-31G*) to 5.34 kcal/mol (mPWPW/6-31G*). The MUE averaged over 26 theoretical levels used in the calculations for 112 aqueous ions is 2.9, 3.1, and 4.6 kcal/mol for SM8AD, SM8, and SMD, respectively. Thus, SM8AD produces on average slightly more accurate solvation energies for 112 aqueous ions tested in the present study.

For 220 ionic data in acetonitrile, dimethyl sulfoxide, and methanol, the MSE in the SM8AD free energies of solvation calculated using 26 theoretical levels is between −2.93 (M05-2X/DZVP) and −1.69 kcal/mol (M05-2X/MIDI!6D). The MSE in the SM8 free energies of solvation for the same data set is between −3.09 (M05-2X/DZVP) and −1.59 kcal/mol (M05-2X/MIDI!6D). The MSE in the SMD free energies of solvation for the same data set is between −0.70 (HF/6-31G*) and 1.86 kcal/mol (mPWPW/6-31G*). The MUE averaged over 26 theoretical levels used in the calculations for 220 ionic data in the three nonaqueous solvents 4.5, 5.1, and 4.2 kcal/mol for SM8AD, SM8, and SMD, respectively.

In general, the SM8AD model demonstrates satisfactory performance with any of the 26 theoretical levels listed in Table 2 for which it has been parametrized. The quality of the model does not substantially depend on the choice of density functional, and it can be recommended for the use with any density functional or the Hartree−Fock method. The SM8AD performance worsens when the model is used with the basis sets DZVP and 6-31B* and with basis sets that contain diffuse functions (6-31+G*, 6-31+G**). For instance, the MUE in the SM8AD free energies of solvation calculated for 274 neutral solutes in water using M05-2X with the CM4 charge model is 0.67 (6-31+G**), 0.69 (6-31+G*), 0.77 (DZVP), and 0.87 kcal/mol (6-31B*). The MUE in the SM8AD free energies of solvation calculated for the same data set using M06-2X with the CM4M charge model is 0.68 (6-31+G**), 0.69 (6-31+G*), 0.73 (DZVP), and 0.73 kcal/mol (6-31B*). The best performance for SM8AD can be achieved if it is used with MIDI!, MIDI!6D, 6-31G*, 6-31G**, and cc-pVDZ basis sets. Tables 9−12 give a breakdown of the errors in solvation calculations using SM8AD/CM4/M05-2X/6-31G* by solute class. For 274 neutral data in water, the MSE ranges from −1.74 to 0.87 kcal/mol (Table 9). For 2 072 neutral data in 90 organic solvents, the MSE ranges from −1.72 to 0.97 kcal/mol (Table 10). For 143 transfer energies, the MSE ranges from −1.35 to 0.94 kcal/mol (Table 11). For 124 cationic data (Table 12), the smallest error is observed for cations in DMSO (MSE = 0.1, MUE = 0.7 kcal/mol), and the largest error is observed for cations in acetonitrile (MSE = 3.1, MUE = 5.6 kcal/mol). The errors for the sulfur-containing cations in acetonitrile are somewhat large, but this discrepancy is mainly attributed to the H_3S^+ cation which is systematically (by ∼16 kcal/mol) undersolvated by the SM8AD model with respect to the experimental solvation free energy ΔG_s^o(expt.) = −100.2 kcal/mol. However, the SM8AD error for H_3S^+ is substantially smaller than the corresponding SM8 error. For 208 anionic data (Table 12), the smallest error is observed for anions in water (MSE = 1.4, MUE = 3.0 kcal/mol), and the largest error is observed for anions in DMSO (MSE = −4.6, MUE = 5.2 kcal/mol).

Other theoretical levels for which the SM8AD model has been tested include those listed in Table 3 (see the Supporting Information), though we did not use these theoretical levels in the SM8AD parametrization. The SM8AD model applied with M05-2X/6-31B** and the CM4 charge model to 274 aqueous solutes substantially undersolvates these compounds (MSE = 0.89 kcal/mol, MUE = 1.14 kcal/mol). The MUE in the free energies of solvation computed with SM8/CM4/M05-2X/6-31B** and SMD/M05-2X/6-31B** for the same data set are even larger, reaching 1.27 and 1.23 kcal/mol, respectively. The use of M06-2X/6-31B** instead of M05-2X/6-31B** slightly reduces these errors, resulting in 0.99, 1.13, and 0.93 kcal/mol for SM8AD, SM8, and SMD, respectively, the CM4M charge model being used with SM8AD and SM8. We also tested SM8AD and SM8 with the cc-pVTZ basis set for which we do not have a charge model; we used partial atomic charges from Löwdin population analysis (LPA) in this case. The two GB models used with cc-pVTZ perform poorly in comparison with the SMD model,[20] which is a density-based solvation model that does not require partial atomic charges (the MUEs for 274 aqueous data are equal to 2.46, 2.59, and 0.68 kcal/mol, respectively, for SM8AD, SM8, and SMD). The use of partial atomic charges obtained from Löwdin population analysis results in poor performance for SM8AD and SM8, even in the case of smaller basis sets such as 6-31G*. Indeed, the MUE in the SM8AD/LPA/6-31G* free energies of solvation calculated for 274 aqueous data is overly large (2.15 kcal/mol). Adding diffuse functions to the basis results in even larger errors in the SM8AD/LPA calculations, though the use of RLPA[51] instead of LPA can reduce the error (MUE = 1.46 kcal/mol for 274 aqueous data with SM8AD/RLPA/6-31+G**). As it was previously indicated,[18] the use of GB solvation models with partial atomic charges from population analyses (class II charges) results in less accurate solvation energies than with CM4 charges and other comparably reliable class IV charges.

In addition, we tested the SM8AD model using the semiempirical electronic structure methods AM1 and PM3 combined with the CM2[97] and CM3[98] partial atomic charges (see the Supporting Information). The charge models CM2[97] and CM3[98] were specifically parametrized for the use with AM1 and PM3. For 274 aqueous data, the MSE in the SM8AD free energies of solvation calculated with AM1/CM2, AM1/CM3, PM3/CM2, and PM3/CM3 varies from −1.75 to −2.20 kcal/mol. For 66 aqueous data, the corresponding MSE varies from −5.2 to −4.2 kcal/mol. For five sulfonylureas in water, the MSE varies from −18.9 to −15.1 kcal/mol. Similar errors were obtained for SM8/AM1/CM2, SM8/AM1/CM3, SM8/PM3/CM2, SM8/PM3/CM3 as well as for SMD/AM1 and SMD/PM3. These errors are substantially larger than those obtained, for instance, with SM8AD/CM4/M05-2X/6-31G*. Unfortunately, the applicability of the SM8AD model parameters developed in the present study for the use with any density functional or the Hartree−Fock method cannot be extended to use with AM1 and PM3, and the SM8AD model would require a special parametrization of its Coulomb radii and atomic surface tension coefficients to be used with AM1 and PM3. The same is true for SM8

and SMD. The earlier continuum models SM5.42/AM1 and SM5.42/PM3, which were parametrized for the use with AM1 and PM3, demonstrate good performance on 274 aqueous neutral data (MUE = 0.61 kcal/mol) and on 2 072 nonaqueous neutral data (MUE = 0.53−0.54 kcal/mol), and somewhat poorer performance on 112 aqueous ions (MUE = 4.2−4.9 kcal/mol) and on 220 nonaqueous ions (MUE = 6.7−6.9 kcal/mol).

Overall, the SM8AD errors are typically smaller than those of SM8 and SMD in many cases when all the three models perform well. However, the difference in the SM8AD, SM8, and SMD errors is usually smaller than the estimated uncertainty of the corresponding experimental targets (the uncertainty ranges from 0.1 to 1.9 kcal/mol for neutrals,[17,60,61] and it is about 3 kcal/mol for ions[17]), therefore, one can assume that all the three models do equally well. This indicates the fact that the CDS formalism used in the parametrization of SM8AD, SM8, and SMD is able to account for most of the nonelectrostatic solvation effects as well as for the deviations of the electrostatics from the assumed bulk model (which is different in all the models) due to the inexactness of the solvent permittivity model including the assumed values for intrinsic Coulomb radii, the uncertainties in the treatment of solute charge outside the solute cavity,[2,4,40,103] as in the case of SMD, and the inexactness of the solute charge model as in the case of SM8 and SM8AD.

In Table 13, we compare the free energies of solvation for 71 aqueous solutes calculated using SM8AD with those calculated using the IEFPCM algorithm[39−42] as implemented in Gaussian 03[78] and the Poisson-Boltzmann (PB) self-consistent reaction field solver as implemented in Jaguar.[104−106] For both electrostatic and nonelectrostatic contributions, we accept the defaults of these programs. Thus, the Gaussian 03 calculations include not only electrostatics but also cavitation, dispersion, and repulsion.[75] The molecular cavities in the IEFPCM/Gaussian 03 calculations were constructed using the united atom Hartree−Fock (UAHF) scheme[107] for atomic radii that is a recommended method for predicting solvation free energies with PCM according to the Gaussian 03 manual.[78] We have also tested another united atom scheme called UA0[108] that is a simplified united atom implementation based on the universal force field radii, available for the full periodic table.[109] The PB/Jaguar calculations employ atomic radii that depend on typing certain functional groups in a solute molecule.[105]

The MUE in the PB/Jaguar solvation free energies of 71 aqueous solutes (Table 13) is 1.5 times larger than the corresponding SM8AD error, whereas the MUE in the IEFPCM/UAHF/Gaussian 03 calculations for the same data set is twice as large. The solvation energies calculated using IEFPCM/UA0/Gaussian 03 are much less accurate (MSE = 9.9, MUE = 9.9, RMSE = 11.1 kcal/mol) and many of them are in disagreement with those calculated using IEFPCM/UAHF. Table 14 shows individual contributions to the free energies of solvation such as polarization (G_P), electronic relaxation (G_E), cavitation (G_C), dispersion (G_D), and repulsion (G_R) calculated by IEFPCM/Gaussian 03 using UA0, UAHF, and Bondi's radii for four compounds selected out

Table 13. Standard-State Free Energies of Solvation (kcal/mol) Calculated for a Subset of 71 Aqueous Data using SM8AD, IEFPCM/Gaussian 03, and PB/Jaguar with M05-2X/6-31G*[a]

solute name[b]	exp	SM8AD	IEFPCM/UAHF Gaussian 03	PB Jaguar
1,1-diacetoxyethane	−5.0	−7.6	−3.5	−6.3
1,1-diethoxyethane	−3.3	−2.7	−2.7	−4.1
1,2-diethoxyethane	−3.5	−3.3	−4.2	−4.5
1,2-dinitroxypropane	−5.0	−1.6	−4.7	−5.5
1,4,5,8-tetraminoanthraquinone	−8.9	−17.3	−11.6	−18.3
1-amino-4-anilinoanthraquinone	−7.4	−11.0	−4.9	−11.7
1-amino-anthraquinone	−8.0	−9.7	−5.5	−9.1
2-butyl nitrate	−1.8	−0.1	−0.7	−1.6
4-amino-4′-nitroazobenzene	−11.2	−12.1	−8.2	−10.0
alachlor	−8.2	−5.5	−1.7	−8.4
aldicarb	−9.8	−7.6	−5.7	−9.1
ametryn	−7.7	−9.5	−7.7	−12.3
azinphos methyl	−10.0	−10.0	−4.7	−12.4
benefin	−3.5	−0.7	2.9	−2.6
bensulfuron	−17.2	−24.7	−21.0	−35.1
bis(2-chloroethyl) ether	−4.2	−3.1	−4.1	−4.2
bromacil	−9.7	−10.4	−5.8	−11.3
butyl nitrate	−2.1	−0.1	−1.7	−2.0
captan	−9.0	−6.3	−4.7	−7.0
carbaryl	−9.5	−9.9	−6.9	−8.5
carbofuran	−9.6	−11.2	−7.1	−8.1
carbophenothion	−6.5	−3.4	1.4	−4.7
chlordane	−3.4	−3.2	5.5	−1.3
chlorfenvinphos	−7.1	−7.5	2.2	−6.6
chlorimuron ethyl	−14.0	−16.4	−9.4	−25.5
chloropicrin	−1.5	−0.2	2.6	1.0
chlorpyrifos	−5.0	−3.5	4.8	−3.6
dialifor	−5.7	−8.6	−0.1	−12.0
diazinon	−6.5	−7.1	0.4	−9.2
dicamba	−9.9	−8.5	−4.9	−6.7
diethyl propanedioate	−6.0	−6.4	−4.7	−6.0
dimethoxymethane	−2.9	−3.4	−3.8	−4.7
dinitramine	−5.7	−3.9	−0.5	−6.1
dinoseb	−6.2	−9.2	−5.8	−9.7
endosulfan alpha	−4.2	−6.0	0.8	−8.2
endrin	−5.5	−6.3	0.1	−3.6
ethion	−6.1	−4.0	2.9	−10.6
ethylene glycol diacetate	−6.3	−8.1	−4.3	−6.0
ethylene glycol mononitrate	−8.2	−5.4	−8.5	−8.7
glycerol triacetate	−8.8	−11.3	−5.9	−10.0
heptachlor	−2.6	−2.6	5.4	−0.5
imidazole	−9.8	−8.0	−10.1	−10.6
isobutyl nitrate	−1.9	0.2	−0.6	−1.9
isophorone	−5.2	−4.4	−2.9	−5.9
lindane	−5.4	−4.2	−5.0	−4.8
malathion	−8.2	−8.8	0.0	−6.2
m-bis(trifluoromethyl) benzene	1.1	0.7	1.5	−0.6
methomyl	−10.7	−10.2	−7.5	−9.5
metsulfuron methyl	−15.5	−18.9	−11.9	−25.9
N,*N*,4-trimethylbenzamide	−9.8	−7.4	−4.2	−8.3
N,*N*-dimethyl-*p*-methoxybenzamide	−11.0	−9.2	−6.0	−17.3
nitralin	−8.0	−6.0	−0.8	−16.1
nitroglycol	−5.7	−1.1	−5.5	−5.4
nitroxyacetone	−6.0	−3.4	−5.4	−6.3
parathion	−6.7	−5.3	−0.2	−9.6
pebulate	−3.6	−2.8	−0.8	−5.0
phenyl formate	−3.8	−3.9	−3.9	−3.3
phorate	−4.4	−2.9	−0.4	−6.5
pirimor	−9.4	−10.3	−3.5	−12.0
profluralin	−2.5	−0.9	4.4	−3.5
prometryn	−8.4	−8.4	−7.0	−11.9
propanil	−7.8	−9.1	−5.9	−8.9
pyrazon	−16.4	−13.2	−7.8	−12.6
simazine	−10.2	−12.0	−10.4	−14.8
sulfometuron methyl	−20.3	−19.8	−12.5	−29.3
terbacil	−11.1	−9.3	−6.1	−11.0
terbutryn	−6.7	−8.8	−7.2	−15.9
thifensulfuron	−16.2	−18.5	−13.0	−26.8
trichlorfon	−12.7	−9.6	−5.0	−13.7
trifluralin	−3.3	−0.5	4.2	−2.9
vernolate	−4.1	−3.5	−1.2	−6.4
	MSE	0.2	3.5	−1.8
	MUE	1.8	3.8	2.7
	RMSE	2.3	4.7	4.3

[a] IUPAC names and other known names for these compounds are given in the Supporting Information. MSE denotes mean signed error; MUE denotes mean unsigned error, and RMSE denotes root mean squared error. [b] Sorted in alphabetical order.

Table 14. Various Contributions (kcal/mol) to the Free Energies of Solvation for Selected Molecules Calculated using IEFPCM/Gaussian 03 and Various Schemes for Coulomb Radii[a]

model	G_P	G_E	ΔG_{EP}	G_C	G_D	G_R	G_{CDR}	ΔG_s^o
				Lindane				
UA0	−12.73	2.04	−10.69	26.45	−23.46	2.34	5.34	−5.35
UAHF	−12.99	1.98	−11.01	27.25	−25.31	4.11	6.05	−4.97
Bondi	−13.64	2.28	−11.36	28.78	−27.74	4.56	5.61	−5.75
				Pyrazon				
UA0	−18.81	3.51	−15.31	27.59	−21.66	1.79	7.72	−7.59
UAHF	−14.23	2.49	−11.74	26.23	−26.52	4.18	3.89	−7.84
Bondi	−23.05	4.93	−18.12	28.49	−27.43	4.01	5.07	−13.05
				Pirimor				
UA0	−7.86	1.33	−6.53	39.27	−19.01	0.59	20.84	14.31
UAHF	−9.94	1.64	−8.30	32.07	−33.89	6.65	4.83	−3.47
Bondi	−16.97	3.56	−13.41	38.91	−31.65	4.33	11.59	−1.82
				Prometryn				
UA0	−12.88	1.79	−11.09	39.54	−21.97	1.10	18.67	7.58
UAHF	−13.64	1.72	−11.92	33.28	−35.13	6.82	4.96	−6.96
Bondi	−16.35	2.48	−13.87	40.44	−34.17	4.86	11.12	−2.74

[a] The table shows contributions to the standard-state free energies of solvation in water (ΔG_s^o) calculated by IEFPCM/Gaussian 03 using UA0, UAHF, and Bondi's radii for four compounds selected out of the 71 compounds presented in Table 13. The first two compounds (sorted by name in alphabetical order) have the smallest deviation in the UA0 and UAHF values of ΔG_s^o, whereas the remaining two compounds have the largest deviation. The electronic structure method used for these calculations is M05-2X/6-31G*.

of 71 aqueous solutes presented in Table 13. The first two compounds (lindane, pyrazon) have the smallest deviation in the free energies of solvation calculated using UA0 and UAHF, whereas the remaining two compounds (pirimor, prometryn) have the largest deviation. In the last two cases, the deviation between the UA0 and UAHF total solvation energies is dominated by the deviation between the corresponding nonelectrostatic terms, though the bulk electrostatic contributions are roughly the same. For instance, the G_D term for prometryn in water calculated using UA0 is 13 kcal/mol larger than G_D(UAHF), contributing to the unphysically positive value of the total free energy of solvation for this solute (7.58 kcal/mol with UA0 versus −6.96 kcal/mol with UAHF, Table 14). Thus, the electrostatic and nonelectrostatic terms are not separately meaningful in a quantitative sense, and the validity of a model can be judged by the usefulness of the whole model in predicting and correlating experimental observables but not by any supposed rigor in the electrostatic or nonelectrostatic parts of the formulation.[110] We also note that all the compounds with overly positive total solvation energies obtained using UA0 have three or more methyl groups, indicating a possible problem with the CH_3 parameters used by the UA0 united atom model. The poor performance by IEFPCM models may be caused in part by the difficulty of separately estimating the cavity, dispersion, and repulsion contributions. Table 14 shows that adding these contributions leads to considerable cancellation. In contrast, SM8AD, SM8, and SMD directly model the sum (the CDS term). To make this more clear, Table 15 shows, for the same four molecules as in Table 14, the P, E, EP, and CDS contributions of SM8AD, SM8, and SMD.

To conclude this section, we will make an additional comparison of our solvation models (SM8AD, SM8, SMD) with the IEFPCM/Gaussian 03 model that uses the UAHF scheme for atomic and group radii by considering four additional solutes (acetic acid, benzaldehyde, ethanol, nicotinamide) solvated by three solvents, in particular benzene, methylene chloride, and water (Table 16). These solutes

Table 15. Various Contributions (kcal/mol) to the Free Energies of Solvation for Selected Molecules Calculated using SM8AD, SM8, and SMD[a]

model	G_P	G_E	ΔG_{EP}	G_{CDS}	ΔG_s^o
			Lindane		
SM8AD	−4.62	0.83	−3.80	−0.38	−4.18
SM8	−3.91	0.64	−3.28	0.26	−3.02
SMD	−11.65	1.95	−9.70	2.47	−7.23
			Pyrazon		
SM8AD	−17.14	5.28	−11.86	−1.34	−13.20
SM8	−14.65	3.96	−10.69	−0.63	−11.32
SMD	−21.43	4.64	−16.79	4.85	−11.94
			Pirimor		
SM8AD	−15.46	3.37	−12.09	1.79	−10.30
SM8	−14.08	2.80	−11.27	0.63	−10.64
SMD	−16.95	3.65	−13.30	6.46	−6.84
			Prometryn		
SM8AD	−9.06	1.67	−7.39	−0.96	−8.35
SM8	−9.94	1.74	−8.20	−1.03	−9.23
SMD	−13.95	1.84	−12.11	3.70	−8.41

[a] The table shows contributions to the standard-state free energies of solvation in water (ΔG_s^o) calculated by SM8AD, SM8, and SMD for compounds presented in Table 14. The electronic structure method used for these calculations is M05-2X/6-31G*.

represent major classes of chemical compounds with various functionalities, and the set of solvents is chosen to span a range of dielectric constants and solvent properties. Table 16 also contains gas-phase and liquid-phase dipole moments and compares the former to experiment.[111] SM8AD, SM8, and SMD agree well between themselves in predicting the solvation free energies for any of these solutes with the difference in ΔG_s^o lying between 0 and 1.5 kcal/mol, whereas the individual (bulk electrostatic and nonbulk electrostatic) components to the free energy of solvation vary more significantly. For instance, the difference in ΔG_{EP} as well as in ΔG_{CDS} calculated by SM8 and SMD for any solute in water is about 4 kcal/mol or larger. SM8AD, SM8, and SMD agree with IEFPCM/Gaussian 03 better for aqueous solutes than for nonaqueous ones. The IEFPCM/Gaussian 03 free energies of solvation for the nonaqueous solutes are usually

Table 16. Standard-State Free Energies of Solvation (kcal/mol) and Dipole Moments (debye) for Selected Solutes in Benzene, Methylene Chloride, and Water Calculated using SM8AD, SM8, SMD, and IEFPCM/Gaussian 03[a]

	gas	benzene				methylene chloride				water			
model	μ	ΔG_{EP}	G_{CDS}	ΔG_s^o	μ	ΔG_{EP}	G_{CDS}	ΔG_s^o	μ	ΔG_{EP}	G_{CDS}	ΔG_s^o	μ
						Acetic Acid							
SM8AD	1.59	−1.48	−2.62	−4.10	1.74	−3.03	−1.62	−4.65	1.86	−7.52	0.90	−6.62	2.04
SM8	1.59	−1.38	−2.82	−4.20	1.73	−2.73	−1.96	−4.70	1.84	−6.04	−0.32	−6.36	1.99
SMD	1.59	−2.03	−1.79	−3.82	1.74	−4.95	−0.33	−5.27	1.89	−10.03	3.83	−6.20	2.15
IEFPCM	1.59	−1.61	0.68	−0.93	1.70	−3.08	0.10	−2.98	1.78	−7.92	1.35	−6.57	1.92
exp	1.70 ± 0.03			−4.02								−6.70	
						Benzaldehyde							
SM8AD	3.32	−1.71	−4.81	−6.52	3.62	−3.19	−4.00	−7.18	3.92	−5.58	1.25	−4.32	4.38
SM8	3.32	−1.68	−4.87	−6.55	3.63	−3.08	−4.22	−7.30	3.92	−4.92	0.58	−4.34	4.26
SMD	3.32	−2.61	−3.98	−6.59	3.69	−5.53	−2.73	−8.26	4.08	−8.79	4.28	−4.51	4.67
IEFPCM	3.32	−1.24	0.82	−0.42	3.62	−2.41	0.11	−2.30	3.89	−5.40	2.17	−3.23	4.26
exp												−4.02	
						Ethanol							
SM8AD	1.71	−0.57	−2.82	−3.38	1.79	−1.32	−2.46	−3.78	1.87	−4.13	−0.74	−4.87	1.97
SM8	1.71	−0.72	−2.61	−3.33	1.80	−1.51	−2.27	−3.78	1.87	−3.60	−1.19	−4.79	1.95
SMD	1.71	−1.42	−1.61	−3.02	1.85	−3.54	−0.94	−4.48	2.03	−7.46	2.43	−5.04	2.30
IEFPCM	1.71	−1.12	−0.08	−1.20	1.82	−2.26	−0.75	−3.01	1.93	−5.88	0.34	−5.54	2.14
exp	1.69 ± 0.03			−3.42				−3.82				−5.01	
						Nicotinamide							
SM8AD	2.07	−4.19	−3.94	−8.13	2.16	−7.41	−3.63	−11.03	2.25	−11.32	−2.25	−13.57	2.51
SM8	2.07	−4.86	−3.01	−7.86	2.17	−8.34	−2.80	−11.14	2.25	−10.70	−2.16	−12.86	2.38
SMD	2.07	−4.95	−2.26	−7.21	2.24	−10.07	−1.17	−11.24	2.46	−14.91	2.80	−12.11	2.84
IEFPCM	2.07	−2.58	0.62	−1.96	2.21	−5.04	−0.34	−5.38	2.36	−12.08	1.70	−10.38	2.61

[a] The electronic structure method used for these calculations is M05-2X/6-31G*. The IEFPCM/Gaussian 03 calculations use the UAHF scheme for atomic and group radii. The ΔG_{EP} term refers to the bulk electrostatic contribution to the free energy of solvation ΔG_s^o. The G_{CDS} term refers to the cavity dispersion solvent structure component as defined by SM8AD, SM8, and SMD, and it corresponds to the cavity dispersion repulsion (G_{CDR}) component as defined by IEFPCM/Gaussian 03. The quantity of μ refers to the dipole moment, in all cases the calculated value is calculated from the electron density. Experimental values of the free energies of solvation and the gas-phase dipole moments are taken from refs 59 and 111, respectively.

much less negative than the corresponding SM8AD, SM8, SMD, and available experimental values. Note that, according to the Gaussian 03 output, IEFPCM/Gaussian 03 by default scales the UAHF radii used in these calculations differently for different solvents, using the scaling factor of 1.2 for water and 1.4 for benzene and methylene chloride.

5. Discussion of Model Physics

In this section we will examine some general advantages in using the improved electrostatic algorithm based on the generalized Born approximation with asymmetric descreening (see Section 2) as incorporated now in the new SM8AD solvation model instead of the traditional generalized Born Coulomb field approximation.

Grycuk has tested[23] the GB/CF approximation against the Kirkwood model[22] applied to biopolymer electrostatics for the case of a spherical biopolymer (in this case the Kirkwood model is equivalent to the model based on the Poisson equation for electrostatics). In a simple case when a single charge is placed inside a spherical cavity embedded in a dielectric continuum, the GB/CF model provides the exact polarization energy due to the charge only if the charge is located in the center of the spherical cavity.[23] If the charge is located near the dielectric boundary, the effective Born radius is overestimated by up to a factor of 2, and therefore, the resulting polarization energy can be underestimated by the same factor.[23] As suggested by Grycuk,[23] the use of the corrected Born radius (i.e., eq 8 instead of eq 6) allows one to effectively reduce the errors of conventional GB/CF models related to the deficiency of the Coulomb field approximation.

The observed discrepancy between the GB/AD polarization energies and the polarization energies obtained by solving the NPE can be attributed to the possible inaccuracy of any of these models. Although the NPE electrostatics is considered as the standard in this discussion, methods that solve the NPE may have uncertainties in the bulk electrostatic part due to the portion of the solute charge that lies outside the cavity and the assumed way in which the permittivity changes at and near the solute−solvent boundary.[2,4,40,103] On the other hand, even the GB/AD model does not eliminate all the deficiencies of the GB approximation, including the oversimplified treatment of charge distributions by replacing the continuous charge density of the solute by a set of atom-centered partial charges for all stages of the calculation. With that preface, the rest of this section uses the working hypothesis that IEFPCM is an accurate standard for a given set of radii.

Table 17 compares the polarization energies calculated by using the GB/CF and GB/AD approximations and by solving the NPE for bulk electrostatics in a medium with $\varepsilon = 78.3$ for the sodium-doped fullerene cation $Na^+@C_{60}$ (a system with nearly spherical symmetry) and the *n*-butylammonium cation *n*-$CH_3(CH_2)_3NH_3^+$ (a nonspherical system). The NPE was solved by the IEFPCM algorithm[39−42] as implemented in Gaussian 03[78] with the user-defined intrinsic atomic Coulomb radii and with the default tessellation settings. The GB/CF and GB/AD approximations were evaluated using a

Table 17. Polarization Energies (kcal/mol) Calculated using the GB/CF and GB/AD Approximations and the Nonhomogeneous Poisson Equation for Bulk Electrostatics[a]

	Coulomb radius				G_P		
charge model[b]	H	C	N	Na	GB/CF	GB/AD	NPE
			$Na^+@C_{60}$				
delocalized[c]		1.57		3.55	−33.5	−35.6	−37.0[d]
localized[c]		1.57		3.55	−33.7	−33.7	
delocalized[e]		2.04		3.55	−30.2	−32.2	−33.2
localized[e]		2.04		3.55	−30.5	−30.5	
			n-$CH_3(CH_2)_3NH_3^+$				
delocalized[c]	1.02	1.57	1.61		−71.0	−76.3	−81.9
localized[c]	1.02	1.57	1.61		−92.4	−100.2	
delocalized[e]	1.44	2.04	1.86		−58.9	−63.6	−62.7
localized[e]	1.44	2.04	1.86		−72.7	−79.2	

[a] The electronic structure method used for these calculations is M05-2X/6-31G*. The molecular geometry was optimized at the mPW1PW/3-21G level for $Na^+@C_{60}$ and the mPW1PW/MIDI! level for n-$CH_3(CH_2)_3NH_3^+$. [b] We used either the delocalized CM4 partial atomic charges on H, C, N, and Na obtained using the unpolarized (gas-phase) wave function or the localized partial atomic charges which were defined as follows: the Na atom in $Na^+@C_{60}$ has a charge of +1, and each of the three hydrogen atoms in the NH_3 group in n-$CH_3(CH_2)_3NH_3^+$ has a charge of +1/3, the charges on all other atoms in these molecules are assigned to zero. [c] The electrostatic cavity was defined by superpositions of the nuclear-centered spheres corresponding to the SM8 values of intrinsic atomic Coulomb radii for H, C, and N, and the value of the Na radius equal to the half of the largest C−C distance in $Na^+@C_{60}$ unless noted otherwise. The Na radius was chosen in the way to exclude the space inside of the C_{60} cavity from the dielectric continuum. [d] The $Na^+@C_{60}$ cavity in this case was approximated by a sphere centered on the Na atom with the radius equal to 5.12 Å that is the half of the largest C−C distance plus the SM8 radius for C. [e] The electrostatic cavity was defined by superpositions of the nuclear-centered spheres corresponding to 1.2 times the Bondi values of intrinsic atomic Coulomb radii for H, C, and N and the value of the Na radius equal to the half of the largest C−C distance in $Na^+@C_{60}$. The Na radius was chosen in a way to exclude the space inside of the C_{60} cavity from the dielectric continuum.

locally modified version of Gaussian 03, in particular MN-GSM.[79] In all three cases we used the M05-2X/6-31G* unpolarized gas-phase wave function to obtain the continuous charge density for NPE calculations and to evaluate the CM4 class IV partial atomic charges[17] based on the charge density for GB calculations. The dielectric boundary was built to precisely enclose a superposition of nuclear-centered spheres with intrinsic Coulomb radii ρ_Z, which depend only on the atomic numbers of the atoms (Z). We tested two sets of the ρ_Z values corresponding to the SM8 radii[18] and the van der Waals radii of Bondi[54] scaled by a factor of 1.2. The other details of these calculations are given in footnotes to Table 17.

Table 17 indicates large differences among the polarization energies calculated by GB/CF, GB/AD, and NPE for both $Na^+@C_{60}$ and n-$CH_3(CH_2)_3NH_3^+$. In the case of a nearly spherical molecule with the charge located near the center of the sphere such as in $Na^+@C_{60}$, one might have expected that the GB/CF and GB/AD methods should agree among themselves and should agree well with the NPE. However, this is not the case for $Na^+@C_{60}$ because there is some charge transfer between the Na^+ cation and the C_{60} shell, and the transferred charge is located near the dielectric boundary rather than in the center of the molecular cavity. The latter circumstance makes the GB/CF approximation agree less closely than GB/AD with IEFPCM. The G_P value for $Na^+@C_{60}$ varies from −33.5 (GB/CF) to −35.6 (GB/AD) kcal/mol (with the SM8 radii used to construct the $Na^+@C_{60}$ electrostatic cavity), whereas for the situation in which all the charge was localized on the Na atom, the value of G_P was found to be −33.7 kcal/mol. The former calculation, with realistic charges, is labeled "delocalized" in Table 17, and the latter is labeled "localized". In fact, the GB/AD approximation is expected to be more realistic than the traditional GB/CF approach for almost any real solute other than a monatomic system (in which case the two models converge according to eq 4) because there is almost always one or more charged groups or atoms in the solute molecule that are exposed to the solvent, i.e., located near the dielectric boundary.

One advantage of GB models is their lower computational cost compared to that of the cost of NPE solvers. The use of the GB/AD approximation instead of GB/CF allows one to reproduce the NPE results more closely and extend the applicability of generalized Born models to a wider class of solutes with no additional cost.

6. Summary

We have presented a new self-consistent reaction field universal continuum solvent model based on the GB/AD approximation introduced by Grycuk.[23] The new model is called Solvation Model 8 with Asymmetric Descreening (SM8AD). "Universal" denotes its applicability to solvation in water or any nonaqueous 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). Water is treated as a special solvent that is given its own set of model parameters. The SM8AD model is applicable to any charged or uncharged solute or supersolute. This model was parametrized over 26 combinations of electronic structure methods and basis sets with the use of CM4[17] and CM4M class IV partial atomic charges.[48] Namely, we used nine density functionals (M05-2X, M05, M06-2X, M06, M06-HF, M06-L, mPWPW, mPW1PW, and B3LYP), the Hartree−Fock method, and nine basis sets for which the charge models CM4 and CM4M are available (MIDI!, MIDI!6D, 6-31G*, 6-31+G*, 6-31+G**, 6-31G**, cc-pVDZ, DZVP, or 6-31B*).

The SM8AD model was tested against the earlier SM8 model based on the GB/CF approximation and the density-based continuum solvent model SMD over a set of 2 892 solvation data, including 345 free energies of solvation for neutral solutes in water, 2 072 free energies of solvation for neutral solutes in 90 nonaqueous solvents, 143 transfer free energies for neutral solutes between water and 15 organic solvents, and 332 free energies of solvation for ions in acetonitrile, dimethyl sulfoxide, methanol, and water. The number of solvation energy calculations performed in this testing totals to 75 192 for each of the three models. The mean unsigned error averaged over 26 theoretical levels for 2 560 solvation data for neutral solutes is 0.6, 0.7, and 0.8 kcal/mol for SM8AD, SM8, and SMD, respectively. The

mean unsigned error averaged over 26 theoretical levels for 332 free energies of solvation for ions is 4.0 (SM8AD), 4.4 (SM8), and 4.3 kcal/mol (SMD).

Acknowledgment. This work was supported by the Office of Naval Research under Grant N 00014-05-01-0538 and the National Science Foundation (Grant CHE06-10183 and Grant CHE07-04974). Computational resources were provided by Minnesota Supercomputing Institute.

Supporting Information Available: Two thousand four hundred seventeen reference solvation free energies and 143 reference transfer energies for neutral solutes in the SM8AD training set; reference free energies for 112 selectively clustered ions in water; and 220 unclustered ions in acetonitrile, DMSO, and methanol (part I); complementary tables with the SM8AD errors in the solvation free energies for neutral compounds (part II); the functional forms of atomic surface tensions used by SM8AD (part III). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Politzer, P.; Truhlar, D. G. In *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Politzer, P., Truhlar, D. G., Eds.; Plenum: New York, 1981; p 1.

(2) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(3) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(4) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(5) Hoijtink, G. J.; de Boer, E.; van der Meij, P. H.; Weijland, W. P. *Recl. Trav. Chim. Pays−Bas Belg.* **1956**, *75*, 487.

(6) Peradejordi, F. *Cah. Phys.* **1963**, *17*, 393.

(7) Klopman, G. *Chem. Phys. Lett.* **1967**, *1*, 200.

(8) Tapia, O. In *Quantum Theory of Chemical Reactions*; Daudel, R., Pullman, A., Salem, L., Viellard, A., Eds.; Wiley: London, 1981; Vol. 2, p 25.

(9) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164.

(10) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(11) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

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

(13) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117. Errata. *J. Chem. Phys.* **1999**, *111*, 5624 and **2000**, *113*, 3930.

(14) Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *288*, 293.

(15) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9.

(16) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532.

(17) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(18) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.

(19) Dolney, D. M.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2000**, *21*, 340.

(20) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378.

(21) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.

(22) Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351.

(23) Grycuk, T. *J. Chem. Phys.* **2003**, *119*, 4817.

(24) Tjong, H.; Zhou, H.-X. *J. Phys. Chem. B* **2007**, *111*, 3055.

(25) Tjong, H.; Zhou, H.-X. *J. Chem. Phys.* **2007**, *126*, 195102.

(26) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.

(27) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348.

(28) Pokala, N.; Handel, T. M. *Protein Sci.* **2004**, *13*, 925.

(29) Wojciechowski, M.; Lesyng, B. *J. Phys. Chem. B* **2004**, *108*, 18368.

(30) Sigalov, G.; Scheffel, P.; Onufriev, A. *J. Chem. Phys.* **2005**, *122*, 094511.

(31) Tanizaki, S.; Feig, M. *J. Chem. Phys.* **2005**, *122*, 124706.

(32) Schnieders, M. J.; Ponder, J. W. *J. Chem. Theory Comput.* **2007**, *3*, 2083.

(33) Mongan, J.; Svrcek-Seiler, W. A.; Onufriev, A. *J. Chem. Phys.* **2007**, *127*, 185101.

(34) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693.

(35) Cai, W.; Xu, Z.; Baumketner, A. *J. Comput. Phys.* **2008**, *227*, 10162.

(36) Bardhan, J. P. *J. Chem. Phys.* **2008**, *129*, 144105.

(37) Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2009**, *30*, 673.

(38) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1996**, *61*, 8720.

(39) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(40) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.

(41) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506.

(42) Tomasi, J.; Mennucci, B.; Cancès, E. *J. Mol. Struct. (Theochem)* **1999**, *464*, 211.

(43) Ben-Naim, A. *Solvation Thermodynamics*; Plenum: New York, 1987; p 4.

(44) Born, M. *Z. Phys.* **1920**, *1*, 45.

(45) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1995**, *16*, 422.

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

(47) Tuñón, I.; Ruiz-López, M. F.; Rinaldi, D.; Bertrán, J. *J. Comput. Chem.* **1996**, *17*, 148.

(48) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2046.

(49) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.

(50) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365.

(51) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *PhysChemComm* **2002**, *5*, 117.

(52) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(53) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

(54) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(55) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. *J. Chem. Soc., Perkin Trans.* **1989**, *2*, 699.

(56) Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73.

(57) Abraham, M. H. *J. Phys. Org. Chem.* **1993**, *6*, 660.

(58) 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, The Netherlands, 1994; p 83.

(59) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database version 2009*; University of Minnesota: Minneapolis, MN, 2009. http://comp.chem.umn.edu/mnsol (accessed Jun 20, 2009).

(60) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769.

(61) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501.

(62) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066.

(63) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 408.

(64) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 4538.

(65) Winget, P.; Dolney, D. M.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvent Descriptor Database version 1999*; University of Minnesota: Minneapolis, MN, 1999. http://comp.chem.umn.edu/solvation/mnsddb.pdf (accessed Jun 20, 2009).

(66) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(67) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(68) Li, J.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1998**, *99*, 192.

(69) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(70) Fast, P. L.; Sánchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.

(71) Boresch, S.; Archontis, G.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 25.

(72) Smith, P. E.; Van Gunsteren, W. F. *J. Phys. Chem.* **1994**, *98*, 13735.

(73) Pethica, B. A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6253.

(74) Latimer, W. M.; Pitzer, K. S.; Slansky, C. M. *J. Chem. Phys.* **1939**, *7*, 108.

(75) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1996**, *255*, 327.

(76) Hyun, J.-K.; Ichiye, T. *J. Phys. Chem. B* **1997**, *101*, 3596.

(77) Babu, C. S.; Lim, C. *Chem. Phys. Lett.* **1999**, *310*, 225.

(78) 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. *Gaussian03, Revisions C.01, C.02, D.02, and E.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(79) Marenich, A. V.; Olson, R. M.; 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 2009*; University of Minnesota: Minneapolis, MN, 2009.

(80) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

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

(82) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2009**, *113*, 5806.

(83) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(84) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(85) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(86) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(87) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(88) Becke, A. D. *Phys. Rev. A:At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(89) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785.

(90) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(91) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(92) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(93) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

(94) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.

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

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

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

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

(99) Marenich, A. V.; Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *GESOL - version 2008*, University of Minnesota, Minneapolis, 2008.

(100) 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.

(101) Marenich, A. V.; Higashi, M.; Olson, R. M.; Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Zhu, T.; Hawkins, G. D.; Chuang, Y.-Y.; Fast, P. L.; Lynch, B. J.; Liotard, D. A.; Rinaldi, D.; Gao, J.; Cramer, C. J.; Truhlar, D. G. *GAMESSPLUS - version 2009−2*, University of Minnesota, Minneapolis, 2009.

(102) Schmuckler, M. E.; Barefoot, A. C.; Kleier, D. A.; Cobranchi, D. P. *Pest Manage. Sci.* **2000**, *56*, 521.

(103) Baldridge, K.; Klamt, A. *J. Chem. Phys.* **1997**, *106,* 6622.

(104) 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.

(105) 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.

(106) *Jaguar 6.5, Release 112*; Schrödinger, Inc.: Portland, OR, 2005.

(107) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(108) Barone, V.; Improta, R.; Rega, N. *Theor. Chem. Acc.* **2004**, *111*, 237.

(109) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.

(110) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 877.

(111) *CRC Handbook of Chemistry and Physics*; Lide, D. R., Ed.; Taylor and Francis: Boca Raton, FL, 2008−2009, Vol. 89. http://www.hbcpnetbase.com (accessed Jun 20, 2009).

CT900312Z

JCTC Journal of Chemical Theory and Computation

Coarse-Grained MD Simulations and Protein−Protein Interactions: The Cohesin−Dockerin System

Benjamin A. Hall and Mark S. P. Sansom*

Department of Biochemistry & Oxford Centre for Integrative Systems Biology, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

Received March 25, 2009

Abstract: Coarse-grained molecular dynamics (CG-MD) may be applied as part of a multiscale modeling approach to protein−protein interactions. The cohesin−dockerin interaction provides a valuable test system for evaluation of the use of CG-MD, as structural (X-ray) data indicate a dual binding mode for the cohesin−dockerin pair. CG-MD simulations (of 5 μs duration) of the association of cohesin and dockerin identify two distinct binding modes, which resemble those observed in X-ray structures. For each binding mode, ca. 80% of interfacial residues are predicted correctly. Furthermore, each of the binding modes identified by CG-MD is conformationally stable when converted to an atomistic model and used as the basis of a conventional atomistic MD simulation of duration 20 ns.

Introduction

Coarse-grained molecular dynamics (CG-MD) simulations have been used in a number of simulation studies of lipid bilayers and related systems.[1−4] More recently they have been extended with some success to simple membrane−peptide and membrane−protein systems.[5,6] The latter studies have included simulations of protein−protein interactions within membranes.[7,8] It is therefore of interest to explore whether such approaches can be applied to protein/protein interactions *outside* of a membrane environment. This is of relevance both in the context of computational studies of protein docking in general[9−11] and also in the context of wishing to develop multiscale biomolecular simulations[12−14] to study protein complexes.

The cohesin−dockerin system[15] provides a good test case for the application of CG-MD to protein−protein interactions, as crystallographic studies indicate a degree of complexity in the interaction, with the possibility of a dual binding mode.[16] This complex, which forms a key recognition element of the cellulosome,[17] provides a model system for macromolecular assembly. Subunits within the larger complex recognize each other through the interaction of type 1 dockerins and various cohesins. X-ray structures of the cohesin−dockerin complex from *Clostridium thermocellum*[15,16] suggest a role for plasticity in the protein−protein interaction. Comparisons of two structures indicate that the complex can be seen to bind in two modes, related to one another by a ∼180° rotation. The interactions defined in the crystal structures are composed of both packing of hydrophobic residues and of a hydrogen bonding network along two α-helices (related by a pseudo-2-fold axis) on the dockerin binding face.

Here we describe the application of CG-MD simulations to the encounter and interactions between dockerin and cohesin. These simulations are compared for the wild-type protein and for a mutant of dockerin[15,16] at the interaction interface. The CG-MD simulations of dockerin/cohesin encounter are also compared with those with simulations of the intact complex as seen in the two X-ray structures. We show that CG-MD is able to reproduce key aspects of the interaction between these two proteins. We also explore the conversion of a CG-MD generated model of the cohesin−dockerin complex to an atomistic representation, exemplifying a multiscale approach[14] to simulation of protein−protein interactions.

Methods

Coarse-Grained Simulations. Coarse-grain models of proteins were generated from their X-ray structures using procedures described previously[6,18] using a modified version

* Corresponding author phone: +44 1865 613306; fax: +44 1865 613238; e-mail: mark.sansom@bioch.ox.ac.uk.

10.1021/ct900140w CCC: $40.75
Published on Web 08/20/2009

Table 1. Summary of Simulations

simulation[a]	protein coordinate source	starting configuration
WT_{dock}	1OHZ	subunits separated
$WT_{X\text{-}ray}$	1OHZ	crystal structure
MUT_{dock}	2CCL	subunits separated
$MUT_{X\text{-}ray}$	2CCL	crystal structure

[a] For each simulation setup, four individual simulations of duration 5 μs were performed, differing in their initial random velocities.

of procedure described by Marrink and colleagues.[2] Briefly, a 4:1 mapping of non-H atoms to coarse grain particles is used. Interparticle interactions are modeled as Lennard-Jones interactions between 4 classes of particles (2 of which are divided into subtypes to reflect hydrogen bonding). Electrostatics interactions are treated Coulombically. The protein fold is maintained using an elastic network model[19] with a cutoff between backbone particles of 7 Å and a force constant of 10 kJ mol^{-1} $Å^{-2}$. Lennard-Jones interactions were shifted to zero between 9 and 12 Å, and electrostatic interactions were shifted to zero between 0 and 12 Å. CG simulations were performed with GROMACS 3.3.1. (www.gromacs.org).[20]

To perform a docking simulation the two binding partners (cohesin and dockerin) in a cubic simulation box (length 100 Å) at a distance of 40 Å apart, equal to the sum of their respective radii of gyrations multiplied by 1.2, plus the cutoff distance for interactions. The proteins were randomly oriented with respect to one another, and four separate simulations were performed, each of duration 5 μs. The system temperature was 348 K, maintained using a Berendsen thermostat[21] ($\tau_T = 1$ ps). Pressure was coupled with a Berendsen barostat at 1 bar ($\tau_P = 1$ ps). Simulations of the 1OHZ and 2CCL structures were performed from their crystal structure with the same representation.

Atomistic Simulations. Following a CG simulation, 20 representative structures (based on cluster analysis) from the last 0.5 μs in thesis of a trajectory were used along with the initial X-ray structures of component proteins from 1OHZ to generate 25 structures using MODELER (http://www.salilab.org/modeller/).[22] The resultant model structure of the cohesin−dockerin complex was used as the starting point for a conventional atomistic MD simulation using GROMACS and the GROMOS96 force field.[23] The structure was solvated with SPC water (10,500 waters in a (70 Å)^3 simulation box) and counterions, energy minimized for 100 steps (using the steepest descent algorithm), and equilibrated by a 0.5 ns protein position-restrained) simulation followed by a 20 ns unrestrained simulation. The temperature was 300 K and was coupled using a Berendsen thermostat ($\tau_T = 0.1$ ps). Long range electrostatic interactions were treated with particle mesh Ewald.[24] Analysis of simulations and visualization used VMD.[25]

Results

Progress of CG-MD Simulations. Four sets of simulations were performed, as summarized in Table 1. These correspond to the wild-type (WT; PDB id 1OHZ) and mutant (MUT; PDB id 2CCL) structures. Note that the mutant has two changes of side chain (S45A, T46A) on the surface of one of the interaction site helices (α3) of dockerin, resulting in a change in the mode of interaction within the crystal structure.[16] For each of WT and MUT, simulations were performed starting with either the two subunits separated (and in random orientations relative to one another) or with the two subunits in the complex found in their respective crystal structure.

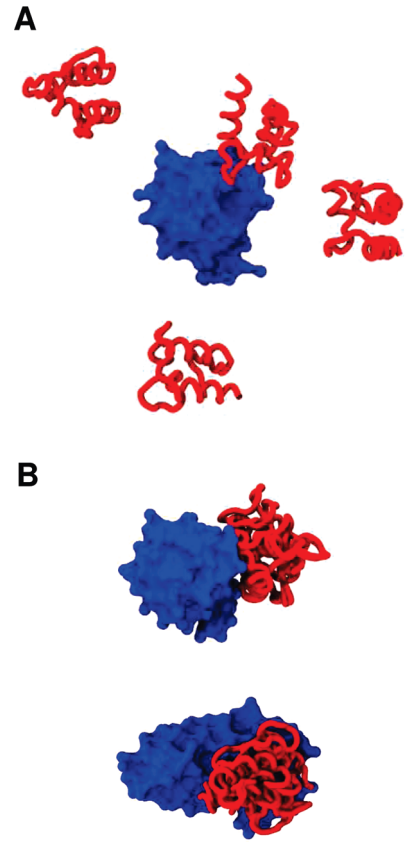


Figure 1. Progress of the four CG-MD simulations (WT_{dock}) of the cohesin−dockerin interaction. **A** Initial positions of dockerin molecules (red; Cα trace) relative to cohesin (blue; surface). **B** Two perpendicular views of the final (5 μs) positions of the four dockerin molecules (red; superimposed structures, one from each simulation) relative to cohesin (blue).

The start and end configurations of four 5 μs CG-MD simulations of the WT_{dock} cohesin−dockerin encounter are shown in Figure 1. In each case the dockerin finds the same 'site' on the surface of the cohesin molecule. The cohesin usually attaches to this site within 100 ns of simulation time, and the final binding mode is achieved within ∼1 μs (although it should be noted that the scaling of CG time to real time is somewhat uncertain[2,18]). It is evident that in all four simulations the same broadly defined 'site' on the cohesin is occupied by the dockerin. This is the same site of interaction as is seen in both of the X-ray structures (1OHZ and 2CCL) which define the two binding modes of dockerin on cohesin.[16] Indeed in comparable simulations of the encounter of cohesin with the mutated dockerin (MUT_{dock})

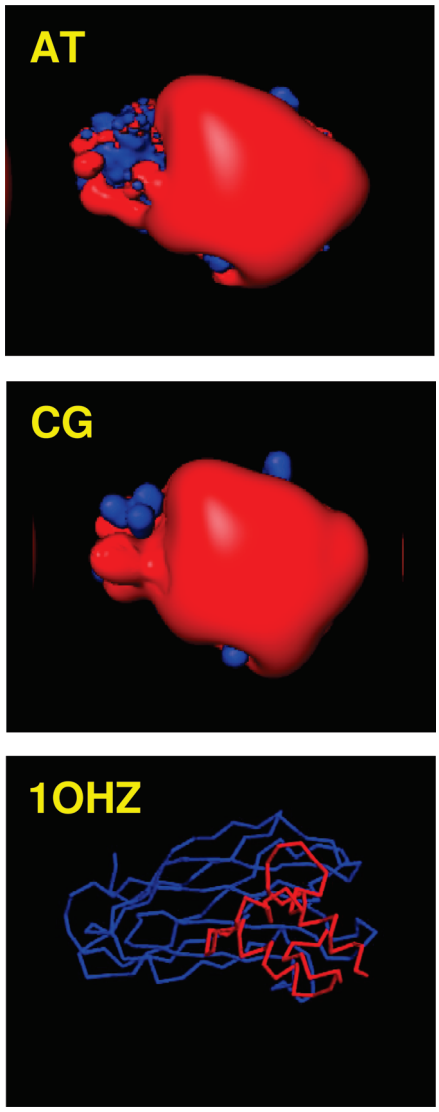


Figure 2. Isopotential surfaces (contours at ±3 kT, red = negative, blue = positive) from Poisson−Boltzmann electrostatics calculations (using APBS[26]) for atomistic (AT) and coarse-grained (CG) models of cohesin. A Cα trace representation of the cohesin/dockerin complex (1OHZ) in approximately the same orientation is shown for reference.

the same interaction site is seen in the two simulations for which a complex is formed within 5 μs (see below for further details).

Comparison of surface electrostatic potentials (calculated using APBS;[26] Figure 2) from CG models of cohesin and dockerin with those from the corresponding all atom structures reveals that the coarse-graining does not qualitatively alter the overall pattern of protein surface electrostatics between the AT and the CG model. The binding surface of cohesin has a distinct region of negative potential, which interacts with a complementary positive surface on dockerin.

The nature of the protein−protein interaction surface yielded by the CG-MD docking can be examined in more detail by calculation of the fraction of correct residues in the interface (F*IR* - see Table 2 for details) and by examination in more detail of the patterns of contact residues (Figure 3).

Table 2. Fraction of Residues in the Interface for WT_{dock} Simulations[a]

comparison structure	protein	F*IR*				
		Sim1	Sim2	Sim3	Sim4	mean (SD)
1OHZ	cohesin	0.85	0.87	0.87	0.90	0.87 (0.02)
	dockerin	0.57	0.84	0.61	0.86	0.72 (0.15)
2CCL	cohesin	0.76	0.76	0.74	0.80	0.77 (0.03)
	dockerin	0.51	0.35	0.50	0.38	0.44 (0.08)

[a] F*IR* = (number of correct residues in interface)/(total number of residues in interface). The values given are averages over the last microsecond of each simulation. Residues within the interface are defined using a cutoff distance between partner proteins of 7.5 Å. The 'correct' residues are scored according to either the 1OHZ or 2CCL structures of the cohesin−dockerin interface.

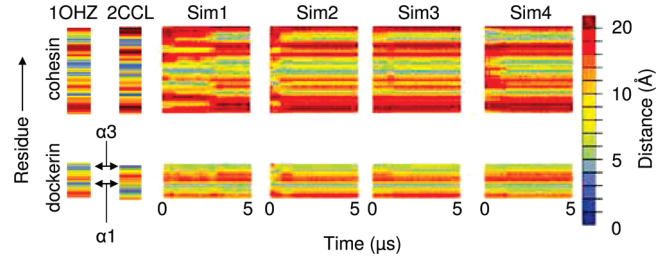


Figure 3. Fingerprint analysis of the time dependence of the cohesin−dockerin contacts a function of time for the WT_{dock} simulations. Blues are closer contacts, reds more distant. The corresponding fingerprints of the contacts in the two crystal structures (1OHZ and 2CLL) are given to the left of the time dependent fingerprints of the four CG-MD simulations. (Interaction fingerprints are calculated from the shortest distance between a residue in one binding protein and any residue in its partner. This is plotted against time for each simulation.)

The 'correct' residues for the cohesin−dockerin interface can be defined using either the WT (1OHZ) or MUT (2CCL) structures (which have very similar residues in their interfaces). Thus, for each simulation and protein two F*IR* values can be defined (Table 1). It should be noted that when evaluating F*IR* values for CG models the reduction in particle number coupled with the increase in particle size in coarse-graining requires a cutoff of 7.5 Å to be used, rather than a more typical value of e.g. 5 Å used to evaluate inter-residue contacts in an AT model (see Supporting Information, Figure S1). From these it can be seen that scoring against either X-ray structure, between 80% and 90% of the cohesin interface residues are present in the CG-MD generated structures from the WT_{dock} simulations. For dockerin, between ∼50% (scored against 2CCL) and ∼75% (scored against 1OHZ) of the correct interfacial residues are present in these CG-MD simulations. rmsd analyses and clustering reinforces the identification of 2 distinct end points in the simulations, resembling the two crystal conformations.

The patterns of interacting residues can be identified through the course of the simulation via residue contact 'fingerprints' (Figure 3). Examination of these fingerprints for the two X-ray structures (1OHZ and 2CCL) shows that similar residues are involved in both modes of binding (but see below). Examination of the time-dependent simulations reveals some changes in the initial contact pattern over the first 1−2 μs of a simulation, followed by a pattern of interaction which remains constant until the end of the

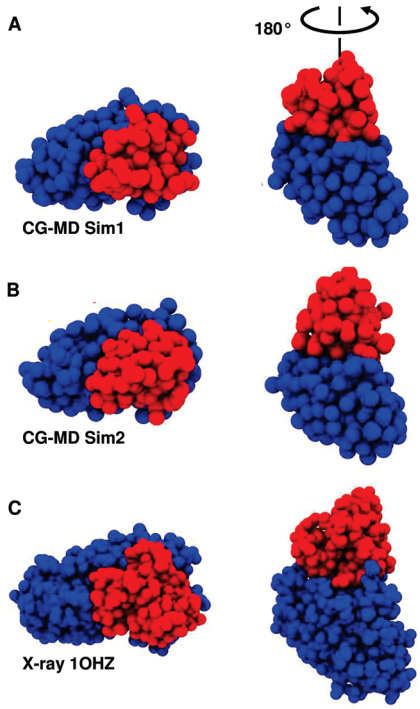


Figure 4. Identification of alternative models of interactions. Orthogonal views of the cohesin (blue)−dockerin (red) complex from **A** CG-MD WT_{dock} Sim1, **B** CG-MD WT_{dock} Sim2, and **C** X-ray structure 1OHZ. In **A** the ∼180° rotation about an axis perpendicular to the protein−protein interface which relates a 1OHZ-like to a 2CCL-like complex is indicated.

simulation. The final patterns are similar for the four WT_{dock} simulations (and in particular the main contacts are for the α1 and α3 helices of dockerin) and in turn resemble those in the two X-ray structures. The MUT_{dock} simulations match the two X-ray structures in two of the four simulations, in one simulation matching the 2CCL structure better than in any of the WT_{dock} simulations (as revealed by the rmsd values, see Supporting Information, Figure S3B). The other two MUT_{dock} simulations yielded non-native binding interactions.

Two Interaction Modes. Visualization of the structure yielded by CG-MD WT_{dock} reveals a complication. There are two distinct orientations of the dockerin molecule while interacting with the same site on the cohesin. These correspond to an ∼180° rotation of the dockerin relative to cohesin about an axis normal to the protein−protein interface (compare Figure 4A and B). This can be compared to the two X-ray structures (1OHZ and 2CCL) which are related by a similar transformation, reflecting the approximate 2-fold symmetry of the structure of dockerin, which relates helices α1 and α3, i.e. the two helices which interact with the cohesin binding site. Again, simple visualization suggests that the CG-MD structures can be classified as either 1OHZ-like (Sim1 and Sim3, Figure 5A) or 2CCL-like (Sim2 and Sim4, Figure 5B) in terms of the orientation of the dockerin relative to the cohesin, even though both modeled (and indeed X-ray) complexes share the same interfacial residues. Interestingly, in the two mutant simulations (MUT_{dock}) which yielded complexes, Sim4 gave a 1OHZ-like structure and Sim1 a 2CCL-like structure, whereas the remaining pair gave non-native docks (see Supporting Information, Figure S3).

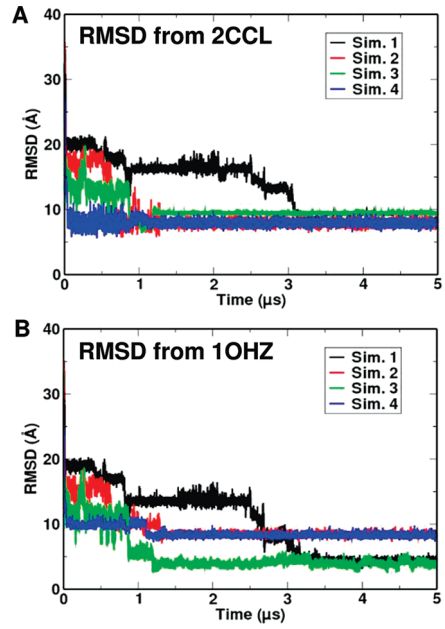


Figure 5. Comparison with X-ray structures. For each of the four WT_{dock} self-assembly simulations, Cα RMSDs are evaluated relative to the two crystal structures, namely **A** 2CCL and **B** 1OHZ. From this it can be seen that Sim1 and Sim3 yield structures close to 1OHZ, while Sim2 and Sim4 yield structures closer to 2CCL.

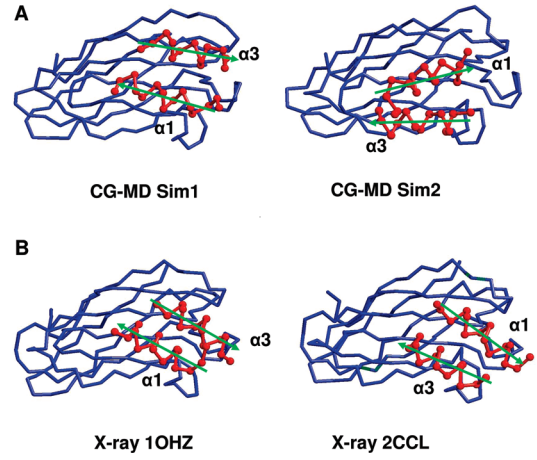


Figure 6. **A** Final (5 μs) structure of the cohesin (blue)-dockerin (red) complex from CG-MD simulations WT_{dock} Sim1 and Sim2, showing the Cα trace of the complete cohesin domain but just the two interaction helices (α1 and α3) of the dockerin domain.. The directions of the two interaction helices are shown by green arrows. **B** Equivalent diagrams for the 1OHZ and 2CCL X-ray structures of the complex.

This can be seen more clearly if one focuses on just the α1 and α3 interaction helices of dockerin (Figure 6). Comparison of e.g. the end structures of simulations WT_{dock} Sim1 and Sim2 reveals two different orientations of α1 and α3 on the cohesin binding site. This closely mirrors the similar two orientations seen in the wild-type (1OHZ) and mutant (S45A-T46A, in α3; 2CCL) structures. Thus it would seem that not only is CG-MD able to predict the correct binding interface for the cohesin−dockerin interaction but also is able to reproduce the dual binding mode revealed by comparison of a wild-type and mutant crystal structure.

This may be analyzed in more detail via calculation of the Cα particle root-mean-square deviations (RMSDs) for the WT_{dock} simulations relative to the two possible structures of the complex (i.e., 1OHZ and 2CCL; Figure 5). Those simulations (Sim1 and Sim3) which converge to a 1OHZ-like structure are within 5 Å of the original 1OHZ crystal structure. They also give mean F_{NAT} values of 0.21 and 0.26 over the last microsecond (compared to the Sim2 and Sim4, which give 0.07; all calculated with a 7.5 Å cutoff). This is comparable to the drift from this structure seen in the $1OHZ_{Xray}$ simulation (see Supporting InformationFigure S2), which corresponds to the drift observed in simulations starting from the crystal structure. Likewise, the rmsd of the 2CCL-like structures is also similar to that observed in the $2CCL_{Xray}$ simulations. In addition, cluster analysis supports the convergence of Sim1 and Sim3 to a 1OHZ-like common structure and of Sim2 and Sim4 to a 2CCL-like common structure (data not shown).

Atomistic MD Simulations. In order to facilitate further comparison with the X-ray structures, and also to explore the (short time scale) conformational stability of the model complexes generated by CG-MD, the latter were converted to atomistic models and used as the starting point for short (20 ns) conventional MD simulations. Two CG complexes were used as the starting point for AT-MD simulations. One (from WT_{dock} Sim1) was a 1OHZ-like complex; the other (from WT_{dock} Sim2) was a 2CCL-like complex, as discussed above. We note that AT simulations have been used previously to refine and explore flexible docking of proteins and that in these studies some rearrangements in protein−protein interactions were seen on a nanosecond time scale.[27] However, we realize that large scale rearrangements of docked structures are unlikely to occur on a 20 ns time scale. Thus, the AT-MD simulations largely serve to 'relax' the CG-MD generated models.

The individual proteins were conformationally stable (over 20 ns) in both AT-MD simulations. Thus, for the Sim1 complex, the Cα RMSDs of the individual proteins are ∼2 Å for the cohesin and ∼3 Å for the dockerin (Figure 7A). Similar behavior is observed for the Sim2 derived structure. Furthermore, the two Ca^{2+} ions bound by the EF hands of the dockerin remained stably bound over the course of the 20 ns. (Note that these ions were not modeled explicitly in the CG representation). Thus, the individual proteins remain conformationally stable after the CG to AT conversion.

It is also of interest to examine the energetic relaxation (Figure 7B) of the complex over the course of the AT-MD simulation. From this analysis it can be seen that, as anticipated, there is a fast initial relaxation (over ∼2 ns), but no further major changes. This suggests 'local' relaxation of the CG-generated model but no substantial changes in conformation over the course of the AT-MD simulation.

By comparing the Cα rmsd of the model complex vs the two X-ray structures (Figure 7C) it can be seen that the structure of the complex from Sim1 compares well with the 1OHZ X-ray structure, with an overall Cα rmsd of ∼3 Å. This does not change significantly over the course of the AT-MD simulation and is within the drift typically observed in atomistic simulations from *crystal* structures. By way of

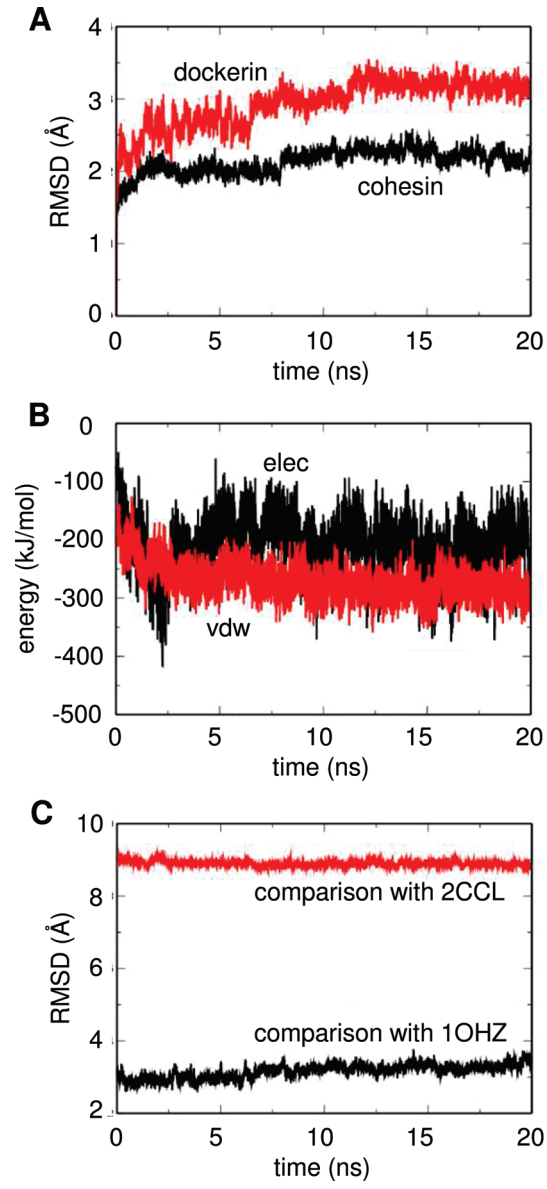


Figure 7. Atomistic simulation of the cohesin−dockerin complex generated from CG-MD WT_{dock} Sim1. **A** Cα RMSDs relative to starting model for cohesin and dockerin separately (fitted on cohesin). **B** Potential energy of interaction vs time. **C** Cα RMSDs of the complete complex vs the two X-ray structures vs time.

comparison, the rmsd vs 2CCL is ∼9 Å, supporting our classification of Sim1 having yielded a 1OHZ-like complex. For Sim2 the rmsd vs 1OHZ is 8 Å and that vs 2CCL is ∼6 Å. Again there is little change over the course of 20 ns AT-MD. Thus, we may conclude that although conversion to atomistic resolution and subsequent MD simulation allows some local relaxation of the proteins and there interface, as one might expect no substantial changes are seen on this relatively short time scale.

One may also analyze the AT-MD simulations via the metrics used in e.g. the CAPRI protein docking assessment.[9] F*NAT* is defined as the (number of correct contacts)/(number of contacts in the crystal) assessed using a 5 Å cutoff. For the AT-MD simulation based on Sim1 F*NAT* (compared to 1OHZ) is ∼0.25. The F*IR* values are essentially unchanged from the CG-MD simulations. Thus, CG-MD has generated

an acceptable structure of the 1OHZ-like complex. For Sim2 the comparison is not as promising - its F*NAT* is <0.1 (compared to 2CCL), and the F*IR* values are again the same as in the CG-MD model. However, one should remember that the 2CCL X-ray structure is of a mutant rather than wild-type protein.

Discussion

In this study we demonstrate the application of CG-MD to a complex protein−protein interaction problem involving a dual binding mode of the two proteins. Within the limitations of a CG representation, the approach is successful in revealing both binding modes[16,28] of the cohesin−dockerin complex.

A multiscale approach (in which the CG model of the complex was converted to an atomistic model) confirmed the conformational stability of the resultant complex, albeit in relatively short time scale MD simulations. This multiscale strategy has some potential for modeling protein−protein interactions. The CG model allows protein flexibility to be addressed efficiently, while conversion to an atomistic model enables validation and refinement of the resultant structures. Further exploration of this method with a wider set of test cases (Hall and Sansom, unpublished results) will enable refinement of the approach e.g. via refinement of CG models for proteins.[29]

It is of interest that CG-MD predicted the 1OHZ binding mode better than the 2CCL mode. However, we note that in the experimental studies a mutation of key interaction residues (S45A-T46A, in the α3 helix of dockerin) was required to promote the 2CCL binding mode. In particular, the 2CCL-like structure yielded by CG-MD is to be related to the (mutant) crystal structure by a ∼10 Å translation. Of course, the crystal structure also represents the protein complex at a temperature of 110 K, and it is possible that the interaction is more plastic in the wild-type under physiological conditions. However, a similar orientation has been observed in the 2CCL-like mode seen in the A47S/F48T mutant of the *Cl. cellulolyticum* (PDB id 2 VN5) cohesin−dockerin complex.[28] CG-MD simulations of the 2CCL crystal structure (MUT_{Xray}) do indicate a somewhat 'softer' interface for the S45A, T46A dockerin mutant, clustering less clearly than the $WT_{X\text{-}ray}$ simulation and drifting further away from the crystal structure (see Supporting Information, Figure S2).

Protein−protein interactions in general, and the cohesin−dockerin interaction in particular, have been studied using a wide range of computational approaches to protein−protein docking (see refs 10 and 11 for recent reviews), and indeed the 1OHZ cohesin−dockerin complex was used in the CAPRI (round 4) assessment of protein docking (http://www.ebi.ac.uk/msd-srv/capri/capri.html) both as a direct protein−protein dock (T12) and as dock between the experimental cohesin structure and a homology model of dockerin (T12).[10]

Our results compare favorably with those of a number of protein docking algorithms applied to 1OHZ in round 4 of CAPRI, including ZDOCK,[30] *Hex*,[31] and HADDOCK.[32] Thus ZDOCK yielded a F*IR* of 0.84 and HADDOCK of ∼0.8 compared to ∼0.8 for CG-MD (see above). Interestingly, it was noted for e.g. HADDOCK that several predicted docks were for a ∼180° rotated (i.e., 2CCL-like) interaction of the two proteins. Scoring of these docking methods against the 2CCL interaction mode have not been reported.

In summary, we have described CG-MD as an extension of the use of MD simulations for docking (e.g., refs 27 and 33), related to a number of other coarse-grained approaches to study protein−protein interactions (e.g., refs 34 and 35). It will be of interest to further develop the multiscale approach applied here to the cohesin−dockerin interaction to a range of other protein−protein interactions, perhaps by combining high throughput approaches (to improve sampling) with treatment of the initial protein−protein encounter by e.g. Brownian dynamics simulation[36] followed by CG- and AT-MD.

Acknowledgment. This work was supported by grants from the BBSRC and MRC. Our thanks to Oliver Beckstein and Peter Bond for helpful discussions.

Supporting Information Available: Figures S1−S3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. A coarse grain model for phospholipid simulations. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

(2) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(3) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

(4) Stevens, M. J. Coarse-grained simulations of lipid bilayers. *J. Chem. Phys.* **2004**, *121*, 11942–11948.

(5) Lindahl, E.; Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425–431.

(6) 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.

(7) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.

(8) Psachoulia, E.; Bond, P. J.; Fowler, P. W.; Sansom, M. S. P. Helix-helix interactions in membrane proteins: coarse grained simulations of glycophorin helix dimerization. *Biochemistry* **2008**, *47*, 10503–105012.

(9) Janin, J.; Henrick, K.; Moult, J.; Ten Eyck, L.; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 2–9.

(10) Gray, J. J. High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 183–193.

(11) Bonvin, A. M. J. J. Flexible protein-protein docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194–200.

(12) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(13) Ayton, G. A.; Noid, W. G.; Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(14) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. Multiscale methods for macromolecular simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.

(15) Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. Cellulosome assembly revealed by the crystal structure of the cohesin−dockerin complex. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13809–13814.

(16) Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C.; Gilbert, H. J. Evidence for a dual binding mode of dockerin modules to cohesins. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3089–3094.

(17) Doi, R. H.; Kosugi, A. Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nature Rev. Microbiol.* **2004**, *2*, 541–551.

(18) Bond, P. J.; Sansom, M. S. P. Insertion and assembly of membrane proteins via simulation. *J. Am. Chem. Soc.* **2006**, *128*, 2697–2704.

(19) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80*, 505–515.

(20) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(21) 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. Chem. Phys.* **1984**, *81*, 3684–3690.

(22) Sali, A.; Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.

(23) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(24) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(25) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* 1996*14*, 33–38.

(26) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.

(27) Wang, T.; Wade, R. C. Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. *Proteins: Struct., Funct., Bioinf.* **2003**, *50*, 158–169.

(28) Pinheiro, B. A.; Proctor, M. R.; Martinez-Fleites, C.; Prates, J. A. M.; Money, V. A.; Davies, G. J.; Bayer, E. A.; Fontes, C.; Fierobe, H. P.; Gilbert, H. J. The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. *J. Biol. Chem.* **2008**, *283*, 18422–18430.

(29) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(30) Wiehe, K.; Pierce, B.; Mintseris, J.; Tong, W. W.; Anderson, R.; Chen, R.; Weng, Z. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 207–213.

(31) Mustard, D.; Ritchie, D. W. Docking essential dynamics eigenstructures. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 269–274.

(32) van Dijk, A. D. J.; de Vries, S. J.; Dominguez, C.; Chen, H.; Zhou, H. X.; Bonvin, A. Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 232–238.

(33) Smith, G. R.; Fitzjohn, P. W.; Page, C. S.; Bates, P. A. Incorporation of flexibility into rigid-body docking: Applications in rounds 3−5 of CAPRI. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 263–268.

(34) Zacharias, M. ATTRACT: Protein-protein docking in CAPRI using a reduced protein model. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 252–256.

(35) Basdevant, N.; Borgis, D.; Ha-Doung, T. A coarse-grained protein-protein potential derived from an all-atom force field. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.

(36) Spaar, A.; Dammer, C.; Gabdoulline, R. R.; Wade, R. C.; Helms, V. Diffusional encounter of barnase and barstar. *Biophys. J.* **2006**, *90*, 1913–1924.

CT900140W

JCTC Journal of Chemical Theory and Computation

Computational Screening of Rhodopsin Mutations Associated with Retinitis Pigmentosa

Angelo Felline,[†] Michele Seeber,[†] Francesco Rao,[‡] and Francesca Fanelli*,[†]

Dulbecco Telethon Institute and Department of Chemistry, via Campi 183, 41100 Modena, Italy, and Laboratoire de Chimie Biophysique/ISIS 8, Universitè Louis Pasteur, allee Gaspard Monge, 67000 Strasbourg, France

Received March 26, 2009

Abstract: Retinitis pigmentosa (RP) refers to a group of debilitating, hereditary disorders that cause severe visual impairment in as many as 1.5 million patients worldwide. Rhodopsin mutations account for >25% of the autosomal dominant form of the disease (ADRP). Forty artificial and ADRP-associated mutations located in the second extracellular loop (EL2) that folds into a twisted β-hairpin were screened through replica exchange molecular dynamics (REMD) simulations using the FACTS implicit solvent model. According to in vitro experiments, ADRP-linked mutants fail to express at the plasma membrane and/or to reconstitute with 11-*cis*-retinal, indicative of variable defects in protein folding and/or stability. The computational protocol was first probed on the protein G C-terminal β-hairpin, proving the effectiveness of the implicit solvent model in reproducing the free energy landscape of β-hairpin formation. Eight out of the 40 EL2 mutants resulted in misfolding effects on the native β-hairpin structure, consistent with in vitro evidence that they all share severe impairments in folding/expression. Five mutants displayed moderate misfolding attitudes, whereas the remaining 27 mutants, overall characterized by milder effects on rhodopsin expression, did not perturb significantly the conformational behavior of the native β-hairpin but are expected to exert variably disturbing effects on the native interactions of the loop with the chromophore and/or the surrounding receptor domains. Collectively, the results of this study add structural insight to the poorly resolved biochemical behavior of selected class II ADRP mutations, a fundamental step toward an understanding of the atomistic causes of the disease.

1. Introduction

Retinitis pigmentosa (RP) is a group of hereditary human diseases that are characterized by progressive retinal degeneration due to death of the rod photoreceptor cells, the vertebrate photoreceptors dedicated to dim light vision.[1−3] Patients affected by RP display nyctalopia (night blindness), progressive loss of peripheral and, eventually, central vision and the characteristic accumulation of intraretinal pigment deposits, from which the disease gets its name. Despite the high genetic heterogeneity of the RP syndrome, over 120 point mutations have been discovered in the gene of rhodopsin, the visual pigment molecule of rod cells that generates a detectable electrical response following light capture.[4] Although some of the rhodopsin mutations cause autosomal recessive retinitis pigmentosa (ARRP), the vast majority cause the autosomal dominant form (ADRP) of the pathology (collected in part in the rhodopsin mutation database at http://www.retina-international.com/sci-news/rhomut.htm).[1−3]

Rhodopsin, the cornerstone of family A of G protein coupled receptors (GPCRs),[5] is a transmembrane receptor protein expressed in the retina and composed of a protein (opsin) and a chromophore. Opsin is an up-and-down bundle of seven transmembrane (TM) helices linked to three intracellular (IL) and three extracellular (EL) loops as well

* Corresponding author. Tel: +39 059 2055114. Fax: +39 059 37353. E-mail: fanelli@unimo.it.

[†] Dulbecco Telethon Institute and Department of Chemistry.

[‡] Universitè Louis Pasteur.

10.1021/ct900145u CCC: $40.75
Published on Web 07/31/2009

Table 1. Computational and in Vitro-Determined Indices Concerning Spontaneous and Artificial Mutants of Rhodopsin EL2

mutant[a]	D&K[b](%)	P&R[c](%)	Sut[d](%)	HB1−4_{avg}[e]	rmsd[f]	El2 misfolding[g]	fold[h]	expr[j]	retinal[i]	ref[k]
WT				77.44	88.66					
R177C*	58.80	44.46	**44.79**[l]	44.79	53.76	+++	nd	nd	no	18
R177K	63.19	53.33	**59.79**	59.79	69.46	+	≈	≈	≈	18
R177Q*	**52.40**	53.46	52.39	52.40	57.48	++	≈	≈	≈	18
$Y178C_{ADRP}$	74.16	**72.14**	57.29	72.14	86.49	−	nd	no	no	37, 38
$Y178N_{ADRP}$	70.92	62.49	**64.81**	64.81	80.49	+	nd	nd	nd	39
$P180A_{ADRP}$				38.70	34.32	+++	no	nd	nd	40
$G182S_{ADRP}$	**70.35**	71.76	65.73	70.35	81.82	−	nd	−	nd	41, 42
$Q184P_{ADRP}$				67.97	78.49	+	nd	nd	nd	43
C185S	**66.44**	79.25	66.29	66.44	78.69	+	≈	≈	≈	44, 45
$S186P^*_{ADRP}$				52.54	61.88	++	nd	no	no	38, 41
$S186W^*_{ADRP}$	76.46	53.16	**74.50**	74.50	90.98	+	nd	nd	nd	46
C187A				67.55	81.22	+	−	nd	−	45
$C187Y_{ADRP}$	**71.59**	70.02	73.99	71.59	92.09	−	nd	nd	no	45
$G188E_{ADRP}$	70.05	**67.79**	65.57	67.79	80.43	+	nd	−	nd	41
$G188R^{**}_{ADRP}$	72.53	74.77	**72.83**	72.83	88.22	+	no	no	no	38, 40, 41, 47
$D190A^*_{ADRP}$				43.95	51.4	+++	−	nd	−	18, 47, 48
D190C		51.72	**45.60**	45.60	53.68	+++	nd	nd	no	18
D190E	**77.72**	80.14	75.38	77.72	89.55	−	nd	nd	no	18
$D190G^*_{ADRP}$				40.88	49.03	+++	nd	no	no	18, 37
$D190N^*_{ADRP}$	60.09	**56.79**	45.35	56.79	68.64	++	≈	≈	≈	18, 38
$D190Y^*_{ADRP}$	−	**43.11**	53.26	43.11	50.73	+++	nd	no	no	18, 38

[a] Spontaneous (ADRP) and artificial mutants of rhodopsin EL2. A single asterisk means that the energy of the folded state is about 0.5 RT units higher than that of the wild type, whereas double asterisks mean that the lowest energy basin is shifted in between two and three interstrand H-bonds. [b] Average interstrand H-bond probability (HB1−4_{avg}) derived from simulation on the mutated side chain rotamer from the D&K library.[29] [c] HB1−4_{avg} index derived from simulation on the mutated side chain rotamer from the P&R library.[30] [d] HB1−4_{avg} index derived from simulation on the mutated side chain rotamer from the Sut library.[31] [e] Selected HB1−4_{avg} index. [f] Fraction of native-like structures, i.e., those characterized by a C_α-rmsd ≤2 Å from the native structure. [g] Predicted misfolding effect, based upon REMD simulations. In detail, the symbols "+++", "++", "+", and "−" stand, respectively, for misfolding, moderately misfolding, low misfolding, and non-misfolding. [h] Preservation of the native fold by mutation, according to in vitro experiments. The symbols "no", "−", "≈", and "nd" stand, respectively, for unfolded, misfolded, folded like the wild type, and not determined. [i] Maintenance of the native expression by mutation. The symbols "no", "−", "≈", and "nd" stand, respectively, for not expressed, expressed lower than the wild type, expressed like the wild type, and not determined. [j] Preservation of the native retinal binding ability by mutation. The symbols "no", "−", "≈", and "nd" stand, respectively, for no binding, lower binding than the wild type, wild type-like binding, and not determined. [k] Source of information concerning in vitro data. [l] In bold are the HB1−4_{avg} indices selected as the closest to the average value from three independent simulations.

as to an extracellular N-term and an intracellular C-term.[6] In the dark, inactive state of rhodopsin, the chromophore is 11-*cis*-retinal, forming a Schiff base with K296 in H7 ("H" stands for helix). Absorption of a photon provides rhodopsin with the energy to form the active state. Three phases of the activation process can be distinguished: (1) light-induced cis−trans isomerization of the retinal, (2) thermal relaxation of the retinal−protein complex, and (3) the late equilibria that are affected by the interaction of rhodopsin with the G protein.[7,8] The latter state, metarhodopsin II (Meta II or MII), is in equilibrium with Meta I (MI), which derives from the lumirhodopsin state (LUMI) following a shift of the protonated Schiff base (PSB) from the counterion E113 (in H3) to E181 (in EL2).[7,9]

The spectrum of biochemical and cellular properties of rhodopsin mutations associated with ADRP is quite wide and includes six different classes.[1−3]

ADRP rhodopsin mutations are essentially located in the N-term, EL1, EL2, and the seven-helix bundle. With respect to the extracellular domains, pathogenic mutations concentrate essentially in EL2 (i.e., the S176-T198 sequence), which contains a highly stable twisted β-hairpin that lays alongside the retinal chromophore. Very recent solid-state NMR determinations support the high stability of this rhodopsin segment that seems to change position, rather than conformation, following photoactivation of the pigment.[10] Moreover, five out of the nine amino acid residues predicted to participate in the stability core of rhodopsin by either one or two different computational methods,[11] i.e., R177$^{(1)}$, P180$^{(4)}$, Q184$^{(8)}$, C185$^{(9)}$, S186$^{(10)}$, and C187$^{(11)}$ (each amino acid in the peptide is labeled by two numbers; the first number is the sequential one, whereas the number in parentheses indicates the position of the amino acid residue in the 14-residue β-hairpin), belong to EL2, thus emphasizing the fundamental role of this loop in rhodopsin folding.[11]

This study is part of an ambitious project aimed at structurally characterizing, through molecular simulations, the majority of spontaneous rhodopsin mutations. Within this project, computational protocols are defined ad hoc on the basis of the structural localization and the biochemical classification, if any, of each mutation. In this respect, the 40 mutations considered in this study lay in the β-hairpin portion of EL2 and comprise 15 ADRP mutants, the majority of which falls in the biochemical class II, as they fail to express at the plasma membrane and/or to reconstitute with 11-*cis*-retinal, indicative of variable defects in protein folding and/or stability (Tables 1 and 2 and references therein) The simulated mutants include also all the 19 possible substitutions of E181$^{(5)}$, comprising one ADRP mutation (i.e., E181K) (Table 2).

The effects of the 40 natural and artificial mutations on the structural stability of rhodopsin EL2 were, therefore, studied through parallel replica exchange molecular dynamics (REMD) simulations using the FACTS (fast analytical continuum treatment of solvation) implicit solvent model implemented in the CHARMM biomolecular simulation

Table 2. Computational and in Vitro-Determined Indices Concerning Spontaneous and Artificial Mutants of Rhodopsin E181

mutant[a]	D&K[c](%)	P&R[d](%)	Sut[e](%)	HB1−4$_{avg}$[f]	rmsd[g](%)	EL2 misfolding[i]	expr.[j](%)	λ_{max}[j] dark (nm)	λ_{max}[k] light (nm)	HA[l] react (min)	MII[m] half-life (min)
WT				77.44	88.66		100	501	382	5440 ± 170	12.5 ± 0.5
E181A*				66.37	80.82	+	30−50	499	386	60.1 ± 4.0	17.3 ± 0.7
E181C	72.68	**70.70**[n]	66.97	70.70	80.41	−	30−50	499	383	21.7 ± 2.0	16.7 ± 0.5
E181D	47.50	**56.27**	60.76	56.27	61.09	+	>50	497	383	1733 ± 357	7.7 ± 0.2
E181F	58.29	78.98	**72.45**	72.45	81.52	−	30−50	501	375	8.4 ± 0.4	52 ± 18
E181G*				42.74	43.84	+++	30−50	500	384	20.1 ± 4.3	6.5 ± 0.1
E181H*	**62.35**	67.33	58.65	62.35	76.17	+	30−50	497	388	24.7 ± 5.0	10 ± 2.0
E181I	**76.99**	75.19	79.09	76.99	86.99	−	30−50	501	384	1.8 ± 0.2	9.0
E181K*[b]$_{ADRP}$	45.97	70.70	**70.28**	70.28	82.54	+	<20	nd	nd	nd	nd
E181L*	70.85	**70.37**	63.26	70.37	81.47	+	30−50	502	384	nd	9.6 ± 0.9
E181M	54.19	**73.94**	75.28	73.94	86.24	−	>50	500	383	5.0 ± 0.6	15.4 ± 0.7
E181N	**65.91**	54.07	71.38	65.91	76.63	+	30−50	500	384	25.1 ± 3.0	11.0 ± 1.0
E181P**[b]				43.11	52.07	+++	<20	nd	nd	nd	nd
E181Q	65.89	**74.41**	77.90	74.41	90.13	−	>50	508/5	386	280 ± 9.0	5.2 ± 0.2
E181R*[b]	65.42	**65.13**	63.18	65.13	77.41	+	<20	nd	nd	nd	nd
E181S*	53.33	**54.53**	55.26	54.53	64.12	++	30−50	500	383	23.3 ± 1.7	13.0 ± 0.1
E181T*	42.77	**51.22**	68.86	51.22	62.73	++	30−50	502	383	6.0 ± 0.6	17.3 ± 0.7
E181 V	68.88	61.43	**61.99**	61.99	73.34	+	30−50	489	383	5.7 ± 0.7	12.0 ± 0.5
E181W	60.61	**61.93**	74.25	61.93	68.97	+	30−50	502	376	111 ± 4.0	27.5 ± 1.0
E181Y*	65.62	**73.27**	73.50	73.27	91.61	+	30−50	501	392	9.8 ± 0.8	6.9 ± 0.1

[a] Spontaneous (ADRP) and artificial mutants of E181.[36,41,49] A single asterisk means that the energy of the folded state is about 0.5 RT units higher than that of the wild type, whereas double asterisks mean that the lowest energy basin is shifted in between two and three interstrand H-bonds. [b] Mutant did not bind 11-*cis*-retinal to form a stable pigment.[36] [c] Average interstrand H-bond probability (HB1−4$_{avg}$) derived from simulation on the mutated side chain rotamer from the D&K library.[29] [d] HB1−4$_{avg}$ index derived from simulation on the mutated side chain rotamer from the P&R library.[30] [e] HB1−4$_{avg}$ index derived from simulation on the mutated side chain rotamer from the Sut library.[31] [f] Selected HB1−4$_{avg}$ index. [g] Fraction of native-like structures, i.e. those characterized by a C_α-rmsd ≤2 Å from the native structure. [h] Predicted misfolding effect, based upon REMD simulations. In detail, the symbols "+++", "++", "+", and "−" stand, respectively, for misfolding, moderately misfolding, low misfolding, and non-misfolding. [i] Level of expression compared to the wild type; "nd" stands for not determined.[36] [j] Absorbance wavelength in the dark; "nd" stands for not determined.[36] [k] Absorbance wavelength upon illumination; "nd" stands for not determined.[36] [l] Rates of reaction with hydroxylamine; "nd" stands for not determined.[36] [m] MII decay rates; "nd" stands for not determined.[36] [n] In bold is the representative trajectory of each mutant that was selected to produce HB1−4$_{avg}$ values closest to the average value from the three independent simulations.

package.[12] The computational protocol was first probed and optimized on the protein G C-terminal amino acid peptide, which is known to fold into a stable 4:4 type β-hairpin and has been widely used as a model system to test REMD-based folding protocols.[13−17]

This study describes the strategy employed to individuate and differentiate EL2 rhodopsin mutations that would affect the intrinsic stability of the native β-hairpin from mutations expected to variably impair native contacts between the loop and the surrounding receptor domains.

2. Computational Details

2.1. Structural Analysis of EL2 and Its Mutation Sites. EL2 of rhodopsin (i.e., the S176-T198 sequence) folds into a highly stable twisted β-hairpin (i.e., the ^{177}RYIPEGMQCSCGID^{190} sequence) that makes extensive contacts with the other extracellular domains, on one side, and with the retinal chromophore, on the other one (Figure 1). In particular, it forms a four-stranded β-sheet with the N-terminal tail of the receptor protein, thus making a plug that shields the chromophore from the extracellular (intradiscal) solvent.[6] A cysteine residue from this loop, C187$^{(11)}$, is engaged in a disulfide bridge with C110 on the N-terminal end of H3, thus contributing to the stability of rhodopsin. The turn of the EL2 β-hairpin corresponds to the ^{182}GMQC^{185} amino acid stretch, whereas the N-terminal and C-terminal strands are, respectively, made by the ^{178}YIPE^{181} and ^{186}SCGI^{189} amino acids stretches. The EL2 amino acids directly involved in interactions with retinal belong to the C-terminal strand and include only S186$^{(10)}$, G188$^{(12)}$, and I189$^{(13)}$. In folded rhodopsin, R177$^{(1)}$ and D190$^{(14)}$, the first and last amino acids in the β-hairpin, respectively, are involved in a salt bridge interaction expected to contribute to the stability of the loop.[18]

The 40 EL2 mutations screened in this study comprise 15 ADRP-linked and 25 artificial mutations (Tables 1 and 2). The ADRP-linked mutations concern nonconservative mutations of Y178$^{(2)}$ in cysteine and asparagine, P180$^{(4)}$ in alanine, E181$^{(5)}$ in lysine, G182$^{(6)}$ in serine, Q184$^{(8)}$ in proline, S186$^{(10)}$ in proline and tryptophan, C187$^{(11)}$ in tyrosine, G188$^{(12)}$ in glutamate and arginine, and D190$^{(14)}$ in glycine, asparagine, and tyrosine (Tables 1 and 2), which represent the majority of the amino acids of the EL2 β-hairpin. In contrast, the 25 artificial mutations include cysteine, lysine, and glutamine substitutions for R177$^{(1)}$, serine substitution for C185$^{(9)}$, and 18 different amino acid substitutions for E181$^{(5)}$ (i.e., all the possible natural amino acid substitutions except for the ADRP-linked lysine substitution). The E181$^{(5)}$ amino acid residue, which points toward the center of the retinal polyene chain, is, indeed, the most investigated residue in the loop. Photochemistry studies indicated the involvement of such amino acid in the counterion switch during the photoactivation of rhodopsin.[9] In fact, according to earlier studies, such a switch was suggested to occur by the proton transfer from E181$^{(5)}$, protonated in the dark state, to E113 (in H3), through an evolving H-bond (HB) network formed primarily with residues of EL2.[9] In fact, in the crystal structures of dark rhodopsin,[19] as well as of the BATHO and LUMI

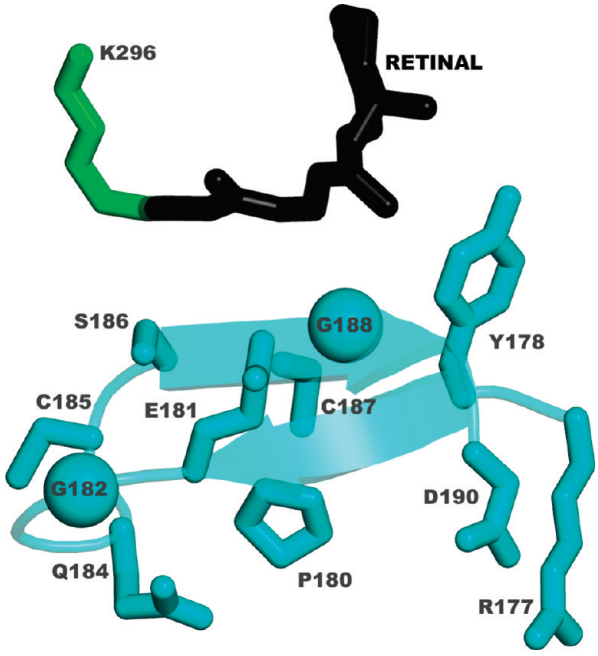


Figure 1. Drawing of the retinal and the EL2 β-hairpin extracted from the 1U19 crystal structure of dark rhodopsin. The retinal chromophore in its 11-cis conformation linked to K296 as a Schiff base is drawn as sticks. The retinal is black, the covalently bound K296 is green, and the EL2 cyan. The side chains of the β-hairpin amino acids targeted by in silico mutagenesis are represented as sticks. In this respect, the positions of G182 and G188 are indicated by two spheres positioned on the C_α-atoms. Drawings were done by means of PyMOL 0.99 (http://pymol.sourceforge.net/).

photointermediates,[20,21] this anionic amino acid is hydrogen bonded to both Y192 (in EL2) and Y268 (in H6) and is also involved in a water-mediated HB with both S186$^{(10)}$ and E113 (in H3) (Figure 1). More recent studies suggest that a change of the PSB counterion from the dark state to MI would not necessarily require a proton transfer.[22]

Collectively, the structural analysis of the EL2 β-hairpin in the context of folded rhodopsin in its dark state shows that seven of the spontaneous or artificial mutation sites considered in this study, R177$^{(1)}$, P180$^{(4)}$, G182$^{(6)}$, Q184$^{(8)}$, C185$^{(9)}$, S186$^{(10)}$, and C187$^{(11)}$, belong to the extracellular−TM interface of the retinal binding pocket or to the interface between EL2 and the surrounding domains.

2.2. REMD Simulation Setup: The GB1 Model System. The computational protocol was set on the β-hairpin from the C terminus (i.e., G44-E56 segment) of Streptococcal protein G (PDB code 2gb1, i.e., ACE-^{41}GEWTYDDATK-TFTVTE^{56}-NME, herein named as GB1), patched with acethyl and methyl-amino groups at the N- and C-terminals. The finally selected setup consists of REMD simulations by means of the CHARMM force field (in all-atom mode),[23] using the FACTS implicit solvent model.[12] REMD simulations were carried out using an in-house developed program that calls internally the CHARMM package.[24,25] The algorithm is implemented in C language with the use of the message-passage-interface (MPI) libraries.

A total of 20 replicas were simulated by Langevin dynamics with a friction coefficient of 5.0 ps^{-1} and temperature values spanning the interval from 270 to 690 K (270, 283, 298, 313, 328, 345, 363, 381, 400, 421, 442, 465, 488, 513, 539, 566, 594, 625, 656, and 690 K). Each replica was thermalized at its respective temperature for 20 ps with a time step of 2 fs. REMD sampling was carried out for a total of 5 ns using a time step of 2 fs. Transitions between adjacent temperatures were attempted every 0.2 ps and protein configurations were saved every 0.4 ps, giving a total of 0.25 million configurations. This setup was selected following a significant number of trials. The latter included tests of (a) an alternative implicit solvation model, i.e., GBSW;[26] (b) a number of different temperature sets; (c) longer REMD sampling, i.e., 10 and 20 ns; (d) more frequent coordinate saving, i.e., every 0.2, 0.1, and 0.05 ps; and (e) different friction coefficients, i.e., 2.0, 1.0, and 0.5 ps^{-1}.

Both the charged and neutral form of the peptide, i.e. protonated at E42, D46, and E56, were considered, leading to 256 and 259 atoms per replica, respectively. The results of simulations on the neutral and charged peptide are comparable, although the ones on the neutral forms are more realistic. Therefore, the results herein presented refer to simulations on the neutralized peptide.

The selected REMD setup is the best compromise between speed of simulation and consistency with NMR determinations as well as explicit water simulations.[14] In this respect, prolonging REMD simulations from 5 to 20 ns did not change significantly the distribution of conformational states. In fact, the average fraction of native contacts at 313 K following 5, 10, and 20 ns REMD simulations are, respectively, 67.2, 68.1, and 68.0. Thus, selection privileged the shortest simulation time. Similarly, denser sampling did not change the outcome. In fact, coordinate saving every 0.4, 0.2, 0.1, and 0.05 ps led, respectively, to 67.2, 68.5, 68.5, and 68.5 fractions of native contacts. On the same line, friction coefficients did not affect the outcome of simulations. Indeed, simulations at friction coefficients equal to 0.1, 0.5, 1.0, 2.0, and 5.0 ps^{-1} resulted, respectively, in 67.8, 63.7, 65.1, 68.6, and 67.2 fractions of native contacts. Final selection concerned a friction coefficient of 5 ps^{-1}, as it was closer than the others to the viscosity of water (10 ps^{-1} < γ < 100 ps^{-1}) although still in the low viscosity regime.[27]

As native contacts stabilizing the GB1 peptide during trajectory analysis we took the following 26 pairs, by considering a distance cutoff of 7.6 Å between the side chain geometric centers, according to previous reports:[28] 4−5, 5−6, 8−9, 15−16, 1−3, 2−4, 3−5, 4−6, 5−7, 6−8, 9−11, 11−13, 12−14, 13−15, 6−9, 7−10, 6−11, 5−11, 4−11, 5−12, 5−13, 3−12, 4−13, 2−13, 3−14 and 2−15.

2.3. REMD Simulation Setup: The Rhodopsin EL2. The computational protocol set on the GB1 peptide was then extended to the EL2 β-hairpin of rhodopsin, a 14 amino acid peptide from R177$^{(1)}$ to D190$^{(14)}$ extracted from the crystal structure of dark rhodopsin (PDB code 1U19)[19] and patched with acethyl and methyl-amino groups at the N- and C-terminals, respectively. Comparative REMD simulations were carried out on the wild type and on 40 mutated forms (Tables 1 and 2). The wild type was simulated both in the charged and neutral states, i.e., carrying E181$^{(5)}$ in its protonated form. Controversial data support both protonation

and deprotonation of E181$^{(5)}$ in the dark state, whereas the deprotonated (charged) form would characterize the active states starting from MI.[9,22] The results shown in this study refer to simulations with protonated E181$^{(5)}$.

For wild type EL2, the total number of atoms per replica was 218. For each replacing amino acid, three different input rotamers were subjected to REMD simulations. These starting conformations were assigned according to the Dunbrack and Karplus (D&K),[29] Ponder and Richards (P&R),[30] and Sutcliffe (Sut)[31] rotamer libraries.

The transition acceptance ratio was around 45%.

2.4. REMD Simulation Analyses. For both GB1 and EL2 β-hairpins, the energy landscape or potential of mean force (PMF) was calculated from the normalized population densities as previously described[15]

$$\mathrm{PMF} = -\log P(X_1, X_2)$$

where $P(X_1, X_2)$ is the normalized probability as a function of X_1 and X_2, and X_1 and X_2 are parameter sets describing the peptide conformations. In this study, such parameters are the native β-sheet hydrogen bonds (H-bonds) and the geometric radius of gyration of the hydrophobic core (RgCore). In detail, for the 2GB1, the seven native β-sheet hydrogen bonds are E42:N−H/T55:O, T55:N−H/E42:O, T44:N−H/T43:O, T43:N−H/T44:O, D46:N−H/T51:O, T51:N−H/D46:O, and K50:N−H/D47:O. A hydrogen bond was counted if the distance between the O and the N atoms was less than 3.5 Å, and the angle formed by the three atoms (N, H, and O) was larger than 150°. Furthermore, the RgCore was computed on the side chain atoms of the four hydrophobic residues W43, Y45, F52, and V54, to allow for comparisons with the results of previous computational studies.[13,14,16] For rhodopsin EL2, the five native β-sheet hydrogen bonds are D190:N−H/R177:O, I179:N−H/G188:O, G188:N−H/I179:O, E181:N−H/S186:O, and Q184:N−H/E181:O. Rg calculations were limited to the hydrophobic core amino acids rather than to the whole peptide. The RgCore was, thus, computed on the side chain atoms of the following four residues: Y178, P180, C187, and I189.

Cluster analysis of the REMD trajectories was based on the QT clustering algorithm[32] implemented in the Wordom software.[33] In this case study, the algorithm first calculated the C_α-atom root mean square deviation (C_α-rmsd) for each superimposed pair of frames and then it computes the number of neighbors for each frame by using a threshold C_α-rmsd. The frame with the highest number of neighbors is considered as the center of the first cluster. All the neighbors of this configuration are removed from the ensemble of configurations to be counted only once. The center of the second cluster is then determined in the same way as the first cluster, and this procedure is repeated until no more clusters can be found.

3. Results

3.1. The C-Terminal β-Hairpin of Protein G as a Model System. The computational protocol was set on the C terminal β-hairpin of Streptococcal protein G (i.e., the G44-E56 segment herein named as GB1), extensively used as a model system to probe in silico β-hairpin folding protocols. REMD simulations in implicit solvent, spanning 20 temperatures from 270 to 690 K, produced a free energy contour map at 313 K characterized by a wide unique global energy minimum corresponding to the native-like state, i.e., with five β-sheet H-bonds and a RgCore around 5.92 Å (Figure 2). The latter index was computed on the side chain atoms of W43$^{(3)}$, Y45$^{(5)}$, F52$^{(12)}$, and V54$^{(14)}$. These results overlap significantly with those of explicit water simulations by Zhou and co-workers.[14] Along this line, the probability of finding all five interstrand HBs satisfied reaches the maximum value at temperatures between 270 and 313 K, whereas it drops at temperature values above 520 K (Figure 3A). In this respect, the inner interstrand H-bonds (HB3, HB4, and HB5, Figure 3A) are more persistent than the most external ones (HB1 and HB2, Figure 3A). The HBs that involve the turn (HB6 and HB7), in particular HB7, are the less persistent ones. In summary, at 313 K, the rank order of each H-bond probability (expressed as percentages with respect to the total number of frames) is HB4 > HB5 > HB3 > HB2 > HB6 > HB1 > HB7 (the values being 87.26, 85.62, 82.03, 65.73, 62.15, 47.21, and 14.3, respectively). Interestingly, this rank order is overlapping with that from explicit water simulations.[14] The average HB probability at 313 K from our simulations is 63.47%, slightly higher but characterized by a higher decrease rate compared to that from explicit water simulations.[14] In fact, according to our data, such an index decreases to around 0 at 690 K (i.e., 0.27), whereas it never drops to zero following explicit water simulations.[14] In line with the trend of the HB probability, the β-hairpin population, as accounted for by the fraction of native contacts (see the experimental procedures for its definition), is 67.87% at 270 K, 67.52% at 298 K, and 67.22% at 313 K (Figure 3B). These data are similar to those from explicit water simulations with the OPLSAA all-atom force field, which found populations of native contacts of about 71% at 270 K and 66% at 310 K.[14] Thus, near the biological temperature, our simulations, similar to simulations in explicit water,[14] found a β-hairpin population of the GB1 peptide in reasonable agreement with in vitro experiments. Indeed, NMR data found a population of about 80% at 270 K, 50% at 300 K, and 40% at 310 K (as shown above, the fraction of native contacts from computational experiments does not change significantly in the 270−313 K range). The highest discrepancies with NMR data concern temperatures higher than 313 K. Indeed, the temperature-dependent decrease rate of the fraction of native contacts from our simulations is quite slower than that from NMR determinations, though faster than that from explicit water simulations.[14] In fact, according to our computations, the fraction of native contacts decreases significantly after 420 K, reaching the lowest value of 26.89% at 690 K (Figure 3B). In contrast, according to NMR data, the population of native β-hairpin is already 0 around 360 K, whereas the fraction of native contacts from explicit water simulations is still above 35% at 690 K.[14] In line with the trend of the fraction of native contacts, the fraction of native-like structures, i.e., those characterized by a C_α-rmsd $\leq$2 Å from the native structure, is 56.4% at 313 K, whereas it

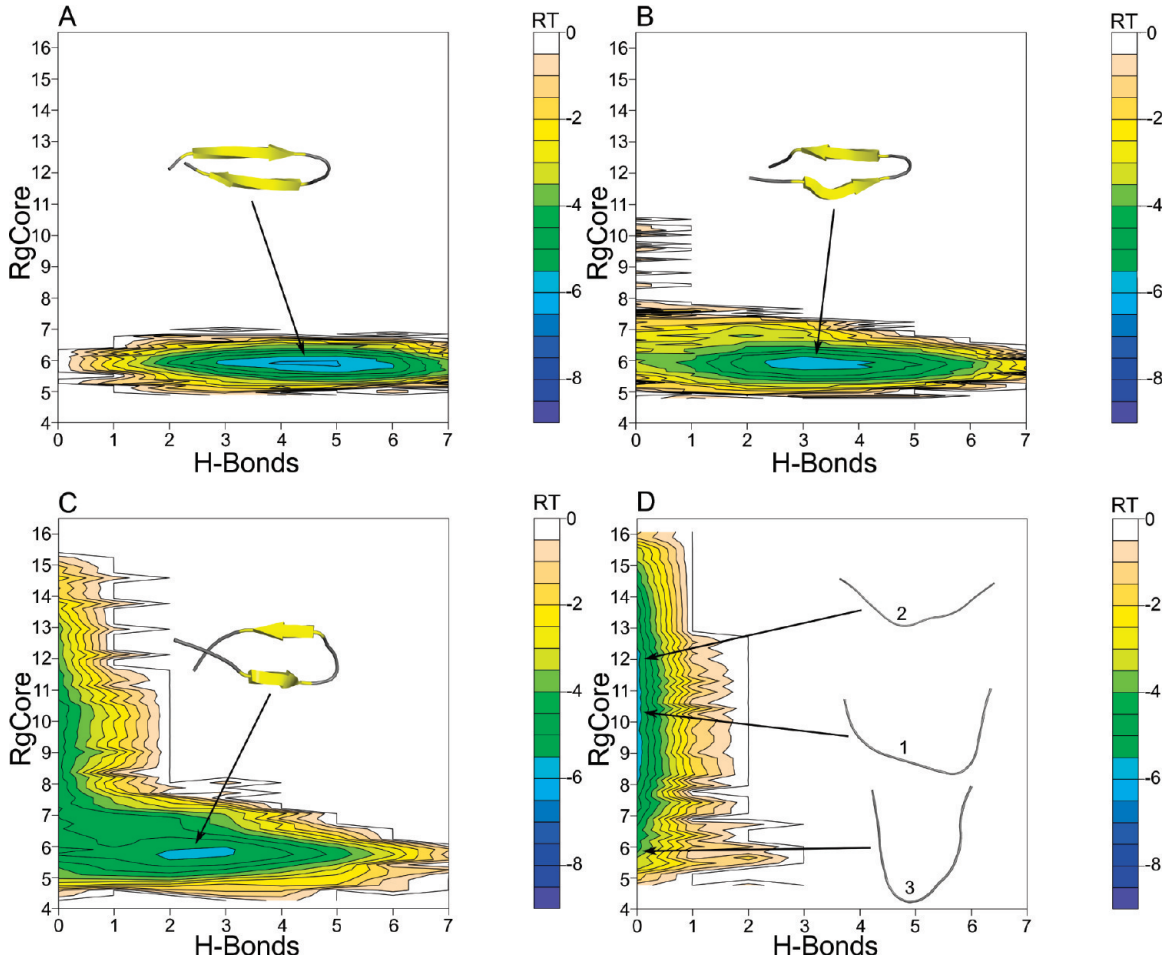


Figure 2. Free energy contour maps at different temperatures of the GB1 peptide folding versus the number of native β-sheet hydrogen bonds and the hydrophobic core radius of gyration (RgCore). The seven native β-sheet hydrogen bonds are E42:N−H/T55:O, T55:N−H/E42:O, T44:N−H/T43:O, T43:N−H/T44:O, D46:N−H/T51:O, T51:N−H/D46:O, and K50:N−H/D47:O. A hydrogen bond was counted if the distance between the O and the N atoms was less than 3.5 Å and the angle formed by the three atoms (N, H, and O) was larger than 150°. The geometric radius of gyration (excluding any mass weighing) of the hydrophobic core was computed on the side chain atoms of the four hydrophobic residues W43, Y45, F52, and V54. The contours are spaced at intervals of 0.5 RT. Cartoon representations of the cluster centers from the lowest energy basins (indicated by arrows) are also shown. In detail, (A) the free energy landscape at 313 K is characterized by one broad energy basin whose representative structure holds five H-bonds and a RgCore = 5.58 Å, (B) the free energy landscape at 400 K is characterized by one broad energy basin whose representative structure holds three H-bonds and a RgCore = 5.73 Å, (C) the free energy landscape at 465 K is characterized by one broad energy basin whose representative structure holds three H-bonds and a RgCore = 5.43 Å, and (D) the free energy landscape at 690 K is characterized by a number of energy basins corresponding to zero H-bonds. The representative structures extracted from these basins hold the following RgCore values: (1) 10.37 Å, (2) 12.04 Å, and (3) 5.96 Å.

progressively decreases with the increase in temperature (Figure 3C).

Collectively, as already discussed in previous studies,[14] high-temperature discrepancies between in vitro and in silico data are shared by computational experiments using different force fields (i.e., CHARMM, OPLSAA, and AMBER) and different treatments of the solvent and may depend on many factors, including force field parameters, water models (for explicit water simulations), the employment of high pressures at high temperatures in constant volume simulations, or the lack of water-density-dependent parameters in implicit solvent simulations.

3.2. REMD Simulations on Rhodopsin EL2. The computational protocol optimized on the GB1 peptide was employed to investigate the effects of 40 point mutations on the structural stability of rhodopsin EL2 (the S176-T198 sequence) that holds a highly stable twisted β-hairpin (the ^{177}RYIPEGMQCSCGID^{190} sequence). The experimental set consists of 15 spontaneously occurring mutants associated with ADRP and 25 artificial mutants (Tables 1 and 2). Comparative REMD simulations were, hence, carried out on the wild type and the 40 mutant forms.

Similar to the GB1 peptide, the free energy contour map of wild type EL2 at 313 K is characterized by a wide unique global energy minimum corresponding to the native-like state, i.e. with four interstrand H-bonds and a RgCore equal to 5.14 Å (Figure 4). This index was computed on the side chains of Y178$^{(2)}$, P180$^{(4)}$, C187$^{(11)}$, and I189$^{(13)}$ from the center (i.e., the most representative frame) of the lowest energy cluster. The lowest energy basin at 313 K progres-

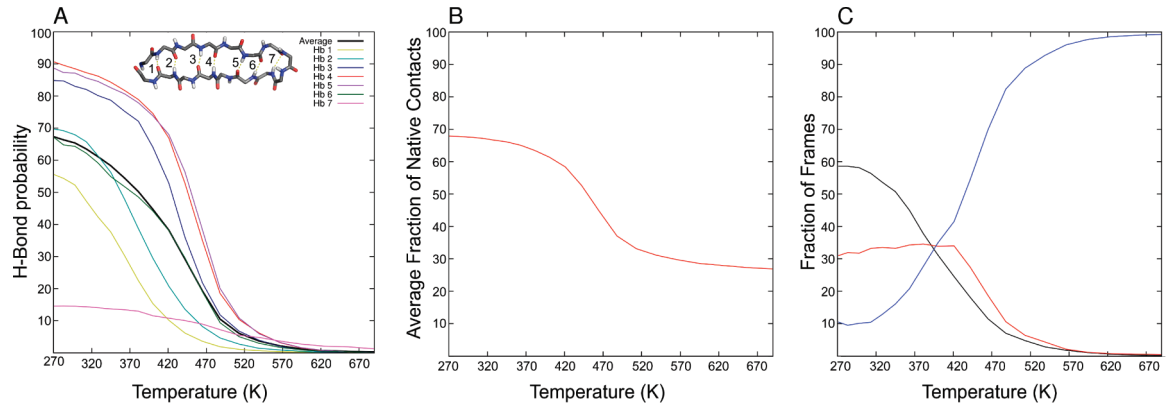


Figure 3. Temperature dependence of structural indices. (A) The temperature dependence of the probability of forming the seven native β-sheet H-bonds (E42:N−H/T55:O (HB1), T55:N−H/E42:O (HB2), T44:N−H/T43:O (HB3), T43:N−H/T44:O (HB4), D46:N−H/T51:O (HB5), T51:N−H/D46:O (HB6), and K50:N−H/D47:O (HB7)) is shown. The curves relative to HB1−7 are colored in yellow, cyan, blue, red, violet, green, and magenta, respectively. The thick black line represents the average probability over all native hydrogen bonds. A stick representation of the GB1 β-hairpin with the seven native β-sheet H-bonds is shown as well. (B) The average fraction of a set of 26 native contacts as a function of the temperature is shown. The set of contacts includes the following pairs: T44:Y45, Y45:D46, A48:T49, T55:E56, G41:W43, E42:T44, W43:Y45, T44:D46, Y45:D47, D46:A48, T49:T51, T51:T53, F52:V54, T53:T55, D46:T49, D47:K50, D46:T51, Y45:T51, T44:T51, Y45:F52, Y45:T53, W43:F52, T44:T53, E42:T53, W43:V54, and E42:T55. A contact was counted if the distance between the side chain geometrical center of the two residues in each pair was less than 7.5 Å. (C) The temperature dependence of the C_α-rmsd from the starting structure is shown. Curves corresponding to C_α-rmsd thresholds $\leq$2 Å, >2 Å and $\leq$3 Å, and $\geq$3 Å are colored in black, red and blue, respectively.

sively moves toward zero β-sheet HBs and higher RgCore values with the increases in temperature (Figure 4).

The distorted wild type EL2 β-hairpin is characterized essentially by four interstrand HBs (HB1, HB2, HB3, and HB4) and one strand-turn HB (HB5) (Figure 5A). Similar to the GB1 β-hairpin, the probability of finding all four interstrand HBs satisfied reaches the maximum value at temperatures between 270 and 320 K, whereas it drops at temperatures above 520 K. Analogously to the GB1 β-hairpin, the inner interstrand HBs (HB2, HB3, and HB4, Figure 5A) are more persistent than the most external one (HB1, Figure 5A). The HB5 strand-turn H-bond shows the lowest persistency. In summary, at 313 K, the rank order of each HB probability (expressed as percentages with respect to the total number of frames) is HB2 > HB3 $\approx$ HB4 > HB1 > HB5 (the values being 82.78, 77.39, 77.15, 72.45, and 3.50, respectively). The average interstrand HB probability (i.e $HB1-4_{avg}$, that considers only HB1, -2, -3, and -4) at 313 K is 77.44% (Table 1). Consistently, the fraction of native-like structures, i.e., those characterized by a C_α-rmsd $\leq$2 Å from the native structure, is 88.66% at 313 K, whereas it progressively decreases with increases in temperature (Figure 5B).

Possible disturbing effects of the 40 EL2 point mutations on the structural features of wild type β-hairpin were, hence, evaluated by comparing the free energy landscapes as well as the probability of interstrand HB formation and the fraction of native-like structures of the mutant trajectories with those of the wild type.

As for the in vitro behavior of the considered mutants, 13 of them are impaired in folding/expression and/or retinal binding (Tables 1 and 2; marked by red color in Figure 6). Furthermore, 17 mutations exert a milder impairing effect on folding/expression (Tables 1 and 2; marked by the yellow color in Figure 6), whereas seven mutations do not significantly change the expression or chromophore binding compared to the wild type (Tables 1 and 2; marked by the green color in Figure 6). Finally, in three cases, the effect of mutations on the structural stability of rhodopsin is unknown (Tables 1 and 2; marked by the gray color in Figure 6).

For the mutants characterized by rotamers on the replacing amino acid side chains, possible different rotameric states, according to the D&K,[29] P&R,[30] and Sut[31] libraries, were probed as simulation inputs. We then selected as a representative trajectory of each mutant the one that produced $HB1-4_{avg}$ values closest to the average value from the three independent simulations (bold numbers in Tables 1 and 2). A strong agreement was generally achieved between the $HB1-4_{avg}$ values from at least two of the three independent simulations.

In silico screening showed a spectrum of mutation-induced changes in the $HB1-4_{avg}$ index and in the fraction of native-like structures characterizing the wild type at 313 K (i.e., 77.44% and 88.66%, respectively; Figure 6). Given the very high correlation between the fraction of native-like structures and $HB1-4_{avg}$ index ($r = 0.977$), we decided to employ the latter as a structural hallmark of mutation effects. Thus, as for $HB1-4_{avg}$, 14 mutants hold wild type-like values (i.e., higher than 70%), 11 mutants hold values between 60% and 70%, seven mutants hold values between 50% and 60%, and eight mutants hold values below 50% (Tables 1 and 2, Figure 6). The latter include the R177C, P180A, E181G, and E181P mutants as well as the nonconservative mutations of D190[14] (i.e., A, C, G, and Y substitutions; Tables 1 and 2, Figure 6).

REMD simulations showed also mutation-induced changes in the free energy landscape characterizing the

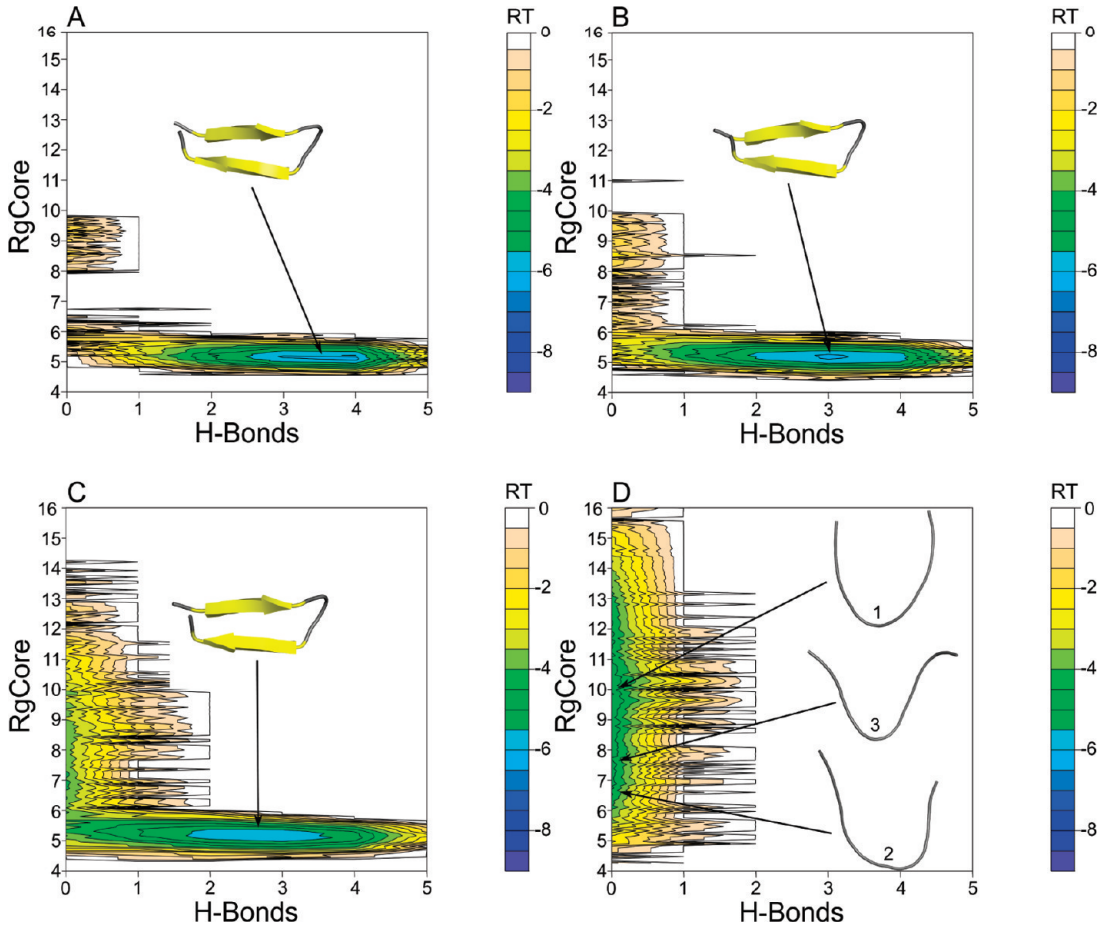


Figure 4. Free energy contour maps at various temperatures of rhodopsin EL2 (177−190 sequence) versus the number of native β-sheet hydrogen bonds and the core radius of gyration (RgCore). The five native β-sheet hydrogen bonds are D190:N−H/R177:O, I179:N−H/G188:O, G188:N−H/I179:O, E181:N−H/S186:O, and Q184:N−H/E181:O. The radius of gyration was computed on the side chain atoms of the following four residues: Y178, P180, C187 and I189. Contours are spaced at intervals of 0.5 RT. Cartoon representations of the cluster centers from the lowest energy basins (indicated by arrows) are also shown. In detail, (A) the free energy landscape at 313 K is characterized by one broad energy basin whose representative structure holds four H-bonds and a RgCore = 5.14 Å, (B) the free energy landscape at 400 K is characterized by one broad energy basin whose representative structure holds two H-bonds and a RgCore = 5.33 Å, (C) the free energy landscape at 465 K is characterized by one broad energy basin whose representative structure holds three H-bonds and a RgCore = 5.4 Å, and (D) the free energy landscape at 690 K is characterized by a number of energy basins corresponding to zero H-bonds. The representative structures extracted from these basins hold the following RgCore values: (1) 9.92 Å, (2) 6.57 Å, and (3) 7.88 Å.

wild type form at 313 K (Figures 6−8, and S1−S7, Supporting Information). Differences between wild type and mutant forms consist of the appearance of alternative higher energy basins at the expense of the native-like state that remains the most populated one. As a consequence, for 17 of the 40 mutants (i.e., marked by an asterisk in Tables 1 and 2) the energy of the folded state is about 0.5 RT units higher than that of the wild type. Furthermore, for the E181P and G188R mutants, the lowest energy basin is shifted in between two and three interstrand HBs, indicative of a misfolding effect of such mutations (marked by two asterisks in Tables 1 and 2). In line with these data, the free energy landscapes of the mutants holding a HB1−4_{avg} index lower than 50% are characterized by the appearance of higher energy basins at zero β-sheet HBs and RgCore values above 6 Å (Figures 7, 8, and S1−S7, Supporting Information). These basins generally correspond to α-helical structures (Figures 7, 8, and S1−S7, Supporting Information).

Collectively, by considering all together the HB1−4_{avg} index and the shapes of the free energy landscapes, we could classify EL2 mutants as misfolding (marked as +++ in Tables 1 and 2), moderately misfolding (marked as ++ in Tables 1 and 2), low misfolding (marked as + in Tables 1 and 2), and non-misfolding (marked as − in Tables 1 and 2). In detail, (a) misfolding mutants are characterized by a HB1−4_{avg} index below 50% that is generally accompanied by a shift in energy or position of the lowest energy basin (i.e., marked by one or two asterisks, respectively, in Tables 1 and 2); (b) moderately misfolding mutants are characterized by a HB1−4_{avg} index between 50% and 60% associated with a shift in energy or position of the lowest energy basin; (c) low misfolding mutants are characterized by a HB1−4_{avg} index between 60% and 70%,

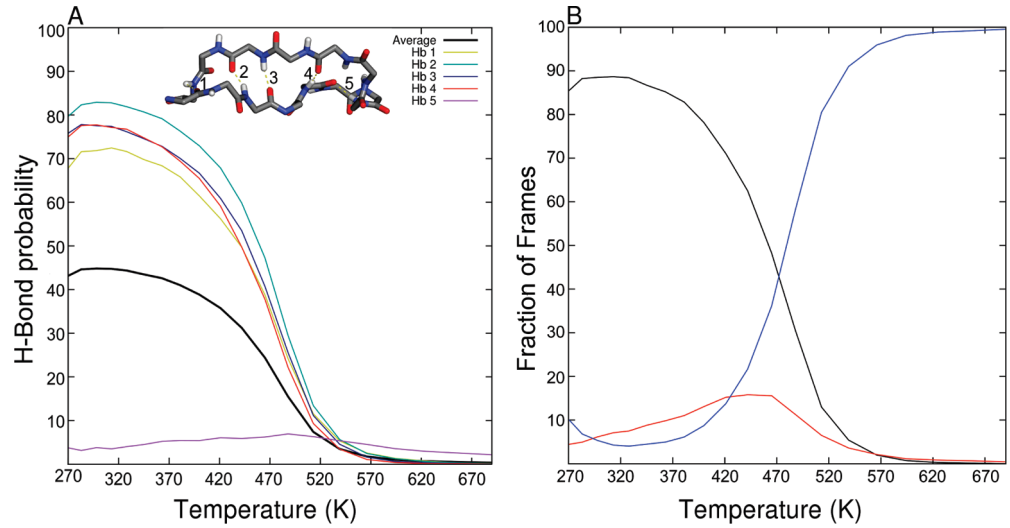


Figure 5. Temperature dependence of the probability of forming individual native β-sheet hydrogen bonds and of the C_α-rmsd for the wild type EL2. (A) The temperature dependence of the probability of forming the five native β-sheet H-bonds (D190:N−H/R177:O (HB1), I179:N−H/G188:O (HB2), G188:N−H/I179:O (HB3), E181:N−H/S186:O (HB4), and Q184:N−H/E181:O (HB5)) is shown. The curves relative to HB1−5 are colored in yellow, cyan, blue, red, and purple, respectively. The thick black line represents the average probability over all native hydrogen bonds. A stick representation of EL2 together with the five native β-sheet hydrogen bonds is shown as well. (B) The temperature dependence of the C_α-rmsd from the starting structure is shown. Curves corresponding to C_α-rmsd thresholds ≤2 Å, >2 Å and <3 Å, and ≥3 Å are colored in black, red, and blue, respectively.

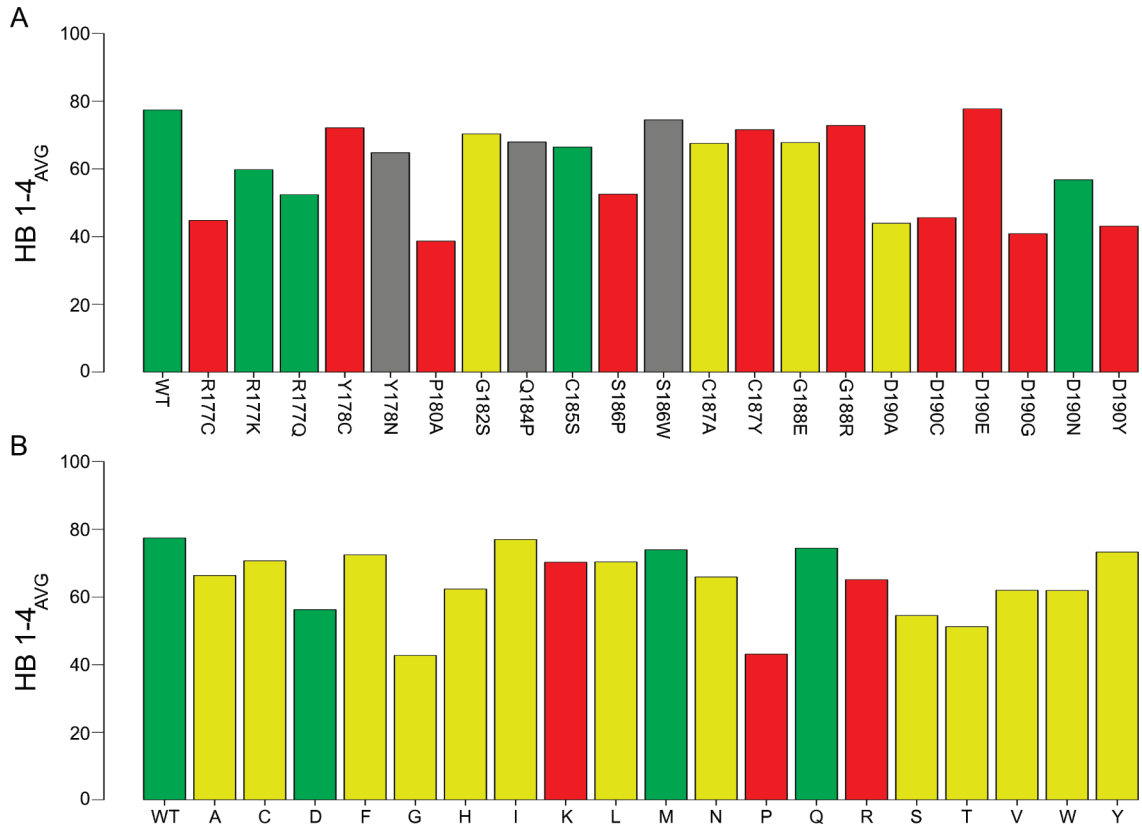


Figure 6. EL2 wild type and mutant average probabilities of interstrand H-bonds (HB1−4_{avg}). The histograms report the average probabilities of the interstrand H-bonds concerning the wild type and all the simulated mutations at EL2 sites other than E181 (A) and all mutations of E181 (B). Color codes refer to in vitro behavior. In detail, (a) red stands for impaired folding/expression and/or retinal binding, (b) yellow indicates a moderate impairing effect on folding/expression and/or chromophore binding, (c) green stands for wild type-like behavior, and (d) gray stands for unknown biochemical effect.

independent of the position and depth of the lowest energy basin, or by HB1−4_{avg} above 70% but associated with shifts in energy and/or position of the lowest energy basins; and, finally, (d) the wild type-like or non-misfolding mutants are characterized by a HB1−4_{avg} index above 70% and the lowest energy basin similar to that of the wild type.

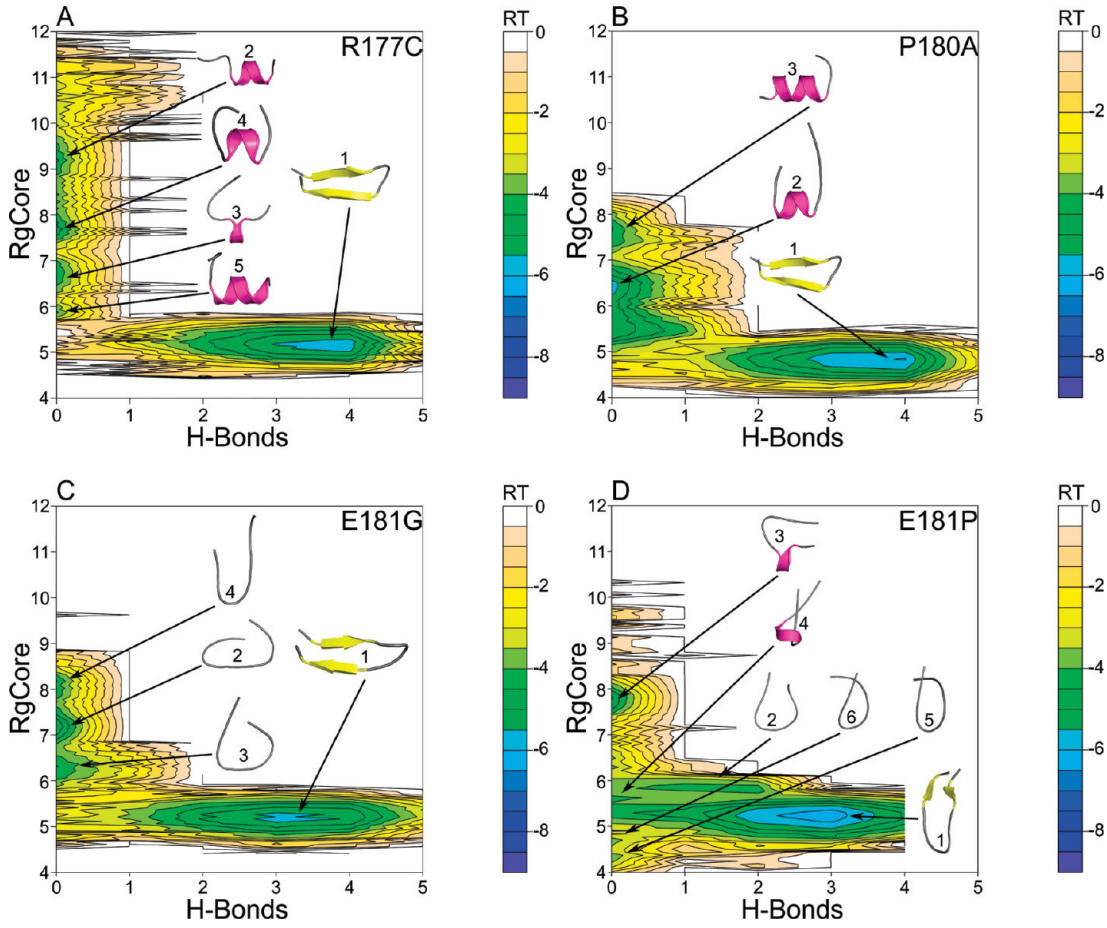


Figure 7. Free energy contour maps at 313 K relative to the R177C (A), P180A (B), E181G (C), and E181P (D) mutants, versus the number of native β-sheet hydrogen bonds and the core radius of gyration (RgCore). The general description of this legend is the same as that of Figure 4. The representative structures from each energy basin are shown as cartoons. In detail, (A) for the R177C mutant five structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a native-like β-hairpin, characterized by four β-sheet H-bonds and RgCore = 5.19 Å. The remaining four structures share zero β-sheet H-bonds and the following RgCore values: (2) 9.30 Å, (3) 6.68 Å, (4) 7.71 Å, and (5) 5.9 Å. (B) For the P180A mutant, three structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a native-like β-hairpin, characterized by four β-sheet H-bonds and RgCore = 4.93 Å. The remaining two structures share zero β-sheet H-bonds and the following RgCore values: (2) 6.46 Å, and (3) 7.59 Å. (C) For the E181G mutant, four structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a misfolded β-hairpin characterized by three β-sheet H-bonds and RgCore = 5.17 Å. The remaining three structures share zero β-sheet H-bonds and the following RgCore values: (2) 7.20 Å, (3) 6.36 Å, and (4) 8.13 Å. (D) For the E181P mutant, six structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a misfolded β-hairpin characterized by three β-sheet H-bonds and RgCore = 5.17 Å; structure 2 holds two β-sheet H-bonds and a RgCore = 5.95 Å. The remaining four structures share zero β-sheet H-bonds and the following RgCore values: (3) 7.75 Å, (4) 5.40 Å, (5) 4.33 Å, and (6) 4.72 Å.

4. Discussion

RP refers to a group of debilitating, hereditary disorders that cause severe visual impairment in as many as 1.5 million patients worldwide.[1−3] Many genes have been associated with RP, and it exhibits extreme heterogeneity in terms of its severity and mode of inheritance. Although 30 RP genes have been recently identified, there were immensely exciting developments in the study of the ADRP form of the disease.[1−3] Rhodopsin mutations account for >25% of ADRP, and ∼100 distinct mutations have been identified throughout the transcript. Mutations have been identified in all of the structural domains of the rhodopsin protein, and although attempts have been made to categorize mutants into six general classes of biochemical defects, many do not fit into predictable groups.

This study is part of a project aimed at structurally characterizing, through molecular simulations, the majority of pathogenic rhodopsin mutations. In this framework, the choice of the approach is dictated by the structural localization and the biochemical effect of a given mutation. The mutations considered in this study concentrate in the structured part of EL2. The crystal structures of dark rhodopsin[6] as well as of the BATHO and LUMI photointermediates[20,21] show that this rhodopsin portion folds into a twisted β-hairpin, whose C-terminal strand forms the "floor" of the chromophore binding pocket (if the receptor is seen in a direction parallel to the membrane surface with the intracellular side on top). Although structure determinations of the isolated EL2 fragment are lacking, very recent solid-state NMR determinations support the structural stabil-

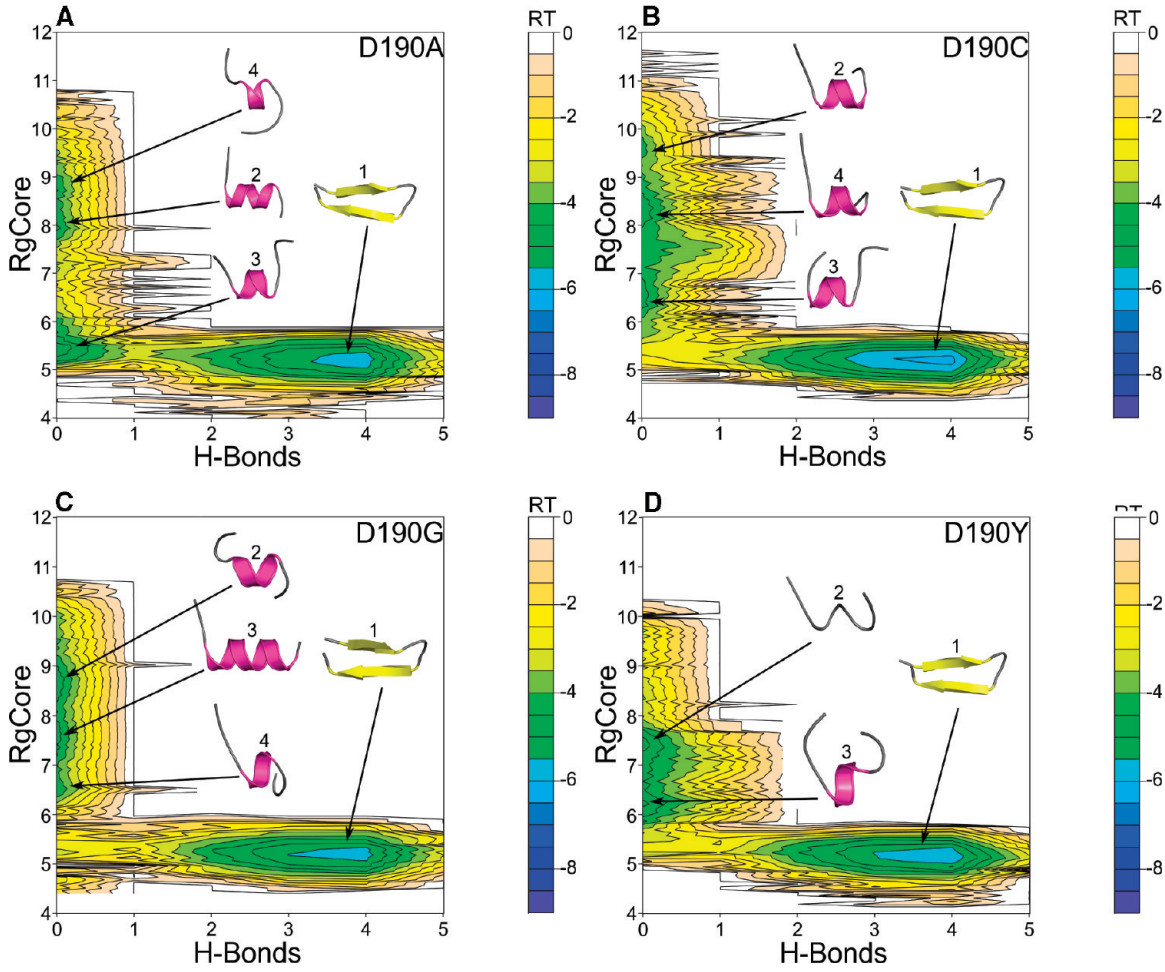


Figure 8. Free energy contour maps at 313 K relative to the D190A (A), D190C (B), D190G (C), and D190Y (D) mutants, versus the number of native β-sheet hydrogen bonds and the core radius of gyration (RgCore). The general description of this legend is the same as that of Figure 4. The representative structures from each energy basin are shown as cartoons. In detail, (A) for the D190A mutant, four structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a native-like β-hairpin, characterized by four β-sheet H-bonds and RgCore = 5.19 Å. The remaining three structures share zero β-sheet H-bonds and the following RgCore values: (2) 8.19 Å, (3) 5.29 Å, and (4) 8.95 Å. (B) For the D190C mutant, four structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a misfolded β-hairpin, characterized by three β-sheet H-bonds and RgCore = 5.44 Å. The remaining three structures share zero β-sheet H-bonds and the following RgCore values: (2) 9.68 Å, (3) 6.38 Å, and (4) 8.21 Å. (C) For the D190G mutant four structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a native-like β-hairpin characterized by four β-sheet H-bonds and RgCore = 5.18 Å. The remaining three structures share zero β-sheet H-bonds and the following RgCore values: (2) 8.80 Å, (3) 7.89 Å, and (4) 6.72 Å. (D) For the D190Y mutant three structures have been extracted: the structure extracted from the lowest energy basin (1) corresponds to a native-like β-hairpin characterized by four β-sheet H-bonds and RgCore = 5.28 Å. The remaining two structures share zero β-sheet H-bonds and the following RgCore values: (2) 7.32 Å and (3) 6.10 Å.

ity of the loop, by showing that during rhodopsin activation EL2 changes position but not conformation.[10] Furthermore, the web server BETAHAIRPRED (http://triton.rmn.iqfr.csic.es/software/behairpredv1.0/behairpred.htm) predicts the ^{177}RYIPEGMQCSCGID^{190} sequence as prone to form a β-hairpin, suggesting that such a conformational propensity is intrinsic to the primary sequence of the peptide, independent of the environment. These data support the strategy to investigate the potential effects of 40 mutations on the structural stability of the native β-hairpin taken off the protein context. The aim was pursued by comparative parallel REMD simulations using the FACTS implicit solvent model.[12] In line with reduction of the molecular system, the choice of an implicit over an explicit solvent model was dictated by the need to implement a fast in silico screening approach. Indeed, in this study, computational screening consisted of 101 independent REMD simulations, starting from different input structures, which included two different prototropic forms of E181 for the wild type form and three different rotameric states for all mutated side chains, excluding alanine, glycine, and proline.

The computational protocol was set on the GB1 peptide, a model system for computational experiments on β-hairpin folding. In agreement with previous explicit water simulations,[14] the most populated state of GB1 at 313 K resulted in the native β-hairpin. Moreover, the average fraction of native contacts at 313 K is quite overlapping with the results of explicit water simulations and does not vary in the

270−313 K temperature range (i.e., 67.87% and 67.22%, respectively). Near the biological temperature range, the fraction of native contacts from our experiments is in reasonable agreement with those found by NMR determinations, whereas the highest discrepancies occur at temperatures higher than 313 K. These findings are in line with previous results of computational experiments on the same model system.[14] It is worth noting, however, that temperatures higher than 313 K go beyond the computational screening of rhodopsin mutations, the main goal of this study.

The computational protocol optimized for the GB1 peptide was thus extended to the R177-D190 structured rhodopsin fragment, toward the building of a fast in silico screening tool for structure-based reclassification of selected ADRP mutations. The latter, indeed, fall essentially in class II according to poorly defined biochemical behaviors, characterized by more or less pronounced impairment in receptor folding/expression and/or retinal binding (Tables 1 and 2).

Similar to the GB1 peptide, the native state of wild type EL2 prevailed at 313 K. The probability of native interstrands HBs, i.e., the $HB1-4_{avg}$ index, and the shape of the free energy landscape were, thus, employed as primary structural hallmarks of the native state in the comparative mutational analysis. These hallmarks were variably perturbed following REMD simulations of the 40 naturally occurring and artificial EL2 mutants considered in this study, resulting in a structural classification of mutational effects in misfolding, moderately misfolding, low misfolding, and non-misfolding (Tables 1 and 2). According to this classification and consistent with in vitro evidence of a detrimental effect on folding/expression and/or chromophore binding, misfolding mutations include nonconservative substitutions of the first and last amino acids in the loop, i.e. R177C and D190G, -A, -C, and -Y, as well as the P180A, E181G, and -P mutants. For these mutants, the $HB1-4_{avg}$ index is, indeed, below 50% due to increases in the population of non-native states compared to the wild type. Furthermore, the lowest energy basin is shifted toward a lower number of β-sheet HBs or higher energy values. For the majority of these mutants, alternative conformational states essentially include one- or two-turn α-helices characterized by zero β-sheet HBs and RgCore values above 6 Å (Figures 7 and 8). The formation of α-helix turns, while compatible with EL2 bridging H4 and H5, as demonstrated by the crystallographic structures of the homologous β1- and β2-adrenergic receptors,[34,35,34,35] is expected to perturb the native interactions of the loop with the surrounding domains of rhodopsin. In contrast, the E181G and -P mutants are essentially characterized by misfolded forms of the native β-hairpin rather than α-helix turns. The misfolding effect of nonconservative mutations of $R177^{(1)}$ or $D190^{(14)}$ is likely due to a disruption of the native interstrand salt bridge between the two residues, expected to stabilize the native β-hairpin based also upon previous in vitro investigations.[18] Differently from the $R177^{(1)}$ and $D190^{(14)}$ mutants, the misfolding effect of P180A, E181G, and -P may be due to the elimination or introduction of proline or glycine residues that would perturb the native backbone conformational behavior of EL2.

Furthermore, five mutants show moderate misfolding properties, whereas the majority of the mutants, 27 over 40 (68%), show a poor or absent misfolding effect (Tables 1 and 2). With respect to this large subset of mutants, we speculate that their structural effects are to introduce nonnatural disulfide bridges or to perturb the EL2−TM and/or EL2−retinal interface rather than the intrinsic folding properties of EL2.

Fifteen of the 27 low or non-misfolding mutants concern $E181^{(5)}$, which was subjected to all possible amino acid substitutions (Table 2). In vitro mutational analysis of this highly studied glutamate shows that replacements with lysine, arginine, and proline result in totally impaired receptor expression (Table 2).[36] The remaining 16 mutants expressed and bound 11-*cis*-retinal to form pigments. Such in vitro data suggest that $E181^{(5)}$ does not contribute significantly to spectral tuning of the ground state of rhodopsin but rather affects the environment of the retinylidene Schiff base in the active MII photoproduct (Table 2).[36]

Consistent with the results of previous analyses,[36] we could not find any significant linear correlation between in vitro data on the $E181^{(5)}$ mutants (Table 2) and a significant number of descriptors of the physicochemical properties of the amino acids, including hydrophobicity and hydrophilicity parameters, size descriptors, volume and surface area values, solution properties, and chromatographic properties, as well as polarity and polarizability indices (results not shown).

The results of REMD simulations, consistent with in vitro data, suggest that the effects of these mutations should be ascribed to perturbations in the network of interactions mediated by such glutamate rather than to a disruption of the native EL2 β-hairpin.

5. Summary

In this study, 40 rhodopsin mutations, 15 ADRP-linked and 25 artificial, all located in EL2, were screened by REMD simulations. The results of the screening constitute the start of a systematic structure-based reclassification of ADRP mutations.

Eight out of 40 EL2 mutants resulted in strong misfolding effects on the native β-hairpin, consistent with in vitro evidence that they all share severe impairments in folding/expression and/or retinal binding. Four of these misfolding mutants, i.e. P180A, and D190A, -G, and -Y, are associated with ADRP. Moreover, five residues displayed moderate misfolding effects and they include two ADRP-linked mutants, i.e. S186P and D190N. The remaining 27 mutants, including nine ADRP-linked mutants and overall characterized by milder effects on rhodopsin expression, did not perturb significantly the conformational behavior of the native β-hairpin. Thus, the computational screening could individuate and differentiate EL2 rhodopsin mutations that would affect the intrinsic stability of the native β-hairpin from mutations expected to variably impair native contacts between the loop and the surrounding receptor domains.

We, therefore, predict that for six out of the 15 ADRP-linked mutants, the structural determinants of the disease are mutation-induced misfolding effects on EL2. A misfolded EL2, being part of the stability core, is expected to undermine

the stability of rhodopsin, consistent with the impaired folding/expression observed for these mutants.

The extensive computational screening carried out in this study relies on strong comparative bases and takes advantage of the use of a fast and effective implicit solvent model. Within a comparative framework, possible overestimations of the native state ensembles can be neglected, as they are expected to be equally shared by wild type and mutant forms and to not affect predictions. The latter, indeed, profit by the internal consistency that characterize any comparative approach aimed at highlighting differences/similarities rather than absolute values/behaviors.

The results of this study add structural insight to the poorly resolved biochemical behavior of selected class II ADRP mutations, a fundamental step toward an understanding of the atomistic causes of the disease.

Acknowledgment. This study was supported by a Telethon-Italy grant no. S00068TELU (To F.F.).

Supporting Information Available: Additional analysis plots (Figures S1−S7). This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Chuang, J. Z.; Vega, C.; Jun, W.; Sung, C. H. Structural and functional impairment of endocytic pathways by retinitis pigmentosa mutant rhodopsin−arrestin complexes. *J. Clin. Invest.* **2004**, *114*, 131–40.

(2) Kennan, A.; Aherne, A.; Humphries, P. Light in retinitis pigmentosa. *Trends Genet.* **2005**, *21*, 103–10.

(3) Mendes, H. F.; van der Spuy, J.; Chapple, J. P.; Cheetham, M. E. Mechanisms of cell death in rhodopsin retinitis pigmentosa: implications for therapy. *Trends Mol. Med.* **2005**, *11*, 177–85.

(4) Burns, M. E.; Arshavsky, V. Y. Beyond counting photons: Trials and trends in vertebrate visual transduction. *Neuron* **2005**, *48*, 387–401.

(5) Fanelli, F.; De Benedetti, P. G. Computational modeling approaches to structure−function analysis of G protein-coupled receptors. *Chem. Rev.* **2005**, *105*, 3297–3351.

(6) Palczewski, K. G protein-coupled receptor rhodopsin. *Annu. Rev. Biochem.* **2006**, *75*, 743–67.

(7) Okada, T.; Ernst, O. P.; Palczewski, K.; Hofmann, K. P. Activation of rhodopsin: New insights from structural and biochemical studies. *Trends Biochem. Sci.* **2001**, *26*, 318–24.

(8) McBee, J. K.; Palczewski, K.; Baehr, W.; Pepperberg, D. R. Confronting complexity: The interlink of phototransduction and retinoid metabolism in the vertebrate retina. *Prog. Retin. Eye Res.* **2001**, *20*, 469–529.

(9) Yan, E. C.; Kazmi, M. A.; Ganim, Z.; Hou, J. M.; Pan, D.; Chang, B. S.; Sakmar, T. P.; Mathies, R. A. Retinal counterion switch in the photoactivation of the G protein-coupled receptor rhodopsin. *Proc Natl Acad Sci U S A* **2003**, *100*, 9262–7.

(10) Ahuja, S.; Hornak, V.; Yan, E. C.; Syrett, N.; Goncalves, J. A.; Hirshfeld, A.; Ziliox, M.; Sakmar, T. P.; Sheves, M.; Reeves, P. J.; Smith, S. O.; Eilers, M. Helix movement is coupled to displacement of the second extracellular loop in rhodopsin activation. *Nat. Struct. Mol. Biol.* **2009**, *16*, 168–75.

(11) Rader, A. J.; Anderson, G.; Isin, B.; Khorana, H. G.; Bahar, I.; Klein-Seetharaman, J. Identification of core amino acids stabilizing rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7246–51.

(12) Haberthur, U.; Caflisch, A. FACTS: Fast analytical continuum treatment of solvation. *J. Comput. Chem.* **2008**, *29*, 701–715.

(13) Dinner, A. R.; Lazaridis, T.; Karplus, M. Understanding beta-hairpin formation. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068–73.

(14) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–6.

(15) Garcia, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins* **2001**, *42*, 345–54.

(16) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* **2004**, *56*, 310–21.

(17) Evans, D. A.; Wales, D. J. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **2004**, *121*, 1080–90.

(18) Janz, J. M.; Fay, J. F.; Farrens, D. L. Stability of dark state rhodopsin is mediated by a conserved ion pair in intradiscal loop E-2. *J. Biol. Chem.* **2003**, *278*, 16982–91.

(19) Okada, T.; Sugihara, M.; Bondar, A. N.; Elstner, M.; Entel, P.; Buss, V. The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J. Mol. Biol.* **2004**, *342*, 571–83.

(20) Nakamichi, H.; Okada, T. Crystallographic analysis of primary visual photochemistry. *Angew. Chem., Int. Ed.* **2006**, *45*, 4270–3.

(21) Nakamichi, H.; Okada, T. Local peptide movement in the photoreaction intermediate of rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12729–34.

(22) Ludeke, S.; Beck, M.; Yan, E. C.; Sakmar, T. P.; Siebert, F.; Vogel, R. The role of Glu181 in the photoactivation of rhodopsin. *J. Mol. Biol.* **2005**, *353*, 345–56.

(23) MacKerell, A. D. J.; Bashford, D.; Bellott, M.; Dunbrack, R. L. J.; 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, T. D.; Prodhom, B.; Reiher, W. E. I.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiòrkiewicz-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.

(24) Rao, F.; Caflisch, A. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **2003**, *119*, 4035–4042.

(25) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys.* **2004**, *121*, 10748–10756.

(26) Im, W.; Lee, M. S.; Brooks, C. L., III. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691–702.

(27) Zagrovic, B.; Pande, V. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. *J. Comput. Chem.* **2003**, *24*, 1432–6.

(28) Klimov, D. K.; Thirumalai, D. Mechanisms and kinetics of beta-hairpin formation. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544–9.

(29) Dunbrack, R. L., Jr.; Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **1993**, *230*, 543–74.

(30) Ponder, J. W.; Richards, F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987**, *193*, 775–91.

(31) Sutcliffe, M. J.; Hayes, F. R.; Blundell, T. L. Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. *Protein Eng.* **1987**, *1*, 385–92.

(32) Heyer, L. J.; Kruglyak, S.; Yooseph, S. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res.* **1999**, *9*, 1106–15.

(33) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. Wordom: A program for efficient analysis of molecular dynamics simulations. *Bioinformatics* **2007**, *23*, 2625–2627.

(34) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–1265.

(35) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486–91.

(36) Yan, E. C.; Kazmi, M. A.; De, S.; Chang, B. S.; Seibert, C.; Marin, E. P.; Mathies, R. A.; Sakmar, T. P. Function of extracellular loop 2 in rhodopsin: glutamic acid 181 modulates stability and absorption wavelength of metarhodopsin II. *Biochemistry* **2002**, *41*, 3620–7.

(37) Sung, C. H.; Schneider, B. G.; Agarwal, N.; Papermaster, D. S.; Nathans, J. Functional heterogeneity of mutant rhodopsins responsible for autosomal dominant retinitis pigmentosa. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 8840–4.

(38) Kaushal, S.; Khorana, H. G. Structure and function in rhodopsin. 7. Point mutations associated with autosomal dominant retinitis pigmentosa. *Biochemistry* **1994**, *33*, 6121–8.

(39) Souied, E.; Gerber, S.; Rozet, J. M.; Bonneau, D.; Dufier, J. L.; Ghazi, I.; Philip, N.; Soubrane, G.; Coscas, G.; Munnich, A. Five novel missense mutations of the rhodopsin gene in autosomal dominant retinitis pigmentosa. *Hum. Mol. Genet.* **1994**, *3*, 1433–4.

(40) Iannaccone, A.; Man, D.; Waseem, N.; Jennings, B. J.; Ganapathiraju, M.; Gallaher, K.; Reese, E.; Bhattacharya, S. S.; Klein-Seetharaman, J. Retinitis pigmentosa associated with rhodopsin mutations: Correlation between phenotypic variability and molecular effects. *Vision Res.* **2006**, *46*, 4556–67.

(41) Sung, C. H.; Davenport, C. M.; Nathans, J. Rhodopsin mutations responsible for autosomal dominant retinitis pigmentosa. Clustering of functional classes along the polypeptide chain. *J. Biol. Chem.* **1993**, *268*, 26645–9.

(42) Colley, N. J.; Cassill, J. A.; Baker, E. K.; Zuker, C. S. Defective intracellular transport is the molecular basis of rhodopsin-dependent dominant retinal degeneration. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3070–4.

(43) Dryja, T. P.; McEvoy, J. A.; McGee, T. L.; Berson, E. L. Novel rhodopsin mutations Gly114Val and Gln184Pro in dominant retinitis pigmentosa. *Invest Ophthalmol. Vision Sci.* **2000**, *41*, 3124–7.

(44) Kono, M.; Yu, H.; Oprian, D. D. Disulfide bond exchange in rhodopsin. *Biochemistry* **1998**, *37*, 1302–5.

(45) Hwa, J.; Reeves, P. J.; Klein-Seetharaman, J.; Davidson, F.; Khorana, H. G. Structure and function in rhodopsin: further elucidation of the role of the intradiscal cysteines, Cys-110, -185, and -187, in rhodopsin folding and function. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1932–5.

(46) Rüther, K.; von Ballestrem, C. L.; Müller, A.; Kremmer, S.; Eckstein, A.; Apfelstedt-Sylla, E.; Gal, A.; Zrenner, E. In *Degenerative Diseases of the Retina*; Anderson, R. E., LaVail, M. M., Hollyfield, J. G., Eds.; Plenum Press: New York, 1995, pp 303−312.

(47) Liu, X.; Garriga, P.; Khorana, H. G. Structure and function in rhodopsin: correct folding and misfolding in two point mutants in the intradiscal domain of rhodopsin identified in retinitis pigmentosa. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4554–9.

(48) Doi, T.; Molday, R. S.; Khorana, H. G. Role of the intradiscal domain in rhodopsin assembly and function. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 4991–5.

(49) Dryja, T. P.; Hahn, L. B.; Cowley, G. S.; McGee, T. L.; Berson, E. L. Mutation spectrum of the rhodopsin gene among patients with autosomal dominant retinitis pigmentosa. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 9370–4.

CT900145U

JCTC Journal of Chemical Theory and Computation

Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles

Christopher L. McClendon,[†,‡] Gregory Friedland,[†,‡,§,||] David L. Mobley,[†,⊥] Homeira Amirkhani,[○,+] and Matthew P. Jacobson*[,†]

Graduate Group of Biophysics, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, Department of Bioengineering and Therapeutic Sciences University of California, San Francisco, California, 94158, and Department of Physics, Sharif University of Technology

Received April 15, 2009

Abstract: Allostery describes altered protein function at one site due to a perturbation at another site. One mechanism of allostery involves correlated motions, which can occur even in the absence of substantial conformational change. We present a novel method, "MutInf", to identify statistically significant correlated motions from equilibrium molecular dynamics simulations. Our approach analyzes both backbone and side chain motions using internal coordinates to account for the gear-like twists that can take place even in the absence of the large conformational changes typical of traditional allosteric proteins. We quantify correlated motions using a mutual information metric, which we extend to incorporate data from multiple short simulations and to filter out correlations that are not statistically significant. Applying our approach to uncover mechanisms of cooperative small molecule binding in human interleukin-2, we identify clusters of correlated residues from 50 ns of molecular dynamics simulations. Interestingly, two of the clusters with the strongest correlations highlight known cooperative small-molecule binding sites and show substantial correlations between these sites. These cooperative binding sites on interleukin-2 are correlated not only through the hydrophobic core of the protein but also through a dynamic polar network of hydrogen bonding and electrostatic interactions. Since this approach identifies correlated conformations in an unbiased, statistically robust manner, it should be a useful tool for finding novel or "orphan" allosteric sites in proteins of biological and therapeutic importance.

Introduction

Originally, allosteric proteins were those where multiple subunits achieved cooperative binding through ligand-mediated shifts in conformational equilibria. Nowadays, allostery is broadly defined as any case in which an event at one site on a protein or complex impacts function, dynamics, or distribution of conformations of another site (for recent reviews see[1,2]). This broader definition includes single-domain proteins as well as proteins or complexes where cooperativity occurs without substantial conformational change. Given this broader definition, it has been suggested that allostery is a property of many proteins,[3−5] but is only relevant when a localized event precipitates a change in function. Recently, there has been renewed interest in uncovering allosteric mechanisms of protein regulation and in discovering new allosteric sites, which are of significant interest in biological mechanisms of protein regulation and as novel sites for drug discovery.[6−8]

* To whom correspondence should be addressed. E-mail: matt.jacobson@ucsf.edu.

[†] Graduate Group of Biophysics.

[‡] Department of Pharmaceutical Chemistry.

[§] California Institute of Quantitative Biosciences and Department of Bioengineering and Therapeutic Sciences.

[||] Current address: Sandia National Laboratories, Joint Bioenergy Institute.

[○] Sharif University of Technology.

[⊥] Current address: Department of Chemistry, University of New Orleans.

[+] Current address: Department of Bioengineering, University of California, San Francisco.

10.1021/ct9001812 CCC: $40.75
Published on Web 08/12/2009

Typically, sites are identified as allosteric after mutational, structural, and thermodynamic characterization with allosteric protein, peptide, or small-molecule modulators, which are frequently found serendipitously.[9] As such, there has been much interest in computational approaches to identify novel allosteric sites. One of the most extensively used approaches has been the Statistical Coupling Analysis pioneered by Ranganathan and co-workers,[10−13] where pairs of residues that tend to be mutated together in multiple sequence alignments suggest coupling between protein sites. This approach has recently been used to engineer a novel allosteric network by combining predicted allosteric pathways from a light sensor and an enzyme.[14] However, this approach requires large multiple sequence alignments and the predicted couplings may or may not be relevant to particular proteins in the alignment.[15] Alternative methods to identify allosteric networks using sequence comparisons have also been described.[16]

Other computational methods to study allosteric mechanisms and identify potential sites for allosteric regulation focus on a protein's structure and dynamics. Cooper and Dryden showed that the free energy of cooperativity could be separated into two terms: one that accounts for changes in the protein's conformational distribution (i.e., by population shifts), and one that accounts for changes in the amplitudes or frequencies of protein vibrational motions. One approach to studying allostery is to focus on protein vibrations[17−20] around a static structure, often by a coarse-grained normal-mode analysis,[21−24] in which case allosteric effects of perturbations can be calculated analytically. However, these approaches are unable to capture the anharmonic and multiwell nature of flexible degrees of freedom in proteins. Another approach is to infer groups of residues important for a given allosteric process by analyzing structures trapped in different conformations.[25−27]

Dynamical approaches to studying allostery generate an ensemble of structures and then analyze the ensemble using cross-correlations,[28] contact correlations,[29] principal components,[29] or local unfolding correlations.[30] One widely adopted approach uses a quasi-harmonic metric for correlations that assumes an "average" structure.[28,29,31−33] This approximation may be appropriate for small backbone fluctuations but may not aptly describe conformational changes that involve basin-hopping, such as loop or side-chain motions. To overcome this quasi-harmonic limitation, Lange and Grubmuller[34,35] used a mutual information method to account for both quasi-harmonic and anharmonic correlations in atoms' motion in Cartesian space. Still other methods introduce mechanical perturbations and monitor the subsequent motions of residues.[36,37] These approaches can detect substantial population shifts or structural changes following the induced local perturbations, as the added energy facilitates barrier crossing.

Our MutInf approach for identifying allosteric networks quantifies correlations between the conformations of residues in different sites. We use an entropy-based approach to analyze ensembles of protein conformers, such as those from molecular dynamics or Monte Carlo simulations. The method is applicable even in cases where conformational changes are subtle, for example, when the coupling is mostly entropic in nature.[38,39] Unlike the approaches described above, our approach uses internal coordinates and focuses on dihedral angles, which are responsible for most low-frequency motions, in order to capture correlated changes in side chain rotamers, a highly anharmonic type of correlated motion. The most closely related previously published method is a study that examined side-chain correlations using a mutual information metric and Monte Carlo simulations of side-chains[40] on a set of fixed protein backbones.

Our MutInf method builds upon and extends previous work by (1) directly connecting correlated conformations to the molecular configurational entropy, (2) incorporating more robust entropy estimators, (3) correcting for undersampling using data from multiple simulations, (4) testing statistical significance to filter out correlated motions that are not significant, and (5) analyzing both backbone and side chain torsions, which are frequently coupled.[41,42] The theoretical underpinnings of our approach are described in detail in Methods. Briefly, we use second-order terms from the configurational entropy expansion, the mutual information,[43] to identify pairs of residues with correlated conformations in an equilibrium ensemble. In calculating mutual information, it does not matter whether two residues move at the same time or whether one moves, and then the other; what counts is whether these residues' conformational distributions are correlated. In this work, we use the terms correlated motions and correlated conformations interchangeably.

Because we look for correlated conformations in an unbiased, statistically robust manner, we believe that MutInf will be a useful tool in the discovery of novel, "orphan" allosteric sites, where endogenous protein or small molecule allosteric modulators have yet to be discovered. As a proof-of-principle, we used our approach to identify correlations between the conformations of protein residues lining two small-molecule binding sites in human interleukin-2 (IL-2). This single-domain protein exhibits cooperative ligand binding without substantial conformational change, and to date no follow-up work has been done to uncover the mechanism for this cooperativity. We discuss the rationale behind our approach and compare its strengths and weaknesses to those of other methods and then discuss the mathematical details of our method and our novel results on IL-2.

Methods

Theoretical Underpinnings of the Model. When an equilibrium ensemble of states is altered by small perturbations, the fluctuation−dissipation theorem relates equilibrium fluctuations to the system's response, which will be proportional to equilibrium pair-correlations of the degrees of freedom and linear in the applied perturbations. This linear response theory suggests that external forces, such as those due to ligand binding, cause the largest indirect changes in the degrees of freedom that are most correlated (at equilibrium) with those directly perturbed

by the external forces. As has been previously noted,[29] this also means that the response to small perturbations involves the same fluctuation pathways activated by random solvent collisions at equilibrium. Elastic network models have identified a correspondence between low-frequency normal modes and pathways used in several protein conformational changes,[22,44] suggesting that correlations observed in equilibrium simulations may propagate perturbations in structure and/or dynamics due to ligand or protein binding.

A perturbation at one site can couple to another site directly, through electrostatic or steric interactions, indirectly, through solvent reordering, or through a network of residues with correlated conformations. When the conformation at one site depends on the conformation at another site, the sites' conformations are correlated. When the conformations are correlated, perturbations at one site can cause population shifts in conformations at other sites. Correlated conformations are then signals that can be used to identify allosteric mechanisms and predict new sites for allosteric inhibition by proteins or small molecules.

Our MutInf approach uses equilibrium molecular dynamics simulations to identify correlations in residues' conformations, from which functional coupling between sites is inferred. Approaches such as ours that infer allostery from equilibrium simulations assume that the allosteric phenomena of interest (i.e., ligand binding, protein binding, protonation state changes, etc.) make perturbations to the energy landscape that are relatively small, that is, at most a few kT. For example, proteins and ligand binders with fast on-rates will satisfy this assumption, while proteins and ligand binders with slow on-rates may not. Furthermore, equilibrium approaches that infer allostery assume that there are no large barriers to conformational changes required for allostery. If such barriers existed, they would prevent pairs of residues from sampling relevant correlated shifts in conformation when perturbations of interest are applied. Along these lines, these equilibrium approaches also assume that there is sufficient sampling along the degrees of freedom relevant to the allosteric phenomena, so that productive or "on-pathway" correlated motions can be observed.

To quantify correlations between residues' conformations from equilibrium simulations, we take advantage of a connection between information theory and thermodynamics. Inspired by the use of mutual information by Killian, Kravitz, and Gilson in calculating configurational entropies from conformational ensembles using internal coordinates,[43] we use second-order terms from the configurational entropy expansion, i.e. the mutual information, to identify pairs of residues with correlated conformations. This approach directly and quantitatively connects correlations in conformation to configurational entropy. Using internal coordinates to calculate the mutual information has the two-fold advantage of (1) capturing the rotameric, flipping, and gear-like nature of correlated side-chains and (2) removing potentially spurious correlations that can arise due to structural alignment. The latter effect occurs because minimization of the rms error in aligning structures in Cartesian space can yield correlated displacements in many atoms' positions as some atoms are fit better than others. An overview of our approach is presented in Scheme 1.

Scheme 1. Schematic of the MutInf Approach for Identifying Correlated Residue Conformations[a]

[a] This shows how the observed mutual information is statistically filtered and corrected before being summed over residues pairs. The resulting matrix is then clustered as in a microarray experiment to identify groups of residues showing similar patterns of correlations.

In applying entropy and mutual information to studying allostery, we sought to obtain a measure of the statistical significance of our results and filter out noisy and artifactual correlations. To accomplish this, we extended established methods for calculating entropies and examining correlations via mutual information to handle finite sample sizes, to incorporate data from multiple simulations, to account for the variability between simulations, and to correct for the fact that multiple simulations do not, in practice, represent independent samples of the macromolecular ensemble.

Calculation of Mutual Information. The configurational space of a molecule can be described in a standard Cartesian coordinate system or in an internal coordinate system of bond lengths, bond angles, and torsion angles (BAT).[43] For proteins, key torsion angles include the φ, Ψ, and ω torsion angles of the protein backbone and the χ torsion angles of the amino acid side-chains. In the present work, we consider only the φ, Ψ, and heavy-atom χ torsion angles (only the first χ angle for proline) and neglect changes in bond lengths, bond angles and omega backbone torsion angles, as we believe that the dynamics of the first three are the most relevant to describing motions of biological importance.[43]

Small sample sizes are notoriously challenging for entropy and mutual information-based approaches, so we use robust estimators and correct for bias using simulated data.

Configurational Entropy Expansion and Correlations Between Degrees of Freedom. We wish to quantify correlations between residues' torsions. Following the works of Matsuda[45] and of Killian, Kravitz, and Gilson,[43] we connect correlated torsions to thermodynamics using an

expansion of the molecular configurational entropy into terms over single torsions, pairs of torsions, etc. The total torsional entropy is given by

$$S_{\text{conf}} = \sum_{i}^{n} \int_{0}^{2\pi} p(\phi)\ln p(\phi)\mathrm{d}\phi - \sum_{i}^{n}\sum_{j\neq i}^{n} \int_{0}^{2\pi}\int_{0}^{2\pi} p(\phi_1,\phi_2)\ln \frac{p(\phi_1,\phi_2)}{p(\phi_1)p(\phi_2)}\mathrm{d}\phi_1\mathrm{d}\phi_2 + \ldots - \ldots \quad (1)$$

where indices i and j are residues' torsions, and n is the number of torsions (φ, Ψ, and χ torsion angles in the present work). The second-order term here represents a sum of the mutual information of each pair of torsions. The mutual information describes correlations between degrees of freedom and gives a measure of how much information about one degree of freedom is gained by knowledge about another.[46] Because the mutual information values are terms in the entropy, which is related to free energy, the mutual information in eq 1 is in units of kT. The mutual information has been a popular, distribution-free analysis method, and more recently has been used in the context of molecular conformational ensembles.[40,43]

As an example, consider the distributions of the χ_1 torsion angles for two side chains. For concreteness, we use an example of two aromatic residues in close proximity from our molecular dynamics simulations of interleukin-2 (Figure 1). The expected joint distribution of these torsion angles under the null hypothesis (Figure 1A) of independence is merely the outer product of the marginal distributions. However, the joint distribution from the observed simulations (Figure 1B) shows that these torsion angles are correlated (I = 0.203 kT), because a cross-peak (indicated by a gray box) appears in the simulations that would not be expected if these torsions were independent.

In practice, we compute the mutual information, I, between two degrees of freedom as the difference between the self-entropies and the joint entropy, using the relation, $I = S(1) + S(2) - S(1,2)$ and a corrected histogram entropy estimate[47] over adaptive partitions[46]

$$I = \sum_{i=1}^{r}\frac{n_i}{N}\left(\ln N - \Psi(n_i) - \frac{(-1)^{n_i}}{n_i+1}\right) + \sum_{j=1}^{s}\frac{n_j}{N}\left(\ln N - \Psi(n_j) - \frac{(-1)^{n_j}}{n_j+1}\right) - \sum_{j=1}^{rs}\frac{n_{ij}}{N}\left(\ln N - \Psi(n_{ij}) - \frac{(-1)^{n_{ij}}}{n_{ij}+1}\right) \quad (2)$$

where r and s are the number of marginal bins, n_i, n_j, and n_{ij} are the histogram counts, N is the number of data points, and Ψ is the digamma function. Adaptive partitions make efficient use of discrete bins, preserve correlations between variables, and normalize each joint distribution to a reference distribution in which marginal counts are as uniform as discretization allows.[46] In this work, we used 24 bins per dimension. Adaptive partitions also enable accurate mutual information values to be calculated whether torsional motions are large or small. Note also that we account for the 2-fold symmetry in the χ_2 angle of Asp, Phe, and Tyr and in the χ_3 angle of Glu.

The histogram entropy estimator above assumes that histograms are populated by a Poisson process ($n_{ij} \ll N$) and so is especially appropriate for sparse joint histograms. It

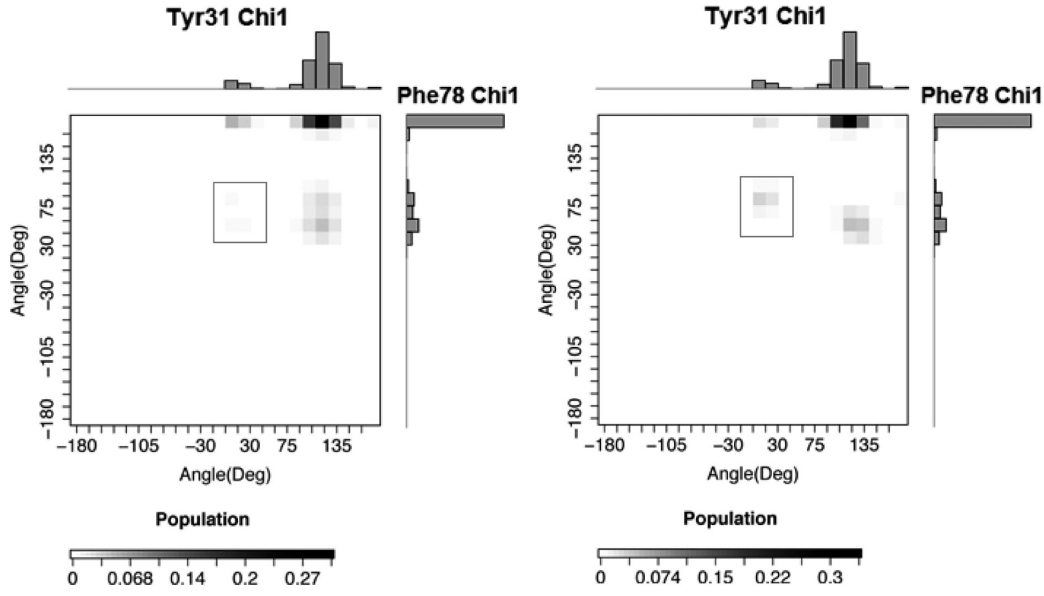


Figure 1. Joint distributions of correlated torsions are different from what would be expected if they were independent. (A) Distributions of two χ_1 torsion angles are shown along with the joint distribution expected if they were independent (which the product of the marginal probabilities). (B) Distributions of the same two χ_1 torsion angles are shown along with the observed joint distribution from molecular dynamics simulations. Gray boxes highlight a cross-peak with substantial height in the observed simulations (B) but with negligible height under the null hypothesis of independence (A).

also implicitly includes finite bin and data size corrections used in other discrete entropy estimators.[48] As a statistic for examining correlations between variables, the mutual information (with corrections discussed below) is far more robust against small sample sizes than the χ^2 statistic, which assumes $n_{ij} \geq 5$. While the nearest-neighbor approach[49] could have been used instead to compute these integrals, it can require $N \approx 50\,000$ data points[50] or more to yield a converged estimate of the mutual information for pairs of torsions. Nearest-neighbor approaches are accurate for very large data sets but have biases under finite sample sizes that depend on the topology of conformational space sampled in simulations.[50] As our goal was to extend mutual information calculations to handle smaller sample sizes, we chose instead to use adaptive partitioning in combination with the corrected histogram mutual information estimate above (eq 2), so that each pair of degrees of freedom could be compared against the same empirically generated reference distribution and evaluated for significance.

Correction for Nonzero Mutual Information in Independent Data Sets. In a number of applications using mutual information, it has been found that samples of two variables that are independent can yield nonzero mutual information in calculations.[46,51−53] We empirically observe the same in simulated data (data not shown), and this is not surprising because errors in estimates of the true mutual information are a consequence of finite samples. To correct for this, one approach is to create P permutations of the original data, so that the marginal probability distributions remain the same, while correlations between the data are scrambled. One can use these permutations to establish a test of significance of the observed mutual information with a null hypothesis of independence versus a one-sided alternative. The approximate p-value is then the percentage of mutual information values from different permutations that are greater than the observed mutual information from the original data.[46,53] Also, the average mutual information for the permuted data, the "independent information", can be subtracted from the observed mutual information to yield the "excess mutual information", a more reliable estimate of the true mutual information.[51,52]

When adaptive partitioning is not used (and hence the marginal densities are not normalized), permutation approaches are inefficient in sampling the distribution of the mutual information under the null hypothesis, because permuted values are likely to fall into bins overrepresented in the marginal densities; adaptive partitioning fixes this inefficiency by normalizing marginal densities without altering correlations between variables. One can apply the permutation approach above to nearest-neighbor estimates as well, as these also will have bias due to finite sample sizes. For example, a combined K-fold resampling/permutation test was found to be useful in conjunction with nearest-neighbor mutual information estimates in feature selection.[53] A major drawback to the permutation approach is that it is computationally demanding in processing time and in memory. Moreover, permutations introduce random error because not all $N!$ permutations can be made.[53]

Instead, since adaptive partitioning is used in this work, we noted that the same distribution of the "independent information" is appropriate for all pairs of degrees of freedom. The distribution of the mutual information for independent variables for given data size N and number of marginal bins r and s has not yet been analytically solved, though in some cases can be empirically fit.[52] However, because an analytical, parametric approach is not available, we perform Monte Carlo sampling to obtain the reference distribution of the "independent information" for all pairs of torsions. With adaptive partitioning, the marginal counts are nearly uniform and in any case are equivalent for different pairs of torsions. Thus, all pairs of torsions will have the same distribution in histogram bin space under the null hypothesis of independence. To construct the reference distribution for a pair of independent torsions, we first make a copy of the marginal distributions for a given pair of torsions (it does not matter which pair we choose). Then, we choose ordered pairs of bin indices at random from these marginal distributions and place them into a 2-D histogram without replacement. The mutual information is calculated according to eq 2 above, and this procedure is repeated 1000 times to create a distribution of the mutual information under the null hypothesis of independence for the given number of data points N and number of bins r.

We use this distribution of "independent information" for a significance test of observed mutual information values, and we subtract the average "independent information" from the observed mutual information to yield the "excess" mutual information; this filters out insignificant mutual information values and corrects for finite sample size bias. Because this analysis empirically generates a distribution under the null hypothesis, the false positive rate for keeping a nonzero mutual information value for torsions that are truly independent is α, the significance level (in our case, 0.01). This false positive rate will be further reduced by consideration of the alternative hypothesis.

Bayesian Filter to Remove False Positives. Most approaches that filter mutual information values using tests of statistical significance do so according to whether the null hypothesis of independence can be rejected using descriptive statistics. One disadvantage of these approaches is that they do not consider the distribution of the mutual information under the alternative hypothesis. In Bayesian statistics, the mutual information is a random variable with a distribution, and the probability that the mutual information is greater than a a given value can be calculated. Approximations to the distribution of the mutual information have been described that account for uncertainties in the estimates of the probability density functions.[54] The first two central moments of the distribution, the expectation $E[I]$ and variance $\mathrm{Var}[I]$ of the true information given the data and prior, are given as follows:

$$E[I] = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\Psi(n_{ij}+1) - \Psi(n_i+1) - \Psi(n_j+1) + \Psi(N+1)) \quad (3)$$

$$\mathrm{Var}[I] = \left| \frac{K - J^2}{N+1} + \frac{M + (r-1)(s-1)\left(\frac{1}{2} - J\right) - Q}{(n+1)(n+2)} \right| + O(n^{-3}),$$
$$J = \sum_{ij} \frac{n_{ij}}{N} \ln \frac{n_{ij}N}{n_i n_j}, K = \sum_{ij} \frac{n_{ij}}{N} \left(\ln \frac{n_{ij}N}{n_i n_j} \right)^2,$$
$$M = \sum_{ij} \left(\frac{1}{n_{ij}} - \frac{1}{n_i} - \frac{1}{n_j} + \frac{1}{N} \right) n_{ij} \ln \frac{n_{ij}N}{n_i n_j}, Q = 1 - \sum_{ij} \frac{{n_{ij}}^2}{n_i n_j} \quad (4)$$

where $n_{ij} = n_{ij}(\text{observed}) + n_{ij}(\text{virtual})$, and the virtual counts come from a noninformative Dirichlet prior ($n_{ij} = 1$ for the uniform prior, which was used in this work). Approximations for the leading order terms for the third and fourth central moments have been reported,[54] and could be used in an Edgeworth expansion to approximate the distribution, but for robustness, we chose instead to simply use a Gaussian with the above mean and variance, which fit reasonably well to simulated data in a model system.[54] We then use this approximate distribution of the mutual information to calculate $P(I < E[I_{\text{ind}}])$, the probability that the true mutual information is below that expected for independent torsions (calculated using eq 3 averaged over 1000 simulated independent data sets). Pairs of torsions with $P(I < E[I_{\text{ind}}]) > \alpha$ are not significant and are discarded.

Corrections to the Mutual Information Accounting for Incomplete Sampling. To obtain accurate entropies and mutual information values up to second order, simulations must be run many times longer than the slowest autocorrelation and pair correlation times, and data points should represent independent observations. Due to limited computing power, this is rarely the case, and molecules in simulations carry with them some memory of their initial states. For example, consider a salt bridge. Salt bridges can form strong electrostatic interactions, and hence it can take a long time to sample their full conformational space (long autocorrelation time) and even longer to sample all populated pairwise conformations (long pair correlation times). Thus, a salt bridge may retain some memory of its initial conformation, which will fade away on the time scale of the pair decorrelation time (approximately). In practice, we decided to use data from multiple simulations to penalize this kind of undersampling in a novel way.

First, we first aggregate the counts for two degrees of freedom from a set of simulations (sample ensembles) of size n_{sims} and calculate the mutual information for all the simulations taken together. Intuitively, two torsions in different simulations should not be correlated, as they should sample their probability distributions independently. Any nonzero (excess) mutual information between these torsions is a measure of conformational undersampling bias that we can subtract from the mutual information between the torsions for the set of simulations. To correct the mutual information for artifactual correlations caused by incomplete sampling, we calculate the excess mutual information and then subtract the average excess mutual information between two degrees of freedom in different pairs of simulations (when it is positive):

$$I_{i,j}^{\text{sims}} = I_{i,j}^{\text{sims}} - \langle I^{\text{ind}}(N,r) \rangle - \binom{n_{\text{sims}}}{2}^{-1} \sum_{k=1}^{n_{\text{sims}}} \sum_{l \neq k}^{n_{\text{sims}}} I_{i,j}^{k,l} - \langle I^{\text{ind}}(N',r') \rangle \quad (5)$$

Here i and j correspond to the different torsions, l and k are the indices of the pairs of different simulations, and I^{ind} is calculated for a pair of simulations just as the independent information is calculated for a set of simulations using the Monte Carlo recipe above, except that values of $\langle I^{\text{ind}} \rangle$ lower than the standard deviation of I^{ind} are zeroed to reduce noise from this term. For the mutual information between torsions in different simulations, we use half as many bins ($r' = r/2$), because the number of data points N' for the histograms is smaller than the total number of data points from all simulations, N ($N' = N/n_{\text{sims}}$). Significant mutual information values are those that have passed the significance test vs the null hypothesis, the Bayesian filtering using the alternative hypothesis, and whose corrected excess mutual information (eq 5) is greater than zero.

When we consider the mutual information between pairs of residues, we take the sum of the mutual information between pairs of residues' torsions:

$$I_{\text{residues}}(i,j) = \sum_{\substack{k=\phi,\varphi,\chi's \\ \text{residue}(i \neq j)}} \sum_{\substack{l=\phi,\varphi,\chi's \\ \text{residue}(j)}} I_{k,l}^{\text{sims}} \quad (6)$$

This may overestimate the total mutual information between two residues, as we neglect the higher-order terms in eq 1. Inclusion of statistically significant higher order terms (which would require more data points) would further increase the accuracy of the calculated mutual information between pairs of residues. Nonetheless, our results below show that our robust use of second-order terms is a powerful means to identify residues with correlated conformations.

Molecular Dynamics Simulations. Molecular dynamics simulations on interleukin-2 (IL-2), alone and in complex with two ligands, were performed using GROMACS 3.3[55,56] and the AMBER-99φ forcefield.[57] Loops missing atomic coordinates, such as residues 1−5, 75−76, and 99−102 in apo IL-2, were closed using loop prediction via the Protein Local Optimization Program (PLOP[58]). Protonation states of histidine side-chains at pH 7 were given by MCCE:[59,60] we modeled His16 as positively charged (residue name HIP) and His55 and His79 as ε-protonated. Two ligand-bound forms of IL-2 were prepared, with either Ro26−4550 (amino(3-(2-(1-methoxy-1-oxo-3-(4-(phenylethynyl)phenyl)-propan-2-ylamino)-2-oxoethyl)piperidin-1-yl)methaniminium)[61,62] bound to the competitive IL-2Rα site (PDB 1M48) or compound **7c** (1-(3,4-dihydro-1*H*-pyrido[3,4-*b*]indol-2(9*H*)-yl)-2-methoxyethanone) bound to the allosteric site.[63] Compound **7c** was built from PDB 1NBP by modifying the covalent compound **7t** in the crystal structure to the noncovalent compound **7c** in Maestro (Schrodinger, 2007), then using PLOP loop prediction to optimize the loop from residue 29 to 33 and to fill in missing residues between residues 73 to 78, in each case simultaneously optimizing side-chains within 12 Å of the given loop region. These ligands were parametrized for MD by GAFF[64] and assigned AM1-BCC

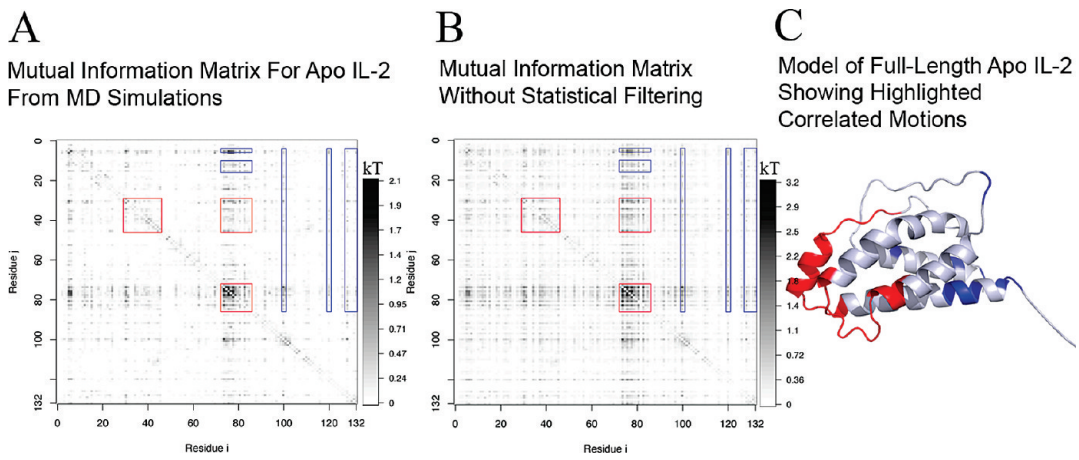


Figure 2. Mutual Information captures significant correlations between residues in human interleukin-2. (A) Mutual information between residues' torsions computed using the present approach, with statistical filtering as detailed in Methods. (B) Same as in A but without any of the aforementioned statistical corrections. (C) The model of full-length human interleukin-2 used in the apo simulations, based on the crystal structure of apo IL-2 (PDB: 1M47). Residues surrounded by red boxes in A are colored red, while residues correlated to these that are surrounded by blue boxes in A are colored blue.

charges.[65,66] Each apo protein or complex was energy minimized in explicit solvent using GROMACS, and five copies of each of the three constructs (apo, competitive site bound, and allosteric site bound) were equilibrated at 300 K (with different random seeds) using constant volume for 10 ps and using a constant pressure of 1 atm for 100 ps using the Berendsen barostat,[67] with hydrogens constrained using the Lincs algorithm.[68] Equilibration of each simulation was followed by a 10 ns production run, with snapshots of the atomic coordinates recorded every 1 ps. Actual lengths of individual simulations ranged between 9.6 ns and 11 ns for technical reasons. RMSDs of the two compound binding sites over the simulations showed that all of the five simulations per apo or small-molecule bound construct were stable (Supporting Information Figure 1).

Ensemble Docking. We clustered MD snapshots from the IL-2 simulations with allosteric compound bound according to the coordinates of residues in the competitive site using QT clustering[69] as provided in GROMACS ("g_cluster-method gromos"). Then, to each cluster representative we docked the Roche competitive site ligand Ro26−4550,[61] which binds cooperatively with the allosteric ligand. This ensemble docking was performed using the XGlide cross-docking script (Schrodinger, 2007, Script Center XGlide v. 1.1.2.6, mmshare v. 16109, using inner and outer grid box lengths of 10 Å and 18 Å, respectively).[70]

Results and Discussion

We applied our MutInf approach to elucidate the mechanism of small molecule binding cooperativity in human interleukin-2. Little is known about how binding of ligand at the IL-2Rα binding site enables binding of a small molecule fragment to a cryptic allosteric site. These ligands bind at least 6.5 Å apart at their closest approach in the predicted ternary complex. Crystal structures of complexes of interleukin-2 with small molecules bound to different sites did not show substantial structural changes at the other sites (maximum Cα rmsd 0.88 Å at the allosteric site for apo PDB 1M47 and competitive-bound PDB 1M48). We therefore hypothesized that allostery and cooperativity in IL-2 arises largely from changes in dynamics and subtle population shifts rather than a major change in the preferred backbone conformation.[28,38,39]

We used molecular dynamics to study correlated motions at the atomic level on a picosecond to nanosecond time scale, and used our MutInf approach to analyze sets of 10 ns trajectories of human interleukin-2, alone and in complex with different small molecule binding partners. Our goal was to show that MutInf can identify significant correlated conformations for functionally important residues in simulations whose lengths and recording frequencies are typical of those in the current literature.

Mutual Information From Molecular Dynamics Identifies Significant Long-Range Correlations. We first analyzed whether an unbiased, whole-protein analysis of correlations between residues in interleukin-2 would be able to identify cooperative sites and the correlations between them from the apo simulations alone. For each pair of residues, we calculate the mutual information as per (eq 6) between all pairs of φ, Ψ, and χ torsion angles for our apo simulations of interleukin-2. The mutual information is reported in units of kT, because of the relationship between mutual information and entropy (eq 1).

When we plotted the statistically significant mutual information between pairs of residues' torsions in IL-2, we found that only a small subset of residue pairs are highly correlated, while many are only marginally correlated (Figure 2A). A substantial part of the present work involved incorporating more robust entropy estimators for calculating the mutual information and filtering out insignificant correlations with the help of empirical or approximated distributions under the null and alternative hypotheses. So, as a control, we plotted the unfiltered mutual information between residues' torsions in Figure 2B. Protocols with and without statistical filtering showed correlation between residues in the loops after helix 1 and between helices 2 and 3 (Figure 2A and B, red boxes on the diagonal and off the diagonal,

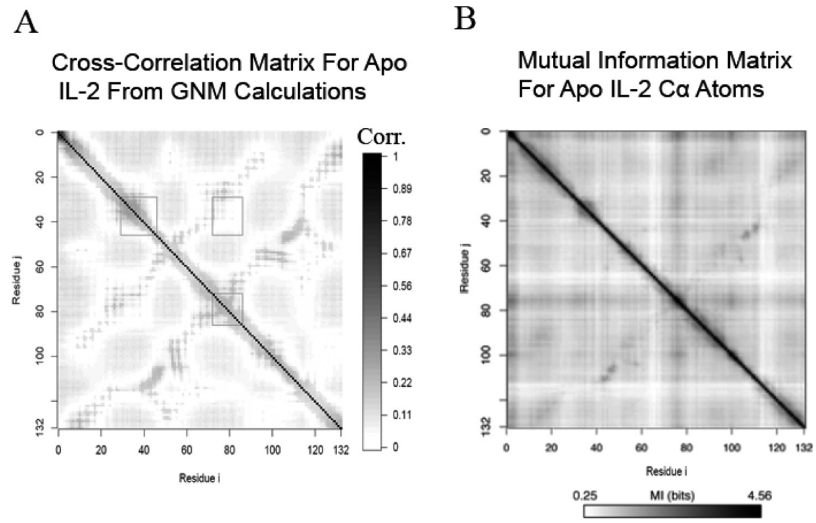


Figure 3. Comparison of pairwise, dynamical correlations between residues computed by alternative methods. (A) Absolute value of the cross-correlation matrix computed using the Gaussian Network Model. (B) Mutual Information between residues' C_α Cartesian coordinates using the approach of Lange and Grubmüller.

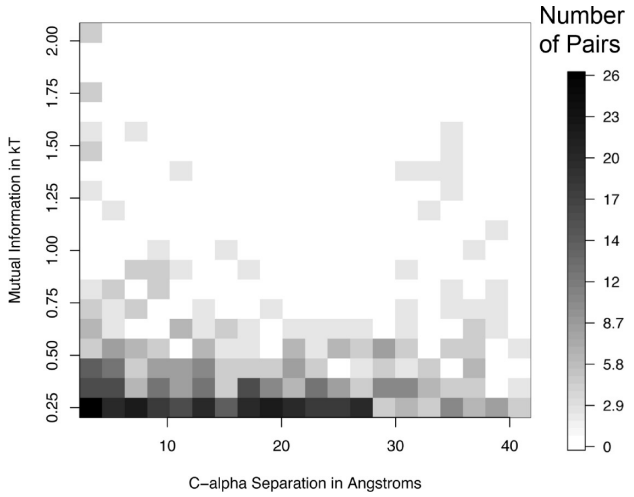


Figure 4. Most of the significant correlations are between distant residues. A 2-D histogram showing the number of pairs of correlated residues vs Cα separation and mutual information value shows that the number and strength of correlated pairs decreases only modestly with distance. For clarity, only those pairs of correlations with a mutual information greater than 0.25 kT are shown.

respectively, and Figure 2C, red residues). Our statistical filtering, however, highlights these and removes background noise; only a subset of the pairwise correlations between residues in these regions make statistically significant contributions to the conformational entropy. Moreover, our MutInf approach identified significant correlations between residues in the loop region between helices 2 and 3 and residues in other regions, in particular residues in the floppy N-terminal tail (Ser6, Thr7) and the beginning of the N-terminal helix (11−15), residues in the loop between helices 2 and 3, and residues in the C-terminal helix (Figure 2A, blue boxes, and Figure 2C, blue residues). The loop between helices 2 and 3 displays significant variability in the different crystal structures of IL-2 and is at least partially disordered in most structures, indicating both that it is flexible and that it can adopt at least several conformations. Residues

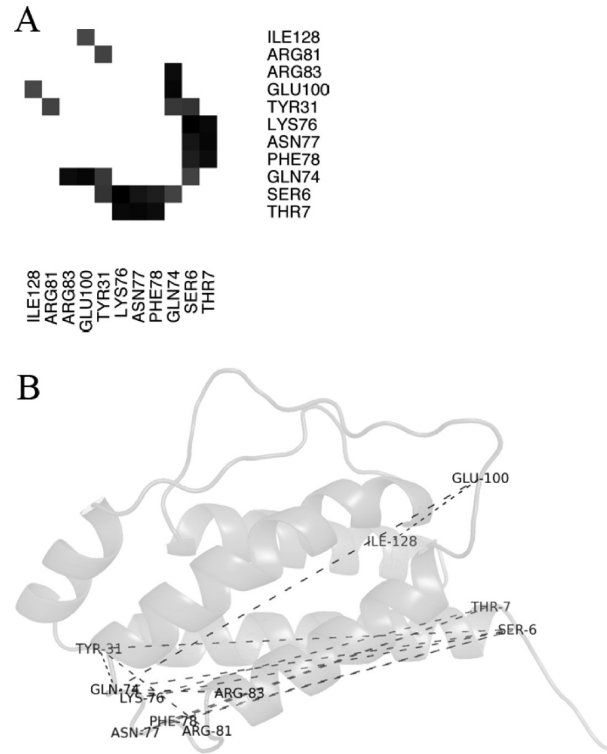


Figure 5. Several distant residues are highly correlated. (A) Correlations greater than kT are shown only for IL-2 residues whose α carbons are separated by more than 5 Å. (B) Dashes connecting each pair of these correlated residues show long-rage correlations across the length of the helical bundle.

showing significant correlations near the C-terminus include two residues in the loop before the C-helix (Glu100 and Thr101), and residues along the C-helix (Arg120, Ile128, and Leu132, proximal to IL-2's negatively charged C-terminus).

We compared our MutInf method to previously reported methods for identifying correlated motions, in particular the Gaussian Network Model (GNM) approach of Bahar and colleagues[71] and the Cartesian mutual information method of Lange and Grubmüller.[34] Both our method and the GNM

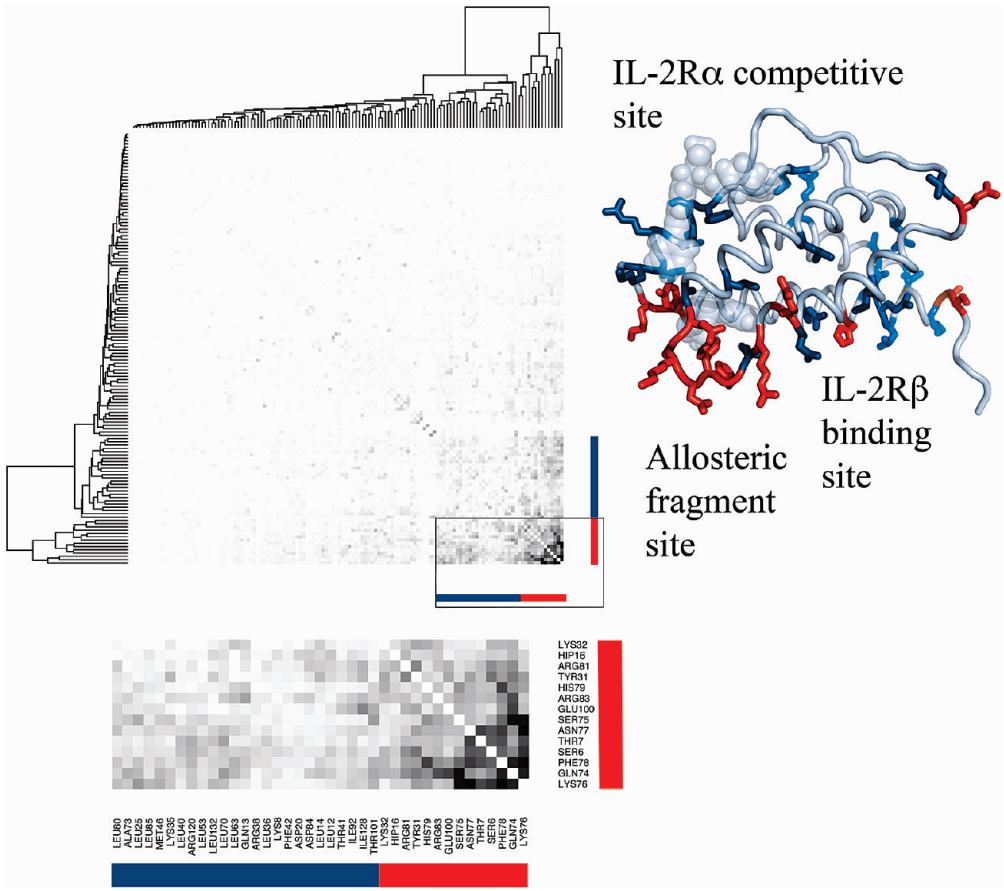


Figure 6. Hierarchical clustering of significant mutual information values identifies allosteric sites. A hierarchically clustered heatmap shows clusters (top left) of residues with similar patterns of mutual information across IL-2 residues. A close-up view highlights numerous significant mutual information values between pairs of residues in two different clusters, red and blue. These red and blue clusters are highlighted in a model of IL-2's ternary complex (right). The strongest cluster (red sticks) chiefly involves a loop enclosing the allosteric fragment's binding site, and this cluster is correlated to a cluster (blue sticks) containing two protein binding sites, the IL-2Rα-receptor-binding/IL-2Rα-inhibitor-binding site and the IL-2Rβ-binding site. The two compound binding sites and the two protein-binding sites are directly correlated through the hydrophobic core (in the blue cluster), through a highly flexible loop (in the red cluster), and crosstalk between these elements, seen in a close-up view of the matrix (bottom).

method suggested correlation between residues in the loop after helix 1 and residues in the loop between helices 2 and 3 (red boxes on the diagonal and off the diagonal, respectively, Figures 2A and 3A). We found that our approach highlighted strong correlations and gave low background noise, while the Cα cross-correlation matrix using a GNM (with 10 Å Cα-Cα cutoff) gave a noisier pattern of correlations, as did Lange and Grubmüller's mutual information method applied to the Cartesian coordinates of Cα atoms (Figure 3B).

Our identification of significant long-range correlations led us to investigate the distribution of correlations between residue pairs as a function of distance between the residues' Cα atoms (Figure 4). As would be expected, the number of weak correlations decreases as the distance between residues increases. However, residues separated by substantial distances have correlations of 0.4 kT or more just as often as residues separated by short distances.

Looking more closely, we see that there are strong couplings between pairs of distant residues (Figure 5). Here, we highlight correlations of magnitude greater than kT between distant residues, namely those with α carbon separations of more than 5 Å. Again, we observe strong couplings between the N-terminus of helix A, Tyr31 at the C-terminal end of helix A, and the adaptive loop region (74, 76−78), as well as between Gln74 and Glu100, and between Glu100 and Ile128 near IL-2's C-terminus. It is not surprising that many of these residues are polar, as molecular dynamics simulations include terms for long-ranged electrostatic and ion-dipole interactions. Gaussian Network Models, on the other hand, do not model such sequence-dependent, long-range interactions, and so it is not surprising that the correlations between distant residues are typically weak. Electrostatic interactions can be both directly and indirectly responsible for long-range correlations in residues' conformations; directly, through Coulomb's law, and indirectly, through a dynamic network of charged and hydrogen-bonding polar residues, and through altering the first-shell water structure around the protein (in simulations with explicit solvent). Unlike charge−dipole or dipole−dipole interactions, where the effective range decreases through averaging over orientations,[72] charge−charge interactions retain their long-range nature even when averaged over orientations.

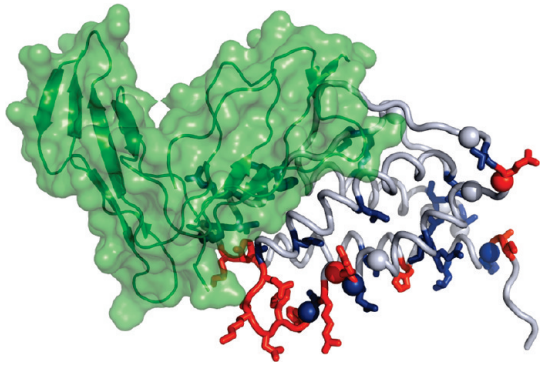


Figure 7. Predicted couplings are consistent with regions perturbed upon IL-2Rα binding. Regions distant from the IL-2Rα receptor binding site that show substantial backbone chemical shift perturbations upon IL-2Rα binding[61] roughly correspond to regions with residues whose conformations are correlated with residues in the IL-2Rα binding site (predominantly residues in the "blue" cluster in Figure 6). Amides on IL-2 whose resonances shifted by more than three linewidths upon IL-2Rα binding or fell below 7% of the original intensity are shown as spheres. Residues from the "red and blue" clusters shown in Figure 6 are colored accordingly. IL-2 is as shown in cartoon and sticks as in Figure 6, while IL-2Rα is shown in green.

Other factors that can give rise to correlated conformations include hydrophobic packing and rigid-body motions of semirigid secondary structure elements, such as α-helices. We note that our approach does not typically show strong correlations within semirigid elements such as α-helices or the central hydrophobic core of the four-helix bundle. Two possible reasons for this are because mutual information values are not normalized quantities and because higher-order correlations are not captured. As the maximum mutual information between two residues is the minimum of their self-entropies (flexibilities), residues that have higher self-entropies (more flexible) can exhibit a greater magnitude of coupling with other residues. This behavior is thermodynamically appropriate because the un-normalized mutual information is related to the configurational entropy (eq 1), and not a normalized quantity. Furthermore, this behavior is consistent with thermodynamic considerations in intrinsically disordered proteins, where disorder in one or both domains serves to optimize allosteric coupling between the sites.[73] Such allosteric coupling does not require a network of interactions linking the sites. In the present study on interleukin-2, flexible resides at either end of a helix showed couplings > kT in some cases while intervening helical residues did not (Figure 5), because mutual information values, unlike correlation coefficients, are not normalized for residues' self-fluctuations. We note that even normalized pairwise correlations (i.e., Figure 3A) do not include the higher-order correlations that would be expected within semirigid elements, and while the Gaussian Normal Mode results (with default cutoffs) show weak couplings between the N-terminus of helix A and Tyr31 at the C-terminal end of helix A, and between the adaptive loop region (74 and 76−78) and Glu100, these are not visibly distinct features (Figure 3A). In any case, as the goal of our approach is to identify correlations between the conformations of functional sites on a protein, it is not necessary to identify all of the residues that indirectly mediate such correlations, though this is an area for future work. We will later examine coupling between two small molecule binding sites in interleukin-2, which are physically connected by flexible loops and side chains, rather than semirigid secondary structure elements.

Hierarchical Clustering Indentifies Dynamic "Hot-Spots" In Interleukin-2. For a site to be suitable for allosteric inhibitor design, it must be both (1) allosteric, causing shape or flexibility changes at other sites,[1] and (2) druggable, having the right shape and hydrophobicity for drug-like small molecules.[74] Here, our goal was to predict which sites, when perturbed (by ligand binding for example), were most likely to alter structure or dynamics at known functional sites. To search for such sites, we wanted to identify groups of residues responsible for correlations between functional sites.

In biological networks such as protein−protein interaction networks, functional connections between various proteins are preferentially mediated by "hubs" that interact with a greater than average number of partners.[75] Similarly, functional connections between various protein sites are thought to be mediated by "hub" residues or clusters of residues.[25,27] We hypothesized that clusters of residues correlated to many other residues, that is, "dynamical hotspots", could be potential sites or mediators for allosteric modulation of other sites. To find such "dynamical hotspots", we performed a hierarchical clustering of the matrix of mutual information values between residues, in analogy to the analysis of microarray data, using the "heatmap" function in R (http://www.r-project.org/). We used a Euclidean distance metric so that residues showing similar patterns of correlations with other residues are clustered together. Interestingly, one cluster of residues emerged with the strongest correlations within cluster members and the strongest correlations to other residues, and was previously found to be an adaptive region that could bind a number of small-molecule fragments as measured by Tethering experiments.[62] Residues in this cluster are colored red in Figure 6 and mostly reside in the flexible loop between helices 2 and 3, with two in the N-terminal floppy tail and one in the flexible C-terminal loop. Because the mutual information between two torsions is less than either of their self-entropies, it is not surprising that flexible residues often have high mutual information with other residues. This red cluster constitutes a "dynamic hotspot" because it is highly correlated to other clusters of residues. Furthermore, as this red cluster is correlated to the blue cluster containing the IL-2Rα inhibitor binding site, our method predicts the red cluster to be a candidate region for allosteric modulation of the IL-2Rα site. Two similar clusters can also be seen when mutual information values from subsets of the five simulations are block-averaged (Supporting Information Figure 2). However, our approach does not yet predict whether such a site would be druggable by small-molecule allosteric modulators or contain "hotspots" of affinity for protein−protein interactions.

Chemical Shift Perturbations Upon Binding Corroborate Predicted Correlated Motions. While direct experimental methods to identify correlated motions by NMR

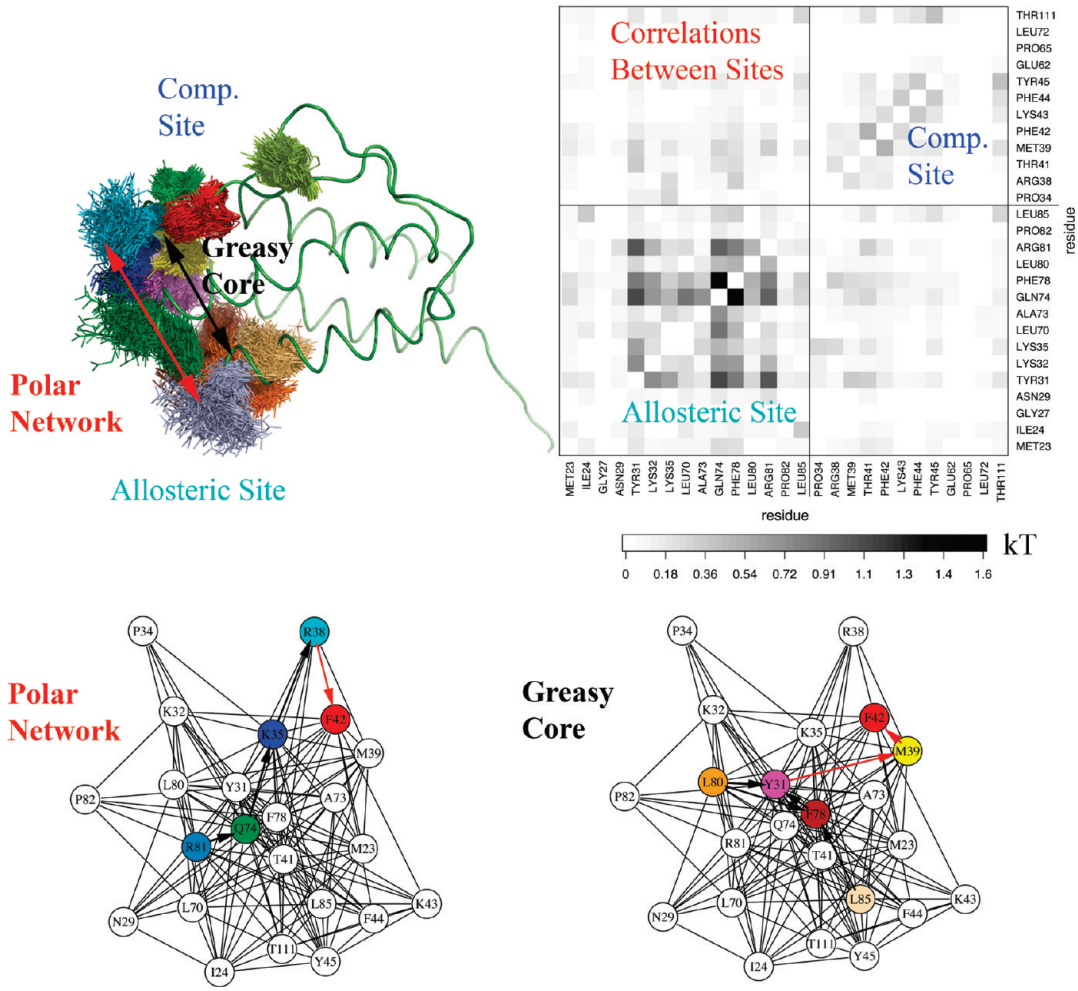


Figure 8. Direct, pairwise correlations couple residues in the IL-2Rα-competitive site (at the IL-2:IL-2Rα interface) to residues in the allosteric fragment-binding site (near the IL-2:IL-2Rβ interface). (Top left) Apo IL-2 is shown in green ribbon while representative conformations of residues showing strong correlations within and between these sites are shown with lines. Overlap between clouds of residues' conformations suggest steric coupling, particularly in the greasy core, from Leu80 (orange) and Ler85 (tan) to Phe78 (brown), to Tyr31 (magenta), to Met39 (yellow), and to Phe42 (red). (Top right) A subset of the full matrix of pairwise correlations reveals direct correlations between residues in the two sites, with the labeled boxes showing correlations within the allosteric site, within the competitive site, and between these two sites. (Bottom) A force-directed network diagram[77] for residues in these sites filtered for correlations of at least 0.05kT shows Phe78, Tyr31, Gln74, and Arg81 as central "hub" residues mediating correlations between the sites.

are limited, one can use chemical shift perturbations and changes in side-chain order parameters of residues outside a binding pocket to identify population shifts in residues within or proximal to allosteric sites that accompany ligand binding. A study by scientists at Roche identified IL-2 residues showing backbone and side-chain amide chemical shift perturbations upon binding IL-2Rα receptor or IL-2Rα-competitive small molecule.[61] For example, Tyr31, Gln74, and Ser75 (in the strong cluster colored "red" in Figure 6) show strong $^{15}N/^{1}H$ chemical shift perturbations following competitive ligand binding, though these residues are not in the competitive binding site. While ring current effects from the ligand's biphenyl group could contribute to these shift perturbations, they are qualitatively consistent with our prediction that these residues' conformations are correlated to the conformation of the IL-2Rα binding site. Furthermore, several residues or regions distal from the IL-2/IL-2Rα interface identified by our approach as highly correlated to the "blue" cluster (encompassing many residues in the IL-2/IL-2Rα interface, PDB 1Z92[76]) showed substantial chemical shift perturbations upon IL-2Rα binding (Figure 7). Unfortunately, resonance overlap restricted the analysis of chemical shift perturbations, notably in most of the flexible loop in the red cluster, so we cannot test our prediction that one would see many perturbations in the flexible loop. It is important to note that none of the nine residues showing insignificant chemical shift perturbations (Asn26, Thr37, Met104, Cys105, Tyr107, Thr113, Ile122)[61] appeared in the red or blue highly correlated clusters.

Communication Between Cooperative Compound Binding Sites Involves a Polar, Solvent-Exposed Network and a Greasy Core. In the previous sections, we used a global description of pairwise couplings to identify putative allosteric sites from correlations between clusters of residues. Presently, we apply our method to study the mechanism of coupling between two given allosteric sites. From our matrix of pairwise correlations between residues in apo-IL2 and

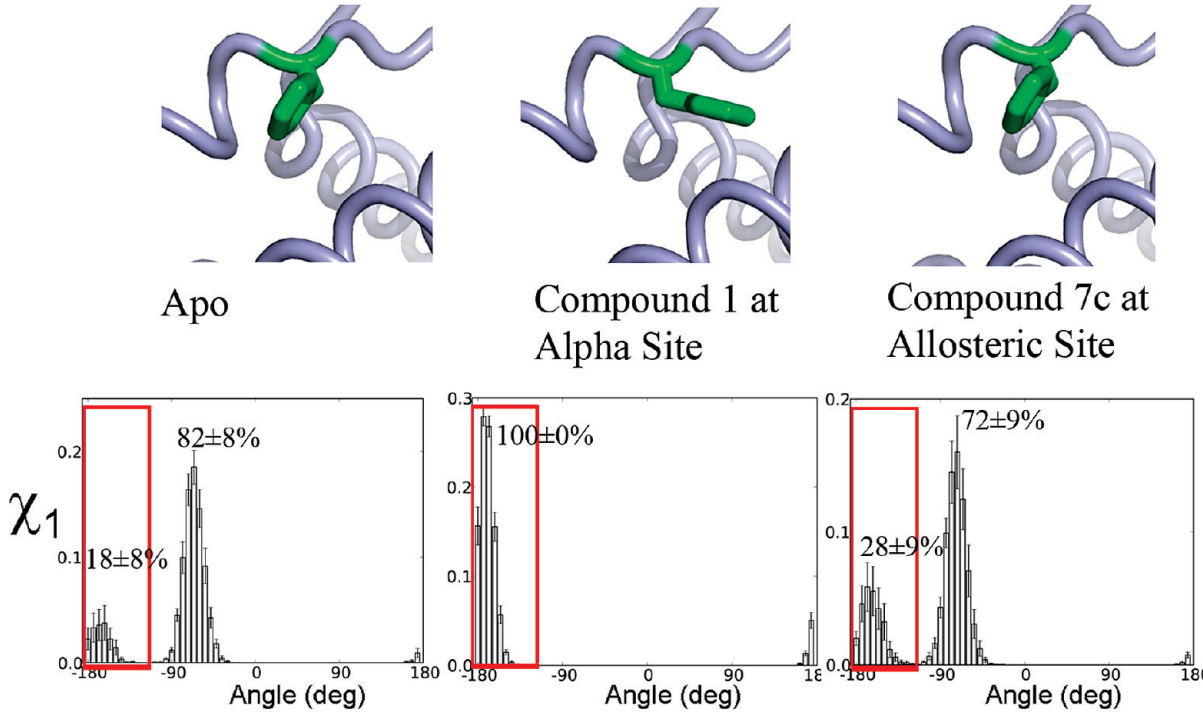


Figure 9. Compound binding to the allosteric site causes a population shift in the conformation of hot-spot residue Phe42 that favors binding compound at the IL-2Rα-competitive site. (Top) Conformations of Phe42 in apo and compound-bound crystal structures (PDB IDs 1M47, 1M48, and 1NBP, respectively). (Bottom) Histograms of Phe42's χ_1 angle from MD simulations. Red boxes highlight the χ_1 population selected by ligand binding at the competitive site.

representative structures from the conformational ensemble, we could infer a structural mechanism by which the two small molecule binding sites might be coupled. In Figure 8 we show matrix elements corresponding to residues in the IL-2Rα-competitive site and residues in the allosteric site, along with representative conformations of these residues. These representative conformers were picked by clustering the MD snapshots according to the rmsd of residues in the red cluster that belonged to the highly flexible loop or were proximal to the N-terminus.

Thermodynamically, the two sites are coupled directly by the off-diagonal gray matrix elements in Figure 8 in the box denoting "Correlations Between Sites". These two sites may also be coupled by higher-order terms involving other residues, which our pairwise analysis does not address (save for the hierarchical clustering which uses patterns of correlation rather than the correlations themselves). From the representative conformations for these residues in Figure 8, the two sites appear to be coupled via a polar network on the protein surface and a greasy core. Two residues are common to both binding sites, namely Lys35 and Met39. The side-chain of Met39, for example, can directly interact with both of the cooperative small molecules, and will be discussed in more detail later. A number of polar side-chains pointing toward the solvent form a network of hydrogen-bonding and electrostatic interactions. In particular, Gln74 (dark green lines) samples a wide swath of conformations, sometimes hydrogen bonding with basic residues in the competitive site (Lys35, dark blue, and Arg38, light blue), and other times hydrogen bonding with basic Arg81 (gray) near the allosteric site. Also, correlations between residues in the greasy core connect hydrophobic surfaces of both compound binding sites; when bound, the allosteric or competitive small molecule would be contiguous with this hydrophobic network. Notably, the matrix elements showing "Correlations Between Sites" indicate that the conformations of residues in the polar network and greasy core are coupled. Thus, a more accurate mechanism for the coupling would be that the two sites are coupled via a polar network, a greasy core, and crosstalk between these elements.

Ligand Binding At Allosteric Site Causes Rotamer Population Shifts That Promote Competitive Site Inhibitor Binding. The preceding analysis suggests that the two sites are connected via correlated motions, and that this could explain the observed allostery. To directly test our hypothesis that the experimentally observed cooperativity between the sites involves subtle population shifts in residues exhibiting correlated motions, we performed additional simulations with a competitive or an allosteric inhibitor bound, and asked whether simulations with the allosteric inhibitor bound would cause population shifts in the competitive site similar to those observed in simulations with the competitive inhibitor. Comparing the crystal structures of apo IL-2 (PDB 1M47) to competitive-site-inhibited IL-2 (PDB 1M48), we note that the motion of two side chains, Met39 and Phe42, opened up a binding groove for the ligand that was closed in the apo structure.

We then asked whether population shifts caused by allosteric ligand binding would help open up the competitive site for competitive ligand binding. To address this question, we examined side-chain torsion angle distribu-

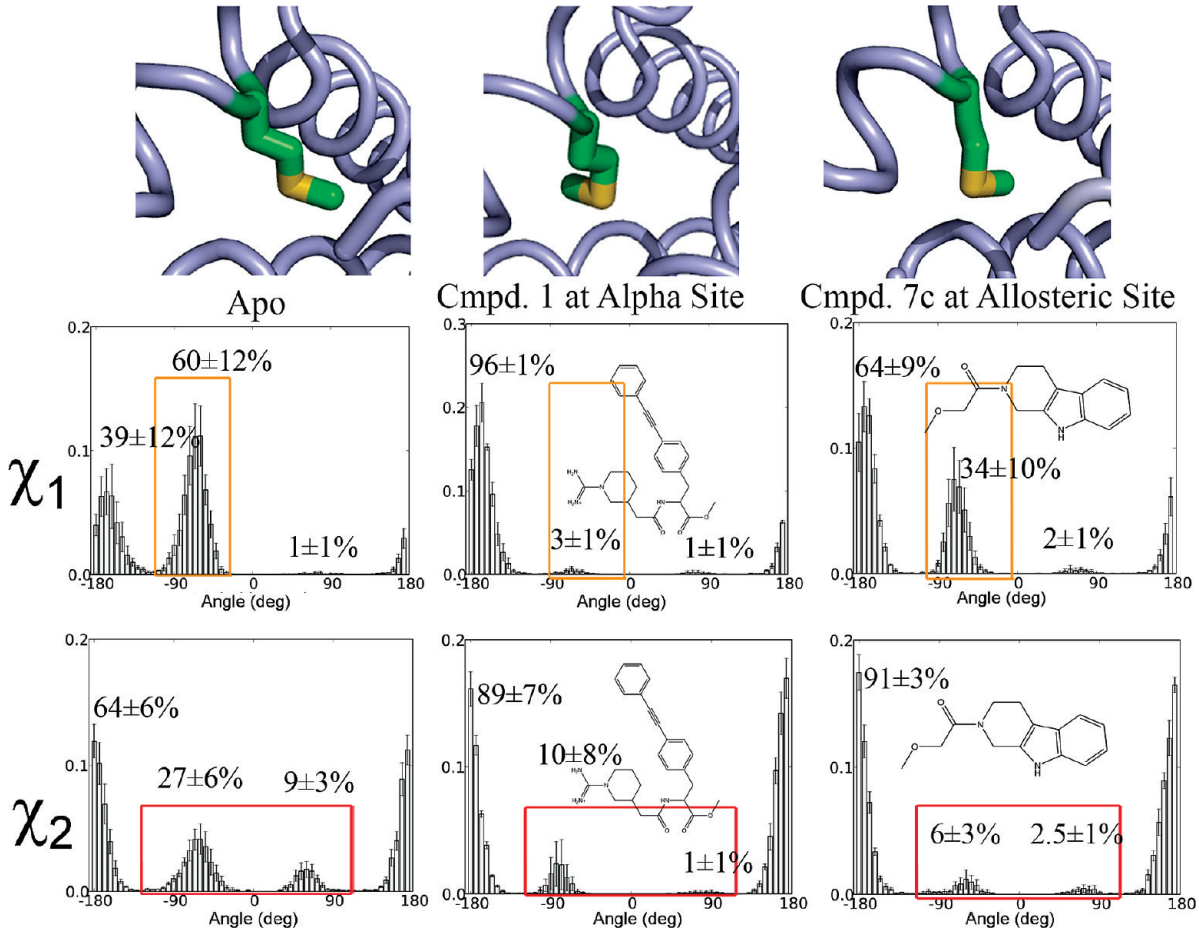


Figure 10. Compound binding to the IL-2Rα site or to the allosteric site selects conformations of Met39 favorable for binding compound at the other site. (Top) Conformations of Met39 in apo and compound-bound crystal structures (PDB IDs 1M47, 1M48, and 1NBP, respectively). (Bottom) Histograms of Met39's χ_1 and χ_2 angles from MD simulations. Orange boxes in χ_1 and red boxes in χ_2 highlight populations suppressed in ligand-bound simulations.

tions for Phe42 and Met39 in simulations with compound at either site and compared these to distributions from the apo simulations (using histogram bins of 6 degrees and 10 ps time intervals). The populations of side chain dihedrals angles in Phe42 and Met39 show substantial differences between the apo protein and the protein bound with competitive and allosteric inhibitors (Figures 9 and 10). Interestingly, the populations observed for the allosteric compound-bound protein are intermediate between apo IL-2 and competitive-site-bound IL-2. Phe42, a hot-spot residue critical for ligand binding and protein binding, adopts a different χ_1 rotamer for ligand binding than it does for protein binding, which is more similar to the apo rotamer (Figure 9). The χ_1 rotamer selected by ligand in competitive-site-bound simulations (100% population) was more populated (89%) in allosteric-bound simulations than in apo simulations (39%), showing a population shift caused by allosteric compound binding.

We predict that Met39 is an important mediator of binding cooperativity because its conformation is correlated to that of Phe42 and because it shows χ_1 and χ_2 population shifts upon allosteric compound binding in the same direction as population shifts from the apo to competitive inhibitor-bound distributions. In crystal structures, Met39 adopts similar conformations in complexes with competitive inhibitor or allosteric inhibitor that both differ from the conformation in the apo structure. Met39 is in an "up" conformation in the apo structure, packed against hot-spot residue Phe42. In competitive inhibitor-bound structures, the side-chain of this Met moves down to slightly enlarge the pocket for a ligand aromatic ring, while in the allosteric-bound structure, the Met side chain moves down to interact weakly with the ligand and fill in part of the hydrophobic pocket opened to accommodate the ligand (Figure 10). Interestingly, this Met39 is not critical for a high-affinity competitive ligand to bind at the competitive site,[78] presumably because mutating it to alanine would simply make that hydrophobic pocket a little larger. However, it is currently unknown whether this residue is required for allosteric ligand binding or for the binding cooperativity, as we predict. Our calculations indicate that cross-talk contributing to cooperativity involves not only the greasy core (of which Met39 and Phe42 are a part) but also the loop between helices 2 and 3 and a polar network involving a number of basic residues on the protein surface. This hypotheses could be further tested by conservative mutations of residues such as Gln74, which is not part of either ligand's binding site, Lys35, whose alkyl tail but not polar head contacts compound 1 in the crystal structure, or Met39, whose

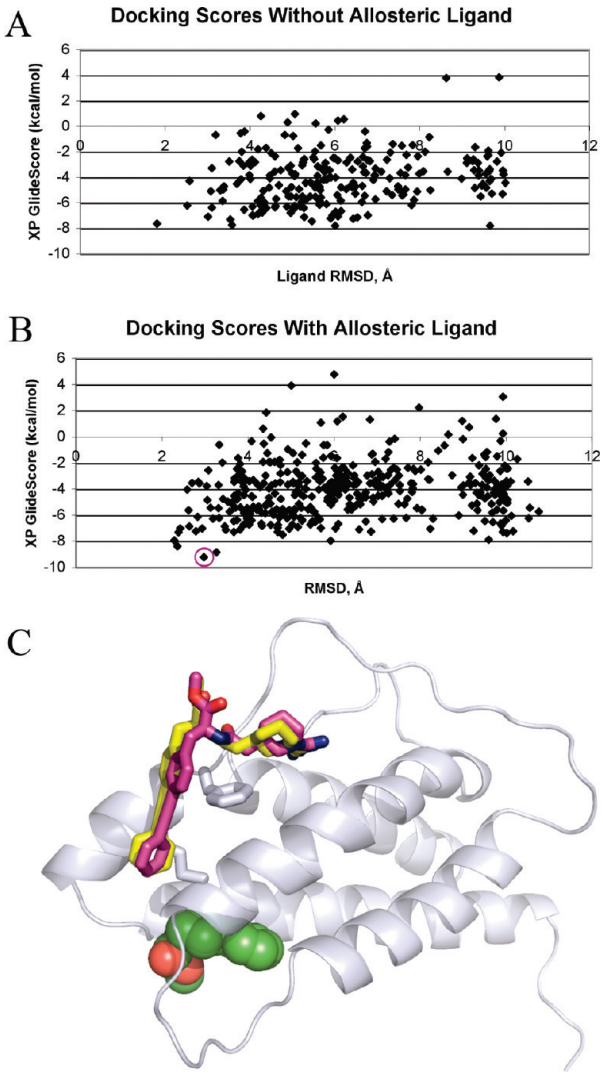


Figure 11. Docking using Glide XP selects a holo-like conformation from an MD ensemble. (A, B) Plots of docking score vs ligand rmsd to the forcefield-minimized cocrystal conformation show that the best-scoring docked poses, from simulations with B but not without A, allosteric compound bound had relatively low rmsd values. The best-scoring pose is circled. (C) Molecular dynamics snapshot from a simulation of IL-2 with bound allosteric fragment is shown with docked (yellow) vs superimposed X-ray (magenta) conformations of a micromolar small molecule inhibitor of IL-2Rα binding. Though the absolute rmsd for this ligand is 2.9 Å (1.6 Å rmsd after fitting), it has a binding mode very similar to that of the crystal ligand.

alanine mutation only shows a slight effect on competitive ligand binding.[79]

Conformational Selection In Silico by Allosteric Ligand. The preceding analysis suggests that binding at the allosteric site can positively modulate binding at the competitive site through changing the distribution of side chain rotamers. To more directly assess the relationship between these relatively subtle changes and ligand binding, we have performed small-molecule docking against snapshots from the MD simulations (Figure 11). Although the scores from docking to MD snapshots cannot be interpreted as accurate binding affinities, we use them as way of qualitatively assessing whether the conformation of the site is appropriate for binding the competitive inhibitor. Because this site consists of many flexible side-chains, rmsd of the competitive binding site residues to the complexed crystal structure was not an appropriate measure of whether the competitive site's conformation was favorable for competitive ligand binding. We found that the best-scoring docked pose roughly superimposes with the crystal ligand (2.8 Å rmsd without fitting, 1.4 Å rmsd with rotational/translational fitting, Figure 11C). The cluster of MD snapshots with the best-scoring docked ligand represented 0.033% of total snapshots. Thus, our relatively short molecular dynamics simulations sampled conformers suitable for binding competitive site ligand at 300 K in the presence of allosteric ligand, enabling us to create a model of the ternary complex (Figure 11C and the Supporting Information).

Conclusions

We have reported novel improvements to mutual information calculations that make them robust enough for relatively short molecular dynamics simulations and have applied our MutInf method to interrogate the mechanism of small molecule binding cooperativity in human interleukin-2. We found better separation of signal from noise in our matrix of correlations between pairs of torsions than in similar matrices that examined backbone C_α correlations. We identified not only local correlations in sequence and in distance space but also long-range correlations. Clustering the matrix of mutual information between residues, we identified a few clusters whose residues showed strong patterns of correlations. Two of these clusters highlighted key functional sites, namely the IL-2Rα-competitive protein interface/inhibitor binding site and a highly flexible loop that has to move to reveal a cryptic binding pocket for the allosteric ligand. Furthermore, we found that the conformations of a number of pairs of residues in these two functional sites were strongly correlated.

As MutInf identified known cooperative binding or functional sites within the top clusters and correlations between them, we believe that this approach will be useful in identifying novel allosteric sites for proteins or for small molecules. For example, we predict potential allostery between the flexible loop surrounding the allosteric compound binding site and the N-terminus of helix 1, the loop region around Glu100, and the C-terminus. Our prediction is further supported by the observation that all of these regions showed significant backbone NMR chemical shift perturbations upon binding of the IL-2Rα receptor.[61] However, the biological roles of these regions are not clear. The C-terminus of IL-2 interacts weakly with the γ_c receptor[75,80,81] ($K_D \approx 0.7$ mM) and independently of IL-2Rα binding. NMR chemical shift perturbations and isoelectric point changes upon addition of methionine to IL-2's N-terminus suggested a potential interaction between the N- and C-termini of apo IL-2 in solution.[82] Intriguingly, Thr3 is a site on human IL-2 that is variably glycosylated;[83] in mice, the N-terminus is longer and displays substantial sequence and glycosylation pattern variability, which in turn impacts IL-2's function in

type I diabetes in mouse models.[84,85] An important caveat is that correlated motions are necessary but not sufficient for biologically relevant allostery.

Though our statistical filtering enables us to find significant correlations in relatively short simulations, in applications where accurate total conformational entropy is desired, multiple longer simulations may be required to obtain absolute total entropy values that all converge to the same value, and higher-order terms might be needed. Nonetheless, our approach is useful in determining which residues or groups of residues show correlated conformations and which residues may mediate crosstalk between functional sites. One caveat is that MutInf focuses on coupled residue conformations rather than on vibrations, and so we may not efficiently capture the role of semirigid elements in mediating correlations between more flexible sites. Though we did not look at motions faster than 1 ps, these are likely not as critical for ligand binding cooperativity in interleukin-2, where the residues linking the sites are primarily in flexible loops and have flexible side-chains. In general, however, such faster-time scale motions can help mediate cooperativity between more flexible sites, and so future work is needed to properly account for these in our approach.

Our calculations suggest that small molecule binding cooperativity in human interleukin-2 involves subtle population shifts and correlated conformations of two binding pockets coupled through a greasy core and a solvent-exposed polar network. New biophysical techniques to directly measure correlated motions by NMR would be useful in testing our predictions about correlated motions that couple allosteric sites.

Acknowledgment. The authors acknowledge J. Wells, M. Gilson, J. Gross, M. Arkin, H. Li, I. Kuntz, M. Kelly, J. Gross, and R. Abel for helpful discussions. The authors also acknowledge M. Hutter for assistance with his Bayesian approach. G.F. thanks T. Kortemme (UCSF) for support, and D.L.M. and H.A. thank Ken Dill (UCSF) for support. C.L.M. thanks J. Nilmeier for inspiration and gratefully acknowledges funding for this work from the UCSF integrated Program in Quantitative Biology fellowship and the UCSF Cancer Research Coordinating Committee. M.P.J. is a consultant to Schrodinger LLC. This work was supported by NIH grant P50-GM082250 to M.P.J.

Supporting Information Available: Figures showing rmsd values for binding site residues in each of the simulations, hierarchical clustering of mutual information values from block-averaged data, and pdb files of the models used in the MD simulations and of the predicted ternary complex. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tsai, C.-J.; del Sol, A.; Nussinov, R. Allostery: Absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.* **2008**, *378* (1), 1–11.

(2) Cui, Q.; Karplus, M. Allostery and cooperativity revisited. *Protein Sci.* **2008**, *17* (8), 1295–1307.

(3) Lindsley, J. E.; Rutter, J. Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (28), 10533–10535.

(4) Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Struct., Funct., Bioinf.* **2004**, *57* (3), 433–443.

(5) Kuriyan, J.; Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **2007**, *450* (7172), 983–990.

(6) Hardy, J. A.; Wells, J. A. Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **2004**, *14* (6), 706–715.

(7) Jeffrey Conn, P.; Christopoulos, A.; Lindsley, C. W. Allosteric modulators of GPCRs: A novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discovery* **2009**, *8* (1), 41–54.

(8) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9* (1), 28–39.

(9) Hardy, J. A.; Wells, J. A. Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **2004**, *14* (6), 706–715.

(10) Shulman, A. I.; Larson, C.; Mangelsdorf, D. J.; Ranganathan, R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* **2004**, *116* (3), 417–429.

(11) Suel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **2003**, *10* (1), 59–69.

(12) Lockless, S. W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999**, *286* (5438), 295–299.

(13) Hatley, M. E.; Lockless, S. W.; Gibson, S. K.; Gilman, A. G.; Ranganathan, R. Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 14445–14450.

(14) Lee, J.; Natarajan, M.; Nashine, V. C.; Socolich, M.; Vo, T.; Russ, W. P.; Benkovic, S. J.; Ranganathan, R. Surface sites for engineering allosteric control in proteins. *Science* **2008**, *322* (5900), 438–442.

(15) Fuentes, E. J.; Gilmore, S. A.; Mauldin, R. V.; Lee, A. L. Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J. Mol. Biol.* **2006**, *364* (3), 337–351.

(16) Page, M. J.; Carrell, C. J.; Di Cera, E. Engineering protein allostery: 1.05 Å resolution structure and enzymatic properties of a Na+-activated trypsin. *J. Mol. Biol.* **2008**, *378* (3), 666–672.

(17) Ming, D.; Wall, M. E. Quantifying allosteric effects in proteins. *Proteins: Struct., Funct., Bioinf.* **2005**, *59* (4), 697–707.

(18) Hawkins, R. J.; McLeish, T. C. B. Coarse-grained model of entropic allostery. *Phys. Rev. Lett.* **2004**, *93* (9), 098104.

(19) Ming, D.; Wall, M. E. Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett.* **2005**, *95* (19), 198103.

(20) Hawkins, R. J.; McLeish, T. C. B. Dynamic allostery of protein α-helical coiled-coils. *J. R. Soc. Interface* **2006**, *3* (6), 125–138.

(21) Zhang, D.; McCammon, J. A. The association of tetrameric acetylcholinesterase with ColQ tail: A block normal mode analysis. *PLoS Comput. Biol.* **2005**, *1* (6), e62.

(22) Chennubhotla, C.; Yang, Z.; Bahar, I. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol. BioSyst.* **2008**, *4* (4), 287–292.

(23) Dengming Ming, M. E. W. Quantifying allosteric effects in proteins. *Proteins: Struct., Funct., Bioinf.* **2005**, *59* (4), 697–707.

(24) Chennubhotla, C.; Bahar, I. Markov propagation of allosteric effects in biomolecular systems: Application to GroEL-GroES. *Mol. Syst. Biol* **2006**, 2.

(25) Daily, M. D.; Upadhyaya, T. J.; Gray, J. J. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Struct., Funct., Bioinf.* **2008**, *71* (1), 455–466.

(26) Daily, M. D.; Gray, J. J. Local motions in a benchmark of allosteric proteins. *Proteins: Struct., Funct., Bioinf.* **2007**, *67* (2), 385–399.

(27) Daily, M. D.; Gray, J. J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.* **2009**, *5* (2), e1000293.

(28) Li, L.; Uversky, V. N.; Dunker, A. K.; Meroueh, S. O. A computational investigation of allostery in the catabolite activator protein. *J. Am. Chem. Soc.* **2007**, *129* (50), 15668–15676.

(29) Bradley, M. J.; Chivers, P. T.; Baker, N. A. Molecular dynamics simulation of the *Escherichia coli* NikR protein: Equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J. Mol. Biol.* **2008**, *378* (5), 1155–1173.

(30) Sayar, K.; Ugur, O.; Liu, T.; Hilser, V.; Onaran, O. Exploring allosteric coupling in the α-subunit of Heterotrimeric G proteins using evolutionary and ensemble-based approaches. *BMC Struct. Biol.* **2008**, *8* (1), 23.

(31) Horstink, L. M.; Abseher, R.; Nilges, M.; Hilbers, C. W. Functionally important correlated motions in the single-stranded DNA-binding protein encoded by filamentous phage Pf3. *J. Mol. Biol.* **1999**, *287* (3), 569–577.

(32) Liu, J.; Nussinov, R. Allosteric effects in the marginally stable von Hippel−Lindau tumor suppressor protein and allostery-based rescue mutant design. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (3), 901–906.

(33) Watney, J. B.; Hammes-Schiffer, S. Comparison of coupled motions in *Escherichia coli* and *Bacillus subtilis* dihydrofolate reductase. *J. Phys. Chem. B.* **2006**, *110* (20), 10130–10138.

(34) Lange, O. F.; Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins: Struct., Funct., Bioinf.* **2006**, *62* (4), 1053–1061.

(35) Lange, O. F.; Grubmüller, H.; de Groot, B. L. Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions. *Angew. Chem., Int. Ed.* **2005**, *44* (22), 3394–3399.

(36) Ming, D.; Cohn, J.; Wall, M. Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct. Biol.* **2008**, *8* (1), 5.

(37) Ho, B. K.; Agard, D. A. Probing the flexibility of large conformational changes in protein structures through local perturbations. *PLoS Comput. Biol.* **2009**, *5* (4), e1000343.

(38) Cooper, A.; Dryden, D. T. F. Allostery without conformational change. *Eur. Biophys. J.* **1984**, *11* (2), 103–109.

(39) Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* **2006**, *13* (9), 831–838.

(40) Lenaerts, T.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L.; Schymkowitz, J.; Rousseau, F. Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure. *BMC Struct. Biol.* **2008**, *8* (1), 43.

(41) Smith, C. A.; Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **2008**, *380* (4), 742–756.

(42) Friedland, G. D.; Linares, A. J.; Smith, C. A.; Kortemme, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.* **2008**, *380* (4), 757–774.

(43) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **2007**, *127* (2), 024107−024116.

(44) Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (20), 7664–7669.

(45) Matsuda, H. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E* **2000**, *62* (3), 3096.

(46) Steuer, R.; Kurths, J.; Daub, C. O.; Weise, J.; Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **2002**, *18* (Suppl 2), S231–240.

(47) Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A* **1988**, *128* (6−7), 369–373.

(48) Roulston, M. S. Estimating the errors on measured entropy and mutual information. *Phys. D* **1999**, *125* (3−4), 285–294.

(49) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comput. Chem.* **2008**, *29* (10), 1605–1614.

(50) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.* **2007**, *28* (3), 655–668.

(51) Karchin, R.; Kelly, L.; Sali, A. Improving functional annotation of non-synonomous SNPs with information theory. *Pac. Symp. Biocomput.* **2005**, 397–408.

(52) George Shackelford, K. K. Contact prediction using mutual information and neural nets. *Proteins: Struct., Funct., Bioinf.* **2007**, *69* (S8), 159–164.

(53) Francois, D.; Rossi, F.; Wertz, V.; Verleysen, M. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing* **2007**, *70* (7−9), 1276–1288.

(54) Hutter, M.; Zaffalon, M. Distribution of mutual information from complete and incomplete data. *Comp. Stat. & Data Anal.* **2005**, *48* (3), 633–657.

(55) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7* (8), 306–317.

(56) Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(57) Sorin, E. J.; Pande, V. S. Exploring the helix−coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **2005**, *88* (4), 2472–2493.

(58) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., and Bioinf.* **2004**, *55* (2), 351–367.

(59) Alexov, E. G.; Gunner, M. R. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **1997**, *72* (5), 2075–2093.

(60) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophys. J.* **2002**, *83* (4), 1731–1748.

(61) Emerson, S. D.; Palermo, R.; Liu, C.-M.; Tilley, J. W.; Chen, L.; Danho, W.; Madison, V. S.; Greeley, D. N.; Ju, G.; Fry, D. C. NMR characterization of interleukin-2 in complexes with the IL-2Rα receptor component, and with low molecular weight compounds that inhibit the IL-2/IL-Rα interaction. *Protein Sci.* **2003**, *12* (4), 811–822.

(62) Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. Binding of small molecules to an adaptive protein−protein interface. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (4), 1603–1608.

(63) Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R. Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2. *Biochemistry (Moscow)* **2003**, *42* (21), 6475–6483.

(64) Junmei, W.; Romain, M. W.; James, W. C.; Peter, A. K.; David, A. C. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.

(65) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132.

(66) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623.

(67) 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. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(68) Berk, H.; Henk, B.; Herman, J. C. B.; Johannes, G. E. M. F. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(69) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Gunsteren, W. F. v.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38* (1−2), 236–240.

(70) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196.

(71) Yang, L.-W.; Rader, A. J.; Liu, X.; Jursa, C. J.; Chen, S. C.; Karimi, H. A.; Bahar, I. oGNM: online computation of structural dynamics using the Gaussian Network Model. *Nucl. Acids Res.* **2006**, *34* (Suppl 2), W24–31.

(72) Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry & Biology*; Garland Science: New York, 2002.

(73) Hilser, V. J.; Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (20), 8311–8315.

(74) Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L.; Chen, W.; Moghaddam, S.; Gilson, M. K.; Kim, K.; Inoue, Y. A synthetic host-guest system achieves avidin-biotin affinity by overcoming enthalpy entropy compensation. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (52), 20737–20742.

(75) Wang, H.; Kakaradov, B.; Collins, S. R.; Karotki, L.; Fiedler, D.; Shales, M.; Shokat, K. M.; Walther, T.; Krogan, N. J.; Koller, D. A complex-based reconstruction of the S. cerevisiae interactome. *Mol. Cell. Proteomics* **2009**, M800490–MCP800200.

(76) Rickert, M.; Wang, X.; Boulanger, M. J.; Goriatcheva, N.; Garcia, K. C. The structure of interleukin-2 complexed with its α receptor. *Science* **2005**, *308* (5727), 1477–1480.

(77) Fruchterman, T. M. J.; Reingold, E. M. Graph drawing by force-directed placement. *Software: Pract. Experience* **1991**, *21* (11), 1129–1164.

(78) Thanos, C. D.; DeLano, W. L.; Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (42), 15422–15427.

(79) Thanos, C. D.; DeLano, W. L.; Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (42), 15422–15427.

(80) Liparoto, S. F.; Myszka, D. G.; Wu, Z.; Goldstein, B.; Laue, T. M.; Ciardelli, T. L. Analysis of the role of the interleukin-2 receptor gamma chain in ligand binding. *Biochemistry* **2002**, *41* (8), 2543–2551.

(81) Stauber, D. J.; Debler, E. W.; Horton, P. A.; Smith, K. A.; Wilson, I. A. Crystal structure of the IL-2 signaling complex: paradigm for a heterotrimeric cytokine receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (8), 2788–2793.

(82) Endo, S.; Yamamoto, Y.; Sugawara, T.; Nishimura, O.; Fujino, M. The additional methionine residue at the N-terminus of bacterially expressed human interleukin-2 affects the interaction between the N- and C-termini. *Biochemistry* **2001**, *40* (4), 914–919.

(83) Robb, R. J.; Smith, K. A. Heterogeneity of human T-cell growth factor(s) due to variable glycosylation. *Mol. Immunol.* **1981**, *18* (12), 1087–1094.

(84) Podolin, P. L.; Wilusz, M. B.; Cubbon, R. M.; Pajvani, U.; Lord, C. J.; Todd, J. A.; Peterson, L. B.; Wicker, L. S.; Lyons, P. A. Differential glycosylation of interleukin 2, the molecular basis for the NOD Idd3 type 1 diabetes gene. *Cytokine* **2000**, *12* (5), 477–482.

(85) Sgouroudis, E.; Albanese, A.; Piccirillo, C. A. Impact of protective IL-2 allelic variants on CD4+Foxp3+ regulatory T cell function in situ and resistance to autoimmune diabetes in NOD mice. *J. Immunol.* **2008**, *181* (9), 6283–6292.

CT9001812

JCTC Journal of Chemical Theory and Computation

Assembly of Viral Membrane Proteins

J. Krüger and W. B. Fischer*

Institute of Biophotonics, School of Medical Science and Engineering, National Yang-Ming University, 155, Section 2, Li-Nong Street, Taipei 112, Taiwan

Received April 16, 2009

Abstract: The generation of computational models is an alternative route to obtain reliable structures for the oligomeric state of membrane proteins. A strategy has been developed to search the conformational space of all possible assemblies in a reasonable time, taking symmetry considerations into account. The methodology tested on M2 from influenza A, shows an excellent agreement with established structures. For Vpu from HIV-1 a series of conformational distinct structures are proposed. For the first time a structural model for a fully assembled transmembrane part of 3a from SARS-CoV is proposed.

Introduction

Membrane proteins represent a huge challenge in terms of experimental and computational structure generation. As the proteins are located at the lipid−protein interface their structure is adapted to this special environment. This special environment has to be taken care of by modern structural biological techniques including computational methods.

The biophysical properties of the lipid membrane are imposed by the topology of its constituents and generate a hydrophobic core flanked by two hydrophilic slabs, the hydrophilic headgroup regions. Despite its complexity this environment confines the dynamics of the proteins mostly into two dimensions. The confinement reduces the conformational search space in computational methods by one dimension and allows for efficient sampling being a highly exhaustive search otherwise.

Membrane proteins from viruses, such as M2 from influenza A,[1−3] Vpu from HIV-1,[4,5] and the more recently discovered 3a protein from SARS-Co virus,[6] which are known to homo-oligomerize, are used to develop a strategy to generate plausible assemblies on an atomic level.

Plausible oligomers for the transmembrane (TM) part of Vpu[7] had been suggested using a global search protocol.[8] In the protocol a limited number of structures are generated and subject to a simulated annealing and energy minimization procedure allowing a significant rearrangement of the initial structures. In a similar approach bundles for M2,[9] Vpu,[10−13] and the monomeric part of p7 from HCV[14] had been generated using simulated annealing combined with short molecular dynamics (MD) simulations. A recent study on M2 from influenza A, glycophorin A. and phospholamban employed a replica exchange approach starting from 16 distinct structures using an implicit membrane approximation.[15] Although the structure optimization algorithm is significantly more sophisticated, still only a partial sampling of the conformational space is possible. The relevance of the monomer conformation for the total energetic was examined among other aspects in a study on the glycophorin dimer.[16] With 324 distinct initial structures only a partial coverage of the conformational space can be assumed. The work of Bowie and co-workers on M2 from influenza A, glycophorin A, and phospholamban as well as other TM proteins evaluates the interaction between two initial TM helices with meticulous Monte Carlo simulations.[17−19] The models are then duplicated around a central symmetry axis to generate larger assemblies.

Mentioned methods probe only a limited number of bundle conformations. In the approach described in this study the search is extended to cover a fine grained range of distances, helical rotation, and variation in tilt angle covering the whole conformational space of the assembly. With this method several hundreds of thousand conformers are obtained for which the potential energy is then calculated and the bundles are ranked accordingly. M2 is taken as a test case to 'validate' the quality of the approach. The study includes a first structural model of the 3a protein from SARS-CoV which has been proposed to have three membrane spanning parts.[6]

* Corresponding author phone: +886 - (0)2 2826 - 7394; fax: +886 - (0)2 2823 - 5460; e-mail: wfischer@ym.edu.tw.

10.1021/ct900185n CCC: $40.75
Published on Web 08/12/2009

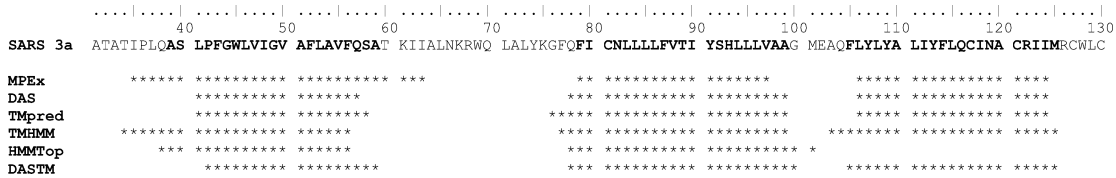


Figure 1. Prediction of the TM parts of 3a from SARS-CoV using different secondary structure prediction programs. The highlighted bold residues reflect the consensus sequence and are used for MD simulations and assembly.

The method can be easily adapted to generate any other membrane protein assembly and thus opens the door for extensive use also in high throughput approaches in proteomics.

Computational Methods

Secondary Structure Prediction and Monomer Modeling. The following ideal helices of $M2_{23-43}$ (SDPLVVAA^{30} SIIGILHLIL^{40} WIL) (see also refs 20 and 21), Vpu_{1-32} (MQPIPIVAIV^{10} ALVVAIIIAI^{20} VVWSIVIIEY^{30} RK),[13,22,23] 3a-$TM1_{39-59}$ (AS^{40} LPFGWLVIGV^{50} AFLAVFQSA), 3a-$TM2_{79-99}$ (FI^{80} CNLLLLFVTI^{90} YSHLLLVAA), and 3a-$TM3_{105-125}$ (FLYLYA^{110} LIYFLQCINA^{120} CRIIM)[6] were generated with backbone dihedrals of $\phi = -65°$ and $\psi = -39°$ using the program MOE (Molecular Operation Environment, www.chemcomp.com).

For the prediction of the TM parts of 3a from SARS-CoV different secondary structure prediction programs were used (Figure 1): Membrane Protein Explorer (MPEx, http://blanco.biomol.uci.edu/mpex/[24]), Dense Alignment Surface prediction of TM regions in proteins (DAS, www.enzim.hu/DAS/DAS.html[25]), TMpred (prediction of transmembrane regions and orientations, www.ch.embnet.org/software/TMPRED_form.html[26]), TMHMM (prediction of transmembrane helices, www.cbs.dtu.dk/services/TMHMM/[27]), and HMMTop (prediction of transmembrane helices and topology of proteins, www.enzim.hu/hmmtop/index.html[28]). They were used with their default setting, and no further adjustments were made. For technical details and algorithm description please refer to the cited literature.

Monomer Equilibration. Prior to any assembly all monomers have been simulated for 10 ns in a fully hydrated POPC bilayer to achieve well equilibrated monomers and to confirm helical stability. The topology for the lipid bilayer (POPC (16:0−18:1 diester PC, 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine) was created on the basis of the parameters of Chandrasekhar et al.[29] The stability of the bilayer was confirmed by a 70 ns MD simulation.[30]

The monomers were inserted into the POPC bilayer, and a stepwise energy minimization and equilibration protocol was used.[30]

All MD simulations were carried out under GROMACS 3.3.2 with the Gromos96 (ffG45a3) force field. The temperature of the peptide, lipid, and the water molecules were separately coupled to a Berendsen thermostat with a coupling time of 0.1 ps. Full isotropic pressure coupling was applied with a coupling time of 1.0 ps and a compressibility 4.5e-5 bar^{-1}. Long range electrostatics were calculated using the particle-mesh Ewald (PME) algorithm with grid dimensions of 0.12 nm and interpolation order 4. Lennard-Jones and short-range Coulomb interactions were cut off at 1.4 and 0.8 nm, respectively.

Assembly. For each of the TM parts of the individual proteins the starting structure for the assembly was the average structure of a principal component analysis (PCA) over the backbone atoms of the whole 10 ns equilibrations. PCA was carried out using the program g_covar from the GROMACS-3.3.2 package. The overall rotational and translational motions were removed by fitting the peptide structure of each time frame to the starting structure.

The following sequences for each of the TM parts were used for the assembly: $M2_{23-43}$ (SDPLVVAA^{30} SIIGILHLIL^{40} WIL), Vpu_{8-26} (AIV^{10} ALVVAIIIAI^{20} VVWSIV), 3a-$TM1_{39-59}$ (AS^{40} LPFGWLVIGV^{50} AFLAVFQSA), 3a-$TM2_{79-99}$ (FI^{80} CNLLLLFVTI^{90} YSHLLLVAA), and 3a-$TM3_{105-125}$ (FLYLYA^{110} LIYFLQCINA^{120} CRIIM). As partial unwinding and strong interaction of the N- and C-terminal residues with the lipid headgroups occurs it was required to shorten the TM parts for the assembly to focus on the main helical core. The truncated residues were not explicitly blocked or protonated and kept neutral.

The helical backbone structure is aligned along the *z*-axis. The absolute rotational orientation was irrelevant for the following steps, but for each data set the same orientation was used to retain its consistence. The homo-oligomeric assembly was considered to be symmetrical toward the central pore axis (C4, C5 symmetry). Multiple copies of the starting helix were placed in the xy-plane with respect to interhelical distance, relative rotational angle, and tilt toward the *z*-axis (here also the membrane normal). The construction of either a trimer, tetramer, or pentamer followed basic geometry with interhelical separation angles of 120°, 90°, and 72°, respectively. The influence of the crossing point, here the point where the xy-plane cuts the starting structure, was also evaluated. To cover weak and tight packing interhelical distances in the range from 8 to 12.5 Å were sampled. Due to symmetry all monomers were rotated around their own helical axis in the same sense with respect to the central pore axis. In the case of hetero-oligomers, e.g. 3a from SARS-CoV it was necessary to sample separate rotation angles for each monomer. As there was no absolute orientation of the monomers with respect to the angle, it was chosen arbitrarily but always in the same way to retain the consistence of the specific data set. A further simplification was to use only one uniform interhelical distance for the 3a trimer.

After each positioning, the side chain atoms were reconstructed with a relative orientation considered as the most probable by the rotational library integrated in MOE. After an energy minimization of not more than 5 steps of either/and steepest descend and conjugated gradient the potential

energy was evaluated according to the united-atom Engh-Huber force field[31] in vacuum without any solvent or lipid present (see the Supporting Information).

To sample the whole conformational space of the bundles each of the degrees of freedom was varied stepwise (interhelical distance 0.05 Å, rotational angle 2°, and tilt 4°). The actual step width for each degree of freedom was evaluated and adjusted with preliminary runs to balance accuracy and performance. For M2 it was possible to limit the angle search to 120° since His-37 and Trp-41 play an important role in the proton conductance through the pore and have to face inward. The tilt search was restricted to 32 to 42° (positive and negative) and a distance restraint between His-37(N_δ) and Trp-41(C_γ) of 3.9 Å was applied, due to experimental evidence.[32,33] For Vpu and especially 3a less data were available so a more extensive search had to be carried out. Depending on constrains on the search space hundreds of thousand different conformers were created each characterized by the set of three or more degrees of freedom and the corresponding individual energy value. In this way for M2 147620, for Vpu 343900, and for 3a 3686058 conformers were generated. The small step size guaranteed high accuracy in determining local minima on the complex high dimensional energy landscape of the assembly process. Further details of the algorithm including a detailed Entity-Relationship-Model (ERM)[34] are available in the Supporting Information.

The simulations were run on a DELL Precision 490n workstation and on facilities of the Paderborn Center for Parallel Computing PC^2 (http://wwwcs.uni-paderborn.de/pc2/). Plots and pictures were generated using xmgrace, VMD, POV-Ray, and MOE.

Results

The multistep method is driven by the full exploration of the conformational space of the assembly of the TM parts. The steps can be described as following:

(i) **TM Prediction:** based on either experimental results or a series of secondary structure prediction programs.

(ii) **Equilibration:** 10 ns MD simulations of the monomer in a fully hydrated lipid bilayer and generation of an averaged structure based on PCA analysis.

(iii) **Assembly:** selection of the core TM spanning part for the bundle assembly and sampling the whole conformational space along the essential degrees of freedom.

Transmembrane Prediction for 3a. Prior to assembly trials the length and amino acid composition of all TM parts of the protein has to be known. The TM parts of M2 and Vpu are experimentally described in the literature (for reviews see refs 35 and 36). 3a from SARS-CoV has been newly discovered, and structural data are still lacking. Therefore a series of TM prediction programs all basing on different algorithms has been used (Figure 1). All predict three TM parts (TM1, TM2, and TM3) of various lengths. The only exception is HMMTop which predicts only the first two parts. The consensus length of the TM parts is calculated to be of 21 amino acids for each of the parts and predicted to have the following sequence: 3a-$TM1_{39-59}$ (AS^{40} LPFGWLVIGV^{50} AFLAVFQSA), 3a-$TM2_{79-99}$ (FI^{80} CNLLLLFVTI^{90} YSHLLLVAA), and 3a-$TM3_{105-125}$ (FLYLYA^{110} LIYFLQCINA^{120} CRIIM).

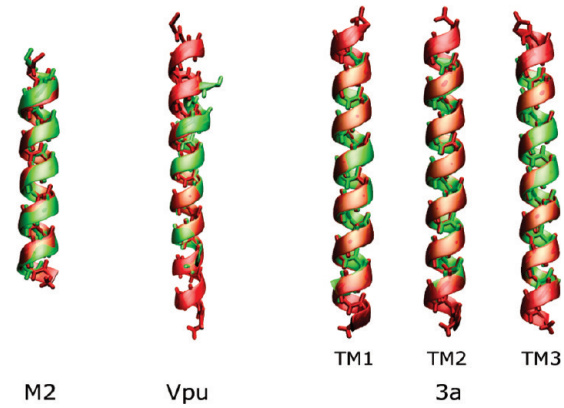


Figure 2. Overlay of the initial ideal helices (red) and the corresponding monomers (green) after MD equilibration which are then used for the assembly protocol.

Assuming a helical secondary structure the length of the predicted TM parts would correspond to a length of 34 Å which is slightly shorter than a typical lipid thickness of DPPC or POPC with 36 Å. As the protein can develop significant tilt angles it may be the case that further residues interact directly with the lipid, specifically the lipid headgroup.

Equilibration. The monomeric proteins of M2, Vpu, and 3a (each of the three membrane spanning parts separately) had been embedded as ideal helices in a fully solvated POPC bilayer and equilibrated for 10 ns. A principal component analysis (PCA) was carried out on each of the data sets. The eigenvectors of the covariance matrices of positional fluctuation give the direction, while eigenvalues quantify the magnitude of the fluctuation. The average structures derived by this method, which are used for the assembly later, reveal that the helical motif remains intact (Figure 2). Some bending and in the case of Vpu also the development of a kink can be observed (Figure 2, see also ref 30). The deviations from the idealized α-helical starting structure are minor but are expected to have an impact on the packing during pore formation. The Root Mean Square Deviation (rmsd) between the starting structure and the averaged PCA structure based on the C_α-atoms lies within the following range: 0.98 (M2), 0.78 (Vpu), 0.30 (3a TM1), 0.43 (3a TM2), and 0.29 (3a TM3).

At the N- and C-termini minor unwinding can occur, due to strong interaction of polar/charged residues with the lipid headgroup (data not shown). To avoid clashes and artificial bumps during the assembly stage only the core portion of each peptide is used. For the 32 residue Vpu peptide the first and last 6 residues are being omitted, finally using 18 residues. M2 and the three membrane spanning parts of 3a have been used in their original length of 20 or 21 residues.

Assembly. The assembly of multiple monomers to form a pore structure has been carried out in the simulation package MOE. Based on its Scientific Vector Language (SVL) existing functions of MOE have been combined that the monomers can be placed in a defined way around the coordinate origin. The distance between the monomers,

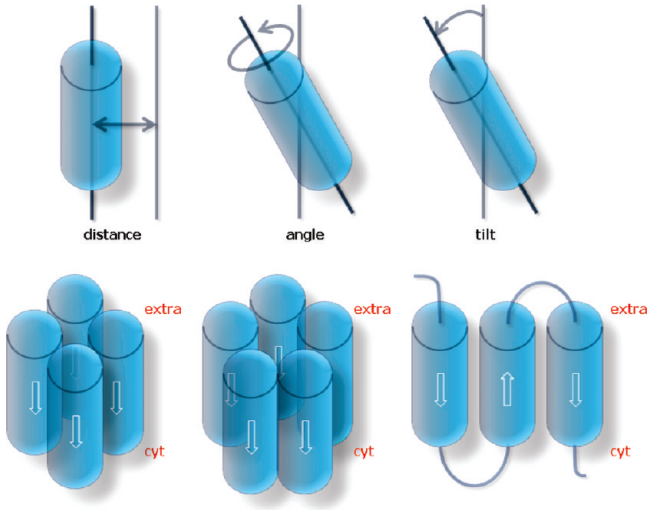


Figure 3. Distance, rotation (angle), and tilt of the TM parts are varied systematically (upper row) to generate e.g. homo tetra- or pentameric bundles of M2 and Vpu and to pack the monomeric trimer of 3a (lower row). White arrows indicate the direction toward the C terminus.

the angle, and the tilt relative to the membrane normal are varied systematically as described in materials and methods (Figure 3).

A major aspect of this approach is to consider homo-oligomers as symmetrical toward their central pore axis (e.g., C4, C5 symmetry). Moreover the dynamics of the monomers are limited to the two-dimensional plane of the lipid bilayer. Therefore it is sufficient to use two degrees of freedom (angle and tilt) to describe the rigid body rotations instead of the usual three Euler angles. These assumptions drastically narrow the search space and enable the creation of reasonable results in an acceptable sampling time, usually hours up to a few days.

Furthermore at this stage it is possible to include restraints, as far as they are known from e.g. NMR experiments, to narrow the search even more.

Distance, Angle, Tilt, and Crossing Point. The distances between packed helices in TM proteins usually show values around 10 Å which is within the range of the 8 to 12.5 Å covered in this study. As each protein has its own characteristics, it is often possible to restrict the distance search to a smaller portion. In the case of the 3a heterotrimer of SARS only one uniform distance was used to simplify the protocol. Preliminary tests have shown that variation in the distance between the three helices is below 0.1 Å and therefore insignificant (data not shown).

The angle for the rotation of each monomer around its own helical axis is sampled for full 360°. Only one value per conformation has to be covered for homo-oligomers, as due to symmetry all monomers are oriented in the same way toward the central pore axis. In cases like M2 from influenza, it is possible to narrow the search space significantly, as it is known from experiments which residues have to be pore lining. His-37 and Trp-41 have been found to play an important role in the proton conductance through the pore[32,33] by facing inward into the lumen of the pore. This allows narrowing the search by 2/3 to 120°. In the case of hetero-oligomers for each nonsymmetrical monomer it is required to sample an individual angle, e.g. 3a from SARS.

The tilt describes the orientation of the helical axis toward the membrane normal. As membrane proteins can develop significant tilts up to 50° it is also required to sample this dimension of the conformational space in a sufficient way. One has to distinguish between left handed helices with negative tilt values and right handed with positive tilt values.

Moving the crossing point did not show major influence on M2 or Vpu assembly. Moving it up or down by 2 Å did not affect the position or the depth of minima in the energy landscape. Considering the symmetric pressure profile in lipid bilayers it seems to be very likely that proteins also have to show similar symmetry. Extreme values for the crossing point would lead to a tepee-like conformation, which is unsymmetrical toward the bilayer and only could be created by the presence of rigid extra membrane parts enforcing such an asymmetry. For the proteins studied in this paper this is considered to not be the case.

Influenza-M2. The calculated data for the assembled pore models of M2 indicate one dominant conformation independent of the usage of a distance restraint (Figure 4). This structure is compared to the two available M2 TM part structures derived from NMR data: 1NYJ, which is described as a presumably closed state of M2,[37] and 2H95, which represents the open state with a bound channel blocking inhibitor.[38]

Comparison of the monomers of the structures reveals a high degree of structural overlap based on the calculated rmsd with respect to the C_α atoms in Å (Table 1). The values are in the range of 0.483 to 0.773. The comparison with the 'closed' structure (1NYJ) shows lower rmsd values of 0.483 and 0.528 than with the 'open' structure (2H95) with rmsd values of 0.773 and 0.735. Internal comparison of the restraint and unrestraint data shows a rmsd value as low as 0.420. It is noteworthy that the two NMR structures differ by 0.789, the highest value in this data set.

All rmsd values for the complete bundles are higher (Table 1). Comparison of the computational generated models and the NMR based structures reveals that the computational models match better with the open ligated NMR model (2.676 with restraints, 2.086 without restraints) than with the closed unligated model (12.765 with restraints, 12.550 without restraints). The difference between the two computed models is the lowest with a value of 1.731 and between the two NMR models the highest with 13.136.

A slight bend found in M2 in the NMR data is reproduced by the computational derived monomer after the 10 ns MD equilibration in POPC (compare Figure 2, further data not shown).

Comparing the equilibrated monomers with experimental structures (1NYJ and 2H95) excellent overlap and very low rmsd values can be observed. When comparing the complete tetrameric pores the values are significantly higher. Taking the size and topology into account, it has to be concluded that rmsd values above 10.0 indicate significant structural differences but not necessarily indicating a dramatically different topology (Figure 5). Although the C_α-rmsd between 1NYJ and the computational models is relatively high with

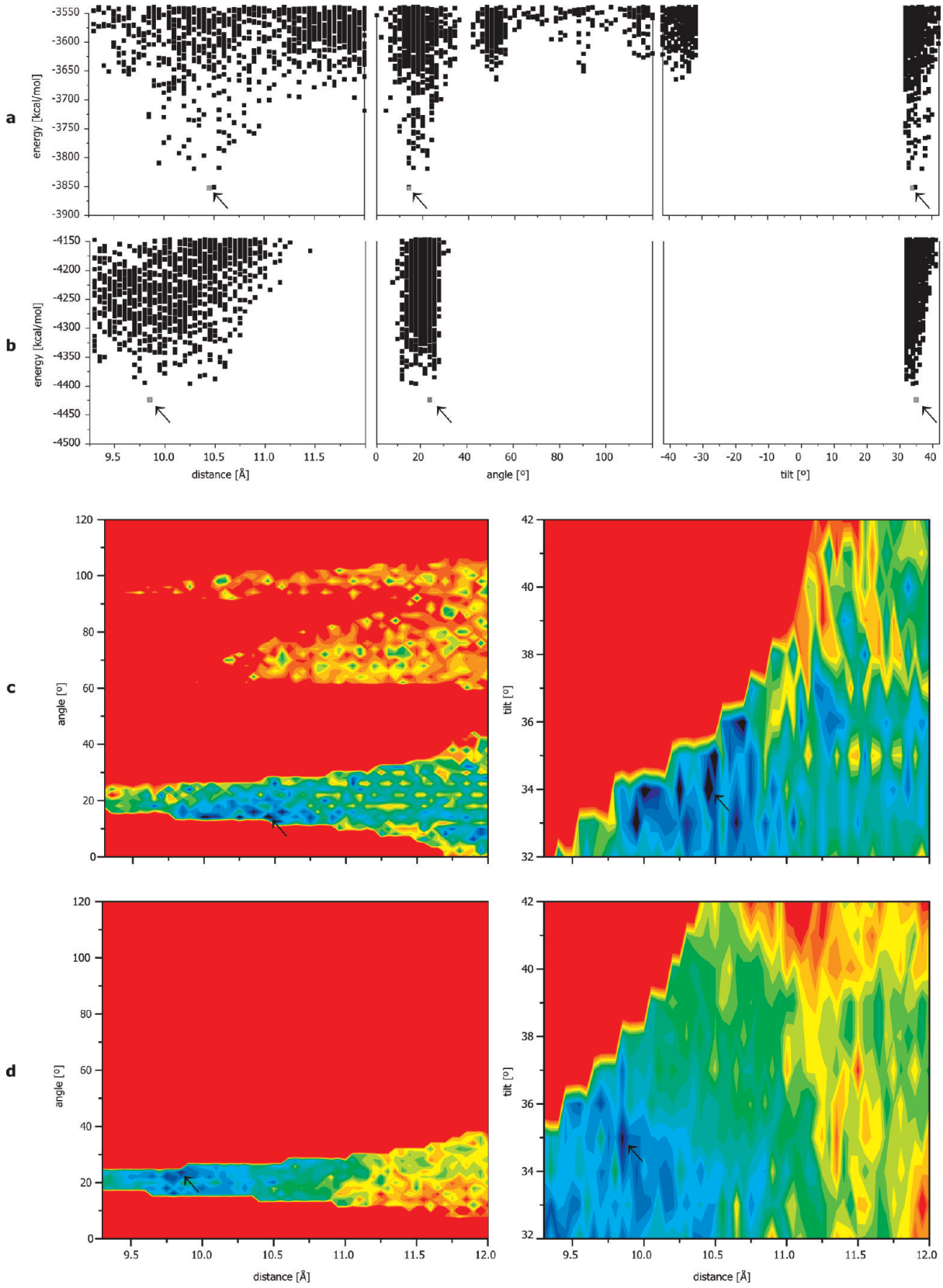


Figure 4. Accumulated energy plots for M2 (a) without restraints and (b) with distance restraint of 3.9 Å between His-37(N_δ) and Trp-41(C_γ). The distance restraint sharpens the energy surface. Comparison of the energy landscapes for the best ranked conformation (arrow) (c) without and (d) with side chain distance restraint for M2 from influenza. The coloring scheme is similar for both cases but covers different energy ranges.

12.765, or respectively 12.550 without constraints, the orientation and shape of the pore are in good agreement. The similarity is even better when compared to the open NMR structure 2H95 (figure not shown). On the molecular scale all amino acids are in places which are supported by experimental evidence and hypothesis.

Table 1. Crosswise Comparison of the C_α-RMSD between Experimental (1NYJ,[37] 2H95[38]) and Computational Structures for the Monomers (Upper Panel) and the Tetrameric Bundles (Lower Panel) of M2 from Influenza

		rmsd monomer		
1NYJ	-			
2H95	0.789	-		
w.[a] restraint	0.483	0.773	-	
no restraint	0.528	0.735	0.420	-
	1NYJ	2H95	w.[a] restraint	no restraint
		rmsd tetramer		
1NYJ	-			
2H95	13.136	-		
w.[a] restraint	12.765	2.676	-	
no restraint	12.550	2.086	1.731	-
	1NYJ	2H95	w.[a] restraint	no restraint

[a] w = with.

The energy contour plots for the three degrees of freedom, distance, angle and tilt are shown in Figure 4. The gap of data in the energy/tilt plot (Figure 4a,b, right graphs) results from the restriction of the search space to 32 to 42° due to experimental evidence.[32,33] The minimum for positive tilt values representing right handed bundle assemblies is dominant. In the angle/energy plot four minima can be identified, with one of them containing a structure with the lowest energy of −3860 kJ/mol. Since the other minima are at significantly higher values and located in more narrow minima, they can be rejected as reasonable 'low-energy' structures. The first rank structure (−3860 kJ/mol) is found at an interhelical distance of 10.5 Å, a rotational angle of 16° with a tilt of 35°. For the bundle with restraints of 3.9 Å between His-37(N_δ) and Trp-41(C_γ) the first rank structure has an energy of −4425 kJ/mol, interhelical distance of 9.8 Å, and angle and tilt values of 25° and 35°. As the application of a restraint alters the potential energy of these bundles in terms of the force field, the resulting single point energies for the restraint and unrestraint bundle cannot be compared directly. It can be stated that the restraint model allows a slightly tighter packing.

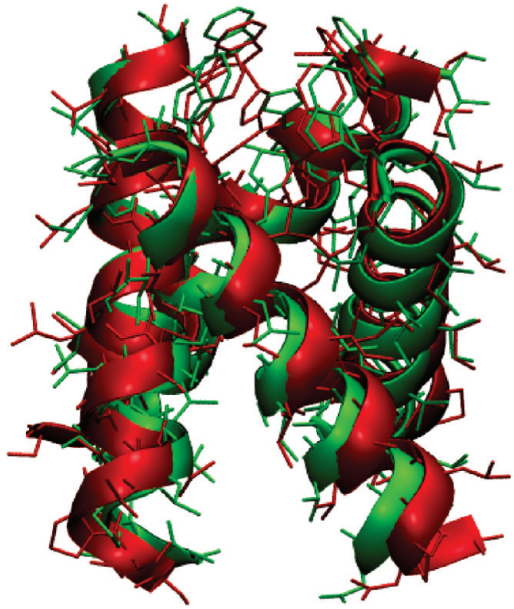


Figure 5. Superposition of 1NYJ (red[37]) with the best rated model according to the assembly protocol (green).

The three-dimensional color coded energy contour map shows that the energy is only favorable within a narrow range of the angle (Figure 4c,d). This is shown by the sharp valley of lower energy values indicated from green to deep blue. For the tilt the green/blue area representing the low energy structures is less sharply defined as for the angle but clearly localizes all low energy structures on the same spot.

HIV1-Vpu. For the different energy plots in Figure 6 several minima are observed each corresponding to a different structure. Especially for the angle values (Figure 6a, middle plot) a characteristic pattern with several minima (around 76, 164, 188, 192, and 326°) is observed. The differences in energy are about 75 kcal/mol between the five best structures, each standing representative for the conformers clustering around it (Figure 6b). These five structures adopt tilts of −28, −16, −4, 16, and 24° (Figure 6a, right plot). In contrast to these large differences the interhelical distance between these conformations is relatively small and ranges from 8.55 to 10.30 Å (Figure 6a, left plot).

With respect to the angle the energy surface expresses narrow valleys which are separated by high energy barriers (Figure 6c). In order to pass from one valley to another, the interhelical distance has to be changed by more than 2 Å. With respect to the tilt the low energy regions are wider and shallower, covering a larger range of the tilt (Figure 6c). Thus, changing the tilt is possible over a larger range by only slightly changing the interhelical distance. It is therefore suggested that a possible gating mechanism is rather via changing the tilt than the angle.

Based on the analysis of the various energy plots the five best structures are shown in Figure 6b. Solely based on the energy it is not possible to favor one structure over the other. Also considering structural aspects like the minimum pore radius or side chain hydrophilicity (data not shown) does not lead to any preference as illustrated by the following examples: Model 1 has the bulky tryptophan's facing into the pore, but with nothing obstructing their move ability to potentially function as gate. In model 4, hydrophilic Ser-24 are facing the lumen of the pore, but they form hydrogen bonds with neighboring carbonyl backbone oxygen's making it unlikely that they take part in any gating mechanism. Thus, further functional analysis is necessary to evaluate the bundles.

SARS-CoV-3a. *The Monomeric Subunit.* Currently, there is no structural information available on an atomistic level. For the present investigation each of the membrane spanning part is considered to be helical. Throughout the 10 ns simulation the helicity of the individual helices remains intact.

In contrast to M2 and Vpu three different angles have to be considered, since there is no symmetry due to the different amino acid composition of each membrane spanning part (TM1, TM2, and TM3). The plot distance versus energy shows an almost linear decrease in energy when the helices approach each other (Figure 7a, first plot). In the range 10.4 to 10.7 Å a minimum is observed. Plotting the angle TM1 versus energy, similar to Vpu several minima, distinct from each other are observed with a dominant minimum at 150° (Figure 7a, second plot). For angle TM2 a single minimum around 0° is observed (Figure 7a, third plot). Rotating TM3

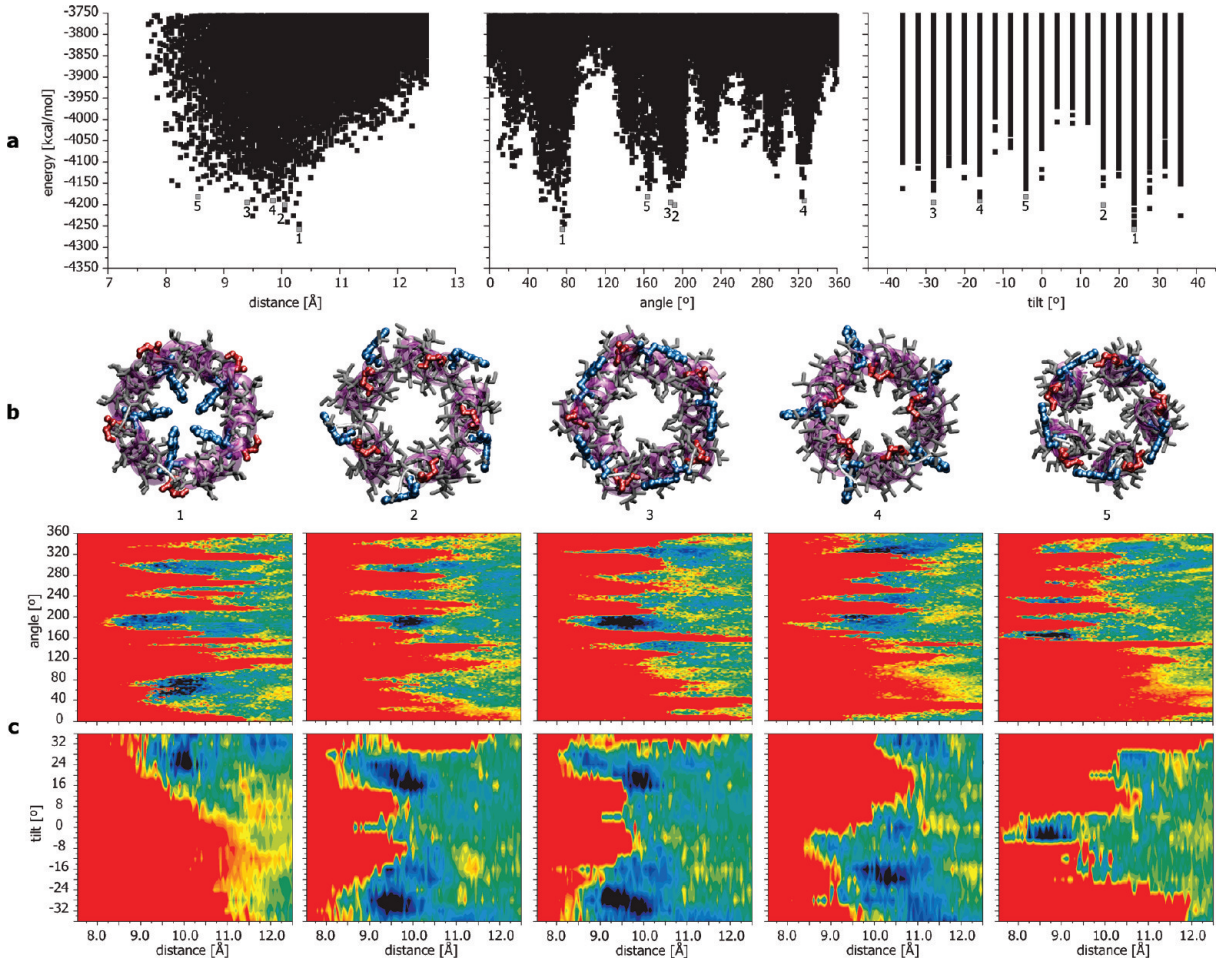


Figure 6. (a) Accumulated plots for distance, rotation, and tilt for Vpu from HIV-1. The contour of the plots indicates the probability of a 'good' conformation. (b) Structural model of the five best models. Trp-23 is highlighted in blue and Ser-24 in red. (c) Energy landscapes for the five most probable conformers of Vpu form HIV-1.

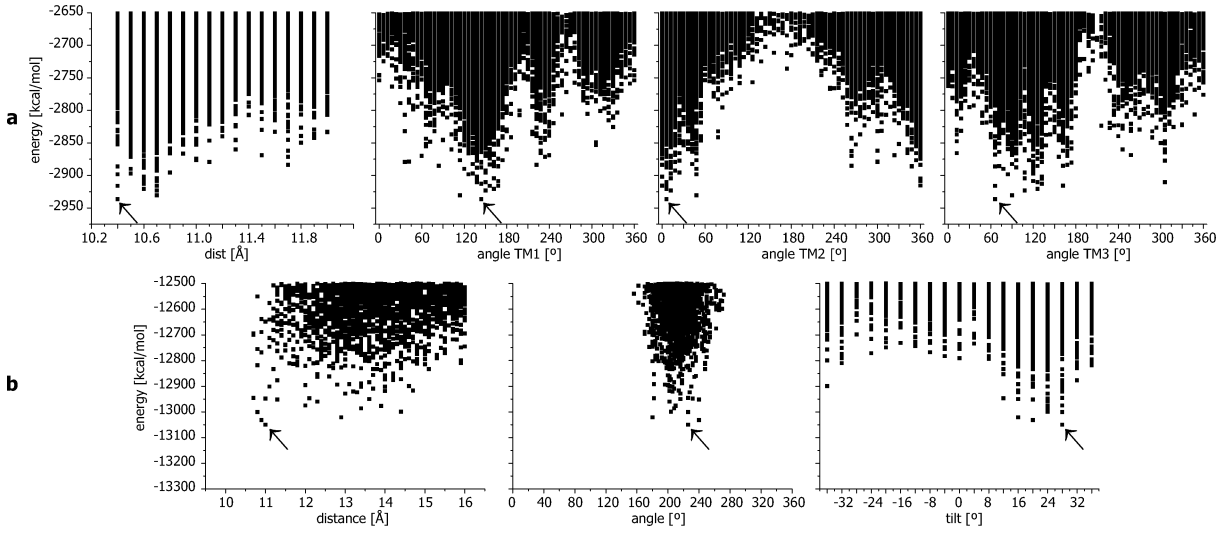


Figure 7. (a) Accumulated plots for distance and rotation for each of the three membrane spanning parts for the monomer of 3a from SARS-CoV. The contour of the plots indicates the probability of a 'good' conformation. (b) Energy plots for the assembly of four subunits of SARS-CoV-3a forming a full pore. Only one clear minimum with respect to the angle is observed. The energy values for the best ranked model (see Figure 8) are marked with an arrow.

proposes a range from 50 to 150° which results in low energy structures (Figure 7a, fourth plot).

Analyzing the dependency on the tilt it has been observed that only small values of 0 ± 2° occur for the low energy structures (plot not shown). Larger values as observed for M2 and Vpu seem to be unlikely as they would decrease the tight packing for the three membrane spanning parts of the monomer.

In summary, it appears that the angle values of TM1 150°, TM2 0°, and TM3 50−150° define the location of the global

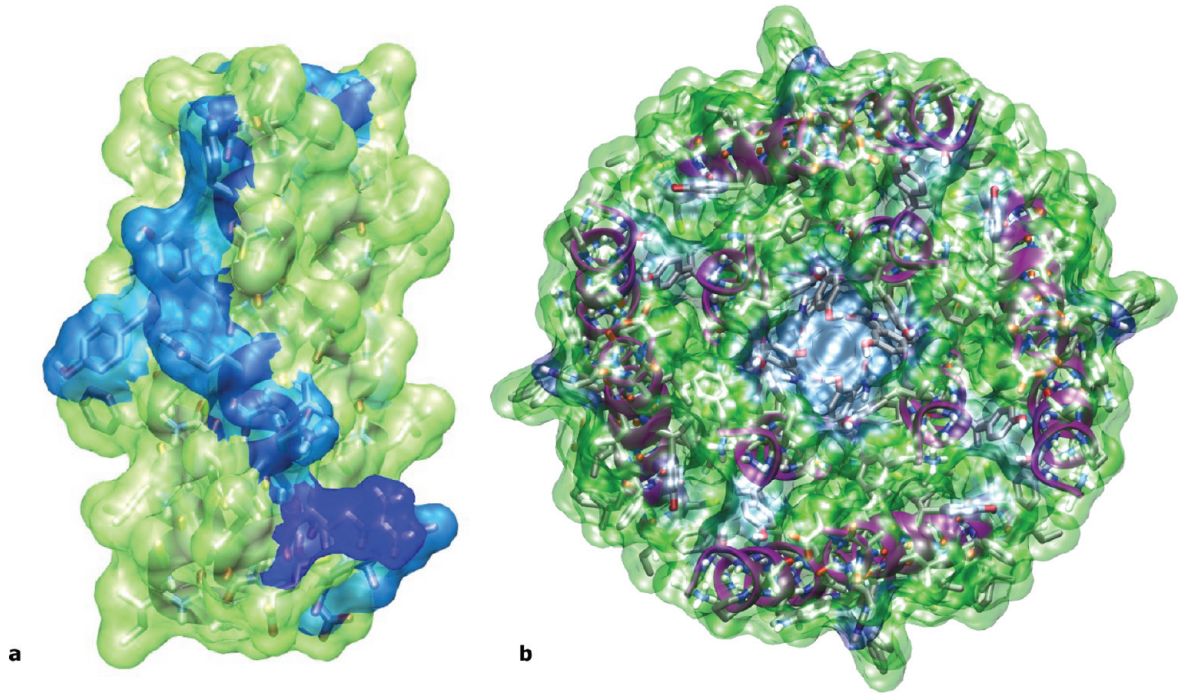


Figure 8. (a) Color coded representation of the best structure of the membrane spanning part of 3a form SARS-CoV. Polar and hydrophilic residues are shown with a light blue surface, ionic residues are shown with a dark blue surface, and hydrophobic residues are shown in green. The hydrophilic stripe (Tyr-109, Tyr-113, Gln-116, Asn-119, and Arg-122) along the structural model can be considered as putative pore lining. (b) Structural model of the fully assembled pore of SARS-CoV-3a. The hydrophilic stripe observed on one subunit lines the pore.

minimum for the monomer of 3a from SARS. The global minimum structure shows a clustering of hydrophilic residues alongside of the monomer assembly. Mapping of these residues at the outside of the 'best structure' of the monomeric trimer indicates a line of Tyr-109, Tyr-113, Gln-116, Asn-119, and Arg-122 in TM3 forming a hydrophilic stripe stretching over the whole TM part (Figure 8a). Other hydrophilic residues such as Thr-89 and Ser-92 of TM2 are buried within the bundle forming hydrogen bonds with neighboring backbone carbonyl oxygen within the same helix. Rotating these residues toward the outside of the monomeric trimer would result in an unfavorable 'high energy' structure. The hydrophilic stripe could be the pore lining part of the ion channel.

Tetramer. To push the assembly protocol to its limits it has been attempted to assemble four trimeric subunits to form a full structural model of the pore of SARS-CoV-3a. The energy contour plots show one minimum at about 26° for the tilt and at 200° for the angle (Figure 7b). This structure corresponds to a tetramer of SARS-CoV-3a where the outer side of the bundle shows a clustering of hydrophobic amino acids, while the putative pore harbors the hydrophilic residues (Figure 8b). In this model Trp-45 and His-93 form a corona at the cytoplasmic side possibly helping anchoring the bundle in the membrane. In addition a series of hydrophobic phenylalanine point outward, facing the hydrophobic tails of surrounding lipid molecules.

Discussion

Validity of the Approach. The assumption is that viral channel forming proteins are produced in the ER as a single unity. Assembly is a consecutive step which then leads to the functional bundles. Between production and assembly the protein has to adopt an equilibrated monomeric structure which then forms the basis of the assembly. At the current state we assume an axial symmetry toward the center of the pore, which has to be adapted to by the average global minimum structure. Therefore this method considers equilibrium structures and does not offer insight into the kinetic pathways leading to a bundle assemble.

The monomer, built as an ideal helix, is due to a 10 ns MD simulation embedded into a fully hydrated POPC lipid bilayer. This system is considered to be a reasonable representation for a membrane environment, although some experimental measurements were published based on notably different conditions. The equilibration reveals helices which can be significantly bend or kinked.[30] This has been reported for other simulations on viral channel forming proteins[9,39] and has been attributed to the electrostatic interactions at the end of the helixes with the lipid head groups. These highly flexible residues may not reflect the global minimum situation and thus using the 'core' TM part is reasonable searching for the bundle structure.

The assembly can be described as a positioning of the monomers with respect to their backbone atoms and a consequent positioning of the side chains. In order to release stress a short minimization is done and the potential energy is calculated. The spatial resolution of positioning with respect to distance, angle, and tilt is extremely small covering finely grained the potential energy surface.

In another search algorithm,[7,40] similar positioning is done prior to a simulated annealing protocol. In that study the CNS

Software[41,42] has been used including the united-atom OPLS parameter set. Large rearrangements of the protein are allowed, while the sampling of the conformational space is very limited. In a modified version the same software has been used with a combined simulated annealing (SA) and short MD simulation protocol.[9−14,43] The conformational search space was narrowed based on the assumption that hydrophilic residues should face the lumen of the pore. In both approaches, the number of potential bundle candidates has then been restricted to be below 300 and 30, respectively.

The replica exchange approach employed by Brooks and co-workers[15] is also limited in the number of starting conformations but has a largely improved chance to reach local minima. Nevertheless only a small part of the conformational space can be sampled by this method. The role of the monomer conformation for the total energetic was put into focus by the group of Lazaridis.[16] The possibility of alternative conformation was explicitly considered. The SA/MD approach used in this study is based on a rotational angle with a stepwidth of 20°.

To this date it remains unclear how monomeric proteins migrate from the ER to their point of action. It has to be considered that different proteins follow different pathways. Bowie and co-workers followed the two-stage-model[44,45] to generate accurate initial dimeric assemblies. But the interaction interface does not have to remain the same for higher oligomeric assemblies. As it is still not known how these higher oligomeric states are reached on a biological and kinetic level it has to be assumed that significant conformational rearrangement occur. We also follow the two-stage model in this study by carefully equilibrating the monomeric subunits. The brief unconstrained energy minimization of the assemblies implicitly takes rearrangements into account, although no statements about kinetic pathways can be made.

The method presented here has its strength in its fine grained sampling. The energy landscape around the proposed equilibrium conformations of the assemblies is mapped thoroughly. Structure optimization steps as mentioned above could easily be implemented into the protocol but are not desired. In comparison to the present approach, any excessive SA or MD simulation steps would smooth out and thereby obscure the fine structure of the energy landscape, consequently missing out low energy structures. Although the energy minimization used in our approach is very short and might contain only partially relived conformations the high conformational sampling density ensures that the energetically 'best' conformations are identified. The decision to carry out this study with a pure united-atom vacuum force field (Engh-Huber) tries to balance accuracy and computational costs. The usage of implicit membrane models as done by others[15,16] does not represent a provable improvement in accuracy at this stage since it will also not account for specific interactions with the lipid head groups or central water columns filling the pores. Therefore the approach with the least number of assumptions was followed using a reasonable protein vacuum force field.

As shown for M2 the presented strategy delivers results matching experimental findings fairly well. Also for Vpu, bundles are found which are similar to the ones suggested earlier based on aforementioned protocols. Therefore it is concluded that a pre-equilibrated monomeric structure is an important step to achieve a good starting position to work on the assembly and that a small step size in positioning is adequate to cover essential aspects of protein assembly.

The Models Proposed. *M2.* There has been considerable evidence in the literature that His-37 and Trp-41 participate in the proton conductance of M2.[32,46,56] Therefore they have to be accessible to water meaning that any conformation where they point into the surrounding lipid can be safely excluded. Limiting the search by 2/3 is a considerable speed up and avoids the potential risk of creating false positive results. Furthermore the employment of NMR based distance restraints has been probed. It has to be noted that they were measured on an unligated M2 pore. It was found that their usage does not improve the already good quality of the proposed structure in this limited search space. But it can be stated that the usage 'sharpens' the energy landscapes. For other proteins the possibility to use experimentally derived restraints might turn out to be more significant.

Comparison of the monomeric and bundle structures from this study with NMR based structures 1NYJ and especially 2H95 shows an excellent agreement, verifying the validity of the approach (Table 1). Describing the two published structures as closed and open states cannot be justified on the basis of our results. Also with the recently published X-ray structure 3BKD[47] on the level of the monomers a good agreement is observed with rmsd values of 1.330−1.495. For comparison, the rmsd among the individual helices within the crystallographic unit cell of 3BKD spreads over a considerable range (0.214−1.130). This indicates that the bended, helical monomer conformation seems to be stable under various conditions, as it is also found in a recent MD study.[48]

A comparison of the bundles delivers a smaller deviation between the computational and the NMR bundle model 1NYJ (2.676) than with the NMR bundle model 2H95 (12.765). For the later the deviation is in the same range as between the two NMR models (13.136). A comparison of the computational model with the bundles within 3BKD was not carried out as their pore conformation is reminiscent of an open umbrella, which most likely represents an energetically costly conformation within a lipid environment.

The computational model is generated without any structural bias induced by the presence of ligands or crystallization agents and therefore may be seen as a very plausible model. Further studies need to be done to associate a model with an 'open' or 'closed' pore.

Vpu. Models 1, 2, 3, and 5 show the hydrophilic residue Ser-24 pointing to the outside of the bundle or been buried between two subunits. Trp-23, the only other hydrophilic residue in the TM part of Vpu, is facing outward and/or interacting with neighboring subunits. These structures suggest the lumen of the pore to be a widely hydrophobic stretch. Model 4 is similar to the one used so far in MD simulations with the TM part of Vpu,[11,49,50] whereas model 1 corresponds to a model which has been suggested earlier.[7,10] In the former model Ser-24 points into the pore, while in the latter it is Trp-23. At this stage based on the

energy values none of the models can be preferred above any other. This needs further functional *in silico* evaluation such as longer MD simulations to assess the stability of the bundles and simulation of ion permeation through the respective pores. In the present study Ser-24 forms an intrahelical hydrogen bond with the backbone carbonyl oxygen of Ile-20 in all models.

The conformational transition between two models requires variation of the distance, angle, and tilt 'walking on the energy surface'. With respect to the energy plots for the angle (Figure 6c, upper row) a huge energy barrier would have to be crossed while changing the angle and increasing the distance by more than 2 Å to move from one minimum to another. This makes it very unlikely that conformational transitions include huge rotational movements of the individual subunits. Focusing on models 2, 3, and 5 it seems to be more likely that the tilt changes, while the distance and angle would only be varied to a minor degree (Figure 6c, lower row). It can be proposed that models 2 and 3 are ion conducting and can change their conformation toward model 5 a potentially closed state. The models 2 and 3 would represent alternative conducting states, which would be in good agreement with the experimental finding of multiple conductance states for Vpu.[51−53] More than just one model could contribute to the functioning of Vpu. The results further underline the flexibility of the TM part of the protein.[30]

SARS-CoV. The sequence based TM prediction is an established technique. Nevertheless some deviations between the different protocols can be observed. Creating a consensus[54] between the six different techniques used in this study leads to a robust and reliable prediction.

It is noteworthy that the Cys-bridges reported to link the subunits are located in the extramembrane part of the protein, presumably not directly affecting the TM part. The assembly of the monomeric unit results in a profound model with hydrophilic residues clustering on one side. The pore assembly into a putative bundle leads to the first structural model of 3a from SARS with residues such as tyrosines, glutamine, asparagines, and arginine from TM3 lining the lumen. This motif is rather unusual as more commonly serines and tyrosines are suggested for pore lining residues ion in channels.[55]

To screen the 'whole' conformational space of an assembly represents an auspicious approach for both experimentally and computationally based studies. It leads to reliable structural models, helps to avoid structural pitfalls, and opens insights into mechanistic details of the mode of action and can help revealing alternative conformations. The fine grained full search approach is the most direct route for exploring the conformational space of a protein assembly. By simplifying and considering the symmetry of the studied proteins a significant confinement of the search space can be made, without biasing toward a certain result. This enables the resolvability of the search in an acceptable sampling time.

The quality of the constructed structural models does not rank behind any experimental technique. Based on this method alternative configurations of pentameric Vpu have been shown, and a novel pore lining motif is suggested for the bundle model of 3a. Regarding the topology of all of the studied energy landscapes, it has to be concluded that conformational transitions from any open to closed states would have to take place by variation of the tilt and not the angle. As a further quality check it is recommended to do functional studies on the most plausible models suggested, e.g. assessing the bundle stability in a lipid environment with consequent multins MD simulations or to do cross mutations verifying explicit interactions between the monomers in the models.

Acknowledgment. W.B.F. thanks the NYMU and the government of Taiwan for financial support (Aim of Excellence Program). This work was supported by the National Science Council of Taiwan (NSC). J.K. acknowledges a fellowship granted by the Alexander von Humboldt-Foundation and the NSC. We thank the Paderborn Center for Parallel Computing PC^2 (http://wwwcs.uni-paderborn.de/pc2/) for providing computer time.

Supporting Information Available: Detailed description of the placement algorithm including an ERM for all consecutive steps of the assembly. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Winter, G.; Fields, S. *Nucleic Acids Res.* **1980**, *8*, 1965.

(2) Allen, H.; McCauley, J.; Waterfield, M.; Gethering, M. *Virology* **1980**, *107*, 548.

(3) Lamb, R. A.; Choppin, P. W. *Virology* **1981**, *112*, 729.

(4) Strebel, K.; Klimkait, T.; Martin, M. A. *Science* **1988**, *241*, 1221.

(5) Cohen, E. A.; Terwilliger, E. F.; Sodroski, J. G.; Haseltine, W. A. *Nature* **1988**, *334*, 532.

(6) Lu, W.; Zheng, B.-J.; Xu, K.; Schwarz, W.; Du, L.; Wong, C. K. L.; Chen, J.; Duan, S.; Deubel, V.; Sun, B. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12540.

(7) Kukol, A.; Arkin, I. T. *Biophys. J.* **1999**, *77*, 1594.

(8) Adams, P. D.; Arkin, I. T.; Engelman, D. M.; Brünger, A. T. *Nature Struc. Biol.* **1995**, *2*, 154.

(9) Forrest, L. R.; Kukol, A.; Arkin, I. T.; Tieleman, D. P.; Sansom, M. S. *Biophys. J.* **2000**, *78*, 55.

(10) Cordes, F.; Kukol, A.; Forrest, L. R.; Arkin, I. T.; Sansom, M. S. P.; Fischer, W. B. *Biochim. Biophys. Acta* **2001**, *1512*, 291.

(11) Cordes, F. S.; Tustian, A.; Sansom, M. S. P.; Watts, A.; Fischer, W. B. *Biochemistry* **2002**, *41*, 7359.

(12) Lemaitre, V.; Ali, R.; Kim, C. G.; Watts, A.; Fischer, W. B. *FEBS Lett.* **2004**, *563*, 75.

(13) Kim, C. G.; Lemaitre, V.; Watts, A.; Fischer, W. B. *Anal. Bioanal. Chem.* **2006**, *386*, 2213.

(14) Patargias, G.; Zitzmann, N.; Dwek, R.; Fischer, W. B. *J. Med. Chem.* **2006**, *49*, 648.

(15) Bu, L.; Im, W.; Brooks, C. L., III *Biophys. J.* **2007**, *92*, 854.

(16) Mottamal, M.; Zhang, J.; Lazaridis, T. *Proteins* **2006**, *62*, 996.

(17) Kim, S.; Chamberlain, A. K.; Bowie, J. U. *J. Mol. Biol.* **2003**, *329*, 831.

(18) Faham, S.; Yang, D.; Bare, E.; Yohannan, S.; Whitelegge, J. P.; Bowie, J. U. *J. Mol. Biol.* **2004**, *335*, 297.

(19) Bowie, J. U. *Nature* **2005**, *438*, 581.

(20) Duff, K. C.; Kelly, S. M.; Price, N. C.; Bradshaw, J. P. *FEBS Lett.* **1992**, *311*, 256.

(21) Kovacs, F. A.; Cross, T. A. *Biophys. J.* **1997**, *73*, 2511.

(22) Lemaitre, V.; Kim, C. G.; Fischer, D.; Lam, Y. H.; Watts, A.; Fischer, W. B. In *Viral membrane proteins: structure, function and drug design*; Fischer, W. B., Ed.; Kluwer Academic/Plenum Publishers: New York, 2005; Vol. 1, p 187.

(23) Candler, A.; Featherstone, M.; Ali, R.; Maloney, L.; Watts, A.; Fischer, W. B. *Biochim. Biophys. Acta* **2005**, *1716*, 1.

(24) Jaysinghe, S.; Hristova, K.; Wimley, W.; Snider, C.; White, S. H. 2006. http://blanco.bimol.uci.edu/mpex.

(25) Cserzo, M.; Eisenhaber, F.; Eisenhaber, B.; Simon, I. *Protein Eng.* **2002**, *15*, 745.

(26) Hofmann, K.; Stoffel, W. *Biol. Chem.* **1993**, *374*, 166.

(27) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. L. *J. Mol. Biol.* **2001**, *305*, 567.

(28) Tusnády, G. E.; Simon, I. *Bioinformatics* **2001**, *17*, 849.

(29) Chandrasekhar, I.; Kastenholz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; van Gunsteren, W. F. *Eur. Biophys. J.* **2003**, *32*, 67.

(30) Krüger, J.; Fischer, W. B. *J. Comput. Chem.* **2008**, *29*, 2416.

(31) Engh, R. A.; Huber, R. *Acta Crystallogr. Sect. A: Found. Crystallogr.* **1991**, 47.

(32) Hu, J.; Fu, R.; Nishimura, K.; Zhang, L.; Zhou, H.-X.; Busath, D. D.; Vijayvergiya, V.; Cross, T. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 6865.

(33) Wang, J.; Kim, S.; Kovacs, F.; Cross, T. A. *Protein Sci.* **2001**, *10*, 2241.

(34) Chen, P. P.-S. *ACM Trans. Database Systems* **1976**, *1*, 9.

(35) Fischer, W. B.; Sansom, M. S. P. *Biochim. Biophys. Acta* **2002**, *1561*, 27.

(36) *Viral membrane proteins: structure, function and drug design*; Fischer, W. B., Ed.; Kluwer Academic/Plenum Publisher: New York, 2005; Vol. 1, p 291.

(37) Nishimura, K.; Kim, S.; Zhang, L.; Cross, T. A. *Biochemistry* **2002**, *41*, 13170.

(38) Hu, J.; Asbury, T.; Achuthan, S.; Bertram, R.; Quine, J. R.; Fu, R.; Cross, T. A. *Biophys. J.* **2007**, *92*, 4335.

(39) Fischer, W. B.; Forrest, L. R.; Smith, G. R.; Sansom, M. S. P. *Biopolymers* **2000**, *53*, 529.

(40) Kukol, A.; Adams, P. D.; Rice, L. M.; Brunger, A. T.; Arkin, I. T. *J. Mol. Biol.* **1999**, *286*, 951.

(41) Brunger, A. T. *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR*; Yale University Press: New Haven, CT, 1992.

(42) Brunger, A.; Adams, P.; Clore, G.; Gros, W.; Grosse-Kunstleve, R.; Jiang, J.; Kuszewski, J.; Nilges, M.; Pannu, N.; Read, R.; Rice, L.; Simonson, T.; Warren, G. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 905.

(43) Kerr, I. D.; Doak, D. G.; Sankararamakrishnan, R.; Breed, J.; Sansom, M. S. P. *Protein Eng.* **1996**, *9*, 161.

(44) Popot, J.-L.; Engelman, D. M. *Biochemistry* **1990**, *29*, 4031.

(45) Engelman, D. M.; Chen, Y.; Chin, C.-N.; Curran, A. R.; Dixon, A. M.; Dupuy, A. D.; Lee, A. S.; Lehnert, U.; Matthews, E. E.; Reshetnyak, Y. K.; Senes, A.; Popot, J.-L. *FEBS Lett.* **2003**, *555*, 122.

(46) Tian, C.; Tobler, K.; Lamb, R. A.; Pinto, L. H.; Cross, T. A. *Biochemistry* **2002**, *41*, 11294.

(47) Stouffer, A. L.; Acharya, R.; Salom, D.; Levine, A. S.; Di Constanzo, L.; Soto, C. S.; Tereshko, V.; Nanda, V.; Stayrook, S.; DeGrado, W. F. *Nature* **2008**, *451*, 596.

(48) Yi, M.; Cross, T. A.; Zhou, H.-X. *J. Phys. Chem. B* **2008**, *112*, 7977.

(49) Grice, A. L.; Kerr, I. D.; Sansom, M. S. P. *FEBS Lett.* **1997**, *405*, 299.

(50) Moore, P. B.; Zhong, Q.; Husslein, T.; Klein, M. L. *FEBS Lett.* **1998**, *431*, 143.

(51) Schubert, U.; Bour, S.; Ferrer-Montiel, A. V.; Montal, M.; Maldarelli, F.; Strebel, K. *J. Virol.* **1996**, *70*, 809.

(52) Mehnert, T.; Lam, Y. H.; Judge, P. J.; Routh, A.; Fischer, D.; Watts, A.; Fischer, W. B. *J. Biomol. Struct. Dyn.* **2007**, *24*, 589.

(53) Mehnert, T.; Routh, A.; Judge, P. J.; Lam, Y. H.; Fischer, D.; Watts, A.; Fischer, W. B. *Proteins* **2008**, *70*, 1488.

(54) Cuthbertson, J. M.; Doyle, D. A.; Sansom, M. S. P. *Prot. Eng. Des. Sel.* **2005**, *18*, 295.

(55) Akabas, M. H.; Kaufmann, C.; Archdeacon, P.; Karlin, A. *Neuron* **1994**, *13*, 919.

(56) Schnell, J. R.; Chou, J. J. *Nature* **2008**, *451*, 591.

CT900185N

JCTC Journal of Chemical Theory and Computation

Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields

Eva Fadrná,[†] Nad'a Špačková,[‡] Joanna Sarzyñska,[§] Jaroslav Koča,[†] Modesto Orozco,[∥] Thomas E. Cheatham III,[⊥] Tadeusz Kulinski,[§] and Jiří Šponer*,[‡]

National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic, Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61 704 Poznań, Poland, Joint IRB-BSC program on Computational Biology, Institute for Research in Biomedicine, Baldiri Reixac 10-12, 08028 Barcelona, Spain, Barcelona Supercomputing Center, Jordi Girona 29, 08034 Barcelona, Spain, Department of Biochemistry, University of Barcelona, Diagonal 647, 08028 Barcelona, Spain, and Departments of Medicinal Chemistry and of Pharmaceutics and Pharmaceutical Chemistry, University of Utah, 30 South 2000 East, Salt Lake City, Utah 84112

Received April 24, 2009

Abstract: We have carried out a set of explicit solvent molecular dynamics (MD) simulations on two DNA quadruplex (G-DNA) molecules, namely the antiparallel $d(G_4T_4G_4)_2$ dimeric quadruplex with diagonal loops and the parallel-stranded human telomeric monomolecular quadruplex d[AGGG(TTAGGG)$_3$] with three propeller loops. The main purpose of the paper was testing of the capability of the MD simulation technique to describe single-stranded topologies of G-DNA loops, which represent a very challenging task for computational methods. The total amount of conventional and locally enhanced sampling (LES) simulations analyzed in this study exceeds 1.5 μs, while we tested several versions of the AMBER force field (parm99, parmbsc0, and a version with modified glycosidic χ torsion profile) and the CHARMM27 force field. Further, we compared minimal salt and excess salt simulations. Postprocessing MM-PBSA (Molecular Mechanics, Poisson−Boltzmann, Surface Area) free energy calculations are also reported. None of the presently available force fields is accurate enough in describing the G-DNA loops. The imbalance is best seen for the propeller loops, as their experimental structure is lost within a few ns of standard simulations with all force fields. Among them, parmbsc0 provides results that are clearly closest to the experimental target values but still not in full agreement. This confirms that the improvement of the γ torsional profile penalizing the γ trans substates in the parmbsc0 parametrization was a step in the right direction, albeit not sufficient to treat all imbalances. The modified χ parametrization appears to rigidify the studied systems but does not change the ultimate outcome of the present simulations. The structures obtained in simulations with the modified χ profile are predetermined by its combination with either parm99 or parmbsc0. Experimental geometries of diagonal loops of $d(G_4T_4G_4)_2$ are stable in standard simulations on the ~10 ns time scale but are becoming progressively lost in longer and LES simulations. In addition, the $d(G_4T_4G_4)_2$ quadruplex contains, besides the three genuine binding sites for cations in the channel of its stem, also an ion binding site at each stem-loop junction. This arrangement of five cations in the quadruplex core region is entirely unstable in all 24 simulations that we attempted. Overall, our results confirm that G-DNA loops represent one of the most difficult targets for molecular modeling approaches and should be considered as reference structures in any future studies aiming to develop or tune nucleic acids force fields.

10.1021/ct900200k CCC: $40.75
Published on Web 08/19/2009

Introduction

Guanine rich sequences are known to occur in many positions of the genome and are especially common at the chromosome ends, the telomers. Such sequences readily form four stranded guanine quadruplex structures (G-DNA) *in vitro*, and it is believed that the same could happen with telomeric sequences (such as the human one: $d(TTAGGG)_n$),[1] where repetitive double stranded segments are followed by single strand overhangs of the same repeat. Such overhangs can fold back defining a quadruplex, whose formation could protect them from being accessed by reverse transcriptase enzyme telomerase. Telomerase is active in most cancer lines and contributes to their immortality by maintaining the length of the overhang.[2] Thus, it has been suggested that compounds stabilizing G-DNA *in vivo* can act as anticancer drugs.[3,4] There are also other biological or pharmacological roles that have been suggested for G-DNA[5] which together with the increasing number of applications of the quartet structures in nanosciences[6,7] makes G-DNA the most important noncanonical DNA architecture, explaining the very intense research focused in the last years in this molecule.[8−13]

The basic structural element of G-DNA is a quartet involving four cyclically bonded guanines that are interconnected by hydrogen bonds (Figure 1a). The quartets are further stabilized by monovalent ions placed along the central channel of the structure and interacting with O6 atoms of the guanines, compensating the highly electronegative electrostatic potential region in the quartet centers (Figure 1a). Several (usually 2−4) consecutive quartets form a G-DNA stem and the cations fill its channel, being either in planes of the quartets or in the cavities between them. The stem can be created by 1, 2, or 4 separate DNA pieces and thus there are intramolecular (unimolecular), dimeric (bimolecular), or tetramolecular quadruplexes. The adjacent strands of the G-DNA stem may run either in parallel or antiparallel fashion. The guanines are in anti orientation in the all-parallel stems while antiparallel arrangements have to utilize also some syn nucleotides. The stems of bimolecular and intramolecular quadruplexes are supplemented by single stranded (mostly thymine-rich) loops which are formed by the nucleotides interdispersed between the guanine stretches forming the stem strands. The loops can be placed above the planes of the terminal quartets of the stem, and then they can link either adjacent or diagonal guanines, resulting into lateral or diagonal (Figure 1b) loop arrangements. Alternatively, the loops can run across the G-DNA grooves, from top to bottom (or bottom to top) of the quadruplex stem. These are called propeller, groove, or chain-reversal loops (Figure 1c).

G-DNA molecules have been extensively studied by atomic resolution experiments and other experimental approaches that have provided unique insights into many aspects of G-DNA structural, dynamical, and kinetic properties.[5,16−31] One of the amazing features of G-DNA with loops is their enormous structural polymorphism, where a given sequence can adopt multiple folds and often subtle changes in the sequence or environment (such as type of ions) may have large effect on the topology.[19,20,32−39] This G-DNA structural variability may be reminiscent of the complexity of rules governing topologies of nucleic acids junctions.[40]

Theoretical methods have been also widely applied to investigate various aspects of G-DNA,[41−69] for a recent review see ref 65. Among the different theoretical methods for the study of G-DNA atomistic molecular dynamics (MD) with explicit solvent is probably that able to capture with higher accuracy the structure and dynamics of G-DNA in aqueous solution. Unfortunately, when using MD simulations we cannot ignore that they are based on simple empirical force fields which can lead to artifacts in simulations. This means that testing and benchmarking become a crucial step to validate the reliability of any simulation. Earlier studies indicated a very good performance of the simulation technique in studies of G-DNA stems, except for some modest imbalance of the cation positions within the stem.[62] It was observed that the ions look oversized (too large) and avoid in-plane positions in the quartets even when simulated with Na^+. Bifurcated bonding of the quartets was also often noticed, which is a perturbation of the structure compared to the experiments. These problems were tentatively attributed to the lack of polarization term in the pair additive classical force field, which was also confirmed by quantum - chemical calculations showing that the strength of direct cation - G(O6) interactions is underestimated with a too early onset of short-range repulsion. Besides that, the simulations revealed that the cation-stabilized stem is a uniquely rigid molecular assembly and the ions are necessary for its stabilization.[62,66] However, the quadruplex stem is still stable with a reduced number of ions in the channel, which allows a smooth exchange of ions with the bulk solvent. An initially empty stem is capable to attract a bulk ion swiftly,[63] thus, in reality, G-DNA stems should never be left vacant by cations. Alternative topologies of G-DNA stems were found, with shifted (slipped) strands.[62,70] The possibility of such substates was later confirmed by experiments.[71] Further studies demonstrated that guanine to thioguanine substitution significantly sterically destabilizes the stem while inosine causes only its subtle destabilization.[64,66] However, inosine may interfere with the process of G-DNA stem formation, by destabilization of kinetic intermediates that rely on interbase H-bonding, before the ion binding starts to dominate the stabilization.[66,70] Simulations were also used to investigate a wide range of double, triple, and quadruple stranded species that could occur as intermediates during quadruplex stem formation[70] and to analyze the properties of G-DNA under hostile conditions such as vacuum, something relevant to rationalize mass field spectroscopy experiments.[21,61,72]

Our subsequent attempt to in-depth characterize the loop topology of the *Oxytricha nova* $d(GGGGTTTTGGGG)_2$ (or

* Corresponding author phone: +420 5415 17133; fax: +420 5412 12179; e-mail: sponer@ncbr.chemi.muni.cz.
† Masaryk University.
‡ Academy of Sciences of the Czech Republic.
§ Polish Academy of Sciences.
|| Institute for Research in Biomedicine, Barcelona Supercomputing Center, and University of Barcelona.
⊥ University of Utah.

$d(G_4T_4G_4)_2$) quadruplex using simulations indicated inadequacy of the major nucleic acids molecular modeling force field AMBER (versions parm94 - parm99)[73−75] for this particular task.[45] In standard simulations, the diagonal loops were basically stable as taken from the starting experimental structures. However, the ions residing at the stem-loop junction in the X-ray structures were lost. In contrast, locally enhanced sampling (LES)[76,77] molecular dynamics simulations aimed at finding the global loop minimum independently of the starting structure predicted entirely different loop geometry, which was in clear disagreement with the experimental structures. Note that the experimental geometry of the $d(G_4T_4G_4)_2$ loop has been unambiguously determined by independent X-ray and NMR studies which are mutually entirely consistent.[14,30,78] Subsequent free energy computations indicated that the predicted incorrect loop topology is more stable according to the force field than the correct experimental one, further suggesting that the force field is in trouble and that we were not facing a LES-artifact. These negative results pointed out the difficulties in representing loops by current force fields, which have been always parametrized considering canonical helices or highly compact RNA structures. These results suggest loops as an excellent benchmark to test the accuracy of current force fields to describe highly irregular nucleic acid structures.

In the present paper we substantially expand the G-DNA loop calculations, using them as a benchmark of the quality of current force fields to describe unusual DNA structures. Besides the $d(G_4T_4G_4)_2$ dimeric quadruplex with diagonal loops we study also the parallel stranded human telomeric monomolecular quadruplex $d[AGGG(TTAGGG)_3]$ with three propeller loops, as revealed by X-ray crystallography.[15] The present study was motivated by recent refinements in well established force fields, some of them which provide a dramatic improvement in simulations of canonical nucleic acid structures and that were robust in the microsecond time scale.[79] Particularly, we tested the older version of amber force field for nucleic acids, parm99[74] and its parmbsc0 refinement.[79] Further, we employed another version of AMBER force field,[80] which changed the glycosidic torsion profile and which can be combined with either parm99 or parmbsc0 AMBER force fields, and the latest version of the CHARMm force field,[81,82] which is known to produce reasonable trajectories for canonical DNAs (for a recent comparison with parmbsc0 see ref 83), but that has not been much used to study noncanonical DNA structures. Further, we considered also diverse cation parameters to check the importance of counterions parametrization in the calculations. Explicit solvent molecular dynamics simulations are complemented by locally enhanced sampling simulations and postprocessing MM-PBSA (Molecular Mechanics, Poisson−Boltzmann, Surface Area) free energy calculations.[84,85]

Our results reveal that none of the presently available force fields is accurate enough in describing the G-DNA loops. The imbalance is best seen for the propeller loops, as their experimental structures are lost even during standard simulations. Among the force fields, parmbsc0 provides results that are closest to the experimental target values but still not in full agreement. This indicates that the improvement of the γ torsional profile penalizing the γ trans substate was a step in the right direction, albeit not sufficient to treat all imbalances. The modified χ parametrization appears to rigidify the studied systems but does not change the ultimate outcome of the simulations. The structures obtained in simulations with the modified χ profile are predetermined by its combination with either parm99 or parmbsc0. Overall, our results confirm that loops in guanine quadruplex molecules represent a very difficult target for molecular modeling approaches and should be considered as references in any future studies aiming to develop or tune nucleic acids force fields. Properly tuned force fields, designed to reproduce these complex motifs can provide improved description of many other types of nucleic acids. Note nevertheless that the existing variants of the AMBER force field were shown to be successful in the description of a wide range of noncanonical structures, including many complex RNAs,[86,87] and also the G-DNA and i-DNA stems,[65] underlining the unique complexity of the loop simulations.

Methods

Starting Geometry and Initial Model Building. The initial structures were taken from the following experimental structures: the X-ray structure of the $d(G_4T_4G_4)_2$ sequence from the 3′ overhang of the *Oxytricha nova* telomere (NDB: UD0014, PDB: 1JRN, resolution 2.00 Å)[14] and the X-ray structure of the human telomeric quadruplex sequence $d[AGGG(TTAGGG)_3]$ (NDB: UD0017, PDB: 1KF1, resolution 2.10 Å)[15] (Figure 1b,c).

The earlier simulations of the $d(G_4T_4G_4)_2$ quadruplex which are also assessed in this paper were based on the NMR structure (PDB: 156D).[30,88] The NMR structure is, when utilized as starting structure for MD simulations, basically equivalent to the X-ray structure (there are modest structural differences within the same substate of the loop geometry) except for the absence of monovalent ions at the stem-loop junction (see below). The NMR experiment could not capture positions of the ions. Some test simulations of parallel tetramolecular quadruplex $d(G_4)_4$ started from the X-ray structure (PDB: 352D)[26] or the human telomere structure (for $d(G_3)_4$) where propeller loops were deleted.

All simulations started with the structural monovalent ions fully occupying the G-DNA stem (three and two ions for the Oxytricha (OXY) and human telomeric (HT) G-DNA, respectively). In most simulations of $d(G_4T_4G_4)_2$, structural ions were initially also placed at the stem-loop junction based on their experimental positions (initially 5 ions in the structure, see Figure 1b). Also the HT G-DNA X-ray structure shows a monovalent ion above the upper quartet plane that was included in some starting structures. This ion, however, evidently is not an integral part of the quadruplex, since it is sandwiched between two adjacent stems in the crystal structure. This ion is never stable in simulations.

Note that the HT quadruplex sequence is known to adopt variable topologies with different types of loops depending on the experimental conditions.[89−91] However, analysis of the topological variability of this sequence is outside the scope of this study. The X-ray structures are the most suitable

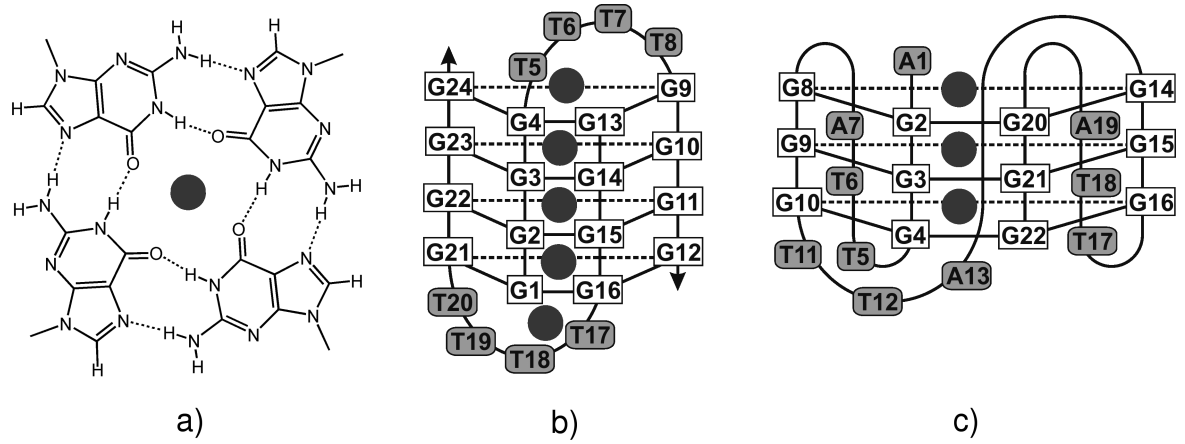


Figure 1. Scheme of a) the guanine quartet and the studied experimental structures of b) the bimolecular antiparallel quadruplex formed by the *Oxytricha nova* telomeric sequence[14] $d(G_4T_4G_4)_2$ with four quartets, two close-to-identical diagonal loops, three stem K^+ ions, and two K^+ stem-junction ions and c) unimolecular parallel quadruplex with three very similar propeller (groove, chain-reversal) loops formed by the human telomeric sequence[15] d[AGGG(TTAGGG)$_3$] with K^+ stem ions. The ion above the upper quartet is a result of crystal packing.

ones for comparison with simulations aimed at force field testing. Since the simulations were unable to keep correct experimental loop structures for the above two G-DNAs, it was not necessary to extend the simulations to other systems.

AMBER Simulations. The structures were prepared by the xleap module of AMBER (adding of hydrogen atoms and neutralizing the system by adding monovalent counterions, either K^+ or Na^+, with numerous control KCl excess salt simulations). The net-neutralizing ion condition leads to a cation concentration of ~0.3 M and ~0.2 M for OXY and HT systems, respectively (the OXY system is more compact and thus the box has less water molecules per nucleotide). When using excess salt we set the cation/anion ratio to be 2:1, resulting in ~0.5 M cation concentration (~0.2 M and ~0.3 M excess salt) for the OXY and HT systems. The solute was embedded in TIP3P water box,[92] which was extended approximately 10 Å in each direction from the solute. Some simulations were done also with the SPC/E water model.[93] Different ion parameters were tested, as specified below. There were ~13000 and ~18000 atoms in OXY G-DNA and HT G-DNA simulations, respectively.

AMBER Force Fields. The AMBER simulations were carried out with the parm99,[74] parmbsc0,[79] and Ode et al.[80] versions of the Cornell et al. force field.[73] Parm94 version[73] was not tested, but its performance is expected to be very similar to the parm99. In contrast, parmbsc0 introduces a substantial modification of the α/γ torsional backbone parameters which is absolutely essential to stabilize B-DNA simulations. Parmbsc0 has been verified extensively by simulations. The Ode et al. force field suggests reparametrization of χ glycosidic torsion and can be combined either with parm94−99 or parmbsc0. The Ode et al. force field has not been tested in simulations so far, except for a few 5 ns runs[80] which we consider as entirely insufficient testing.

Cation Parameters. The pair additive nonpolarizable force field approximation limits the quality of description of the cation - solute interactions. Nevertheless, we tried different cation parametrizations to see if they can affect the results.

We mostly used standard AMBER potassium (radius 2.6580 Å and well depth 0.000328 kcal/mol) and sodium (radius 1.8680 Å and well depth 0.00277 kcal/mol) parameters. We also used K^+ ions with smaller radii (K^+ reduced, radius 2.4 Å and well depth 0.0011 kcal/mol or 2.5 Å and well depth 0.0008 kcal/mol). The reason for the reduction of the ion radii is the observation that the original K^+ ions appear to be oversized (having too large radius) in the G-DNA ion channel, leading sometimes even to expulsions of K^+ out of the channel (see below).

The extent of deficiency of the force field description of the solute - ion interactions is nicely visualized by comparing quantum chemical evaluation of the O6(G)...K^+ interaction energy with force field calculations (Figure 2). The force field underestimates the interaction energy and overestimates the repulsion for shorter O6...K^+ distances. Our parameter adjustment for K^+ (K^+ reduced) was qualitatively based on the quantum-chemical calculations and should be considered as specific for G-DNA simulations. We do not claim that these parameters are better than the original ones for common simulations. It is not possible to simultaneously fully balance all solute - ion and solvent - ion interactions with such simple pair-additive force fields. Obviously our ion parameter adjustment reduces the O6...K^+ repulsion but does not lead to a full agreement since the binding energy remains sharply underestimated due to lack of polarization in the force field. Besides the neglect of polarization, the 6−12 Lennard-Jones force field term is likely excessively repulsive (too steep) in the short-range region.

There have been recent systematic efforts to refine the monovalent ion parameters for AMBER nucleic acids simulations.[94] These efforts, however, were directed to improve the bulk behavior of the ions. As shown in Figure 2, all presently available cation force fields provide essentially similar interaction energy curves with the guanine O6, which exaggerates the short-range repulsion and underestimates the attraction. That is a natural consequence of the pair additive force field which offers just two parameters to be adjusted, the radius and well depth. It is not possible to

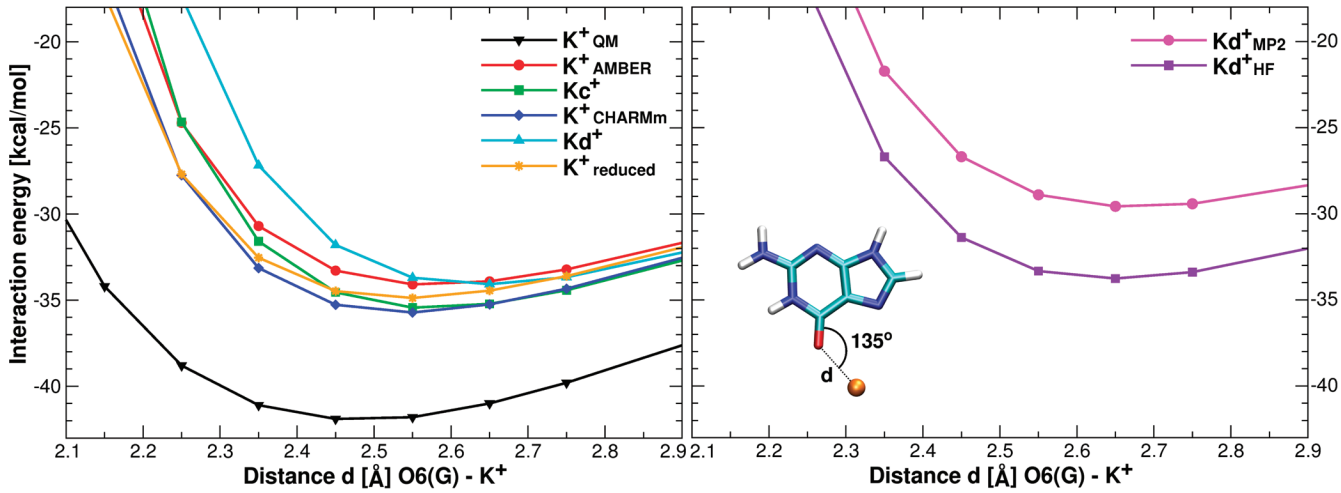


Figure 2. Left - the dependence of the interaction energy between G(O6) and K^+ in a G-DNA like geometry. Black (triangle down), reference QM data with inclusion of electron correlation, Becke3LYP/6-311G(d,p) method corrected for basis set superposition error; red (circle), standard AMBER parameters (2.6580 Å, 0.000328 kcal/mol); blue (diamond), standard CHARMm parameters (1.76375 Å, 0.087 kcal/mol); orange (star), parameters for ions with reduced atomic radii (2.4 Å, 0.0011 kcal/mol); green (square), parameters Kc^+ by Joung and Cheatham[94] (1.7050 Å, 0.1936829 kcal/mol); and cyan (triangle up) parameters Kd^+ by Dang[95] (1.8700 Å, 0.100 kcal/mol). Data in parentheses represent atomic radii and potential well depths. All force fields underestimate the stabilization and exaggerate the short-range repulsion (the optimal O6...K^+ distance and namely the gradient of the energy in the repulsive region). Note that despite the variability of parameters all the force fields cluster in a narrow region rather far from the QM data, indicating that the simple force field function is not sufficient to reproduce the QM data. Specifically, once the ion force fields are tuned to reproduce target condensed phase bulk solvent and ion−ion data, there are no more free parameters to optimize the cation - solute interactions. Comparison for Na^+ would result in a similar picture. Right - comparison of force field calculation with HF (AMBER) ESP charges and ESP charges derived with the inclusion of electron correlation (MP2/aug-cc-pVDZ level, the upper curve).

simultaneously satisfy the ion hydration reference data (or other bulk properties) and direct solute - cation interactions. Interestingly (Figure 2 right), when the force field is combined with MP2 charges instead of the HF ones, the gap between the reference QM data and the force field curve further widens, despite the fact that the MP2 charges should at first sight bring the QM and force field closer to each other (because both computational methods then have electrostatic terms reflecting electron correlation effects). The fact that the HF charge distribution brings the force field calculations closer to the full QM curve than the charge distribution derived with electron correlation is due to partial compensation of errors. The HF charges exaggerate the polarity of the electrostatic potential, and thus the electrostatic attraction between the cation and the guanine is overestimated, partially counterbalancing the missing polarization effects.

In our simulations, we also tested potassium parameters published by Dang[95] (atomic radius 1.8700 Å and well depth 0.100 kcal/mol) and parameters by Joung and Cheatham[94] (atomic radius 1.7050 Å and well depth 0.1936829 kcal/mol). We also carried out some simulations in higher salt conditions using KCl. In these simulations, both potassium and chloride ions were described by Dang parameters which in the case of chloride (radius 2.47 Å, well depth 0.1 kcal/mol) represent standard AMBER parameters.

Standard AMBER Molecular Dynamics Simulations. Molecular dynamics simulations were performed with the Sander module of the AMBER-5.0−9.0 software package.[96−100] The particle mesh Ewald (PME) method[101] was used for a correct treatment of electrostatic interactions. All simulations were run with the SHAKE algorithm[102] (with a tolerance of 0.0005 Å) to constrain covalent bonds involving hydrogens, with periodic boundary conditions, a 2-fs time step, and a temperature of 300 K (Berendsen temperature coupling algorithm with time constant of 0.2 ps[103]). Standard equilibration and production procedures were applied. Analyses of resulting trajectories were performed with ptraj or carnal modules, and the results were visualized with the help of VMD graphic software[104] and X3DNA.[105] The protocol is described in more detail in our recent studies.[106,107]

Locally Enhanced Sampling Molecular Dynamics Simulations (LES). The locally enhanced sampling (LES) method[76,77] was performed with an addles module of AMBER to divide the structure into regions (stem and loops), and each of the loops was split into 5 independent copies. Force field parameters for the copies were adjusted accordingly which lowers the energy barriers. In order to provide an initial "kick" to the 5 copies, the structure was heated to 500 K. Moreover a long relaxation phase appears vital to provide sufficient freedom for the copies to settle in different regions of the conformational space. To allow for this, the temperature was gradually decreased from 500 to 300 K over 1.5 ns (during the first 750 ps the pressure was set to 100 atm), and guanine quartets were maintained with flatwell restraints (R1 = 0.0, R4 = 6.0, RK2 = 5.0, RK3 = 10.0; R2 and R3 depend on the actual distance R between the restrained atoms (R2 = R - 0.5 Å, R3 = R + 0.5 Å)) on the N7...N2 and O6...N1 virtual bonds linking the neighboring guanines. LES simulations were usually followed by standard MD to allow the LES structure to locally relax. For further details about the protocol see ref 45.

Table 1. List of Simulations on $d(G_4T_4G_4)_2$ (Oxytricha) Quadruplex with Diagonal Loops

simulation name	initial structure	ion type[a]	trajectory length and type
		AMBER Simulations with parm99	
OXY^{NMR}	NMR	Na^+	10 ns MD
OXY	X-ray	Na^+	5 ns MD
OXY_K	X-ray	K^+	5 ns MD
OXY^{NMR}_LES	NMR	Na^+	6 ns LES
OXY^{NMR}_LES_MD	OXY^{NMR}_LES end	Na^+	3 ns MD
OXY_SPC_Kd	X-ray	K^+ ions (Dang), SPC waters	50 ns MD
		AMBER Simulations with parmbsc0	
OXY_bsc0_0	X-ray	Na^+	50 ns MD
OXY_bsc0_1	X-ray	Na^+	50 ns MD
OXY_bsc0_2	X-ray	Na^+	50 ns MD
OXY_bsc0_3	X-ray	Na^+	50 ns MD
OXY_bsc0_4	X-ray	Na^+	50 ns MD
OXY_bsc0_K2	X-ray	K^+ ions with radius of 2.4 Å	50 ns MD
OXY_bsc0_Kd	X-ray	K^+ ions (Dang)	50 ns MD
OXY_bsc0_hs_Kd	X-ray	excess salt 0.2 M KCl, K^+ ions (Dang)	50 ns MD
OXY_bsc0_hs_SPC_Kd1	X-ray	excess salt 0.2 M KCl, K^+ ions (Dang), SPC waters	50 ns MD
OXY_bsc0_hs_SPC_Kd2	X-ray	excess salt 0.2 M KCl, K^+ ions (Dang), SPC waters	50 ns MD
OXY_bsc0_Kc	X-ray	K^+ ions (Cheatham)	50 ns MD
OXY_bsc0_hs_Kc1	X-ray	excess salt 0.2 M KCl, K^+ ions (Cheatham)	50 ns MD
OXY_bsc0_hs_Kc2	X-ray	excess salt 0.2 M KCl, K^+ ions (Cheatham)	50 ns MD
OXY_bsc0_LES	X-ray	Na^+	20 ns LES
OXY_bsc0_LES_K2	X-ray	K^+ ions with radius of 2.4 Å	20 ns LES
OXY_bsc0_LES_MD	OXY_bsc0_LES end	Na^+	20 ns MD
OXY_bsc0_LES_MD_K2	OXY_bsc0_LES_K2 end	K^+ ions with radius of 2.4 Å	20 ns MD
		AMBER Simulations with Chi Modification + Either parm99 or parmbsc0	
OXY_chi	X-ray	Na^+	50 ns MD
OXY_bsc0_chi	X-ray	Na^+	50 ns MD
		CHARMm Simulations	
OXY_CHARMm	X-ray	standard CHARMm Na^+ ions	40 ns MD
OXY_CHARMm_mod	X-ray	Na^+ ions with radius of 1.16 Å	20 ns MD
OXY_CHARMm_K	X-ray	standard CHARMm K^+ ions	20 ns MD
PARA_CHARMm[b]	X-ray	standard CHARMm Na^+ ions	10 ns MD

[a] Net neutralizing set of standard AMBER cations and TIP3P water model if not stated otherwise. [b] Test simulation of parallel four-quartet guanine stem $d(G_4)_4$.[26]

MM-PBSA Free Energy Calculations. PB analysis was performed using a modified MM-PBSA procedure.[84,85] Both force field parameters (parm99 and parmbsc0) were employed for the MM part. The Cornell et al. (parm94) charge set, PARSE vdW radii,[108] and a dielectric constant of 1 for DNA were used. The Sander module of AMBER was used for MM energy terms, Delphi software[109] for PB contributions, and Molsurf for calculating SASA. The MD trajectories were examined in 10 ps intervals.

The MM-PBSA energy was calculated with an explicit inclusion of the channel cations (with Na^+ parameters adapted for the free energy computations) as in detail described elsewhere.[70] In order to obtain meaningful numbers for the trajectories in which one cation left the channel the closest solvent ion was considered as a part of the structure. We replaced potassium cations in particular trajectories with sodium cations as we wanted to obtain comparable results. (When attempting to use the potassium ion parameters, the resulting values were far away of the energy interval given by other structures). Since the MM-PBSA calculations are utilized only as a supplementary tool, we did not try to tune the K^+ parameters for free energy computations. MM-PBSA analysis was performed with parm99, also for the $d(G_4T_4G_4)_2$ MD trajectories that were produced with parmbsc0, and vice versa, i.e., we also cross-calculated free energies with the two force fields. The basic free energy trends are the same with both force fields.

CHARMm Simulations. The simulations were performed with the CHARMm code[110] using the CHARMm27 force field for nucleic acids.[81,82] The starting coordinates were the same as for the AMBER simulations. The equilibration protocol started with MD which was applied first to the water molecules only (5 ps) and then to the solvent (water + ions) (25 ps). Then the system was subjected to the several rounds of minimization with gradually reduced harmonic constraints on DNA. The final minimization was performed without any constraints. After that, the whole system was heated from 50 to 300 K in 30 ps by 50 K increments. The Particle Mesh Ewald (PME) method was used for treatment of electrostatic interactions.[101] MD simulations were run with a 2 fs time step and the SHAKE algorithm[102] to constrain all bonds to hydrogens. Two types of sodium ions were tested - either standard sodium ions (with radius 1.3638 Å and well depth 0.0469 kcal/mol) or modified (radius 1.163 Å and well depth 0.21 kcal/mol).[111] For potassium ion the following parameters were employed: radius of 1.7638 Å and well depth 0.0870 kcal/mol. As noted below, however, the results do not depend on these fine details of ion parametrization.

List of Simulations and Abbreviations. More than 1.5 μs of MD and LES trajectories (aggregated time) were run with the above-described protocols, with Na^+, K^+, or KCl ion atmospheres, considering various ion parameters and with

Table 2. List of Simulations on Human Telomere (HT) Quadruplex[15b]

abbreviation	initial structure	ion type[a]	trajectory length
	AMBER Simulations with parm99		
HT	X-ray	Na^+	10 ns MD
HT_K	X-ray	standard K^+	10 ns MD
HT_K2	X-ray	K^+ ions with radius of 2.4 Å	10 ns MD
HT_K3	X-ray	K^+ ions with radius of 2.5 Å	10 ns MD
	AMBER Simulations with parmbsc0		
HT_bsc0	X-ray	Na^+	40 ns MD
HT_bsc0_K2	X-ray	K^+ ions with radius of 2.4 Å	50 ns MD
HT_bsc0_hs_Kd	X-ray	excess salt 0.3 M KCl, K^+ ions (Dang)	50 ns MD
HT_bsc0_LES	X-ray	Na^+	20 ns LES
HT_bsc0_LES_K2	X-ray	K^+ ions with radius of 2.4 Å	20 ns LES
HT_bsc0_LES_MD	HT_bsc0_LES end	Na^+	20 ns MD
HT_bsc0_LES_MD_K2	HT_bsc0_LES_K2 end	K^+ ions with radius of 2.4 Å	20 ns MD
	AMBER Simulations with Chi Modification and Either parmbsc0 or parm99		
HT_chi	X-ray	Na^+	50 ns MD
HT_bsc0_chi	X-ray	Na^+	50 ns MD
HT_bsc0_bsc0+chi	HT_bsc0 end	Na^+	50 ns MD
HT_chi_ LES	X-ray	Na^+	20 ns LES
HT_bsc0_chi_ LES	X-ray	Na^+	20 ns LES
HT_chi_LES_MD	HT_chi_ LES	Na^+	20 ns MD
HT_bsc0_chi_LES_MD	HT_bsc0_chi_ LES	Na^+	20 ns MD
	CHARMm Simulations		
HT_CHARMm	X-ray	standard CHARMm Na^+ ions	10 ns MD
HT_CHARMm_K	X-ray	standard CHARMm K^+ ions	10 ns MD

[a] Net neutralizing set of standard AMBER cations and TIP3P water model if not stated otherwise. [b] In the abbreviations, if not specified otherwise, the simulations were run with parm99 and Na^+ ions. "bsc0", "chi", and "CHARMm" abbreviations indicate that parmbsc0, Ode et al., and CHARMm force fields were used. "LES" and "LES_MD" mean LES simulation and standard simulation that follows LES simulation, respectively. "K", "K2", and "K3" stand for simulations with K^+ having standard 2.66 Å, 2.4 Å, and 2.5 Å radii, respectively. Kc and Kd stand for Cheatham and Dang parameters of K^+ ions, and "hs" marks higher (excess) salt simulations.

five different force field variants. A list of trajectories and abbreviations is given in Tables 1 and 2.

These simulations are abbreviated as OXY or OXY^{NMR} for the X-ray[14] and NMR[30] starting structures. The X-ray structure has five integral potassium ions at the start (three channel and two stem-loop junction K^+). In the simulations with Na^+ ions, coordinates of structural K^+ ions were used for structural Na^+ ions. In the abbreviations, if not stated otherwise, the simulations were run with parm99 and net-neutralizing Na^+ ions. "bsc0", "chi", and "CHARMm" abbreviations indicate that parmbsc0, Ode et al., and CHARMm force fields were used. "LES" and "LES_MD" mean LES simulation and standard simulation that follows LES simulation, respectively. "K", "K2", and "K3" stand for simulations with K^+ having standard 2.66 Å, 2.4 Å, and 2.5 Å radii, respectively. "hs" marks higher (excess) salt simulations, and Kc and Kd stand for Cheatham and Dang parameters of K^+ ions.

Results

Simulations of the $d(G_4T_4G_4)_2$ Quadruplex from Oxytricha (OXY Quadruplex) with Diagonal Loops. *Description of the Structure.* The antiparallel four-quartet bimolecular quadruplex consists of two $d(G_4T_4G_4)$ strands and has two diagonal four-thymidine loops (T5-T6-T7-T8 and T17-T18-T19-T20) - see Figure 1b. Both loops are nearly identical in the X-ray structure[14] with mutual heavy atom rmsd = 0.21 Å. The first and third thymines in each loop are coplanar and connected by T5(O2)...T7(N3) (3.09 Å) and T17(O2)...T19(N3) (2.99 Å) H-bonds, respectively. A notable feature is the presence of K^+ at each stem-loop junction, so that there are three stem (channel) and two stem-loop junction (channel entrance) cations present in the experimental structure. The stem-loop junction ions are coordinated to four O6 atoms of the outer quartet guanines and two O2 atoms of the adjacent thymine basepair (either T5&T7 or T17&T19). This X-ray structure is except for details that are not significant for the simulations in full agreement with other available relevant X-ray[78] and NMR structures.[30,88] Since the experimental top and bottom loops have basically identical topologies, each simulation provides two independent loop trajectories.

Behavior of the Stem in AMBER Simulations. The parm99 force field provides a good description of the G-DNA stem, as consistently shown in all our preceding studies and confirmed also by others (see the Introduction for more details). Good performance of the force field for the stem is confirmed also in the present study and will not be further discussed. Parmbsc0 also provides satisfactory description of the G-DNA stem. We did not make a detailed analysis of the parm99 vs parmbsc0 dynamics of the stem, as no major problems with the stem are indicated.

Redistribution of the Ions. Several earlier conventional 5−10 ns parm99 MD simulations of $d(G_4T_4G_4)_2$ show basically stable loop trajectories.[45] However, the stem-loop junction ions were lost (when initially present) within a very few ns and were not replaced by other bulk ions. In this study we analyze twenty-four 5−50 ns simulations with initially five ions associated with the $d(G_4T_4G_4)_2$ quadruplexes (Table S1). Irrespective of which solute and ion force field parameters were used, the initial experimental configuration with five integral G-DNA ions is unstable. In the

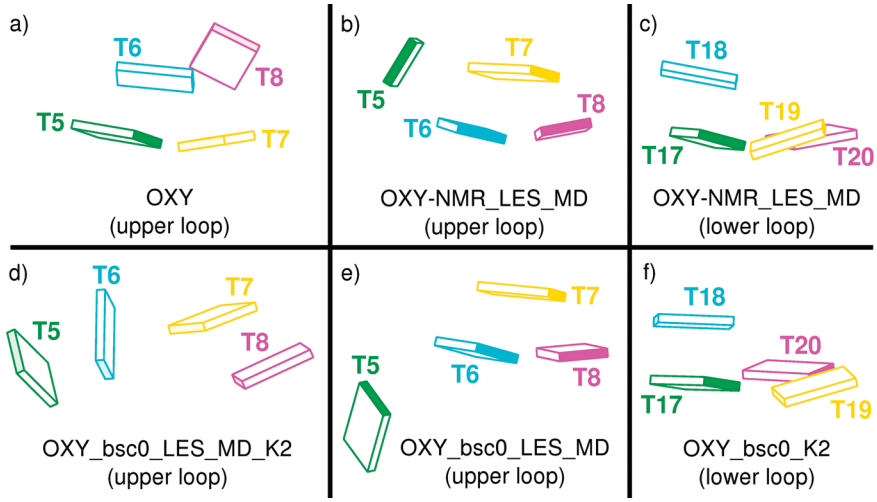


Figure 3. Experimental and three key computed structures of the diagonal four thymidine loops of the Oxytricha quadruplex. a) The experimental structure, b) and c) two LES geometries with parm99, d) entirely unfolded LES geometry with parmbsc0 obtained in one simulation, e) parmbsc0 LES geometry resembling the parm99 LES structure in b), and f) restructured loop in standard parmbsc0 simulation similar to the LES predicted topology in c). Green - T5, T17; cyan - T6, T18; yellow - T7, T19; magenta - T8, T20. Structures c) and f) appear to be achieved by a vast majority of AMBER simulations where the length of the simulation is sufficient to see a transition.

vast majority of simulations, including excess-salt KCl simulations, both stem-loop junction ions left the structure within a very few ns. They were not replaced by any other ion from the bulk or stem channel while three ions remained in the stem (except of one CHARMm simulation, where only two ions remained associated with the quadruplex). In very few simulations an ion was still residing at the end of the simulation in the stem-loop junction position. However, in these cases only three or four ions remained associated with the G-DNA (its central channel area). Therefore, the number of ions was reduced, specifically in the stem cavity adjacent to that bound junction cation. Since the stem-loop junction ions are coordinated also to the outer quartets, it is not surprising that upon reduction of the number of bound cations to less than five some of them can reside in the stem-loop junction. They then provide primary stabilization of the outer quartet. This resembles binding of two ions to the two-quartet stem of thrombin binding aptamer quadruplex, where two K^+ ions are expected to bind to the quartets from the loop area while keeping the central cavity empty.[112]

It is not easy to pinpoint the exact origin of this imbalance in force field description of cation−DNA interaction. It may reflect the underestimation of the direct cation - solute interactions illustrated in Figure 2 (note that the ions are parametrized to provide correct ion-solvation energies, not cation-nucleobase interaction energies) or an overestimation of the ion - ion repulsion. Perhaps the ions in the experimental structure may be stabilized due to rigidification of the experimental X-ray structure or some crystal packing effects. However, it is difficult to believe that this can fully explain the discrepancy between theory and experiment. We can conclude that presently the experimental ion binding to $d(G_4T_4G_4)_2$ cannot be reproduced by simulations.

Loop Structures. As noted above, for the $d(G_4T_4G_4)_2$ quadruplex, the diagonal loop structures were basically stable in earlier standard short simulations with parm99 (Figure 3a), but they were lost in LES simulations with the same force field.[45] The LES-MD structure changed the positions of thymines entirely and formed two new H-bonds in the upper loop, T6(N3)...T8(O4) and T6(O2)...T8(N3), with stacking between the T6...T8 base pair and T7 (Figure 3b). In the second loop rearrangement (Figure 3c) the original base pair is lost, and a new interaction between T17 and T20 is formed with only a single T17(N3)...T20(O4) H-bond. In addition, the methyl groups of T18 and T20 contacted the O2 atom of T17 due to a coplanar arrangement of these three bases - see ref 45 and Figure 3c.

We have thus performed two LES simulations with parmbsc0 (four independent loop trajectories) in which all loops lost their X-ray geometries, similar to the earlier parm99 results. The parmbsc0 LES simulations are much longer than the parm99 LES simulations, and the loops still show no attempts to return to the experimental structure. One loop in the **OXY_bsc0_LES_K2** simulation became completely unfolded (Figure 3d). The three other LES-simulated loops were completely restructured, with the final structures clearly resembling the earlier parm99 LES geometries. Actually, the parmbsc0 and parm99 results are strikingly similar (Figure 3) although the unfolded loop geometry (Figure 3d) was not seen with parm99. However, the parm99 LES simulation was short. Thus, parm99 and parmbsc0 LES data are mutually consistent for the $d(G_4T_4G_4)_2$ loops.

In standard simulations, we concentrated on the parmbsc0 force field, as parm99 was invalidated due to its troubles in B-DNA simulations, rendering parmbsc0 the only AMBER force field that can be used for DNA systems.[79] Our thirteen parmbsc0 simulations gave an unprecedented set of 26 independent 50 ns loop trajectories. Eighteen of them were basically stable with small fluctuations of top thymines (either T8 or T20). In the remaining cases we evidenced formation of a "triadlike" structure in which the top thymine (T8 or T20) formed a close-to-planar arrangement with the original T...T base pair in the particular loop (Figure 3f). This arrangement once formed is stable and clearly resembles the

Table 3. MM-PBSA Free Energy (see Method) in kcal/mol, Averaged over 1−5, 21−25, and 46−50 ns Trajectory Portions

MD simulation	1−5 ns	21−25 ns	46−50 ns
OXY_bsc0_0	−5053 ± 20	−5054 ± 19	−5052 ± 20
OXY_bsc0_1	−5054 ± 20	−5055 ± 20	−5051 ± 20
OXY_bsc0_2	−5050 ± 20	−5054 ± 19	−5050 ± 20
OXY_bsc0_3	−5054 ± 21	−5046 ± 21	−5052 ± 19
OXY_bsc0_4	−5052 ± 20	−5050 ± 19	−5053 ± 20

LES topology shown in Figure 3c. It is the same substate. This suggests that the simulations are slowly converting to different but characteristic geometry, which was anticipated by LES calculations.

Parmbsc0 simulations with standard sodium ions provided us with stable loop structures in almost all cases. Ion modifications in other simulations (see Table 1) were associated with a loss of experimental geometry in 40% of cases. Excess salt simulations with Dang ion parameters kept both loop geometries stable, while Dang ion setting in net-neutralizing salt conditions lost both of them. All three simulations with Cheatham ion parameters showed a formation of "triadlike" structures in one of the loops. However, consideration of these individual cases is not statistically significant and the behavior is most likely incidental. Note (see above) that the modified ion conditions do not stabilize the arrangement with five integral ions in the structures. Table S2 summarizes development of all backbone torsions of both loops in most $d(G_4T_4G_4)_2$ simulations, which gives more detailed insight into the trajectories.

MM-PBSA was used to investigate the energetic origin of the conformational transitions found along some trajectories. However, results summarized in Table 3 for 5 trajectories (four without transition and one with a loop transition; **OXY_bsc0_1**) fail to detect any consistent, statistically significant change in the free energy of the system, which seems quite well converged irrespective of whether or not conformational transitions occur along the trajectory. Thus, the expected free energy change associated with the change of the loop topology appears to be below the threshold which could be detected by MM-PBSA with a confidence, and accordingly the use of MM_PBSA to discriminate between different structural families is not recommended in this particular case.

Supporting Information Tables S4 and S5 contain MM_PBSA data for some other trajectories. The data also do not seem to suggest any easy way to monitor the change of the loop structures by the approximate free energy calculations. Specifically, the free energy of the **OXY_bsc_LES_MD** simulation where both loops are rearranged (−5050 kcal/mol at the end of the standard simulation following the LES run) remains within the range of values in the Table 3, i.e., is not visibly improved. Therefore, we do not use the MM_PBSA data to reach any conclusions in this study (we rely purely on the structural data), and we plan to attempt a more thorough free energy analysis in some subsequent work.

Chi Modification of Force Field. The modification of the AMBER force field by Ode et al.[80] was combined with both parm99 and parmbsc0. We have carried out two 50 ns standard simulations in which the loops were stable and looked stiffer than in parm99 and parmbsc0 standard simulations. Stacking interactions and H-bonds in the loops are all stable. There were perhaps marginally better values of rmsd in the simulated structures with respect to the X-ray structure when we compare **OXY_chi** vs **OXY** and also **OXY_bsc0_chi** vs **OXY_bsc0** simulations (Figure S1).

We did not attempt further simulations for the following reasons. Our simultaneous test simulations (not shown) of B-DNA and Sarcin Ricin 23S rRNA internal loop with the Ode et al. parametrization did not change the results substantially (including the χ angle and helical twist in B-DNA) compared to parmbsc0 (and parm99 for the RNA) simulations. However, the simulated molecules appeared again visually stiffer. The overall impression so far is that the χ-modification apparently slows down transitions (some backbone substates, e.g,, if present) but does not change the ultimate conformational preference. Therefore, it is not surprising that the simulations with modified χ did not reveal any substantial changes of the $d(G_4T_4G_4)_2$ which even with parmbsc0 and parm99 take a time. Meanwhile we obtained better insights into the performance of this force field using the HT G-DNA simulations (see below) and decided that further simulations of $d(G_4T_4G_4)_2$ with the Ode et al. force field are not needed.

Molecular Dynamics with CHARMm. All CHARMm simulations (cf. Table 1) resulted in loop structures that do not resemble the experimental ones. Almost no structural features were kept, including the thymine base pairs and stacking interactions (Figures S2 and S3). Rmsd values were high (Figure S2). During the CHARMm simulations the loops adopted two main conformations (Figure S4). The first conformation is characterized by stacking interactions between T5 and T6 (upper loop) or T17 and T18 (lower loop). In the second conformation T5 stacks on T8 (upper loop) or T17 stacks on T20 (lower loop). Notable is that with CHARMm simulations the X-ray loop structures were lost in standard simulations. In one of the simulations only two ions remained in the channel, while in the other two simulations there were two ions in the stem channel and one ion coordinated to the outer quartet from the loop region (Table S1). This indicates that even the stem behavior is not fully perfect. Therefore, we have carried out a test simulation of an all-parallel four quartet G-DNA stem with CHARMm with Na^+ (last simulation in Table 1). The simulation lost one of the stem ions rather quickly, which was never observed in analogous AMBER simulations (Figure S5). This behavior is not promising, especially when considering that we have used Na^+ cations which should have no steric problems within the stem.

Simulations of the Human Telomeric Monomolecular d[AGGG(TTAGGG)$_3$] Quadruplex (HT Quadruplex) with Propeller Loops. *Description of the Structure.* The structure consists of a common three-quartet guanine stem with two ion cavities and three similar thymine−thymine−adenine propeller loops (T5-T6-A7, T11-T12-A13, and T17-T18-A19) - see Figures 1c and 4. In each loop the adenine is sandwiched between the two thymines but stacks primarily with the first thymine, i.e., there are stacking interactions

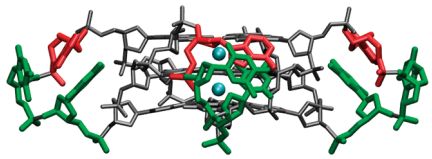


Figure 4. The crystal structure of d[AGGG(TTAGGG)$_3$] (HT) quadruplex; side view. Cyan dots are the channel K^+ ions; loop nucleotides are shown using green (thymine) and red (adenine) licorice model.

between T5 and A7, T11 and A13, and T17 and A19. The other thymines (T6, T12, and T18) are unstacked.

As the loops are conformationally restricted, their backbone torsion angles differ from the canonical DNA α/γ g-/g+ conformation. The first thymine in each loop shows α/γ values of g+/t. The second thymine has α/γ torsions roughly in t/g+ position while the adenine is in the canonical g-/g+ region. Note that the α/γ g+/t substate of the first nucleotides in each loop is the same torsion combination which had to be penalized in the parmbsc0 force field to stabilize B-DNA simulations.[79]

Standard Molecular Dynamics with parm99 Force Field. Guanine stem was stable when simulated with Na^+ and both its channel cations stayed inside the structure. However, with the standard K^+ ion parameters one of the cations left the channel. This most likely is caused by the imbalance in description of solute - cation interactions as described in detail in the Method section. Thus for further simulations we reduced the K^+ radius (see the Method section for further explanation). This stabilized the stem but did not stabilize the loops (see Table 2 for the list of all HT simulations).

Standard MD simulations with the parm99 force field were not capable of keeping the experimental loop structures even on a 10 ns scale, irrespective of the cation parameters. The simulations resulted in a mixture of diverse loop structures, none of them resembling the experiment. Loop geometries were unfolded, and base stacking and other structural signatures of the experimental loops were lost (cf. Figure 5a,b). The loops were swiftly diverging to very diverse conformations, while the experimental one has never been sampled again (see further details in the Supporting Information, including Table S3 summarizing changes of loop backbone angles in all HT simulations). For this reason it was not necessary to make longer simulations or to attempt LES.

Standard Molecular Dynamics and LES Simulations with parmbsc0. In standard parmbsc0 simulations, all three propeller loop structures remained quite close to the experiment even after 50 ns long runs (Figure 5c) and adopted essentially identical geometries. Nevertheless, the agreement with experiment is not perfect for two reasons. The adenine changes its stacking thymine partner from T5 to T6, and the γ trans topology of the first thymine is lost (Figure 6, Table S3).

The loss of the γ trans of the first loop nucleotide is in fact not surprising. Parmbsc0 stabilizes the canonical conformation of α/γ torsional angles compared to g+/t, relative to parm99.[79,113] This is an absolute requirement to achieve stable B-DNA simulations. However, in the HT quadruplex X-ray structure, the first thymine of all three loops (T5, T11, and T17) has γ trans accompanied with the corresponding α torsion in g+. This arrangement was not stable with parmbsc0 simulations (see Figure 6) with both ion types and the backbone flipped basically to the canonical α/γ combination. Some of the loops switched their first thymine to the canonical region quickly; in a few cases it took ~20 ns, but at the end of the standard simulations all nine independent loops (three simulations including one excess-salt, see Tables 2 and S3) lost the γ trans. On the other hand, the remaining loop nucleotides were stabilized by the parmbsc0 force field, sharply contrasting the parm99 behavior. It thus appears that, regarding simulation of this particular loop, the parmbsc0 force field is slightly too canonical though definitely improved over parm99. It allows a quite satisfactory description of this particular loop. At first sight, these simulations might indicate that the α/γ correction of parmbsc0 could be reduced. However, based on the experience accumulated while working on the g+/t B-DNA problem, any significant weakening of the γ trans correction would likely undermine the B-DNA simulations. Note also that (see above) this would hardly improve the loop behavior of the OXY quadruplex where both parm99 and parmbsc0 appear to provide practically identical results.

The modest rearrangement of the loops in parmbsc0 simulations further included creation of a new hydrogen bond between O4 of the first thymine in each loop and a guanine amino group of the central quartet of the stem. The structural change of base stacking in the loops was a consequence of the backbone flip. When the γ torsion flipped from t to g+ value, the α torsion left the g+ arrangement and after some fluctuations adopted a value of ca. $-120°$. This flip brought the above-noted hydrogen bond donors and acceptors close to each other, and the new H-bond was formed. Reduction of the distance between the involved atoms from the starting T(O4) - G(H22) value of 7 Å to the final value <2 Å resulted also in the change of base stacking in the loops. While in the X-ray structure the adenine stacked with the first thymine in each loop, in the final loop structures the adenine stacked with its neighboring thymine in the loop, because the first thymine was rotated and bound to the stem via the newly formed H-bond. All three changes (backbone flip, stacking change, and H-bond formation) appear to be interrelated and in a delicate balance.

Results of excess-salt simulation with Dang parameters for ions are the same as in other parmbsc0 simulations. We again evidenced restacking of the loops, creation of H-bonds between O4 atoms of thymine bases and guanine amino groups of the central quartet of the stem and the loss of γ trans. Thus changing the salt condition is not affecting the simulation outcome significantly, as usual with nucleic acids simulations.[114]

As parmbsc0 provided a rather satisfactory loop description, we performed also extensive LES simulations. LES confirmed that for this loop the parmbsc0 force field is not far from the target structure since the LES simulations localized similar geometries as the standard simulations. The above-mentioned H-bonds between O4 atoms of thymine

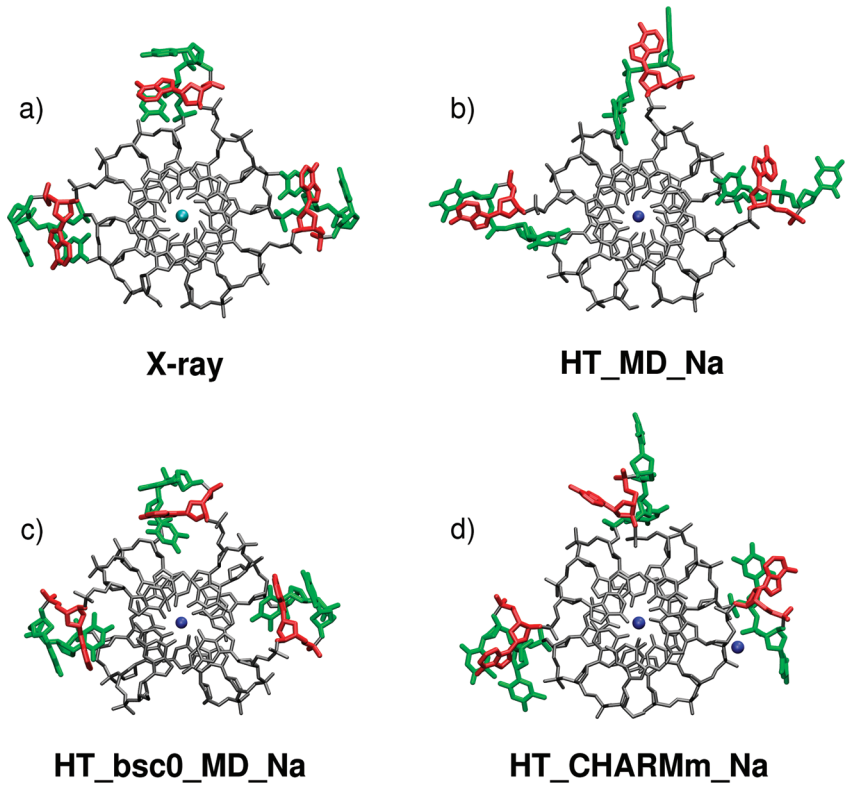


Figure 5. Experimental and simulated structures of the human telomeric quadruplex a) the X-ray structure, b) the structure from parm99 HT simulation, c) the structure from HT_bsc0 simulation, and d) the structure from the HT_CHARMm simulation. Cyan and blue dots are the channel K^+ and Na^+ ions, respectively; loop nucleotides are shown using a green (thymine) and red (adenine) licorice model. Note that in part d) one of the ions left the channel and is seen trapped in the loop region far from the channel. MD structures were averaged over the last 0.5 ns of the trajectory (except for the lost ion in part d).

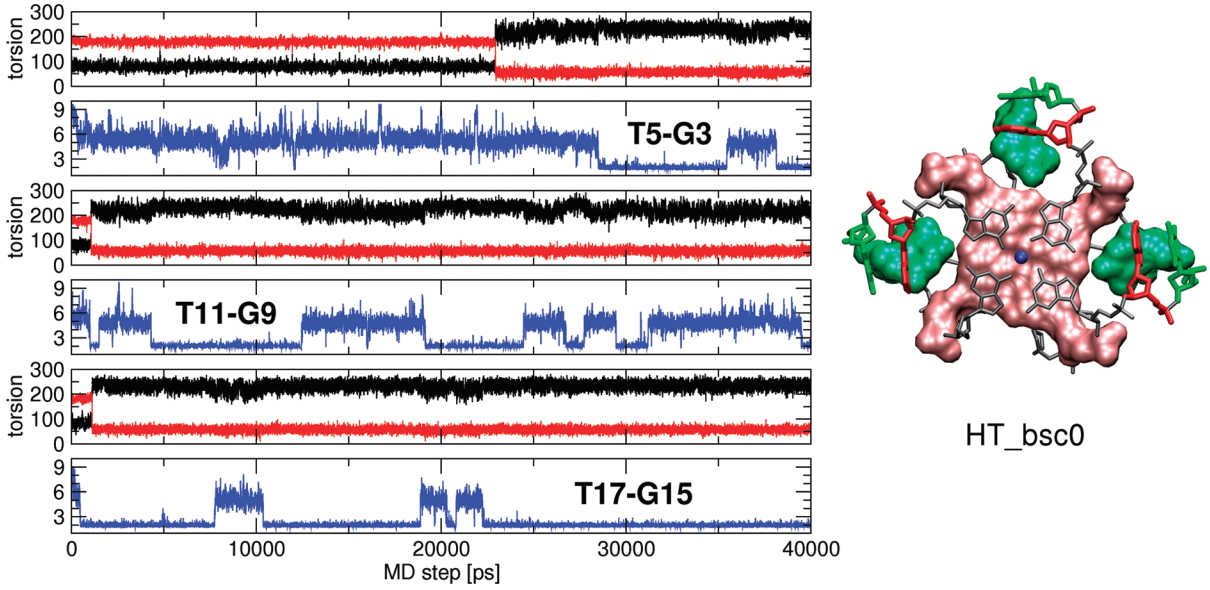


Figure 6. Simulation HT_bsc0 with parmbsc0. a) Time development of α and γ torsions and the loop - stem H-bond of the first thymine in each loop, i.e. T5, T11, and T17. Each pair of graphs corresponds to one loop; the upper graph shows α (black) and γ (red) torsions; the lower graph shows the H-bond between the thymine and the guanine stem. Note that two γ trans states are lost swiftly, the remaining one after more than 20 ns. b) Averaged (last 0.5 ns) structure of HT_bsc0 simulation. The middle guanine quartet and the first thymines in each loop are highlighted by space filling representation.

bases and guanine amino group of the central quartet of the stem were not stably formed within both LES trajectories (**HT_bsc0_LES** and **HT_bsc0_LES_K2**). The **HT_bsc0_LES** trajectory even exhibited the smallest rmsd from all of the trajectories compared to the X-ray structure. Some loops had still γ of the first nucleotide in trans at the end of the LES run. However, the subsequent standard MD trajectories aimed to relax the LES structures show basically again formation of the structure seen in the standard simulations with the H-bonds between the first thymines of the loops

and the stem. The LES simulations give an impression that the parmbsc0 force field prefers the new loop topology as the true global minimum while still sometimes attempting to regain some of the features of the original structure.

Chi Modification of the AMBER Force Field. Simulations run with modified parameters for χ torsions did not bring any breakthrough. The 50 ns **HT_bsc0_chi** simulation resulted in a structure very similar to the **HT_bsc0** simulation. The γ trans of the first thymines in each loop is lost, adenines change their stacking thymine partners, and H-bonds between the first thymine in loop and guanine of the middle quartet are formed. As we noted above, the χ modification appears to rigidify the simulated molecule while not affecting the structures that result from transitions. I.e., we see the same development as without the χ correction, albeit on a longer time scale. The ultimate outcome of the simulation is dictated by whether the χ modification is combined with parm99 or parmbsc0. Even the χ angle values achieved in the simulations with χ modification appear to be unaffected by the χ modification. The modification appears to change the torsional profile mainly in the region between anti and syn nucleotide geometries, which could affect the kinetics and path of the anti to syn transitions; however, further tests will be needed to obtain more insights. When the χ correction was combined with parm99, the loops were significantly destabilized, again as inherent to the parm99 force field without the χ correction. We further took the final structure from **HT_bsc0** simulation, added χ modification and run additional MD. After 50 ns the structure was not changed, and not a single first thymine in the loops switched back to trans. One loop was slightly closer to the original X-ray structure as the H-bond between thymine and guanine was broken, but this is an insignificant observation. In summary, while the χ correction slows down transitions, we see that all simulations are progressing to the same structures as those obtained either with parmbsc0 or parm99 without χ modification, depending on which of them is combined with the χ correction. Therefore, so far we do not see any advantage of using the χ correction, which is also supported by our B-DNA and Sarcin Ricin RNA simulations (unpublished data).

Molecular Dynamics with CHARMm. With the CHARMm force field, the loop geometries appeared as unstable as with parm99, that means in entire disagreement with the experiment and not converging to any common structure (Figure 5d). Thus we do not provide detailed analyses. Even the stem behavior was imperfect. One of the two integral cations left the stem channel at ~2 and ~3 ns for the Na^+ and K^+ simulations, respectively. Such behavior was sometimes observed with AMBER and standard K^+ ions (see Method for discussion) due to the exaggerated short-range repulsion. However, channel ion instability with AMBER has never been seen using a smaller radius for K^+ and any of the Na^+ parameters. The swift loss of the channel ions in CHARMm simulations (already seen for CHARMm OXY and parallel quadruplex, see above) is definitely in disagreement with experiments, and thus no further simulations were attempted.

Discussion and Conclusions

We have carried out an extensive set (more than 1.5 μs in total) of simulations of two guanine quadruplex DNA (G-DNA) molecules: the $d(G_4T_4G_4)_2$ dimeric quadruplex with diagonal loops[14] and also the parallel stranded human telomeric monomolecular quadruplex d[AGGG(TTAGGG)$_3$] with three propeller loops,[15] as revealed by X-ray crystallography. The main aim of the study was to analyze the capability of the explicit solvent molecular dynamics (MD) technique to describe the complicated single stranded loop topologies of these G-DNA molecules.

We have tested five force fields: the parm99 AMBER force field,[74] its recent reparametrization aimed to stabilize B-DNA known as parmbsc0,[79] a combination of both parm99 and parmbsc0 with modification of the χ torsional profile suggested by Ode et al.,[80] and the CHARMm force field[81,82] for nucleic acids. In addition, several ion parameters were used, and net-neutralizing simulations were compared with excess salt ones. The loop behavior does not appear to be dependent on the type of ions, and also the excess salt does not appear to affect the solute behavior on the present time scale. Besides standard simulations, we applied also extensive runs of locally enhanced sampling (LES) dynamics,[76] which is designed to improve sampling of the loop regions, in order to overcome limitations of the short time scale of the simulations. The LES simulations nicely complement the picture emerging from long standard simulations. We see a basic agreement between LES and MD results for both quadruplexes. The study confirms that single stranded hairpin loop topologies represent a major problem for molecular mechanical force fields and that much caution and validation against experimental data is necessary before accepting as real trajectories obtained for these systems.

The $d(G_4T_4G_4)_2$ quadruplex contains, besides the three genuine binding sites for cations in the channel of its stem, also ion binding site at each stem-loop junction. This arrangement of five cations in the quadruplex core region is entirely unstable in all simulations. Most simulations ended up with just three cations in the stem cavities, while there was not a single simulation with five ions inside the structure at the end. The diagonal loops in this structure are stable in short AMBER simulations (all force field variants), while they are lost in CHARMm simulations. In longer AMBER simulations, however, the loops start to convert to a substantially different arrangement, as seen in Figure 3c,f. Analysis of standard and LES simulations give a clear indication that parm99 and parmbsc0 have similar performance for this loop.

The propeller loops of d[AGGG(TTAGGG)$_3$] are very unstable in standard simulations with parm99 and CHARMm, resulting in a diverse mixture of incorrect geometries. Parmbsc0 provides a substantially better description of the -TTA- propeller loops, albeit there are some differences compared with the experimental structure. Namely, γ trans of the first thymine in each loop is lost. This is not surprising, as this force field has been parametrized to penalize γ trans

in order to stabilize B-DNA simulations.[79,113] This also results in some change of stacking partners in the loop. Despite that, this geometry is obtained reproducibly and partially resembles the experimental one.

It is to be noted that the characteristic loop topology of $d(G_4T_4G_4)_2$ which is not reproduced by the force fields has been unambiguously determined by X-ray and solution experiments, with different ion conditions and also with different crystal packing environments.[14,30,78,88] On the other hand, it is well established that the d[AGGG(TTAGGG)$_3$] HT quadruplex would likely fold to an entirely (globally) different topology with an antiparallel (instead of parallel) stem in the presence of Na^+ in solution,[115] as it can adopt multiple topologies. However, this to our opinion is not related to the instability of the loops in our simulations. The overall topological variability in the experiments reflects free energy balance between very different folds. Once a ns-scale simulation starts from one of these folds (parallel with propeller loops in our case) its outcome should not be affected by the fact that some other topology would be more stable. The other topologies are entirely unreachable on the simulation time scale and thus do not interfere with the simulation outcome. The stem would have to be unfolded and refolded to form the alternative topologies. Therefore, once the simulation is confined within a given overall folding arrangement, the simulation should be capable of localizing the appropriate loop geometry provided the force field is appropriately balanced.

The modification of the AMBER glycosidic torsion by Ode et al. does not seem to bring any substantial change of the force field. Its most characteristic feature is rigidification of the simulated structures that slows down the transitions. That is, we see the same development as without the χ correction, albeit on a longer time scale. The ultimate outcome of the simulation is dictated by whether the χ modification is combined with parm99 or parmbsc0. No clear advantage appears then from using Ode's parameters, and the extra rigidification of the system is expected to produce undesired equilibration problems and potential error in the description of the flexibility pattern.

We suggest that the struggle of the force field to deal with the loops is not surprising. Their geometry is a result of a delicate balance of a large number of diverse competing forces including various noncanonical H-bonds between loop bases, different stacking options, specific ion interactions (see the OXY quadruplex), unusual backbone conformations, structural communication between stem and loop not only through the covalent linkage but also via H-bonds or ion interactions, complex solvation effects, and maybe some others. It appears to be very difficult to simultaneously balance all these diverse contributions to obtain the correct loop geometry. It is important to underline that the loops represent a daunting task for the MD simulation technique, and problems of the method with this specific type of nucleic acids architecture does not rule out successful application of the technique to most other types of nucleic acids molecules where it is much easier to obtain a sufficient balance of all the energy contributions. Overall our calculations demonstrate the need for a fine benchmarking of nucleic acids force fields outside from regular structures and well-defined compact arrangements. Ion parameters need to be included in these benchmarks, since for some of these structures they can play a key stabilizing role. Future generation force fields, including polarization corrections,[116,117] might be able to capture the behavior of these complex systems, but, in the meantime, careful checking of simulations in unusual structures such as the quadruplex loops seems necessary to avoid reaching erroneous conclusions.

Acknowledgment. This work was supported by the Ministry of Education of the Czech Republic [grants MSM0021622413 and LC06030], by the Grant Agency of the Academy of Sciences of the Czech Republic [grant numbers 1QS500040581 and IAA400040802], and by grant GA203/09/1476, Grant Agency of the Czech Republic. This work was also supported by the Academy of Sciences of the Czech Republic, grants no. AV0Z50040507 and AV0Z50040702, and by the Spanish Ministry of Science (BIO2006-01602 and Escience consolider Project). J. Sarzyñska acknowledges an access to the Poznań Supercomputing and Networking Centre.

Supporting Information Available: RMSd plots of selected MD simulations of OXY quadruplex, final structures and representative loop geometries of OXY quadruplex from CHARMm simulations, dynamics of channel ions in CHARMm simulations of all-parallel G-DNA stem, table of redistribution of cations in OXY quadruplex simulations, tables of important loops torsion angles for OXY and HT quadruplexes, and tables of free energy estimations for OXY and HT quadruplexes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **2006**, *34*, 5402–5415.

(2) Neidle, S.; Parkinson, G. N. The structure of telomeric DNA. *Curr. Opin. Struct. Biol.* **2003**, *13*, 275–283.

(3) Mergny, J. L.; Mailliet, P.; Lavelle, F.; Riou, J. F.; Laoui, A.; Helene, C. The development of telomerase inhibitors: the G-quartet approach. *Anti-Cancer Drug Des.* **1999**, *14*, 327–339.

(4) Neidle, S.; Read, M. A. G-quadruplexes as therapeutic targets. *Biopolymers* **2000**, *56*, 195–208.

(5) Huppert, J. L. Hunting G-quadruplexes. *Biochimie* **2008**, *90*, 1140–1148.

(6) Alberti, P.; Bourdoncle, A.; Sacca, B.; Lacroix, L.; Mergny, J. L. DNA nanomachines and nanostructures involving quadruplexes. *Org. Biomol. Chem.* **2006**, *4*, 3383–3391.

(7) Davis, J. T. G-quartets 40 years later: From 5′-GMP to molecular biology and supramolecular chemistry. *Angew. Chem., Int. Ed.* **2004**, *43*, 668–698.

(8) Hardin, C. C.; Perry, A. G.; White, K. Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids. *Biopolymers* **2000**, *56*, 147–194.

(9) Huppert, J. L. Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem. Soc. Rev.* **2008**, *37*, 1375–1384.

(10) Lane, A. N.; Chaires, J. B.; Gray, R. D.; Trent, J. O. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.* **2008**, *36*, 5482–5515.

(11) Lane, A. N.; Jenkins, T. C. Structures and properties of multistranded nucleic acids. *Curr. Org. Chem.* **2001**, *5*, 845–869.

(12) Mergny, J. L.; De Cian, A.; Ghelab, A.; Sacca, B.; Lacroix, L. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.* **2005**, *33*, 81–94.

(13) Paramasivan, S.; Rujan, I.; Bolton, P. H. Circular dichroism of quadruplex DNAs: Applications to structure, cation effects and ligand binding. *Methods* **2007**, *43*, 324–331.

(14) Haider, S.; Parkinson, G. N.; Neidle, S. Crystal structure of the potassium form of an Oxytricha nova G-quadruplex. *J. Mol. Biol.* **2002**, *320*, 189–200.

(15) Parkinson, G. N.; Lee, M. P. H.; Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **2002**, *417*, 876–880.

(16) Aboulela, F.; Murchie, A. I. H.; Lilley, D. M. J. NMR-study of parallel-stranded tetraplex formation by the hexadeoxynucleotide d(TG_4T). *Nature* **1992**, *360*, 280–282.

(17) Balagurumoorthy, P.; Brahmachari, S. K. Structure and stability of human telomeric sequence. *J. Biol. Chem.* **1994**, *269*, 21858–21869.

(18) Bardin, C.; Leroy, J. L. The formation pathway of tetramolecular G-quadruplexes. *Nucleic Acids Res.* **2008**, *36*, 477–488.

(19) Crnugelj, M.; Hud, N. V.; Plavec, J. The solution structure of d$(G_4T_4G_3)_2$: a bimolecular G-quadruplex with a novel fold. *J. Mol. Biol.* **2002**, *320*, 911–924.

(20) Dai, J. X.; Carver, M.; Yang, D. Z. Polymorphism of human telomeric quadruplex structures. *Biochimie* **2008**, *90*, 1172–1183.

(21) Gabelica, V.; Rosu, F.; Witt, M.; Baykut, G.; De Pauw, E. Fast gas-phase hydrogen/deuterium exchange observed for a DNA G-quadruplex. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 201–208.

(22) Gros, J.; Rosu, F.; Amrane, S.; De Cian, A.; Gabelica, V.; Lacroix, L.; Mergny, J. L. Guanines are a quartet's best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. *Nucleic Acids Res.* **2007**, *35*, 3064–3075.

(23) Ida, R.; Wu, G. Direct NMR detection of alkali metal ions bound to G-quadruplex DNA. *J. Am. Chem. Soc.* **2008**, *130*, 3590–3602.

(24) Kettani, A.; Bouaziz, S.; Gorin, A.; Zhao, H.; Jones, R. A.; Patel, D. J. Solution structure of a Na cation stabilized DNA quadruplex containing G.G.G.G and G.C.G.C tetrads formed by G-G-G-C repeats observed in adeno-associated viral DNA. *J. Mol. Biol.* **1998**, *282*, 619–636.

(25) Phan, A. T.; Modi, Y. S.; Patel, D. J. Two-repeat Tetrahymena telomeric d(TGGGGTTGGGGT) sequence interconverts between asymmetric dimeric G-quadruplexes in solution. *J. Mol. Biol.* **2004**, *338*, 93–102.

(26) Phillips, K.; Dauter, Z.; Murchie, A. I. H.; Lilley, D. M. J.; Luisi, B. The crystal structure of a parallel-stranded guanine tetraplex at 0.95 angstrom resolution. *J. Mol. Biol.* **1997**, *273*, 171–182.

(27) Rosu, F.; Gabelica, V.; Houssier, C.; Colson, P.; De Pauw, E. Triplex and quadruplex DNA structures studied by electrospray mass spectrometry. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1729–1736.

(28) Sket, P.; Crnugelj, M.; Kozminski, W.; Plavec, J. 15NH_4^+ ion movement inside d$(G_4T_4G_4)_2$ G-quadruplex is accelerated in the presence of smaller Na^+ ions. *Org. Biomol. Chem.* **2004**, *2*, 1970–1973.

(29) Sket, P.; Crnugelj, M.; Plavec, J. Identification of mixed dication forms of G-quadruplex in solution. *Nucleic Acids Res.* **2005**, *33*, 3691–3697.

(30) Smith, F. W.; Feigon, J. Quadruplex structure of Oxytricha telomeric DNA oligonucleotides. *Nature* **1992**, *356*, 164–168.

(31) Zhou, J.; Yuan, G.; Liu, J. J.; Zhan, C. G. Formation and stability of G-quadruplexes self-assembled from guanine-rich strands. *Chem.—Eur. J.* **2007**, *13*, 945–949.

(32) Crnugelj, M.; Sket, P.; Plavec, J. Small change in a G-rich sequence, a dramatic change in topology: New dimeric G-quadruplex folding motif with unique loop orientations. *J. Am. Chem. Soc.* **2003**, *125*, 7866–7871.

(33) Dai, J. X.; Carver, M.; Punchihewa, C.; Jones, R. A.; Yang, D. Z. Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K^+ solution: insights into structure polymorphism of the human telomeric sequence. *Nucleic Acids Res.* **2007**, *35*, 4927–4940.

(34) Li, J.; Correia, J. J.; Wang, L.; Trent, J. O.; Chaires, J. B. Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal. *Nucleic Acids Res.* **2005**, *33*, 4649–4659.

(35) Marathias, V. M.; Bolton, P. H. Determinants of DNA quadruplex structural type: Sequence and potassium binding. *Biochemistry* **1999**, *38*, 4355–4364.

(36) Phan, A. T.; Kuryavyi, V.; Luu, K. N.; Patel, D. J. Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K^+ solution. *Nucleic Acids Res.* **2007**, *35*, 6517–6525.

(37) Phan, A. T.; Luu, K. N.; Patel, D. J. Different loop arrangements of intramolecular human telomeric (3 + 1) G-quadruplexes in K^+ solution. *Nucleic Acids Res.* **2006**, *34*, 5715–5719.

(38) Risitano, A.; Fox, K. R. Inosine substitutions demonstrate that intramolecular DNA quadruplexes adopt different conformations in the presence of sodium and potassium. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2047–2050.

(39) Vorlickova, M.; Chladkova, J.; Kejnovska, I.; Fialova, M.; Kypr, J. Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats. *Nucleic Acids Res.* **2005**, *33*, 5851–5860.

(40) Lilley, D. M. J. Structures of helical junctions in nucleic acids. *Q. Rev. Biophys.* **2000**, *33*, 109–159.

(41) Agrawal, S.; Ojha, R. P.; Maiti, S. Energetics of the human Tel-22 quadruplex-telomestatin interaction: A molecular dynamics study. *J. Phys. Chem. B* **2008**, *112*, 6828–6836.

(42) Arora, A.; Balasubramanian, C.; Kumar, N.; Agrawal, S.; Ojha, R. P.; Maiti, S. Binding of berberine to human telomeric quadruplex - spectroscopic, calorimetric and molecular modeling studies. *FEBS J.* **2008**, *275*, 3971–3983.

(43) Cavallari, M.; Calzolari, A.; Garbesi, A.; Di Felice, R. Stability and migration of metal ions in G4-wires by molecular dynamics simulations. *J. Phys. Chem. B* **2006**, *110*, 26337–26348.

(44) Clay, E. H.; Gould, I. R. A combined QM and MM investigation into guanine quadruplexes. *J. Mol. Graphics Modell.* **2005**, *24*, 138–146.

(45) Fadrna, E.; Spackova, N.; Stefl, R.; Koca, J.; Cheatham, T. E.; Sponer, J. Molecular dynamics simulations of guanine quadruplex loops: Advances and force field limitations. *Biophys. J.* **2004**, *87*, 227–242.

(46) Gu, J. D.; Leszczynski, J. Origin of Na^+/K^+ selectivity of the guanine tetraplexes in water: The theoretical rationale. *J. Phys. Chem. A* **2002**, *106*, 529–532.

(47) Han, H. Y.; Langley, D. R.; Rangan, A.; Hurley, L. H. Selective interactions of cationic porphyrins with G-quadruplex structures. *J. Am. Chem. Soc.* **2001**, *123*, 8902–8913.

(48) Hazel, P.; Huppert, J.; Balasubramanian, S.; Neidle, S. Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.* **2004**, *126*, 16405–16415.

(49) Hazel, P.; Parkinson, G. N.; Neidle, S. Predictive modelling of topology and loop variations in dimeric DNA quadruplex structures. *Nucleic Acids Res.* **2006**, *34*, 2117–2127.

(50) Chowdhury, S.; Bansal, M. A nanosecond molecular dynamics study of antiparallel $d(G)_7$ quadruplex structures: Effect of the coordinated cations. *J. Biomol. Struct. Dyn.* **2001**, *18*, 647–669.

(51) Chowdhury, S.; Bansal, M. G-quadruplex structure can be stable with only some coordination sites being occupied by cations: A six-nanosecond molecular dynamics study. *J. Phys. Chem. B* **2001**, *105*, 7572–7578.

(52) Meng, F. C.; Wang, F. P.; Zhao, X.; Jalbout, A. F. Guanine tetrad interacting with divalent metal ions ($M = Fe^{2+}$, Co^{2+}, Ni^{2+}, Cu^{2+} and Zn^{2+}): A density functional study. *J. Mol. Struct.-Theochem* **2008**, *854*, 26–30.

(53) Meyer, M.; Hocquet, A.; Suhnel, J. Interaction of sodium and potassium ions with sandwiched cytosine-, guanine-, thymine-, and uracil-base tetrads. *J. Comput. Chem.* **2005**, *26*, 352–364.

(54) Meyer, M.; Steinke, T.; Brandl, M.; Suhnel, J. Density functional study of guanine and uracil quartets and of guanine quartet/metal ion complexes. *J. Comput. Chem.* **2001**, *22*, 109–124.

(55) Meyer, M.; Suhnel, J. Density functional study of adenine tetrads with N6-H6. .N3 hydrogen bonds. *J. Phys. Chem. A* **2008**, *112*, 4336–4341.

(56) Ourliac-Garnier, I.; Elizondo-Riojas, M. A.; Redon, S.; Farrell, N. P.; Bombard, S. Cross-links of quadruplex structures from human telomeric DNA by dinuclear platinum complexes show the flexibility of both structures. *Biochemistry* **2005**, *44*, 10620–10634.

(57) Pagano, B.; Mattia, C. A.; Cavallo, L.; Uesugi, S.; Giancola, C.; Fraternali, F. Stability and cations coordination of DNA and RNA 14-mer G-quadruplexes: A multiscale computational approach. *J. Phys. Chem. B* **2008**, *112*, 12115–12123.

(58) Petraccone, L.; Erra, E.; Esposito, V.; Randazzo, A.; Mayol, L.; Nasti, L.; Barone, G.; Giancola, C. Stability and structure of telomeric DNA sequences forming quadruplexes containing four G-tetrads with different topological arrangements. *Biochemistry* **2004**, *43*, 4877–4884.

(59) Read, M. A.; Neidle, S. Structural characterization of a guanine-quadruplex ligand complex. *Biochemistry* **2000**, *39*, 13422–13432.

(60) Ross, W. S.; Hardin, C. C. Iin-induced stabilization of the G-DNA quadruplex - free-energy perturbation studies. *J. Am. Chem. Soc.* **1994**, *116*, 6070–6080.

(61) Rueda, M.; Luque, F. J.; Orozco, M. G-quadruplexes can maintain their structure in the gas phase. *J. Am. Chem. Soc.* **2006**, *128*, 3608–3619.

(62) Spackova, N.; Berger, I.; Sponer, J. Nanosecond molecular dynamics simulations of parallel and antiparallel guanine quadruplex DNA molecules. *J. Am. Chem. Soc.* **1999**, *121*, 5519–5534.

(63) Spackova, N.; Berger, I.; Sponer, J. Structural dynamics and cation interactions of DNA quadruplex molecules containing mixed guanine/cytosine quartets revealed by large-scale MD simulations. *J. Am. Chem. Soc.* **2001**, *123*, 3295–3307.

(64) Spackova, N.; Cubero, E.; Sponer, J.; Orozco, M. Theoretical study of the guanine-6-thioguanine substitution in duplexes, triplexes, and tetraplexes. *J. Am. Chem. Soc.* **2004**, *126*, 14642–14650.

(65) Sponer, J.; Spackova, N. Molecular dynamics simulations and their application to four-stranded DNA. *Methods* **2007**, *43*, 278–290.

(66) Stefl, R.; Spackova, N.; Berger, I.; Koca, J.; Sponer, J. Molecular dynamics of DNA quadruplex molecules containing inosine, 6-thioguanine and 6-thiopurine. *Biophys. J.* **2001**, *80*, 455–468.

(67) Strahan, G. D.; Keniry, M. A.; Shafer, R. H. NMR structure refinement and dynamics of the K^+-$[d(G_3T_4G_3)]_2$ quadruplex via particle mesh Ewald molecular dynamics simulations. *Biophys. J.* **1998**, *75*, 968–981.

(68) Li, H.; Cao, E. H.; Gisler, T. Force-induced unfolding of human telomeric G-quadruplex: A steered molecular dynamics simulation study. *Biochem. Biophys. Res. Commun.* **2009**, *379*, 70–75.

(69) Meyer, M.; Suhnel, J. Interaction of cyclic cytosine-, guanine-, thymine-, uracil- and mixed guanine-cytosine base tetrads with K^+, Na^+ and Li^+ ions - A density functional study. *J. Biomol. Struct. Dyn.* **2003**, *20*, 507–517.

(70) Stefl, R.; Cheatham, T. E.; Spackova, N.; Fadrna, E.; Berger, I.; Koca, J.; Sponer, J. Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates. *Biophys. J.* **2003**, *85*, 1787–1804.

(71) Krishnan-Ghosh, Y.; Liu, D. S.; Balasubramanian, S. Formation of an interlocked quadruplex dimer by d(GGGT). *J. Am. Chem. Soc.* **2004**, *126*, 11009–11016.

(72) Baker, E. S.; Bernstein, S. L.; Gabelica, V.; De Pauw, E.; Bowers, M. T. G-quadruplexes in telomeric repeats are conserved in a solvent-free environment. *Int. J. Mass Spectrom.* **2006**, *253*, 225–237.

(73) 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 2nd generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(74) Wang, J. M.; 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.

(75) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.

(76) Simmerling, C.; Miller, J. L.; Kollman, P. A. Combined locally enhanced sampling and Particle Mesh Ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid. *J. Am. Chem. Soc.* **1998**, *120*, 7149–7155.

(77) Elber, R.; Karplus, M. Enhanced sampling in molecular dynamics - use of the time-dependent Hartree approximation for a simulation of carbon-monoxide diffusion through myoglobin. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.

(78) Horvath, M. P.; Schultz, S. C. DNA G-quartets in a 1.86 angstrom resolution structure of an Oxytricha nova telomeric protein-DNA complex. *J. Mol. Biol.* **2001**, *310*, 367–377.

(79) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817–3829.

(80) Ode, H.; Matsuo, Y.; Neya, S.; Hoshino, T. Force Field Parameters for Rotation Around chi Torsion Axis in Nucleic Acids. *J. Comput. Chem.* **2008**, *29*, 2531–2542.

(81) Foloppe, N.; MacKerell, A. D. 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.

(82) MacKerell, A. D.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257–265.

(83) Perez, A.; Lankas, F.; Luque, F. J.; Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **2008**, *36*, 2379–2394.

(84) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(85) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

(86) Auffinger, P.; Hashem, Y. Nucleic acid solvation: from outside to insight. *Curr. Opin. Struct. Biol.* **2007**, *17*, 325–333.

(87) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. Molecular dynamics simulations of RNA: An in silico single molecule approach. *Biopolymers* **2007**, *85*, 169–184.

(88) Schultze, P.; Smith, F. W.; Feigon, J. Refined solution structure of the dimeric quadruplex formed from the Oxytricha telomeric oligonucleotide d(GGGGTTTTGGGG). *Structure* **1994**, *2*, 221–233.

(89) Phan, A. T.; Patel, D. J. Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: Distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *J. Am. Chem. Soc.* **2003**, *125*, 15021–15027.

(90) Xu, Y.; Noguchi, Y.; Sugiyama, H. The new models of the human telomere d[AGGG(TTAGGG)$_3$] in K^+ solution. *Bioorg. Med. Chem.* **2006**, *14*, 5584–5591.

(91) Ambrus, A.; Chen, D.; Dai, J. X.; Bialis, T.; Jones, R. A.; Yang, D. Z. Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.* **2006**, *34*, 2723–2735.

(92) 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.

(93) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(94) Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(95) Dang, L. X. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether - a molecular dynamics study. *J. Am. Chem. Soc.* **1995**, *117*, 6954–6960.

(96) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.

(97) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, 2004.

(98) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 5*; University of California: San Francisco, 1997.

(99) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California: San Francisco, 1999.

(100) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 7*; University of California: San Francisco, 2002.

(101) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald - an N.logN method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(102) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of Cartesian equations of motion of a system with constraints - molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(103) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(104) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(105) Lu, X. J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

(106) Reblova, K.; Fadrna, E.; Sarzynska, J.; Kulinski, T.; Kulhanek, P.; Ennifar, E.; Koca, J.; Sponer, J. Conformations of flanking bases in HIV-1 RNA DIS kissing complexes studied by molecular dynamics. *Biophys. J.* **2007**, *93*, 3932–3949.

(107) Spackova, N.; Sponer, J. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Res.* **2006**, *34*, 697–708.

(108) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(109) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144–1149.

(110) 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.

(111) Langley, D. R. Molecular dynamic simulations of environment and sequence dependent DNA conformations: The development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* **1998**, *16*, 487–509.

(112) Marathias, V. M.; Bolton, P. H. Structures of the potassium-saturated, 2: 1, and intermediate, 1: 1, forms of a quadruplex DNA. *Nucleic Acids Res.* **2000**, *28*, 1969–1977.

(113) Svozil, D.; Sponer, J. E.; Marchan, I.; Perez, A.; Cheatham, T. E.; Forti, F.; Luque, F. J.; Orozco, M.; Sponer, J. Geometrical and electronic structure variability of the sugar-phosphate backbone in nucleic acids. *J. Phys. Chem. B* **2008**, *112*, 8188–8197.

(114) 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.

(115) Wang, Y.; Patel, D. J. Solution structure of the human telomeric repeat $d[AG_3(T_2AG_3)_3]$ G-tetraplex. *Structure* **1993**, *1*, 263–282.

(116) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. Determination of electrostatic parameters for a polarizable force field based on the classical Drude oscillator. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.

(117) Halgren, T. A.; Damm, W. Polarizable force fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

CT900200K

JCTC Journal of Chemical Theory and Computation

Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition

Xavier Periole,[†,‡,§] Marco Cavalli,[†] Siewert-Jan Marrink,[‡] and Marco A. Ceruso*[,†]

Department of Chemistry and Biochemistry and Institute for Macromolecular Assemblies, The City College of New York, 160 Convent Ave, New York, New York 10031, and Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

Received April 29, 2009

Abstract: Structure-based and physics-based coarse-grained molecular force fields have become attractive approaches to gain mechanistic insight into the function of large biomolecular assemblies. Here, we study how both approaches can be combined into a single representation, that we term ELNEDIN. In this representation an elastic network is used as a structural scaffold to describe and maintain the overall shape of a protein and a physics-based coarse-grained model (MARTINI-2.1) is used to describe both inter- and intramolecular interactions in the system. The results show that when used in molecular dynamics simulations ELNEDIN models can be built so that the resulting structural and dynamical properties of a protein, including its collective motions, are comparable to those obtained using atomistic protein models. We then evaluate the behavior of such models in (1) long, microsecond time-scale, simulations, (2) the modeling of very large macromolecular assemblies, a viral capsid, and (3) the study of a protein−protein association process, the reassembly of the ROP homodimer. The results for this series of tests indicate that ELNEDIN models allow microsecond time-scale molecular dynamics simulations to be carried out readily, that large biological entities such as the viral capsid of the cowpea mosaic virus can be stably modeled as assemblies of independent ELNEDIN models, and that ELNEDIN models show significant promise for modeling protein−protein association processes.

Introduction

Computational modeling of molecular mechanisms of biological processes is a challenging task. It requires models that can reproduce accurately not only the structural and the dynamical properties of all molecular entities involved (i.e., protein receptors, protein effectors, ligands, lipid molecules, aqueous environment) but also the transient intermolecular interactions in which these entities engage and that modulate their various functional states. This task is often complicated further by the size of the biological systems involved and by the time scales over which these functional processes occur.[1]

One way to circumvent these challenges, without sacrificing on the resolution at which the biological system is represented, is to take advantage of experimentally available information.[2−4] For example, the existence of gain-of-function mutants[5,6] within the Gα subunit of the heterotrimeric G protein, transducin, was exploited. It enabled detailed atomistic insights to be obtained via classical molecular dynamics simulations, about the allosteric structural changes that accompany the activation of the G protein

* To whom correspondence should be addressed. Tel: +1-212-650-6035. Fax: +1-212-650-6107. Email: mceruso@sci.ccny.cuny.edu.
† The City College of New York.
‡ University of Groningen.
§ Current address: Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands.

10.1021/ct9002114 CCC: $40.75
Published on Web 08/19/2009

by its cognate transmembrane receptor rhodopsin. This was possible without modeling the receptor, the lipid environment, or their intermolecular interactions with transducin explicitly.[7] Another way to address these challenges is to make use of algorithms that enhance conformational sampling (for recent reviews see[1,8−14]). But the size of the systems that can be studied with such techniques is as limited as for the more classical approaches such as atomistic molecular dynamics simulations.

An approach that has attracted a great deal of interest has been to develop simplified molecular models to reduce the number of degrees of freedom that need to be taken into account.[1,10,15−18] This approach is particularly advantageous because it permits to increase the time-scale of the simulation and the size of the molecular system simultaneously. The challenge for these simplified or coarse-grained (CG) molecular models is to achieve an accurate description of the free energy surface. Transferability of the CG model is also a challenge. Ideally one would like to have a CG model readily applicable in a variety of molecular contexts.[1]

To date, CG models have been developed for a variety of biomolecules including lipids,[19−22] proteins[10,23−28] and DNA.[29−31] Typically a CG model groups atoms into single interaction centers. The degree of coarse-graining can vary from 2−6 atoms to the whole macromolecule and such force fields are usually parametrized following a knowledge-based or a physics-based approach (see refs 1, 10, and 18 for a description of recent CG models). From the latter category, the MARTINI force field,[21,22] compatible with the so-called GROMOS philosophy,[32] is based on the parametrization of a large library of building blocks against experimental thermodynamic data. This approach is particularly valuable because a number of biological phenomena (e.g., protein folding, peptide-membrane binding, or protein−protein association) depend on the degree to which the constituent groups partition between polar and nonpolar environments. In addition because the building blocks can be combined to construct virtually any molecule,[22,28] they offer an immediate solution to the transferability challenge. Finally, another advantage of the MARTINI force field is that its degree of coarse-graining (∼residue-level) is sufficiently detailed to maintain a close tie with experimental approaches that probe the involvement of particular residues in functional mechanisms. The MARTINI model has already been used in a number of protein studies, including the self-assembly of rhodopsins[33] and the gating of mechanosensitive[34] and voltage gated membrane channels.[35]

Elastic Network (EN) models represent another form of coarse-graining (for recent reviews see[36−38]). EN models were first introduced by Tirion[39] as an alternative to classical atomistic normal-mode analysis. By taking the native structure as the minimum of the free energy EN models eliminated the need of an often-costly minimization procedure, but in so doing, EN models introduced an intrinsic bias toward the initial experimental structure. In an EN model, the structure of a macromolecule is described as a network of point masses connected to one another with springs when the distance between the point masses is less than a predefined cutoff distance (R_C). In the simplest form of EN model, the values of the spring force constant, K_{SPRING}, and the cutoff, R_C, are taken to be the same throughout the network. These two parameters, characterize the network, that is, its rigidity and its extent. A number of variants have been proposed throughout the years[40−42] and the effects of various schemes for building the network of connected residues and setting the spring force constants have been evaluated in a number of studies.[43−46] Recently, EN models have attracted a lot of interest because in combination with the rotational-translational block approximation[47−50] the size of the biological systems that can be studied has been increased dramatically (e.g., see refs 51 and 52). But most importantly, the EN approach has shown a remarkable ability to reproduce a number of biologically relevant dynamical properties of macromolecules.[36,38]

EN models have been combined with atomistic (AT) and CG molecular force fields. When used in combination with AT models the main focus has been to enhance or to guide conformational sampling.[53−57] When used in combination with CG models the main goal has been to maintain the structure of the modeled biomolecule. Bond and co-workers have used both a classical EN scaffold ($R_C = 0.7$ nm and $K_{SPRING} = 1000$ kJ mol^{-1} nm^{-2}) and hydrogen-bond based harmonic distance restraints (force constant 1000 kJ mol^{-1} nm^{-2}), when modeling membrane proteins and membrane peptides, respectively.[58,59] Similarly, Periole et al.[33] used a combination of sequential ($C\alpha_i \rightarrow C\alpha_{i+4}$, $C\alpha_i \rightarrow C\alpha_{i+20}$, and $C\alpha_i \rightarrow C\alpha_{i+30}$ with $K = 9250$ kJ mol^{-1} nm^{-2}) and distance-based (between elements of secondary structure with $K = 1250$ kJ mol^{-1} nm^{-2}) harmonic restraints when studying the self-assembly of rhodopsin in membrane models. While a qualitative agreement between AT and CG simulations was found in these CG studies, the effect of using such restraints on the structure and the dynamics of the model remains to be determined.

In the present study, we investigate the possibility of combining both structure-based and physics-based CG models into a unique representation. Specifically, we characterize the structural and dynamical consequences of combining an EN model with the MARTINI force field.[22,28] For the sake of simplicity we focus only on EN scaffolds characterized by two unique parameters: R_C and K_{SPRING}. We use three model proteins, the B1 domain of protein G,[60] the src-SH3 domain,[61] and the villin headpiece subdomain,[62] respectively representing the $\alpha + \beta$ structural class, the all-β structural class, and the all-α structural class, to determine whether optimal and possibly universal values for the EN parameters, R_C and K_{SPRING}, can be identified. The aim is to have the combined EN-CG protein model, referred to as an ELNEDIN protein model from now on, reproduce quantitatively the structural and dynamical properties of the same protein simulated with an atomistic force field. In addition, we evaluate the behavior of an ELNEDIN model in the context of long (microsecond time-scale) molecular dynamics simulations, and large macromolecular assemblies (the capsid of the cowpea mosaic virus[63]). Finally, we evaluate the ability of an ELNEDIN model to be used for the study of protein−protein association processes. This last test was

carried out in the context of the homodimeric four-helix bundle repressor of primer protein, ROP.[64,65]

Methods

Protein Systems. A total of five protein systems were studied. The villin headpiece subdomain (PDB[66] entry 1YRF[62]), the D48G mutant of the α-spectrin SH3 domain (PDB entry 1BK2[61]), and the B1 domain of protein G (PDB entry 1PGB[60]) were used for the comparison of ELNEDIN and atomistic models. The capsid of the cowpea mosaic virus (PDB entry 1NY7[63]), which consists of 60 copies of two proteins containing 190 and 369 residues respectively, was used to evaluate the capacity of ELNEDIN to handle large macromolecular assemblies. And finally a mutant of the homodimeric four-helix bundle repressor of primer protein,[64] with an Ala residue inserted on either side of D31 (PDB entry 1RPO[65]) was used for protein−protein association evaluation.

Atomistic Simulations. All molecular dynamics (MD) simulations were performed using the GROMACS simulation package.[67] Simulations using an atomistic (AT) representation were based on the united-hydrogen GROMOS-43a1 force field[68] for the protein and the SPC water model for the solvent.[69] Each system (see below) was solvated in a rectangular box. The minimum distance between the protein and the edges of the box was initially set to 1.0 nm. The systems were simulated at constant pressure (1 bar) and constant temperature (300 K). Both the temperature and pressure were maintained close to their target values using the Berendsen weak coupling algorithm[70] with time constants $\tau_T = 0.1$ ps and $\tau_P = 1$ ps, for the temperature and pressure respectively. A twin-range cutoff (1.0−1.4 nm) was used for the nonbonded interactions. Interactions within the short-range cutoff (1.0 nm) were evaluated every time step (2 fs), whereas interactions within the long-range cutoff (1.4 nm) were updated every 10 steps together with the pair-list. To correct for the truncation of electrostatic interactions beyond the long-range cutoff a Reaction-Field[71] correction was applied ($\varepsilon = 78$). Bond lengths were constrained using the LINCS algorithm[72] for the protein and the SETTLE algorithm[73] for the water.

After solvation each system was energy-minimized and the solvent relaxed for 10 ps with position restraints (1000 kJ mol^{-1} nm^{-2}) applied to all heavy atoms of the protein. When necessary explicit counterions were added to ensure electroneutrality of the simulation box, and the resulting neutralized system was energy-minimized again. The system was then simulated for 10 ps at constant pressure and temperature, with the restraining potentials applied on Cα atoms only. Finally the system was simulated for 100 ns without any restraints. The last 60 ns were used in the analyses.

ELNEDIN Models and Simulations. The current version of ELNEDIN is based on the version 2.1 of the MARTINI molecular force field.[22,28] However, because we chose to use the position of the Cα atom to place the backbone bead, instead of that of the center of mass of the −N, Cα, C, O− main chain atoms, as in MARTINI, several minor modifications were introduced. These modifications, which only involve changes in the bonded interactions and the structural mapping of aromatic residues from AT to coarse-grained (CG) representation, are detailed in the Supporting Information section. The nonbonded interactions described in MARTINI 2.1 were not modified.

The EN scaffold component of the ELNEDIN models was built only across the CG backbone beads (i.e., Cα atoms). Two backbone beads were linked by a spring with force constant, K_{SPRING}, only if the distance between them in the experimental structure was less than a predefined cutoff value, R_C, and if they were separated by at least two positions in the protein sequence (see Supporting Information for details on how bonded interactions between sequential, $i \rightarrow i + 1$ and $i \rightarrow i + 2$ residues were described). For a given ELNEDIN model, the values of R_C and K_{SPRING} are identical for all pairs of backbone beads. The equilibrium length of a given spring was set to the experimentally observed distance between the two Cα atoms that it connects.

In the simulations using an ELNEDIN model, the temperature and pressure were treated as in the AT simulations, with $\tau_T = 0.5$ ps and $\tau_P = 1.2$ ps. The nonbonded interactions were treated with a switch function from 0.0 to 1.2 nm for the Coulomb interactions, and from 0.9 to 1.2 nm for the Lennard-Jones interactions, conform the standard MARTINI protocol. The integration time-step was set to 20 fs and the neighbor list was updated every 5 steps.

Each ELNEDIN system was solvated in a cubic box with a minimum of 1.2 nm between any protein bead and the edge of the box. After energy-minimization with position restraints (1000 kJ mol^{-1} nm^{-2}) applied to all protein beads, a 50 ps MD simulation using a 1 fs time-step (real times) with the same position restraints was used to relax both the solvent molecules and the protein in the force field. The system was further relaxed with a 1 ns long MD run with a 20 fs time-step and position restraints on the protein "backbone" beads. Finally each system was simulated for production without any restraints. The length of the various production runs is reported in the text.

Note that because of the smoothing of the energy surface in the CG model the time scales are generally faster. Typically a standard conversion factor of 4 is used, corresponding to the effective speed up factor in the diffusion dynamics of CG water compared to real water taking into account that the CG water represents 4 real waters.[21] The CG simulation times reported in the Results are thus effective times (4 × simulation-time) noted with an asterisk (*), unless otherwise stated.

Comparison of ELNEDIN and Atomistic Models. To evaluate the effect of the properties of the EN scaffold on the structural and dynamical properties of a protein modeled using an ELNEDIN representation, the values of R_C and K_{SPRING} were varied systematically in the range 10−10000 kJ mol^{-1} nm^{-2} and 0.6−1.2 nm, respectively. AT simulations were used as benchmark. The comparison of ELNEDIN and AT simulations was based on four structural and dynamical quantities. These quantities were computed from the equilibrated portions of the MD trajectories. The last 60 ns of 100 ns-long MD simulations were used in the case of AT runs,

and the last 15 ns (60 ns*) of 20 ns-long MD simulations for the CG runs. The degree of motion and fluctuation is time-scale dependent, it is therefore important that the length of the trajectories used in the comparison is the same for both models. The four quantities were (1) the time average of the Cα root-mean-square deviation, RMSD, (2) the root-mean-square deviation per residue, RMSD_res, (3) the root-mean-square fluctuation per residue, RMSF_res, and (4) the essential subspace (first ten eigenvectors of the covariance matrix of positional fluctuations; see below and references[7,74−77]). The RMSD and RMSD_res quantify the global and local structural deformation from the experimental structure, while the RMSF and essential subspace characterize the local fluctuations (deviation from the mean) and the direction of the large-amplitude fluctuations of the biomolecule, respectively.

Four similarity indices were defined. The ΔRMSD index was defined as the absolute difference between the average values of RMSD of the protein in the two approaches

$$\Delta\mathrm{RMSD} = |\langle \mathrm{RMSD}\rangle^{\mathrm{AT}}_{\mathrm{last60ns}} - \langle \mathrm{RMSD}\rangle^{\mathrm{ELNEDIN}}_{\mathrm{last60ns*}}|$$

The index of similarity for comparing RMSD_res and RMSF_res values obtained in CG and AT models was defined as

$$\Delta\mathrm{RMSX\_res} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{RMSX\_res}_i^{\mathrm{AT}} - \mathrm{RMSX\_res}_i^{\mathrm{ELNEDIN}})^2}$$

with RMSX representing either RMSD or RMSF, and N the number of residues in a protein.

And finally, the similarity between the essential subspaces obtained from AT and CG models was quantified by computing the root-mean-square inner-product (RMSIP) between the first 10 first eigenvectors in each simulation[76−79]

$$\mathrm{RMSIP} = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\sum_{j=1}^{10}(\eta_i^{\mathrm{AT}}\cdot\eta_j^{\mathrm{ELNEDIN}})^2}$$

where η_i^{AT} and $\eta_j^{\mathrm{ELNEDIN}}$ are the ith and jth eigenvectors obtained from the AT and ELNEDIN simulation, respectively. For a given system and simulation the essential subspace was computed by diagonalizing the covariance matrix of positional fluctuations $[C_{ij}]_{i,j\in\{1,\ldots,3N\}^2}$ whose elements are given by

$$C_{ij} = \langle (q_i - \langle q_i\rangle)\cdot(q_j - \langle q_j\rangle)\rangle$$

where q_i is one of the Cartesian coordinates of one of the Cα atoms in the molecule and $\langle q_i\rangle$ is the corresponding average value over the ensemble of configurations considered for analysis.

Results and Discussion

Components of an ELNEDIN Model. Figure 1 illustrates the various components that make up an ELNEDIN model. The rationale underlying ELNEDIN is to combine a structure-based coarse-grained model, such as an elastic network,[39] with a physics-based coarse-grained (CG) molecular force field to represent a protein. The elastic network acts as a structural scaffold while the force field directs intermolecular interactions. It should be noted that Tozzini and McCammon have followed a similar rationale using Morse potentials to introduce a bias toward the native structure and to supplement a statistics-based force field.[24] Here, we use a simple two-parameter, R_C/K_{SPRING}, elastic network. In addition to this elastic network scaffold, each amino acid is geometrically represented and "typified" (Lennard-Jones and Coulomb potential parameters) according to the MARTINI force field.[22,28] (Slight modifications to the latest version of the MARTINI force field were needed, see Methods, these changes are described in detail in the Supporting Information). It is important to note that backbone beads linked with springs do not interact with each other via nonbonded potentials (Lennard-Jones and Coulomb) because bonded beads are excluded from the nonbonded interaction lists. Thus, the choice of values for R_C and K_{SPRING} will directly affect the extent to which either bonded or nonbonded potentials contribute to the internal dynamics of a given protein model.

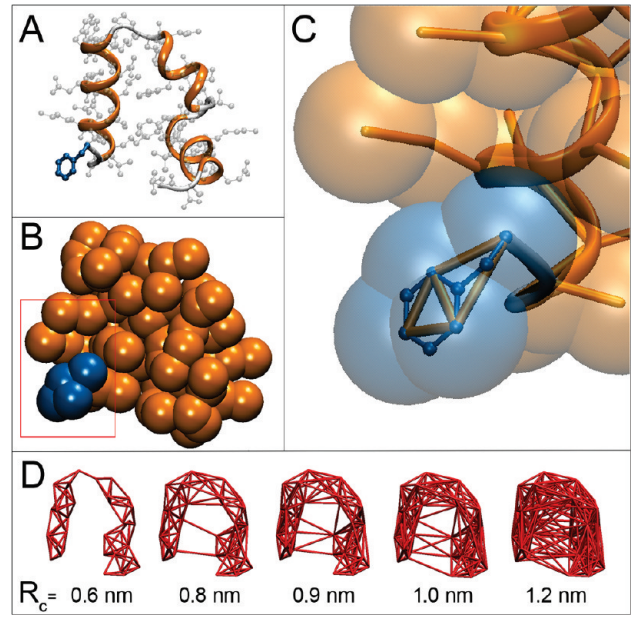


Figure 1. Components of the ELNEDIN model of the villin headpiece subdomain. (A) Ribbon and ball and stick representations of the villin-headpiece subdomain (PDB entry 1YRF[62]). The N-terminal phenylalanine residue is highlighted in blue. (B) CPK representation of a coarse-grained model of the villin-headpiece subdomain based on the modified MARTINI force field. The N-terminal phenylalanine is highlighted in blue. (C) Close-up view of the N-terminal phenylalanine with details of its all atom (blue) and CG (gold) bonding network. (E) Five elastic network scaffolds of the villin-headpiece subdomain built (from left to right) with R_C = 0.6, 0.8, 0.9, 1.0, and 1.2 nm, respectively. All graphics were created using the visualization software VMD.[97]

Influence of the Properties of the Elastic Network on the Structure and Dynamics of a Protein. To evaluate the influence of the scaffold parameters, R_C and K_{SPRING}, on the structure and the dynamics of a protein we first focused on the B1 domain of protein G,[60] a protein whose conformational

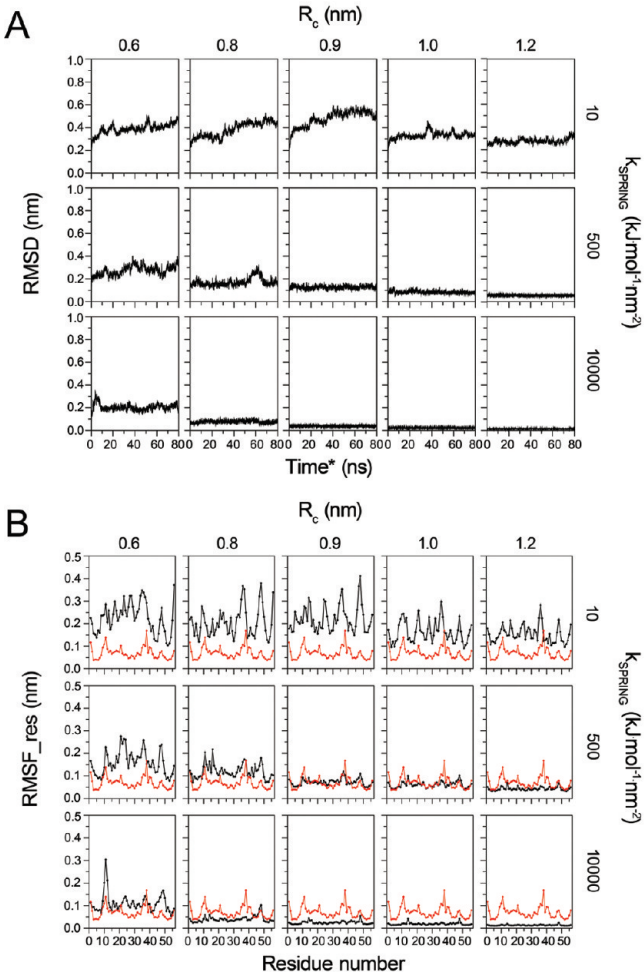


Figure 2. Effect of K_{SPRING} and R_C values on the structure and dynamics of the B1 domain of protein G. (A) Root-mean-square deviation from the experimental structure as a function of effective time. (B) Root-mean-square fluctuation of the backbone bead of each residue as a function of residue number (black curve). The root-mean-square fluctuation of the Cα atoms calculated from an all-atom MD simulation trajectory is shown in red to illustrate the similarities and the differences between the two approaches.

properties we have studied in detail previously.[77] Fifteen different scaffolds were built by varying R_C from 0.6 to 1.2 nm, and K_{SPRING} from 10 to 10000 kJ mol^{-1} nm^{-2}.

The main results of these computational experiments are summarized in Figure 2. The root-mean-square deviation (RMSD) of the backbone beads with respect to their crystallographic position plotted as a function of time showed that the global deformation of the protein decreases when increasing both R_C and K_{SPRING} values (Figure 2A). The deformation of the protein can exceed 0.5 nm with a flexible and undersized scaffold (small values of R_C and K_{SPRING}, upper left panel Figure 2A) and can be as low as 0.01 nm with the stiffest and most extended variant (large values of R_C and K_{SPRING}, lower right panel Figure 2A). Intermediate deformations are observed not only with intermediate values of R_C and K_{SPRING} but also with combinations of short/strong and long/weak values for R_C and K_{SPRING}, respectively. This behavior indicates that R_C and K_{SPRING} compensate each other to maintain the overall structure of the protein.

A similar behavior is observed when monitoring the root-mean-square fluctuation of each residue (RMSF_res) with respect to its average position (Figure 2B). However, it can also be seen that the pattern of the residue-fluctuations (distribution of peaks and valleys across the sequence) varies significantly in some cases between one set of R_C and K_{SPRING} parameters and another. These variations indicate that notwithstanding the compensating effect of the scaffold parameters in relation to overall deformation and amplitude of fluctuations, accurate fluctuation patterns that would agree with experimental B-factors or residue fluctuations computed from atomistic (AT) simulations may only be achieved for specific combinations of the scaffold parameters. This effect can be observed in Figure 2B by comparing the AT (red) and CG (black) fluctuations that show that only the parameters sets $R_C = 0.9$ or 1.0 nm with $K_{SPRING} = 500$ kJ mol^{-1} nm^{-2} provide reasonable overlap between the CG and AT models. To address this issue more in depth, we carried out a systematic comparison between several structural and dynamical properties computed from MD simulations using AT and ELNEDIN representations using three distinct proteins each representing a different structural class.

Comparison of Structural and Dynamical Properties Computed from MD Simulations using AT and ELNEDIN Representations. The B1 domain of protein G, the villin headpiece subdomain, and the α-spectrin SH3 domain, were used for the comparison of AT and ELNEDIN models. These proteins belong to different structural classes: the B1 domain belongs to the $\alpha + \beta$ class, the villin headpiece is an all-α protein, and the SH3 domain belongs to the all-β class. The small size of these proteins allowed relatively long (~100 ns) MD simulations to be performed in an acceptable amount of time with an atomistic (AT) representation. It is important to note that limiting ourselves to small proteins does not affect the generality of our study because ultimately we do not intend to scaffold large proteins with a single EN but with discontinuous or independent elastic networks instead (see below).

For each protein, MD simulations using an AT representation were carried for 100 ns, and MD simulations using an ELNEDIN representation were carried out for 20 ns (corresponding to an effective time of 80 ns*). The parameters for the EN scaffold were varied systematically with R_C (nm) ∈ {0.6, 0.8, 0.9, 1.0, 1.2} and K_{SPRING} (kJ mol^{-1} nm^{-2}) ∈ {10, 50, 100, 200, 500, 1000, 2000, 5000, 10000}. Thus for each protein a total of 45 MD simulations were performed using an ELNEDIN representation.

Four physical quantities were computed from each MD trajectory: (1) the time-average root-mean-square deviation (RMSD) of the backbone beads (Cα atoms), which quantifies the global deformation of the protein with respect to the experimental model, (2) the root-mean-square deviation of the backbone beads per residue (RMSD_res) which quantifies the structural deformation (deviation from the initial structure) of each amino acid, (3) the root-mean-square fluctuation of the backbone beads per residue (RMSF_res) which measures the fluctuation (deviation with respect to the mean position) of each residue, and (4) the large-amplitude collective motions of each protein system. The collective

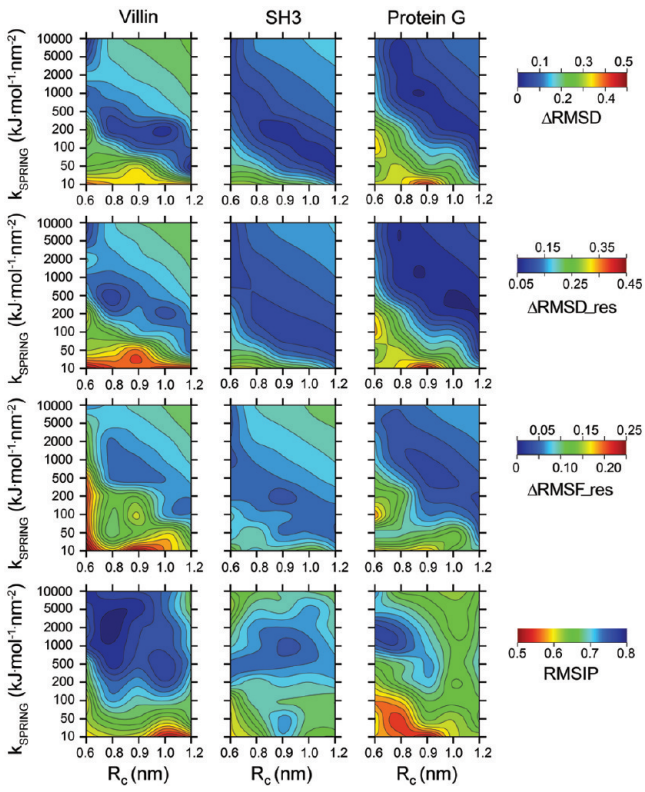


Figure 3. Comparing ELNEDIN and AT representations. The values of ΔRMSD, ΔRMSD_res, ΔRMSF_res, and RMSIP are reported for each of the three model proteins. For all four indices of similarity the color-coding ranges from red to blue, so that regions of low similarity between ELNEDIN and AT models always appear in red, while regions of high similarity are colored in blue. Note that low values for ΔRMSD, ΔRMSD_res, ΔRMSF_res indicate high similarity but that low RMSIP values indicate low similarity.

motions were computed by essential dynamics analysis.[74,75] The comparison between ELNEDIN and AT was quantified using four similarity indices: ΔRMSD, ΔRMSD_res, ΔRMSF_res, and RMSIP (see Methods). The values of these indices are reported on 2D contour maps in R_C/K_{SPRING} space (Figure 3). The values of the indices are color-coded so that the similarity between the ELNEDIN and AT simulations increased from red to blue for all four indices considered. Note that for ΔRMSD, ΔRMSD_res, and ΔRMSF_res the lower the value (minimum = 0 nm) the better the agreement between ELNEDIN and AT models; but for RMSIP the higher the value (maximum = 1) the better the correlation between ELNEDIN and AT models. Similarity indices below ~0.15 nm (cyan to blue) were considered good for ΔRMSD and ΔRMSD_res. The corresponding cutoff values for ΔRMSF_res and RMSIP were set at ~0.075 nm or below, and 0.75 or above, respectively.

The 2D contour maps for the ΔRMSD, ΔRMSD_res and ΔRMSF_res indices confirmed the compensatory relationship between R_C and K_{SPRING}. Indeed, for these three indices, the EN scaffolds which provide the best agreement between ELNEDIN and AT models (blue regions in the corresponding panels) are clearly distributed diagonally across each of the corresponding 2D contour map (Figure 3). This is the case for all three proteins and suggests that the compensatory behavior is independent of the structural class of the protein. This result is consistent with the work of Tirion who reported a similar inverse relationship between these two parameters in order to maximize the agreement between EN and classical normal-mode analysis.[39] It is noteworthy, however, that the width of the diagonal region within which quantitative agreement (cyan to blue contours) between ELNEDIN and AT is achieved, varies from one protein to the other, suggesting that optimal values for R_C and K_{SPRING} may depend on the specific protein.

Such a protein specific behavior is even more marked in the case of the RMSIP index. For this index, the region of R_C and K_{SPRING} values for which the ELNEDIN model agrees with the AT model is confined to a specific perimeter, which is not diagonal (i.e., the effects of R_C and K_{SPRING} do not compensate one another) and whose extent and shape differs for each protein (compare the blue regions across bottom three panels in Figure 3). The reason why a protein specific behavior appears readily in the RMSIP index is that the computation of this latter index is more sensitive to the variation of the measured property (in this case the positional fluctuation) across the sequence than the other indices. Indeed, in the case of the ΔRMSD, ΔRMSD_res and ΔRMSF_res indices these sequence variations are averaged out, and therefore not as readily detectable, although present (see the distribution of peaks and valleys of the residue fluctuations of the B1 domain of protein G in Figure 2B). Thus, taken together, these results suggest that while the compensatory mode in which R_C and K_{SPRING} influence the structural and dynamical properties may be independent of the specific protein being modeled, the region in the R_C/K_{SPRING} space that gives the best quantitative agreement between ELNEDIN and AT models is protein specific.

Nevertheless, a region of R_C/K_{SPRING} space in which ELNEDIN models provide adequate quantitative structural and dynamical agreement with AT models could be delineated (data not shown). A search for a consensus set of parameters across the various indices and protein systems revealed that values within 0.8 and 1.0 nm for R_C and ranging from 500 to 1000 kJ mol^{-1} nm^{-2} for K_{SPRING} could provide adequate quantitative agreement with atomistic simulations. These values are close to the values of R_C = 0.7 nm and K_{SPRING} = 1000 kJ mol^{-1} nm^{-2} used by Bond and co-workers.[58,59] The values are also within the range used in typical EN applications,[39−41,43−45,72,80,81] which range from 0.7 to 1.6 nm for R_C and from 200 to 4000 kJ mol^{-1} nm^{-2} for K_{SPRING}. It is important to note that in EN-based normal-mode analysis the spring force constant is a parameter that is usually adjusted a posteriori in order to match the amplitude of the fluctuations with experimental B-factors for example. In that case the value of the spring force constant only affects the amplitude of the motions, not their directions. EN models are therefore often referred to as single parameter (R_C) models. This is not the case for an ELNEDIN model. Both R_C and K_{SPRING} contribute to the accuracy of the model. This is because, as stated earlier, beads connected via springs do not interact with each other via nonbonded potentials. Thus an increase (respectively decrease) in the value of R_C will

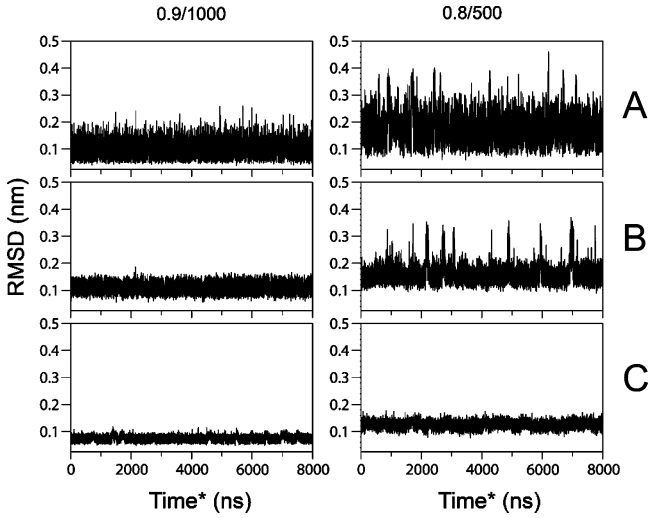


Figure 4. Long time-scale simulations using two distinct scaffolds. The RMSD vs time* of the three protein models relative to their respective experimental structure during long ELNEDIN simulations is shown. (A) The villin headpiece subdomain. (B) The B1 domain of protein G. (C) the src SH3 domain. The values of the R_C (nm) and K_{SPRING} (kJ mol^{-1} nm^{-2}) constants used in the simulations are indicated above the panels: left panel simulations use a 0.9/1000 R_C/K_{SPRING} combination, while the right panel simulations use a 0.8/500 combination.

modify the potential energy in two ways: (1) it will reduce (respectively increase) the number of bead pairs that can interact via nonbonded potentials and (2) increase (respectively reduce) the number of bead pairs that will interact via harmonic springs (K_{SPRING}). This interplay between the elastic network and the nonbonded potentials of the CG force field has the appealing consequence of introducing anisotropy in the description of the protein, akin to that sought in anisotropic network models,[41,46,81] despite the use of an isotropic network.

Long Time-Scale Behavior of ELNEDIN Models. To evaluate the behavior of ELNEDIN models during long MD simulations, each of the three test proteins, the villin headpiece subdomain, the B1 domain of protein G, and the α-spectrin SH3 domain, was simulated for 8 μs*. Two distinct combinations of spring force constant and cutoff radius were used: (i) $R_C = 0.9$ nm, $K_{SPRING} = 1000$ kJ mol^{-1} nm^{-2} (0.9/1000) and (ii) $R_C = 0.8$ nm, $K_{SPRING} = 500$ kJ mol^{-1} nm^{-2} (0.8/500). The time series of the RMSD of each protein from its experimental structure are shown in Figure 4. Overall, the plots indicate that in all cases the native fold is well preserved (average RMSD value ≤0.2 nm). This is expected since we are using an EN scaffold for that very purpose. But the plots also reveal how the quality of the scaffold might influence the ability of a given system to experience transient structural transitions during a simulation. This effect is clear in the case of the villin headpiece subdomain and the B1 domain of protein G, which experience more conformational transitions when modeled with the 0.8/500 scaffold than with the 0.9/1000 one. Visual inspection of the trajectories indicated that these conformational changes involved changes in orientation of secondary structure elements. These transitions had a lifetime on the order of 10−100 ns, and the protein always returned to the native state. This behavior was not as marked for the SH3 domain in which only very small amplitude conformational transitions were observed. The smaller amplitude of the conformational transitions in the SH3 system may be due to its higher degree of compactness with respect to the other systems (e.g., for a given cutoff value SH3 has the largest average number of springs per residue followed by protein G and then villin). Taken together, these data indicate that despite the intrinsic bias that an EN scaffold might introduce toward the native structure it does not preclude changes in the relative orientation of secondary structure elements. However the ability to observe such transitions depends on the choice of EN parameters.

Table 1. Correlation Coefficients between NMR- and CG-Based Residue Fluctuations[a]

R_C/K_{SPRING}	Villin headpiece subdomain	B1 domain of protein G[d]	SH3 domain
0.8/500	0.71[b] − 0.85[c]	0.73	0.75[e]−0.86[f]
0.9/1000	0.80[b]	0.62	0.68[e]

[a] Only Cα atoms were used to compute the fluctuations. [b] The NMR ensemble of structures was taken from entry 2JMO[98] in the PDB. Note that this entry contains a fluorinated residue. [c] The correlation coefficient with respect to the fluctuations computed from a 4 μs*-long simulation with $R_C/K_{SPRING} = 0.8/500$ of the wild-type NMR structure (PDB entry 1VII[99]) is 0.85. [d] The NMR ensemble structures was taken from entry 1GB1[100] in the PDB. [e] The ensemble of NMR structures was taken from entry 1AEY[101] in the PDB. This entry corresponds to the wild-type sequence protein and not the mutant sequence used here. [f] The correlation coefficient with respect to the fluctuations computed from a 4 μs*-long simulation with $R_C/K_{SPRING} = 0.8/500$ of the wild-type structure (PDB entry 1SHG[102]) is 0.86.

To ascertain that the dynamic behavior observed for each protein is inline with experimental data we compared the residue-based fluctuations obtained from each long-time-scale simulation to those computed from the NMR ensemble of the cognate structure deposited in the PDB. The correlation coefficients between the two sets of values are reported in Table 1. Overall the data shows good agreement with experimentally derived fluctuations with correlation coefficients ranging from 0.62 to 0.86. These values indicate that the pattern of fluctuations across the sequence obtained from the simulation was very similar to that obtained from the ensemble of structures satisfying the NMR restraints. Note that the level of agreement between the experimental and ELNEDIN data depends, to a large extent, on the quality of the underlying atomistic simulations from which the optimal choice of EN parameters was inferred.

Application of ELNEDIN to Large Macromolecular Assemblies. We chose to model the viral capsid of the Cowpea Mosaic Virus (CPMV PDB entry 1NY7[63]). The viral capsid of CPMV is a highly symmetrical assembly consisting of 60 copies of two proteins having 190 and 369 residues, respectively. This assembly is almost spherical in shape with a ∼26 nm diameter (Figure 5A). After solvation, the system contained 268,883 CG beads, which would correspond to 2,852,940 atoms if an atomistic representation were used.

Since a single EN scaffold might adversely affect the dynamics of the viral capsid, for example, by interfering with

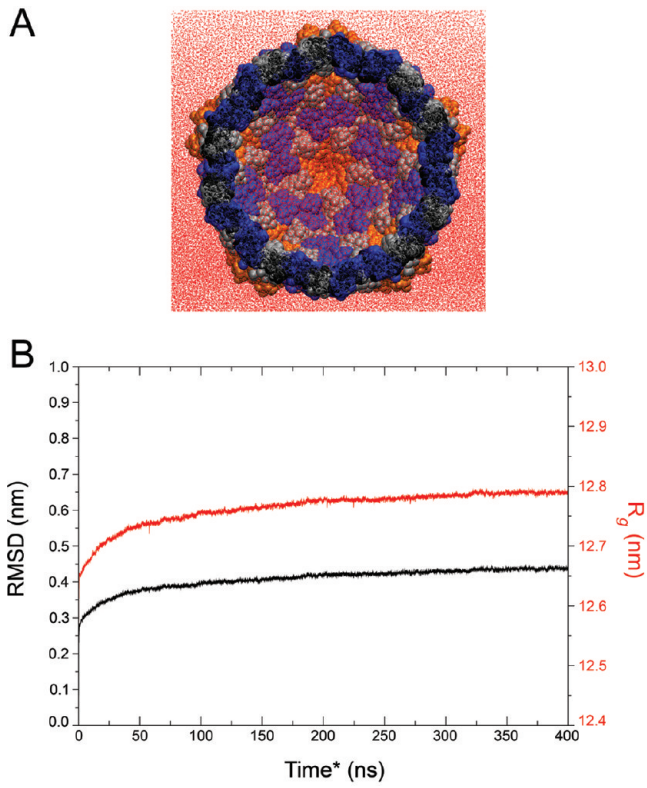


Figure 5. Modeling the Cowpea Mosaic virus (CPMV). (A) The viral capsid of CPMV is shown sliced across the middle. The red dots are solvent molecules; the S viral protein is shown in orange and the L viral protein in gray and blue for aesthetic reasons only. (B) Plot of the RMSD from the initial configuration and radius of gyration (R_g) of the capsid vs time*.

changes in the relative orientations of the protein domains with respect to each other, EN scaffolds were built for each protein domain separately. But each EN scaffold was built using the same values of R_C and K_{SPRING}, that is, 0.9 nm for R_C and 500 kJ mol^{-1} nm^{-2} for K_{SPRING}. Note that beads in distinct proteins interact according to the nonbonded terms of the MARTINI force field. The overall ELNEDIN model was simulated for 400 ns* at 1 atm and 300 K.

The time series of the RMSD and the radius of gyration (R_g) are shown in Figure 5B. Both parameters indicate that the CPMV capsid is structurally stable. The overall RMSD, which remains below ~0.5 nm throughout the simulation, is remarkably small considering the size of the system. Moreover, the viral capsid did not show any symptom of collapse as has been observed for similar systems using an atomistic force field[82] or a more approximate level of coarse-graining than the one used here.[83] This suggests that large macromolecules can be modeled effectively as assemblies of independent ELNEDIN models. One reason for this apparent success might be that maintaining the structural integrity of each subunit independently may have contributed favorably to the overall stability of the viral capsid. This effect may need to be investigated further if reliable mechanistic inferences are sought concerning CPMV. However, this is beyond the scope of the current study. Finally, it is interesting to note that it took about 50 ns for the viral capsid to reach a reasonably stable value for the RMSD and R_g, suggesting that for such large systems the period of relaxation is significantly longer than for smaller proteins.

Application of ELNEDIN to Protein−Protein Association. As a last application, we evaluated the ability of ELNEDIN models to be used for the study of protein−protein association processes. A mutant (PDB entry 1RPO[65]) of the repressor of primer protein (ROP) was used for this test. ROP is a homo dimeric four-helix bundle,[64] each monomer consisting of two antiparallel α-helices. ELNEDIN models were prepared for the native dimer configuration as well as for three other configurations in which the monomers were placed at 0.5 nm (TX = 0.5), 1.0 nm (TX = 1.0), and 1.5 nm (TX = 1.5) from each other by simple translation in the direction perpendicular to the plane defined by the monomer−mononer interface. The relative orientation of the two monomers was not modified. It would have added the difficult task of sampling the conformational space, which is beyond the scope of the present test. For each starting configuration, native, TX = 0.5, TX = 1.0, and TX = 1.5, five independent MD simulations (400 ns*) were carried out by using different sets of initial velocities. Two slightly distinct EN scaffolds were tested: one with R_C = 0.9 nm and K_{SPRING} = 500 kJ mol^{-1} nm^{-2} (0.9/500), and the other with R_C = 1.0 nm and K_{SPRING}= 1000 kJ mol^{-1} nm^{-2} (1.0/1000). Note that while each monomer model used the same R_C and K_{SPRING} values for the EN scaffold, each monomer possessed its own separate scaffold and was thus free to move independently from the other monomer during the MD simulation.

The MD simulations of the native dimer configuration showed that the four-helix bundle could adopt two distinct conformations. These are marked native 1 and native 2 in Figure 6A. The native 2 conformation differs from the native 1 (experimental structure) in the degree by which the monomers are tilted with respect to each other. The significance of this second orientation is not clear but it is noteworthy that similar but less pronounced changes have also been observed in atomistic simulations of ROP.[78] The two native conformations were observed for both EN scaffolds. When simulating the translated systems formation of either the native 1 or native 2 states was taken as a successful reassembly event.

The results of the various monomer association tests were as follows. Monomers that were placed 0.5 nm apart (TX = 0.5 runs) were always able to reassemble into a native structure, and this was true for both EN scaffolds. TX = 1.0 systems were also able to reassemble into a native structure but in this case the more flexible scaffold performed better than the more rigid one: 5 out of 5 successful reassemblies were obtained for the 0.9/500 scaffold vs 2 out of 5 for the 1.0/1000 scaffold. The same trend but with less success overall was observed for the TX = 1.5 systems: 2 out of 5 runs produced a native structure when using the 0.9/500 scaffold vs only 1 out of 5 runs produced a native structure when using the 1.0/1000 scaffold. Runs that failed to produce a native structure after 400 ns* of MD simulation all showed that the monomers had assembled into non-native states burying sometimes as much surface area as the native structure (data not shown). No attempt was made to study

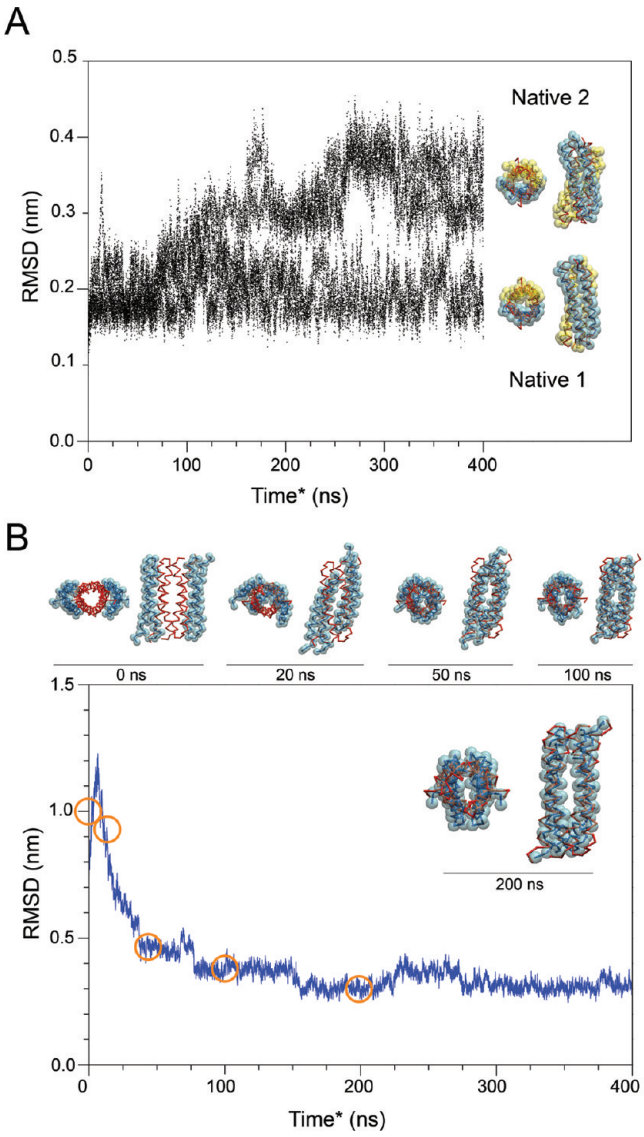


Figure 6. Modeling association of ROP monomers. (A) RMSD plots vs time for the simulations of the native dimer reveal two stable conformational states. (B) Structural intermediates observed in the course of one of the protein association runs and corresponding RMSD trace. The individual monomers (in blue) were initially placed at a distance of 1.5 nm from each other. The trace of the experimental structure is shown in red. In these representations the solvent is omitted for clarity.

these non-native systems, which would be considered false positives in a docking experiment, in more detail.

Taken together, these results show that the quality of the scaffold affects the ability of the monomers to reassemble into a native structure. This is not surprising since the stability of a given structure (be it tertiary[84] or quaternary in this case) is likely to arise from a balance between local/internal and long-range interactions. It is noteworthy that the more successful scaffold is also the more flexible one (0.9/500 vs 1.0/1000). This finding is consistent with the recognized importance of flexibility in molecular recognition[85−87] and underlines the necessity to include an accurate description of a protein's internal dynamics when attempting to predict protein−protein interactions.[88−93]

Visual inspection of the reassembly process in the various successful runs revealed the interesting fact that the sequence of conformational rearrangements that followed the initial encounter and led to the final native structure was not unique. In fact, reassembly proceeded in a different manner in each case. Such multiplicity of pathways has been observed by others[94,95] and has led to the hypothesis that protein association like protein folding also proceeds on a funnel-shaped free-energy landscape.[96] No effort was made to analyze these pathways further since many more simulations would be needed to establish statistically significant results. Nevertheless, for illustration purposes only, one such pathway is depicted in Figure 6B where the time series of the RMSD value of the complex relative to the experimental structure is shown together with representative structures at the different stages of the association process. The figure shows that the two monomers rapidly came into contact with each other. The packing of the core is out-of-register (see side-view at 20 ns*) and is shifted laterally (see top-view) with respect to the native state (shown in red). Following this encounter the process of association took ∼130 ns* and could be described as one monomer slithering along the length of the other monomer until the stable native interface (native 2) was reformed. The complex then remained stable for the following 250 ns* of the simulation with an RMSD value of 0.3 nm from the experimental structure.

Overall the result of these, albeit limited, docking experiments are extremely encouraging and suggests that the combination of a structure-based and a physics-based CG model such as ELNEDIN might provide a valuable approach to the difficult task of predicting protein−protein interfaces and association processes.

Conclusion

We have investigated the possibility to combine a structure-based and physics-based coarse-grained model into a unique representation. An elastic network was mixed with the MARTINI 2.1 force field for the purpose of carrying out molecular dynamics simulations of biological systems. The effect of varying the properties of the EN on the quality of the model was studied systematically. The results of these computational studies show that it is possible to identify appropriate values for the EN parameters, R_C and K_{SPRING}, such that the resulting ELNEDIN model is capable to reproduce simultaneously the global and local deformations of a protein, its residue fluctuations, and its large-amplitude collective motions, as observed in atomistic models. Although, the optimal values for R_C and K_{SPRING} depend on the specific protein studied, we find that values ranging from 0.8 to 1.0 nm for R_C and from 500 to 1000 kJ mol^{-1} nm^{-2} for K_{SPRING} can provide adequate quantitative agreement with atomistic simulations. The results also show that ELNEDIN models are stable enough to allow microsecond time-scale molecular dynamics simulations to be carried out readily. In some cases, transient structural changes corresponding to changes in orientation of elements of secondary structure can be observed during these simulations. But it should be noted that because of the inherent structural bias of EN toward

the reference configuration, the model cannot be expected to produce conformational changes akin to those necessary for a protein to fold. Finally, we find that large biological entities such as the viral capsid of the cowpea mosaic virus can be stably modeled as assemblies of independent ELNEDIN models, and that ELNEDIN models show significant promise for modeling protein association processes.

Abbreviations. AT, atomistic; CG, coarse-grained; EN, elastic network; MD, molecular dynamics; NMR, nuclear magnetic resonance; RMSD, root-mean-square deviation; RMSD_res, root-mean-square deviation per residue; RMSF_res root-mean-square fluctuation per residue; RMSIP, root-mean-square inner-product. CPMV, cowpea mosaic virus; ROP, repressor of primer.

Acknowledgment. The authors acknowledge funding from the CUNY Research Foundation and a CUNY Incentive Collaborative Grants to MAC, as well as from The Netherlands Organization for Scientific Research (NWO) for SJM and XP. The authors also thank the CUNY High performance computing facility (CUNY HPC) and The Netherlands National Computing Facilities (NCF) for allocation of computing time.

Supporting Information Available: Description of the protocol used to extract the parameters for angle bending and bond stretching potentials (other than elastic network bonds) used in ELNEDIN models, description of the structural mapping (from AT to CG) of aromatic residues implemented in ELNEDIN models, and the complete set bonded parameters for amino acid residues as used in this study (note that the nonbonded parameters were taken without modification from the MARTINI force field). This set of parameters should not be considered as a new version of the MARTINI force field but as an alternative to MARTINI-2.1. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(2) de Vlieg, J.; Berendsen, H. J. C.; van Gunsteren, W. F. An NMR-Based Molecular-Dynamics Simulation of the Interaction of the Lac Repressor Headpiece and Its Operator in Aqueous-Solution. *Proteins* **1989**, *6*, 104–127.

(3) Paci, E.; Vendruscolo, M.; Dobson, C. M.; Karplus, M. Determination of a Transition State at Atomic Resolution from Protein Engineering Data. *J. Mol. Biol.* **2002**, *324*, 151–163.

(4) van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. Molecular Simulation as an Aid to Experimentalists. *Curr. Opin. Struct. Biol.* **2008**, *18*, 149–153.

(5) Marin, E. P.; Krishna, A. G.; Sakmar, T. P. Rapid Activation of Transducin by Mutations Distant from the Nucleotide-Binding Site - Evidence for a Mechanistic Model of Receptor-Catalyzed Nucleotide Exchange by G Proteins. *J. Biol. Chem.* **2001**, *276*, 27400–27405.

(6) Marin, E. P.; Krishna, A. G.; Sakmar, T. P. Disruption of the α5 Helix of Transducin Impairs Rhodopsin-Catalyzed Nucleotide Exchange. *Biochemistry* **2002**, *41*, 6988–6994.

(7) Ceruso, M. A.; Periole, X.; Weinstein, H. Molecular Dynamics Simulations of Transducin: Interdomain and Front to Back Communication in Activation and Nucleotide Exchange. *J. Mol. Biol.* **2004**, *338*, 469–481.

(8) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. Computational Techniques for Efficient Conformational Sampling of Proteins. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.

(9) Christen, M.; van Gunsteren, W. F. On Searching in, Sampling of, and Dynamically Moving through Conformational Space of Biomolecular Systems: A Review. *J. Comput. Chem.* **2008**, *29*, 157–166.

(10) Ayton, G. S.; Noid, W. G.; Voth, G. A. Multiscale Modeling of Biomolecular Systems: In Serial and in Parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(11) Elber, R. Long-Timescale Simulation Methods. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151–156.

(12) Tai, K. Conformational Sampling for the Impatient. *Biophys. Chem.* **2004**, *107*, 213–220.

(13) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(14) Schlick, T. Time-Trimming Tricks for Dynamic Simulations: Splitting Force Updates to Reduce Computational Work. *Structure* **2001**, *9*, R45–53.

(15) Flory, P. J. Statistical Thermodynamics of Random Networks. *Proc. R. Soc. A* **1976**, *351*, 351–380.

(16) Levitt, M. A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. *J. Mol. Biol.* **1976**, *104*, 59–107.

(17) Chan, H. S.; Dill, K. A. Intrachain Loops in Polymers—Effects of Excluded Volume. *J. Chem. Phys.* **1989**, *90*, 492–509.

(18) Voth, G. A., Ed. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: New York, 2009.

(19) Smit, B.; Hilbers, P. A. J.; Esselink, K.; Rupert, L. A. M.; van Os, N. M.; Schlijper, A. G. Computer Simulations of a Water/Oil Interface in the Presence of Micelles. *Nature* **1990**, *348*, 624–625.

(20) Saiz, L.; Klein, M. L. Computer Simulation Studies of Model Biological Membranes. *Acc. Chem. Res.* **2002**, *35*, 482–489.

(21) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(22) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The Martini Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(23) Zuckerman, D. M. Simulation of an Ensemble of Conformational Transitions in a United-Residue Model of Calmodulin. *J. Phys. Chem. B* **2004**, *108*, 5127–5137.

(24) Tozzini, V.; McCammon, J. A. A Coarse Grained Model for the Dynamics of Flap Opening in HIV-1 Protease. *Chem. Phys. Lett.* **2005**, *413*, 123–128.

(25) Tozzini, V.; Rocchia, W.; McCammon, J. A. Mapping All-Atom Models onto One-Bead Coarse-Grained Models: General Properties and Applications to a Minimal Polypeptide Model. *J. Chem. Theory Comput.* **2006**, *2*, 667–673.

(26) Pizzitutti, F.; Marchi, M.; Borgis, D. Coarse-Graining the Accessible Surface and the Electrostatics of Proteins for Protein−Protein Interactions. *J. Chem. Theory Comput.* **2007**, *3*, 1867–1876.

(27) Basdevant, N.; Borgis, D.; Ha-Duong, T. A Coarse-Grained Protein−Protein Potential Derived from an All-Atom Force Field. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.

(28) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The Martini Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(29) Tan, R. K.-Z.; Petrov, A. S.; Devkota, B.; Harvey, S. C. Coarse-Grained Models for Nucleic Acids and Large Nucleoprotein Assemblies. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: New York, 2009; Chapter 15, pp 225−235.

(30) Tepper, H. L.; Voth, G. A. A Coarse-Grained Model for Double-Helix Molecules in Solution: Spontaneous Helix Formation and Equilibrium Properties. *J. Chem. Phys.* **2005**, *122*, 124906.

(31) Knotts, T. A. T.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. A Coarse Grain Model for DNA. *J. Chem. Phys.* **2007**, *126*, 084901–084912.

(32) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The Gromos Force-Field Parameter Sets 53a5 and 53a6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(33) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. G Protein-Coupled Receptors Self-Assemble in Dynamics Simulations of Model Bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.

(34) Yefimov, S.; van der Giessen, E.; Onck, P. R.; Marrink, S. J. Mechanosensitive Membrane Channels in Action. *Biophys. J.* **2008**, *94*, 2994–3002.

(35) Treptow, W.; Marrink, S. J.; Tarek, M. Gating Motions in Voltage-Gated Potassium Channels Revealed by Coarse-Grained Molecular Dynamics Simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(36) Bahar, I.; Rader, A. J. Coarse-Grained Normal Mode Analysis in Structural Biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.

(37) Ma, J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure* **2005**, *13*, 373–380.

(38) Tama, F.; Brooks, C. L. Symmetry, Form, and Shape: Guiding Principles for Robustness in Macromolecular Machines. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 115–133.

(39) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905.

(40) Bahar, I.; Kaplan, M.; Jernigan, R. L. Short-Range Conformational Energies, Secondary Structure Propensities, and Recognition of Correct Sequence-Structure Matches. *Proteins* **1997**, *29*, 292–308.

(41) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of Proteins Predicted by Molecular Dynamics Simulations and Analytical Approaches: Application to α-Amylase Inhibitor. *Proteins* **2000**, *40*, 512–524.

(42) Song, G.; Jernigan, R. L. An Enhanced Elastic Network Model to Represent the Motions of Domain-Swapped Proteins. *Proteins* **2006**, *63*, 197–209.

(43) Kundu, S.; Melton, J. S.; Sorensen, D. C.; Phillips, G. N., Jr. Dynamics of Proteins in Crystals: Comparison of Experiment with Simple Models. *Biophys. J.* **2002**, *83*, 723–732.

(44) Sen, T. Z.; Feng, Y.; Garcia, J. V.; Kloczkowski, A.; Jernigan, R. L. The Extent of Cooperativity of Protein Motions Observed with Elastic Network Models Is Similar for Atomic and Coarser-Grained Models. *J. Chem. Theory Comput.* **2006**, *2*, 696–704.

(45) Kondrashov, D. A.; Van Wynsberghe, A. W.; Bannen, R. M.; Cui, Q.; Phillips, G. N. Protein Structural Variation in Computational Models and Crystallographic Data. *Structure* **2007**, *15*, 169–177.

(46) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins. *Biophys. J.* **2008**, *95*, 4183–4192.

(47) Durand, P.; Trinquier, G.; Sanejouand, Y. H. A New Approach for Determining Low-Frequency Normal Modes in Macromolecules. *Biopolymers* **1994**, *34*, 759–771.

(48) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-Block Approach for Determining Low-Frequency Normal Modes of Macromolecules. *Proteins* **2000**, *41*, 1–7.

(49) Li, G.; Cui, Q. A Coarse-Grained Normal Mode Approach for Macromolecules: An Efficient Implementation and Application to Ca^{2+}-ATPase. *Biophys. J.* **2002**, *83*, 2457–2474.

(50) Li, G.; Cui, Q. Analysis of Functional Motions in Brownian Molecular Machines with an Efficient Block Normal Mode Approach: Myosin-II and Ca^{2+}-ATPase. *Biophys. J.* **2004**, *86*, 743–763.

(51) Mitra, K.; Schaffitzel, C.; Shaikh, T.; Tama, F.; Jenni, S.; Brooks, C. L.; Ban, N.; Frank, J. Structure of the E-Coli Protein-Conducting Channel Bound to a Translating Ribosome. *Nature* **2005**, *438*, 318–324.

(52) Tama, F.; Brooks, C. L. 3rd Diversity and Identity of Mechanical Properties of Icosahedral Viral Capsids Studied with Elastic Network Normal Mode Analysis. *J. Mol. Biol.* **2005**, *345*, 299–314.

(53) He, J. B.; Zhang, Z. Y.; Shi, Y. Y.; Liu, H. Y. Efficiently Explore the Energy Landscape of Proteins in Molecular Dynamics Simulations by Amplifying Collective Motions. *J. Chem. Phys.* **2003**, *119*, 4005–4017.

(54) Miyashita, O.; Onuchic, J. N.; Wolynes, P. G. Nonlinear Elasticity, Proteinquakes, and the Energy Landscapes of Functional Transitions in Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 12570–12575.

(55) Tatsumi, R.; Fukunishi, Y.; Nakamura, H. A Hybrid Method of Molecular Dynamics and Harmonic Dynamics for Docking of Flexible Ligand to Flexible Receptor. *J. Comput. Chem.* **2004**, *25*, 1995–2005.

(56) Miller, B. T.; Zheng, W.; Venable, R. M.; Pastor, R. W.; Brooks, B. R. Langevin Network Model of Myosin. *J. Phys. Chem. B* **2008**, *112*, 6274–6281.

(57) Zacharias, M. Combining Elastic Network Analysis and Molecular Dynamics Simulations by Hamiltonian Replica Exchange. *J. Chem. Theory Comput.* **2008**, *4*, 477–487.

(58) Bond, P. J.; Sansom, M. S. Insertion and Assembly of Membrane Proteins Via Simulation. *J. Am. Chem. Soc.* **2006**, *128*, 2697–2704.

(59) 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.

(60) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. Two Crystal Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein G and Comparison with NMR. *Biochemistry* **1994**, *33*, 4721–4729.

(61) Martinez, J. C.; Pisabarro, M. T.; Serrano, L. Obligatory Steps in Protein Folding and the Conformational Diversity of the Transition State. *Nat. Struct. Biol.* **1998**, *5*, 721–729.

(62) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. High-Resolution X-Ray Crystal Structures of the Villin Headpiece Subdomain, an Ultrafast Folding Protein. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7517–7522.

(63) Lin, T. W.; Chen, Z. G.; Usha, R.; Stauffacher, C. V.; Dai, J. B.; Schmidt, T.; Johnson, J. E. The Refined Crystal Structure of Cowpea Mosaic Virus at 2.8 Angstrom Resolution. *Virology* **1999**, *265*, 20–34.

(64) Banner, D. W.; Kokkinidis, M.; Tsernoglou, D. Structure of the Cole1 Rop Protein at 1.7 Å Resolution. *J. Mol. Biol.* **1987**, *196*, 657–675.

(65) Vlassi, M.; Steif, C.; Weber, P.; Tsernoglou, D.; Wilson, K. S.; Hinz, H. J.; Kokkinidis, M. Restored Heptad Pattern Continuity Does Not Alter the Folding of a Four-Alpha-Helix Bundle. *Nat. Struct. Biol.* **1994**, *1*, 706–716.

(66) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(67) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(68) van Gunsteren, W. F.; Daura, X.; Mark, A. E. Gromos Force Field. *Encyclopaedia Comput. Chem.* **1998**, *2*, 1211–1216.

(69) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interactions Models for Water in Relation to Protein Hydration. In Intermolecular Forces, Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331−342.

(70) 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. Chem. Phys.* **1984**, *81*, 3684–3690.

(71) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(72) Hinsen, K. Analysis of Domain Motions by Approximate Normal Mode Calculations. *Proteins* **1998**, *33*, 417–429.

(73) Miyamoto, S.; Kollman, P. A. Settle−An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(74) Garcia, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.

(75) Amadei, A.; Linssen, A. B.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412–425.

(76) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations. *Proteins* **1999**, *36*, 419–424.

(77) Ceruso, M. A.; Amadei, A.; Di Nola, A. Mechanics and Dynamics of B1 Domain of Protein G: Role of Packing and Surface Hydrophobic Residues. *Protein Sci.* **1999**, *8*, 147–160.

(78) Ceruso, M. A.; Grottesi, A.; Di Nola, A. Effects of Core-Packing on the Structure, Function, and Mechanics of a Four-Helix-Bundle Protein Rop. *Proteins* **1999**, *36*, 436–446.

(79) Ceruso, M. A.; Grottesi, A.; Di Nola, A. Dynamic Effects of Mutations within Two Loops of Cytochrome C551 from Pseudomonas Aeruginosa. *Proteins* **2003**, *50*, 222–229.

(80) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.* **1997**, *79*, 3090.

(81) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505–515.

(82) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure* **2006**, *14*, 437–449.

(83) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure* **2006**, *14*, 1767–1777.

(84) Ceruso, M. A.; Weinstein, H. Structural Mimicry of Proline Kinks: Tertiary Packing Interactions Support Local Structural Distortions. *J. Mol. Biol.* **2002**, *318*, 1237–1249.

(85) Yamniuk, A. P.; Vogel, H. J. Calmodulin's Flexibility Allows for Promiscuity in Its Interactions with Target Proteins and Peptides. *Mol. Biotechnol.* **2004**, *27*, 33–57.

(86) Basdevant, N.; Weinstein, H.; Ceruso, M. Thermodynamic Basis for Promiscuity and Selectivity in Protein-Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc.* **2006**, *128*, 12766–12777.

(87) Bakan, A.; Lazo, J. S.; Wipf, P.; Brummond, K. M.; Bahar, I. Toward a Molecular Understanding of the Interaction of Dual Specificity Phosphatases with Substrates: Insights from Structure-Based Modeling and High-Throughput Screening. *Curr. Med. Chem.* **2008**, *15*, 2536–2544.

(88) Deremble, C.; Lavery, R. Macromolecular Recognition. *Curr. Opin. Struct. Biol.* **2005**, *15*, 171–175.

(89) May, A.; Zacharias, M. Accounting for Global Protein Deformability During Protein−Protein and Protein−Ligand Docking. *Biochim. Biophys. Acta* **2005**, *1754*, 225–231.

(90) McCammon, J. A. Target Flexibility in Molecular Recognition. *Biochim. Biophys. Acta* **2005**, *1754*, 221–224.

(91) Bonvin, A. M. J. J. Flexible Protein−Protein Docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194–200.

(92) Totrov, M.; Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.

(93) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. Principles of Flexible Protein−Protein Docking. *Proteins Struct. Funct. Gen.* **2008**, *73*, 271–289.

(94) Camacho, C. J.; Vajda, S. Protein Docking Along Smooth Association Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10636–10641.

(95) Camacho, C. J.; Weng, Z.; Vajda, S.; DeLisi, C. Free Energy Landscapes of Encounter Complexes in Protein−Protein Association. *Biophys. J.* **1999**, *76*, 1166–1178.

(96) Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R. Folding and Binding Cascades: Dynamic Landscapes and Population Shifts. *Protein Sci.* **2000**, *9*, 10–19.

(97) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(98) Cornilescu, G.; Hadley, E. B.; Woll, M. G.; Markley, J. L.; Gellman, S. H.; Cornilescu, C. C. Solution Structure of a Small Protein Containing a Fluorinated Side Chain in the Core. *Protein Sci.* **2007**, *16*, 14–19.

(99) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR Structure of the 35-Residue Villin Headpiece Subdomain. *Nat. Struct. Biol.* **1997**, *4*, 180–184.

(100) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein G. *Science* **1991**, *253*, 657–661.

(101) Blanco, F. J.; Ortiz, A. R.; Serrano, L. ^{1}H and ^{15}N NMR Assignment and Solution Structure of the SH3 Domain of Spectrin: Comparison of Unrefined and Refined Structure Sets with the Crystal Structure. *J. Biomol. NMR* **1997**, *9*, 347–357.

(102) Musacchio, A.; Noble, M.; Pauptit, R.; Wierenga, R.; Saraste, M. Crystal Structure of a Src-Homology 3 (SH3) Domain. *Nature* **1992**, *359*, 851–855.

CT9002114

JCTC Journal of Chemical Theory and Computation

The AGBNP2 Implicit Solvation Model

Emilio Gallicchio,* Kristina Paris, and Ronald M. Levy

Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854

Received May 11, 2009

Abstract: The AGBNP2 implicit solvent model, an evolution of the Analytical Generalized Born plus NonPolar (AGBNP) model we have previously reported, is presented with the aim of modeling hydration effects beyond those described by conventional continuum dielectric representations. A new empirical hydration free energy component based on a procedure to locate and score hydration sites on the solute surface is introduced to model first solvation shell effects, such as hydrogen bonding, which are poorly described by continuum dielectric models. This new component is added to the generalized Born and nonpolar AGBNP terms. Also newly introduced is an analytical Solvent Excluded Volume (SEV) model which improves the solute volume description by reducing the effect of spurious high dielectric interstitial spaces present in conventional van der Waals representations. The AGBNP2 model is parametrized and tested with respect to experimental hydration free energies of small molecules and the results of explicit solvent simulations. Modeling the granularity of water is one of the main design principles employed for the first shell solvation function and the SEV model, by requiring that water locations have a minimum available volume based on the size of a water molecule. It is shown that the new volumetric model produces Born radii and surface areas in good agreement with accurate numerical evaluations of these quantities. The results of molecular dynamics simulations of a series of miniproteins show that the new model produces conformational ensembles in substantially better agreement with reference explicit solvent ensembles than the original AGBNP model with respect to both structural and energetics measures.

1. Introduction

Water plays a fundamental role in virtually all biological processes. The accurate modeling of hydration thermodynamics is therefore essential for studying protein conformational equilibria, aggregation, and binding. Explicit solvent models arguably provide the most detailed and complete description of hydration phenomena.[1] They are, however, computationally demanding not only because of the large number of solvent atoms involved, but also because of the need to average over many solvent configurations to obtain meaningful thermodynamic data. Implicit solvent models,[2] which are based on the statistical mechanics concept of the solvent potential of mean force,[3] have been shown to be useful alternatives to explicit solvation for applications including protein folding and binding,[4] and small molecule hydration free energy prediction.[5]

Modern implicit solvent models[6,7] include distinct estimators for the nonpolar and electrostatic components of the hydration free energy. The nonpolar component corresponds to the free energy of hydration of the uncharged solute, while the electrostatic component corresponds to the free energy of turning on the solute partial charges. The latter is typically modeled treating the water solvent as a uniform high dielectric continuum.[8] Methods based on the numerical solution of the Poisson−Boltzmann (PB) equation[9] provide a virtually exact representation of the response of the solvent within the dielectric continuum approximation. Recent advances extending dielectric continuum approaches have focused on the development of Generalized Born (GB) models,[10] which have been shown to reproduce with good accuracy PB and explicit solvent[7,11] results at a fraction of

* Corresponding author e-mail emilio@biomaps.rutgers.edu.

10.1021/ct900234u CCC: $40.75
Published on Web 07/31/2009

the computational expense. The development of computationally efficient analytical and differentiable GB methods based on pairwise descreening schemes[6,12,13] has made possible the integration of GB models in molecular dynamics packages for biological simulations.[14−16]

The nonpolar hydration free energy component accounts for all nonelectrostatic solute−solvent interactions as well as hydrophobic interactions,[17] which are essential driving forces in biological processes such as protein folding[18−21] and binding.[22−25] Historically the nonpolar hydration free energy has been modeled by empirical surface area models[26] which are still widely employed.[10,27−35] Surface area models are useful as a first approximation; however, qualitative deficiencies have been observed.[29,36−41]

A few years ago we presented the Analytical Generalized Born plus NonPolar (AGBNP) implicit solvent model,[42] which introduced two key innovations with respect to both the electrostatic and nonpolar components. Unlike most implicit solvent models, the AGBNP nonpolar hydration free energy model includes distinct estimators for the solute−solvent van der Waals dispersion energy and cavity formation work components. The main advantages of a model based on the cavity/dispersion decomposition of the nonpolar solvation free energy stem from its ability to describe both medium-range solute−solvent dispersion interactions, which depend on solute composition, as well as effects dominated by short-range hydrophobic interactions, which can be modeled by an accessible surface area term.[40] A series of studies highlight the importance of the balance between hydrophobicity and dispersion interactions in regulating the structure of the hydration shell and the strength of interactions between macromolecules.[43−45] In AGBNP the work of cavity formation is described by a surface area dependent model,[37,46−48] while the dispersion estimator is based on the integral of van der Waals solute−solvent interactions over the solvent, modeled as a uniform continuum.[38] This form of the nonpolar estimator had been motivated by a series of earlier studies[5,37,49−52] and has since been shown by us[38,53−55] and others[39−41,56] to be qualitatively superior to models based only on the surface area in reproducing explicit solvent results as well as rationalizing structural and thermodynamic experimental observations.

The electrostatic solvation model in AGBNP is based on the pairwise descreening GB scheme[13] whereby the Born radius of each atom is obtained by summing an appropriate descreening function over its neighbors. The main distinction between the AGBNP GB model and conventional pairwise descreening implementations is that in AGBNP the volume scaling factors, which offset the overcounting of regions of space occupied by more than one atom, are computed from the geometry of the molecule rather than being introduced as geometry-independent parameters fit to either experiments or to numerical Poisson−Boltzmann results.[14,57−59] The reduction of the number of parameters achieved with this strategy improves the transferability of the model to unusual functional groups often found in ligand molecules, which would otherwise require extensive parametrization.[60]

Given its characteristics, the AGBNP model has been mainly targeted to applications involving molecular dynamics canonical conformational sampling, and to the study of protein−ligand complexes. Since its inception the model has been employed in the investigation of a wide variety of biomolecular problems ranging from peptide conformational propensity prediction and folding,[54,61−63] ensemble-based interpretation of NMR experiments,[64,65] protein loop homology modeling,[55] ligand-induced conformational changes in proteins,[66,67] conformational equilibria of protein−ligand complexes,[68,69] protein−ligand binding affinity prediction,[70] and structure-based vaccine design.[71] The AGBNP model has been reimplemented and adopted with minor modifications by other investigators.[72,73] The main elements of the AGBNP nonpolar and electrostatic models have been independently validated,[39,40,74,75] and have been incorporated in recently proposed hydration free energy models.[76,77]

In this work we present a new implicit solvent model named AGBNP2 which builds upon the original AGBNP implementation (hereafter referred to as AGBNP1) and improves it with respect to the description of the solute volume and the treatment of short-range solute−water electrostatic interactions.

Continuum dielectric models assume that the solvent can be described by a linear and uniform dielectric medium.[78] This assumption is generally valid at the macroscopic level; however, at the molecular level water exhibits significant deviations from this behavior.[1] Nonlinear dielectric response, the nonuniform distribution of water molecules, charge asymmetry, and electrostriction effects[79] are all phenomena originating from the finite size and internal structure of water molecules as well as their specific interactions which are not taken into account by continuum dielectric models. Some of these effects are qualitatively captured by standard classical fixed-charge explicit water models; however others, such as polarization and hydrogen bonding interactions, can be fully modeled only by adopting more complex physical models.[80] GB models make further simplifications in addition to the dielectric continuum approximation, thereby compounding the challenge of achieving with GB-based implicit solvent models the level of realism required to reliably model phenomena, such as protein folding and binding, characterized by relatively small free energy changes.

In the face of these challenges a reasonable approach is to adopt an empirical hydration free energy model motivated by physical arguments[81] parametrized with respect to experimental hydration free energy data.[20] Models of this kind typically score conformations on the basis of the degree of solvent exposure of solute atoms. Historically[82] the solvent accessible surface area of the solute has been used for this purpose, while modern implementations suitable for conformational sampling applications often employ computationally convenient volumetric measures.[83,84] In this work we take this general approach but we retain the GB model component which we believe is a useful baseline to describe the long-range influence of the water medium. The empirical parametrized component of the model then takes the form of an empirical first solvation shell correction function designed so as to absorb hydration effects not accurately described by the GB model. Specifically, as described below, we employ a short-range analytical hydrogen bonding correction

function based on the degree of water occupancy (taking into account the finite size of water molecules) of appropriately chosen hydration sites on the solute surface. The aim of this model is to effectively introduce some explicit solvation features without actually adding water molecules to the system as for example done in hybrid explicit/implicit models.[85,86]

In this work we also improve the description of the solute volume, which in AGBNP1 is modeled by means of atomic spheres of radius equal to the atomic van der Waals radius. The deficiencies of the van der Waals solute volume model have been recognized.[87,88] They stem from the presence of high dielectric interstitial spaces in the solute interior which are too small to contain discrete water molecules. These spurious high dielectric spaces contribute to the hydration of buried or partially buried atoms causing underestimation of desolvation effects. The volume enclosed by the molecular surface (MS), defined as the surface produced by a solvent spherical probe rolling on the van der Waals surface of the solute,[89] represents the region which is inaccessible to water molecules and is often referred as the Solvent Excluded Volume (SEV).[90] The SEV, lacking the spurious high dielectric interstitial spaces, provides a better representation of the low dielectric region associated with the solute. For this reason accurate Poisson−Boltzmann solvers[9,91,92] have employed the SEV description of the solute region.

Despite its clear advantages, the lack of analytical and computationally efficient representation of the SEV have hampered its deployment in conjunction with GB models for molecular dynamics applications. The Generalized Born Molecular Volume (GBMV) series of models[87,93,94] achieve high accuracy relative to numerical Poisson calculations in part by employing the SEV description of the solute volume. The analytical versions of GBMV[93,94] describe the SEV volume by means of a continuous and differentiable solute density function which is integrated on a grid to yield atomic Born radii. In this work we present a model for the SEV that preserves the analytical pairwise atomic descreening approach employed in the AGBNP1 model,[42] which avoids computations on a grid. We show that this approximate model reproduces some of the key features of the SEV while yielding the same favorable algorithmic scaling of pairwise descreening approaches.

This paper focuses primarily on the description and parametrization of the SEV model and the short-range hydrogen bonding function of AGBNP2. In section 2 we present a brief review of the AGBNP1 model, including the electrostatic and nonpolar models, followed by the derivation of the analytical SEV pairwise descreening model and the short-range hydrogen bonding function which are new for AGBNP2. In section 3 we validate the AGBNP2 analytical estimates for the Born radii and atomic surface areas using as a reference accurate numerical evaluations of these quantities. This is followed by the parametrization of the hydrogen bonding function against experimental hydration free energies of small molecules. This section concludes with a comparison between the structural and energetic properties of a series of structured peptides (miniproteins) predicted with the AGBNP2 model and those obtained with explicit solvation. The paper then concludes with a discussion and implications of the results, and with a perspective on future improvements and validation of the AGBNP2 model.

2. Methods

2.1. The Analytical Generalized Born plus Nonpolar Implicit Solvent Model (AGBNP). In this section we briefly review the formulation of the AGBNP1 implicit solvent model, which forms the basis for the new AGBNP2 model. A full account can be found in the original reference.[42] The AGBNP1 hydration free energy $\Delta G_h(1)$ is defined as

$$\begin{aligned}\Delta G_h(1) &= \Delta G_{elec} + \Delta G_{np} \\ &= \Delta G_{elec} + \Delta G_{cav} + \Delta G_{vdW}\end{aligned} \quad (1)$$

where ΔG_{elec} is the electrostatic contribution to the solvation free energy and ΔG_{np} includes nonelectrostatic contributions. ΔG_{np} is further decomposed into a cavity hydration free energy ΔG_{cav} and a solute−solvent van der Waals dispersion interaction component ΔG_{vdW}.

2.1.1. Geometrical Estimators. Each free energy component in eq 1 is ultimately based on an analytical geometrical description of the solute volume modeled as a set of overlapping atomic spheres of radii R_i centered on the atomic positions $\mathbf{r}_i$. Hydrogen atoms do not contribute to the solute volume. The solute volume is modeled using the Gaussian overlap approach first proposed by Grant and Pickup.[95] In this model the solute volume is computed using the Poincaré formula (also known as the inclusion−exclusion formula) for the volume of the union of a set of intersecting elements

$$V = \sum_i V_i - \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} - \ldots \quad (2)$$

where $V_i = 4\pi R_i^3/3$ is the volume of atom i, V_{ij} is the volume of intersection of atoms i and j (second-order intersection), V_{ijk} is the volume of intersection of atoms i, j, and k (third-order intersection), and so on. The overlap volumes are approximated by the overlap integral, $V^g_{12\ldots n}$, available in analytical form (see for example eq 10 of ref 42), between n Gaussian density functions each corresponding to a solute atom:

$$V^g_{12\ldots n} \simeq \int d^3\mathbf{r}\, \rho_1(\mathbf{r})\, \rho_2(\mathbf{r}) \ldots \rho_n(\mathbf{r}) \quad (3)$$

where the Gaussian density function for atom i is

$$\rho_i(\mathbf{r}) = p \exp[-c_i(\mathbf{r} - \mathbf{r}_i)^2] \quad (4)$$

where

$$c_i = \frac{\kappa}{R_i^2} \quad (5)$$

and

$$p = \frac{4\pi}{3}\left(\frac{\kappa}{\pi}\right)^{3/2} \quad (6)$$

and κ is a dimensionless parameter that regulates the diffuseness of the atomic Gaussian function. In the AGBNP1 formulation $\kappa = 2.227$.

Gaussian integrals are in principle nonzero for any finite distances between the Gaussian densities. Although not mentioned in ref 42, to reduce computational cost AGBNP1 includes a switching function that reduces to zero the overlap volume between two or more Gaussians when the overlap volume is smaller than a certain value. If $V^{g}_{12...}$ is the value of the Gaussian overlap volume between a set of atoms, the corresponding overlap volume $V_{12...}$ used in eq 2 is set as

$$V_{12...n} = \begin{cases} 0 & V^{g}_{12...} \le v_1 \\ V^{g}_{12...n} f_w(u) & v_1 < V^{g}_{12...n} < v_2 \\ V^{g}_{12...n} & V^{g}_{12...} \ge v_2 \end{cases} \quad (7)$$

where

$$u = \frac{V^{g}_{12...} - v_1}{v_2 - v_1} \quad (8)$$

$$f_w(x) = x^3(10 - 15x + 6x^2) \quad (9)$$

where, when using van der Waals atomic radii, $v_1 = 0.1$ and $v_2 = 1$ Å^3, and for the augmented radii used in the surface area model (see below), $v_1 = 0.2$ and $v_2 = 2$ Å^3. This scheme sets to zero Gaussian overlap volumes smaller than v_1, leaves volumes above v_2 unchanged, and smoothly reduces volumes between these two limits. It drastically reduces the number of overlap volumes that need to be calculated since the fact that an n-body overlap volume $V_{12...n}$ between n atoms is zero guarantees that all of the $(n + 1)$-body overlap volumes corresponding to the same set of atoms plus one additional atom are also zero. (Note below that the formulation of AGBNP2 employs modified values of v_1 and v_2 to improve the accuracy of surface areas.)

The van der Waals surface area A_i of atom i, which is another geometrical descriptor of the model, is based on the derivative $\partial V/\partial R_i$ of the solute volume with respect to the radius R_i[96]

$$A_i = f_a\left(\frac{\partial V}{\partial R_i}\right) \quad (10)$$

where V is given by eq 2 and

$$f_a(x) = \begin{cases} \dfrac{x^3}{a^2 + x^2} & x > 0 \\ 0 & x \le 0 \end{cases} \quad (11)$$

with $a = 5$ Å^2, is a filter function which prevents negative values for the surface areas for buried atoms while inducing negligible changes to the surface areas of solvent-exposed atoms.

The model further defines the so-called self-volume V'_i of atom i as

$$V'_i = V_i - \frac{1}{2}\sum_j V_{ij} + \frac{1}{3}\sum_{j<k} V_{ijk} + ... \quad (12)$$

which is computed similarly to the solute volume and measures the solute volume that is considered to belong exclusively to this atom. Due to the overlaps with other atoms, the self-volume V'_i of an atom is smaller than the van der Waals volume V_i of the atom. The ratio

$$s_i = \frac{V'_i}{V_i} \le 1 \quad (13)$$

is a volume scaling factor used below in the evaluation of the Born radii.

2.1.2. Electrostatic Model. The electrostatic hydration free energy is modeled using a continuous dielectric representation of the water solvent using the Generalized Born (GB) approximation

$$\Delta G_{\text{elec}} = u_\epsilon \sum_i \frac{q_i^2}{B_i} + 2u_\epsilon \sum_{i<j} \frac{q_i q_j}{f_{ij}} \quad (14)$$

where

$$u_\epsilon = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{w}}}\right) \quad (15)$$

where ϵ_{in} is the dielectric constant of the interior of the solute and ϵ_{w} is the dielectric constant of the solvent; q_i and q_j are the charges of atom i and j, and

$$f_{ij} = \sqrt{r_{ij}^2 + B_i B_j \exp(-r_{ij}^2/4B_iB_j)} \quad (16)$$

In eqs 14−16 B_i denotes the Born radius of atom i which, under the Coulomb field approximation,[57] is given by the inverse of the integral over the solvent region of the negative fourth power of the distance function centered on atom i

$$\beta_i = \frac{1}{B_i} = \frac{1}{4\pi}\int_{\text{solvent}} d^3\mathbf{r}\, \frac{1}{(\mathbf{r} - \mathbf{r}_i)^4} \quad (17)$$

In the AGBNP1 model this integral is approximated by a so-called pairwise descreening formula

$$\beta_i = \frac{1}{R_i} - \frac{1}{4\pi}\sum_{j\neq i} s_{ji} Q_{ji} \quad (18)$$

where R_i is the van der Waals radius of atom i, s_{ji} is the volume scaling factor for atom j (eq 13) when atom i is removed from the solute, and Q_{ji} is the integral (available in analytic form; see Appendix B of ref 42) of the function $(\mathbf{r} - \mathbf{r}_i)^{-4}$ over the volume of the sphere corresponding to solute atom j that lies outside the sphere corresponding to atom i. Equation 18 applies to all of the atoms i of the solute (hydrogen atoms and heavy atoms), whereas the sum over j includes only heavy atoms. The AGBNP1 estimates for the Born radii B_i are finally computed from the inverse Born radii β_i from eq 18 after processing them through the function

$$B_i^{-1} = f_b(\beta_i) = \begin{cases} \sqrt{b^2 + \beta_i^2} & \beta_i > 0 \\ b & \beta_i \le 0 \end{cases} \quad (19)$$

where $b^{-1} = 50$ Å. The filter function eq 19 is designed to prevent the occurrence of negative Born radii or Born radii larger than 50 Å. The goal of the filter function is simply to increase the robustness of the algorithm in limiting cases.

The filter function has negligible effect for the most commonly observed Born radii smaller than 20 Å.

In the AGBNP1 model the scaling factors s_{ji} are approximated by the expression

$$s_{ji} \simeq s_j + \frac{1}{2}\frac{V_{ij}}{V_j} \tag{20}$$

where s_j is given by eq 13 and V_{ij} is the two-body overlap volume between atoms i and j. Also, in the original AGBNP formulation the computation of the scaling factors and the descreening function in eq 18 employed the van der Waals radii for the atoms of the solute and the associated Gaussian densities. These two aspects have been modified in the new formulation (AGBNP2) as described below.

2.1.3. Nonpolar Model. The nonpolar hydration free energy is decomposed into the cavity hydration free energy ΔG_{cav} and the solute−solvent van der Waals dispersion interaction component ΔG_{vdW}:

$$\Delta G_{\mathrm{np}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{vdW}} \tag{21}$$

The cavity component is described by a surface area model[37,46−48]

$$\Delta G_{\mathrm{cav}} = \sum_i \gamma_i A_i \tag{22}$$

where the summation runs over the solute heavy atoms, A_i is the van der Waals surface area of atom i from eq 10, and γ_i is the surface tension parameter assigned to atom i (see Table 1 of ref 42). Surface areas are computed using augmented radii R_i^c for the atoms of the solute and the associated Gaussian densities. Augmented radii are defined as the van der Waals radii (Table 1 of ref 42) plus a 0.5 Å offset. The computation of the atomic surface areas in AGBNP2 is mostly unchanged from the original implementation,[42] with the exception of the values of the switching function cutoff parameters υ_1 and υ_2 of eq 7, which in the new model are set as $\upsilon_1 = 0.01$ Å^3 and $\upsilon_2 = 0.1$ Å^3. This change was deemed necessary to improve the accuracy of the surface areas which in the new model also affect the Born radii estimates through eq 31 below.

The solute−solvent van der Waals free energy term is modeled by the expression

$$\Delta G_{\mathrm{vdW}} = \sum_i \alpha_i \frac{a_i}{(B_i + R_{\mathrm{w}})^3} \tag{23}$$

where α_i is an adjustable dimensionless parameter on the order of 1 (see Table 1 of ref 42) and

$$a_i = -\frac{16}{3}\pi\rho_{\mathrm{w}}\epsilon_{i\mathrm{w}}\sigma_{i\mathrm{w}}{}^6 \tag{24}$$

where $\rho_{\mathrm{w}} = 0.033\,28$ Å^{-3} is the number density of water at standard conditions, and $\sigma_{i\mathrm{w}}$ and $\epsilon_{i\mathrm{w}}$ are the OPLS force field[97] Lennard-Jones interaction parameters for the interaction of solute atom i with the oxygen atom of the TIP4P water model.[98] If σ_i and ϵ_i are the OPLS Lennard-Jones parameters for atom i

$$\sigma_{i\mathrm{w}} = \sqrt{\sigma_i \sigma_{\mathrm{w}}} \tag{25}$$

$$\epsilon_{i\mathrm{w}} = \sqrt{\epsilon_i \epsilon_{\mathrm{w}}} \tag{26}$$

where $\sigma_{\mathrm{w}} = 3.153\,65$ Å and $\epsilon_{\mathrm{w}} = 0.155$ kcal/mol are the Lennard-Jones parameters of the TIP4P water oxygen. In eq 23 B_i is the Born radius of atom i from eqs 18 and 19 and $R_{\mathrm{w}} = 1.4$ Å is a parameter corresponding to the radius of a water molecule.

2.2. The AGBNP2 Implicit Solvent Model. The AGBNP2 hydration free energy $\Delta G_{\mathrm{h}}(2)$ is defined as

$$\Delta G_{\mathrm{h}}(2) = \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{np}} + \Delta G_{\mathrm{hb}} \tag{27}$$

where ΔG_{elec} and ΔG_{np} have the same form as in the AGBNP1 model (eqs 14 and 21−23, respectively). The only major difference is the pairwise descreening model for the Born radii that in AGBNP2 is based on the solvent excluded volume described below rather than the van der Waals volume as in AGBNP1. ΔG_{hb}, described in section 2.2.2, is a novel term for AGBNP2 which represents a first solvation shell correction corresponding to the portion of the hydration free energy not completely accounted for by the uniform continuum model for the solvent. We think of this term as mainly incorporating the effect of solute−solvent hydrogen bonding. As described in detail below, the analytical model for ΔG_{hb} is based on measuring and scoring the volume of suitable hydration sites on the solute surface.

2.2.1. Pairwise Descreening Model Using the Solvent Excluded Volume. When using van der Waals radii to describe the solute volume, small crevices between atoms (Figure 1, panel A) are incorrectly considered as high dielectric solvent regions,[93,99,100] leading to underestimation of the Born radii, particularly for buried atoms. The van der Waals volume description implicitly ignores the fact that the finite size of water molecules prevents them from occupying sites that, even though they are not within solute atoms, are too small to be occupied by water molecules. Ideally a model for the Born radii would include in the descreening calculation all of the volume excluded from water either because it is occupied by a solute atom or because it is located in an interstitial region inaccessible to water molecules. We denote this volume as the solvent excluded volume (SEV). A realistic description of the SEV is the volume enclosed within the molecular surface[89] of the solute obtained by tracing the surface of contact of a sphere with a radius characteristic of a water molecule (typically 1.4 Å) rolling over the van der Waals surface of the solute. The main characteristic of this definition of the SEV (see Figure 2) is that, unlike the van der Waals volume, it lacks small interstitial spaces while it closely resembles the van der Waals volume near the solute−solvent interface. The molecular surface description of the SEV cannot be easily implemented into an analytical formulation. In this section we will present an analytical description of the SEV for the purpose of the pairwise descreening computation of the Born radii, as implemented in AGBNP2, that preserves the main characteristics of the molecular surface description of the SEV.

The main ideas underpinning the SEV model presented here are illustrated in Figure 1. We start with the van der Waals representation of the solute (model A) which presents an

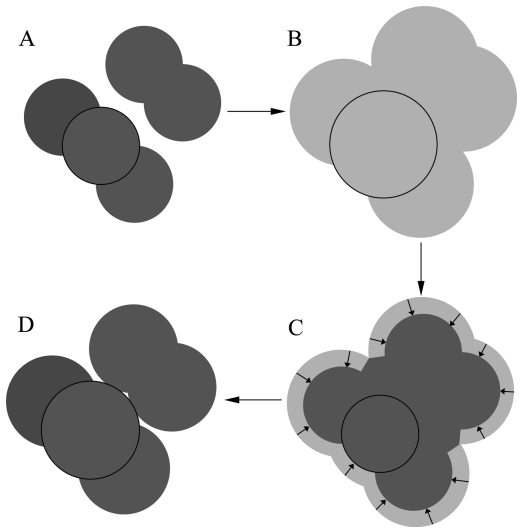


Figure 1. Schematic diagram illustrating the ideas underpinning the model for the solvent excluded volume descreening. Circles represent atoms of two idealized solutes placed in proximity of each other. The van der Waals description of the molecular volume (panel A) leaves high dielectric interstitial spaces that are too small to fit water molecules. The adoption of enlarged van der Waals radii (B) removes the interstitial spaces but incorrectly excludes solvent from the surface of solvent-exposed atoms. The solvent volume subtended by the solvent-exposed surface area is subtracted from the enlarged volume of each atom (C) such that larger atomic descreening volumes are assigned to buried atoms (circled) than exposed atoms (D), leading to the reduction of interstitial spaces while not overly excluding solvent from the surface of solvent-exposed atoms.

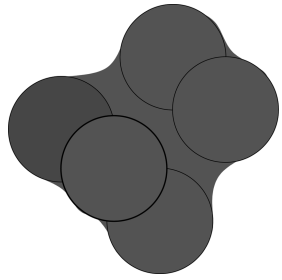


Figure 2. Illustration of the relationship between the van der Waals volume and the solvent excluded volume enclosed by the molecular surface.

undesirable high dielectric interstitial space between the two groups of atoms. Increasing the atomic radii leads to a representation (model B) in which the interstitial space is removed but that also incorrectly excludes solvent from the surface of solvent-exposed atoms. This representation is therefore replaced with one in which the effective volume of each atom in B is reduced by the volume subtended between the solvent-exposed surface of each atom and its van der Waals radius (Figure 1C). This process yields model D in which the effective volume of the most buried atom is larger than those of the solvent-exposed atoms. This SEV model covers the interstitial high dielectric spaces present in a van der Waals description of the solute volume, while approximately maintaining the correct van der Waals volume description of atoms at the solute surface as in the molecular surface description of the SEV (Figure 2).

These ideas have been implemented in the AGBNP2 model as follows. The main modification consists of adopting for the pairwise descreening generalized Born formulation the same augmented van der Waals radii as in the computation of the atomic surface areas. As in the previous model the augmented atomic radii, R_i^c, are set as

$$R_i^c = R_i + \Delta R \tag{28}$$

where R_i is the van der Waals radius of the atom and $\Delta R = 0.5$ Å is the offset. The augmented radii are used in the same way as in the AGBNP1 formulation to define the atomic spheres and the associated Gaussian densities (eqs 3−6). Henceforth in this work all of the quantities (atomic volumes, self-volumes, etc.) are understood to be computed with the augmented atomic radii, unless otherwise specified. In AGBNP2 the form of the expression for the inverse Born radii (eq 18) is unchanged; however, the expressions for the volume scaling factors s_{ji} and the evaluation of the descreening function Q_{ji} are modified as follows to introduce the augmented atomic radii and the reduction of the atomic self-volumes in proportion to the solvent accessible surface areas as discussed above.

The pairwise volume scaling factors s_{ji}, that is the volume scaling factor for atom j when atom i is removed from the solute, are set as

$$s_{ji} = s_j + \frac{V'_{ji}}{V_j} \tag{29}$$

where s_j (defined below) is the volume scaling factor for atom j analogous to eq 13 computed with all the atoms present, and the quantity

$$V'_{ji} = V'_{ij} = \frac{1}{2}V_{ij} - \frac{1}{3}\sum_k V_{ijk} + \frac{1}{4}\sum_{k<l} V_{ijkl} - \ldots \tag{30}$$

subtracts from the expression for the self-volume of atom j all those overlap volumes involving both atoms i and j.

Two differences with respect to the original AGBNP1 formulation are introduced. The first is that s_j is computed from the self-volume after subtracting from it the volume of the region subtended by the solvent-exposed surface between the enlarged and van der Waals atomic spheres of atom j, according to the expression

$$s_j = \frac{V'_j - d_j A_j}{V_j} \tag{31}$$

where A_j is the surface area of atom j from eq 10. Referring to Figure 3, the volume of the subtended region is $d_j A_j$ as in eq 31 with

$$d_j = \frac{1}{3}R'_j\left[1 - \left(\frac{R_j}{R'_j}\right)^3\right] \tag{32}$$

The other difference concerns the V'_{ji} term which in the AGBNP1 formulation is approximated by the two-body overlap volume V_{ij} (see eq 13), the first term in the right-hand side of eq 30. This approximation is found to lack

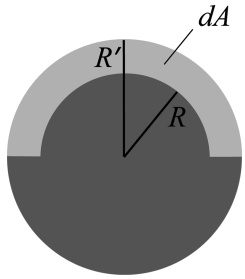


Figure 3. Graphical construction showing the volume subtracted from the atomic self-volume to obtain the surface area corrected atomic self-volume. R is the van der Waals radius of the atom; $R' = R + \Delta R$ is the enlarged atomic radius. dA is the volume of the region (light gray) subtended by the solvent-exposed surface area between the enlarged and van der Waals atomic spheres.

sufficient accuracy for the present formulation given the relative increase in size of all overlap volumes. Therefore in AGBNP2 V'_{ji} is computed including in eq 30 all nonzero overlap volumes after the application of the switching function from eq 7.

In the AGBNP2 formulation the functional form for the pair descreening function Q_{ji} is the same as in the original formulation (see Appendix of ref 42); however, in the new formulation this function is evaluated using the van der Waals radius R_i for atom i (the atom being "descreened") and the augmented radius R_j^c for atom j (the atom that provides the solvent descreening), rather than using the van der Waals radius for both atoms. Thus if the pair descreening function is denoted by $Q(r,R_1,R_2)$, where r is the interatomic distance, R_1 the radius of the atom being descreened, and R_2 the radius that provides descreening, we set in eq 18

$$Q_{ji} = Q(r_{ij}, R_i, R_j^c) \tag{33}$$

The alternative of using enlarged atoms for both atoms and the inclusion of a properly weighted self-descreening term (to take into account the SEV of the atom being descreened) was also tried and judged to be less accurate than eq 33 relative to numerical integration.

2.2.2. Short-Range Hydrogen Bonding Correction Function. In this section we present the analytical model that implements the short-range hydrogen bonding correction function for AGBNP2. The model is based on a geometrical procedure, described below, to measure the degree to which a solute atom can interact with hydration sites on the solute surface. The procedure is as follows. A sphere of radius R_s representing a water molecule is placed in a position that provides near-optimal interaction with a hydrogen bonding donor or acceptor atom of the solute. The position $\mathbf{r}_s$ of this water sphere s is function of the positions of two or more parent atoms that compose the functional group including the acceptor/donor atom:

$$\mathbf{r}_s = \mathbf{r}_s(\{\mathbf{r}_{ps}\}) \tag{34}$$

where $\{\mathbf{r}_{ps}\}$ represents the positions of the set of parent atoms of the water site s. For instance, the water site position in correspondence with a polar hydrogen is

$$\mathbf{r}_s = \mathbf{r}_D + \frac{\mathbf{r}_H - \mathbf{r}_D}{|\mathbf{r}_H - \mathbf{r}_D|} d_{HB}$$

where $\mathbf{r}_D$ is the position of the heavy atom donor, $\mathbf{r}_H$ is the position of the polar hydrogen, and d_{HB} is the distance between the heavy atom donor and the center of the water sphere (see Figure 4). Similar relationships (see the Appendix) are employed to place candidate water spheres in correspondence with hydrogen bonding acceptor atoms of the solute. These relationships are based on the local topology of the hydrogen bonding acceptor group (linear, trigonal, and tetrahedral). This scheme places one or two water spheres in correspondence with each hydrogen bonding acceptor atom (see Table 1).

The magnitude of the hydrogen bonding correction corresponding to each water sphere is a function of the predicted water occupancy of the location corresponding to the water sphere. In this work the water occupancy is measured by the fraction w_s of the volume of the water site sphere that is accessible to water molecules without causing steric clashes with solute atoms (see Figure 4)

$$w_s = \frac{V_s^{free}}{V_s} \tag{35}$$

where $V_s = (4/3)\pi R_s^3$ is the volume of the water sphere and

$$V_s^{free} = V_s - \sum_i V_{si} + \sum_{i<j} V_{sij} - \sum_{i<j<k} V_{sijk} \tag{36}$$

is the "free" volume of water site s, obtained by summing over the two-body, three-body, etc. overlap volumes of the water sphere with the solute atoms. Note that the expression of the free volume is the same as the expression for the self-volume (eq 12) except that it lacks the fractional coefficients 1/2, 1/3, etc. The overlap volumes in eq 36 are computed using radius R_s for the water site sphere (here set to 1.4 Å) and augmented radii R_i^c for the solute atoms. Equation 36 is derived similarly to the expression for the self-volumes by removing overlap volumes from the volume of the water sphere rather than evenly distributing them across the atoms participating in the overlap.

Given the water occupancy w_s of each water sphere, the expression for the hydrogen bonding correction for the solute is

$$\Delta G_{hb} = \sum_s h_s S(w_s; w_a, w_b) \tag{37}$$

where h_s is the maximum correction energy which depends on the type of solute−water hydrogen bond (see Table 1), and $S(w;w_a,w_b)$ is a polynomial switching function which is 0 for $w < w_a$, 1 for $w > w_b$, and smoothly (with continuous first derivatives) interpolates from 0 to 1 between w_a and w_b (see Figure 5). The expression of $S(w;w_a,w_b)$ is

$$S(w;w_a,w_b) = \begin{cases} 0 & w \le w_a \\ f_w\left(\dfrac{w - w_a}{w_b - w_a}\right) & w_a < w < w_b \\ 1 & w \ge w_b \end{cases} \tag{38}$$

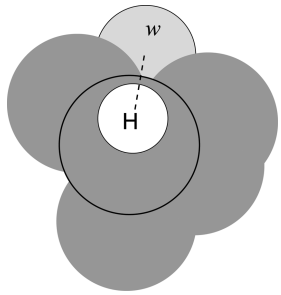


Figure 4. Schematic diagram for the placement of the water sphere (*w*, light gray) corresponding to the hydrogen bonding position relative to the a polar hydrogen (white sphere) of the solute (dark gray). The dashed line traces the direction of the hydrogen−parent heavy atom (circled) bond along which the water sphere is placed. The magnitude of hydrogen bonding correction grows as a function of the volume (light gray) of the water site sphere not occupied by solute atoms.

Table 1. Optimized Surface Tension Parameters and Hydrogen Bonding Correction Parameters for the Atom Types Present in Protein Molecules[a]

atom type	γ (cal/mol/Å^2)	geometry	N_w	h (kcal/mol)
C (aliphatic)	129			
C (aromatic)	120			
H on N		linear	1	−0.25
H on N (Arg)		linear	1	−2.50
H on O		linear	1	−0.40
H on S		linear	1	−0.50
O (alcohol)	117	tetrahedral	2	−0.40
O (carbonyl)	117	trigonal	2	−1.25
O (carboxylate)	40	trigonal	2	−1.80
N (amine)	117	tetrahedral	1	−2.00
N (aromatic)	117	trigonal	1	−2.00
S	117	tetrahedral	2	−0.50

[a] γ is the surface tension parameter, N_w is the number of water spheres, and h is the maximum correction corresponding to each atom type (eq 37). Atom types not listed do not have hydrogen bonding corrections and are assigned $\gamma = 117$ cal/mol/Å^2.

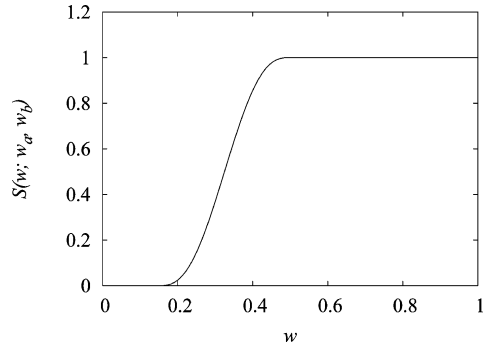


Figure 5. Switching function $S(w;w_a,w_b)$ from eqs 38 and 9 with $w_a = 0.15$ and $w_b = 0.5$.

where $f_w(x)$ is a switching function given by eq 9. In this work we set $w_a = 0.15$ and $w_b = 0.5$. This scheme establishes (see Figure 5) that no correction is applied if more than 85% of the water sphere volume is not water accessible, whereas maximum correction is applied if 50% or more of the water sphere volume is accessible.

2.3. Molecular Dynamics of Miniproteins. We conducted molecular dynamics simulations of what we will refer to as miniproteins (Figure 6), that is, peptides that have been shown to form stable secondary structures in solution: the 23-residue trp-cage peptide of sequence ALQELLGQWLKDGGPSSGRPPPS [Protein Data Bank (PDB) ID 1RIJ],[101] the 28-residue cdp-1 peptide of sequence KPYTARIKGRTFSNEKELRDFLETFTGR (PDB ID 1PSV),[102] and the 28-residue fsd-1 peptide of sequence QQYTAKIKGRTFRNEKELRDFIEKFKGR (PDB ID 1FSD).[103] The structure of trp-cage (see Figure 6) is characterized by a tryptophan side chain enclosed in a cage formed by an α-helix on one side and a proline-rich loop on the other. The cdp-1 and fsd-1 miniproteins (Figure 6) adopt a mixed $\alpha\beta$ conformation and are particularly rich in charged residues. The trp-cage miniprotein was chosen because it has been the target of several computational studies.[104−107] The cdp-1 and fsd-1 peptides were of interest because they showed in preliminary tests with AGBNP1 solvation a significant tendency to deviate from the experimental structures.

Molecular dynamics simulations were conducted for 12 ns starting with the first NMR model deposited in the PDB. The temperature was set to 300 K with the Nosé−Hoover thermostat,[108,109] a molecular dynamics (MD) time step of 2 fs was employed, and covalent bond lengths involving hydrogen atoms were fixed at their equilibrium positions. Backbone motion was restricted by imposing a positional harmonic restraint potential with a force constant of 0.3 kcal/mol/Å^2 on the positions of the Cα atoms, which allows for a range of motion of about 5 Å at the simulation temperature. These restraints are sufficiently weak to allow substantial backbone and side chain motion while preserving overall topology.

Molecular dynamics simulations were conducted with the OPLS-AA potential[97,110] with explicit solvation (SPC water model with 2450, 3110, and 3250 water molecules for trp-cage, cdp-1, and fsd-1, respectively) and with both AGBNP1 and AGBNP2 implicit solvation. The DESMOND program[111] was used for the explicit solvent simulations, and the IMPACT program[15] was used for those with implicit solvation. Identical force field settings were employed in these two programs. The explicit solvent simulations were conducted in the NPT ensemble using the Martyna−Tobias−Klein barostat[112] at 1 atm pressure and employed the smooth Particle Mesh Ewald (PME) method[113] for the treatment of long-range electrostatic interactions with a real-space cutoff of 9 Å. Equilibrium averages and energy distributions were obtained by analysis of the latter 10 ns of saved trajectories. Convergence was tested by comparing averages obtained using the first and second halves of simulation data. Hydrogen bonds were detected using a minimum hydrogen-acceptor distance of 2.5 Å and a minimum donor angle of 120°.

3. Results

3.1. Accuracy of Born Radii and Surface Areas. The quality of any implicit solvent model depends primarily on the reliability of the physical model on which it is based. The accuracy of the implementation, however, is also a critical aspect for the success of the model in practice. This is true in particular for models, such as AGBNP, based on

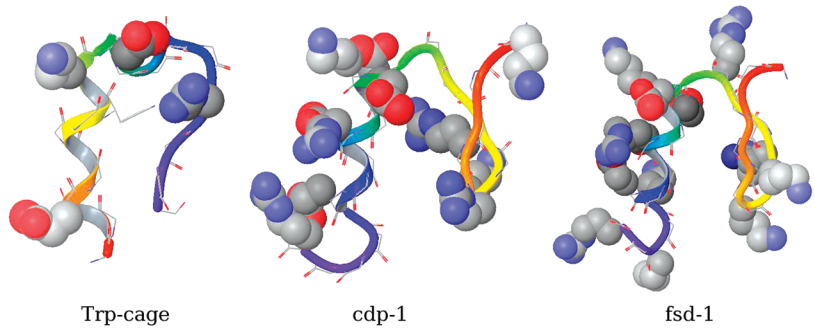


Figure 6. Graphical representations of the NMR structures of the three miniproteins investigated in this work: trp-cage (PDB ID 1RIJ), cdp-1 (PDB ID 1PSV), and fsd-1 (PDB ID 1FSD). In each case the first deposited NMR model is shown. Backbone ribbon is colored from the N-terminal (red) to the C-terminal (blue). Charged side chains are shown in space-filling representation.

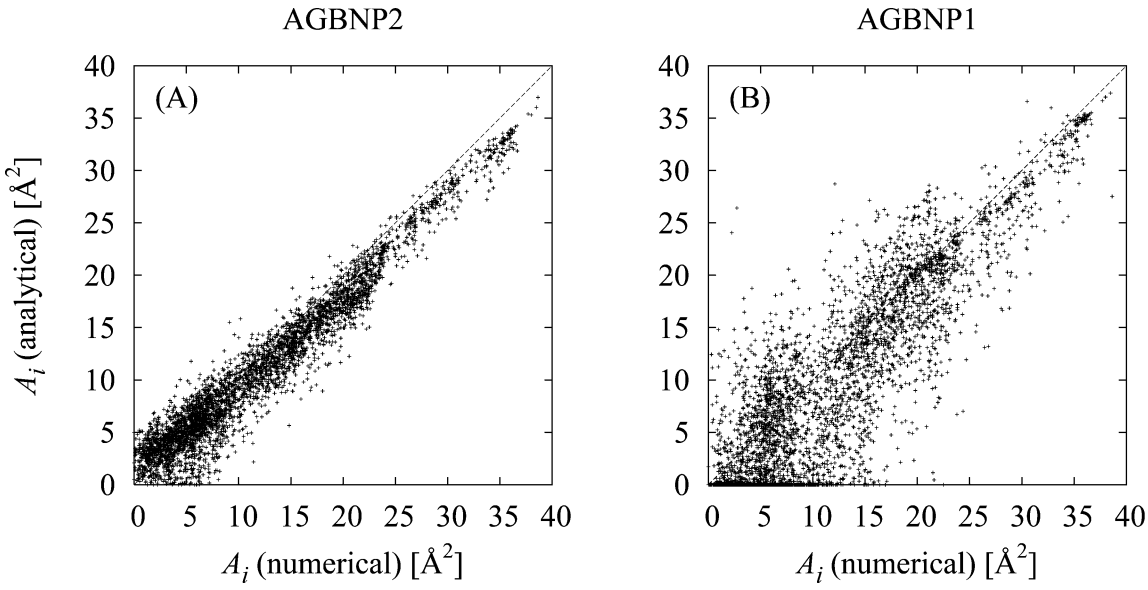


Figure 7. Comparisons between numerical and analytical molecular surface areas of the heavy atoms of the crystal structures (1ctf and 1lz1, respectively) of the C-terminal domain of the ribosomal protein L7/L12 (74 aa) and human lysozyme (130 aa), and of four conformations each of the trp-cage, cdp-1, and fsd-1 miniproteins extracted from the corresponding explicit solvent MD trajectories. (A) Analytical molecular surface areas computed using the present model and (B), for comparison, analytical surface areas computed using the original model from ref 42.

the generalized Born formula. It has been pointed out, for instance, that approximations in the integration procedure to obtain the Born radii may actually be of more significance than the physical approximations on which the GB model is based.[114] It is therefore important to test that the conformational-dependent quantities employed by AGBNP2 are a good approximation to the geometrical parameters that they are supposed to represent. The present AGBNP2 formulation relies mainly on three types of conformational-dependent quantities: Born radii (eq 18), solvent accessible surface areas (eq 10), and solvent accessibilities of hydration sites (eq 38). In this section we analyze the validity of the AGBNP2 analytical estimates for the Born radii and surface areas against accurate numerical results for the same quantities.

We employ the GEPOL program[90] to compute numerically atomic solvent accessible surface areas with a solvent probe diameter of 1 Å, the same probe diameter that defines the solute−solvent boundary in the AGBNP model. Figure 7A shows the comparison between the surface area estimates given by the present formulation of AGBNP and the numerical surface areas produced by GEPOL for a series of native and modeled protein conformations. In Figure 7B we show the same comparison for the surface areas of the original AGBNP1 model. These representative results show that the present analytical surface area implementation, which as described above employs a weaker switching function for the overlap volumes, produces significantly more accurate atomic surface areas than the original model. These improvements in the computation of the surface areas, introduced mainly to obtain more accurate Born radii through eq 31, are also expected to yield more reliable cavity hydration free energy differences.

Figure 8 illustrates on the same set of protein conformations the accuracy of the inverse Born radii, B_i^{-1}, obtained using the AGBNP2 pairwise descreening method using the SEV model for the solute volume described above (eq 18), by comparing them to accurate estimates obtained by evaluating the integral in eq 17 numerically on a grid. The comparison is performed for the inverse Born radii because these, being proportional to GB self-energies, are more reliable accuracy indicators than the Born radii themselves. The grid for the numerical integration was prepared as

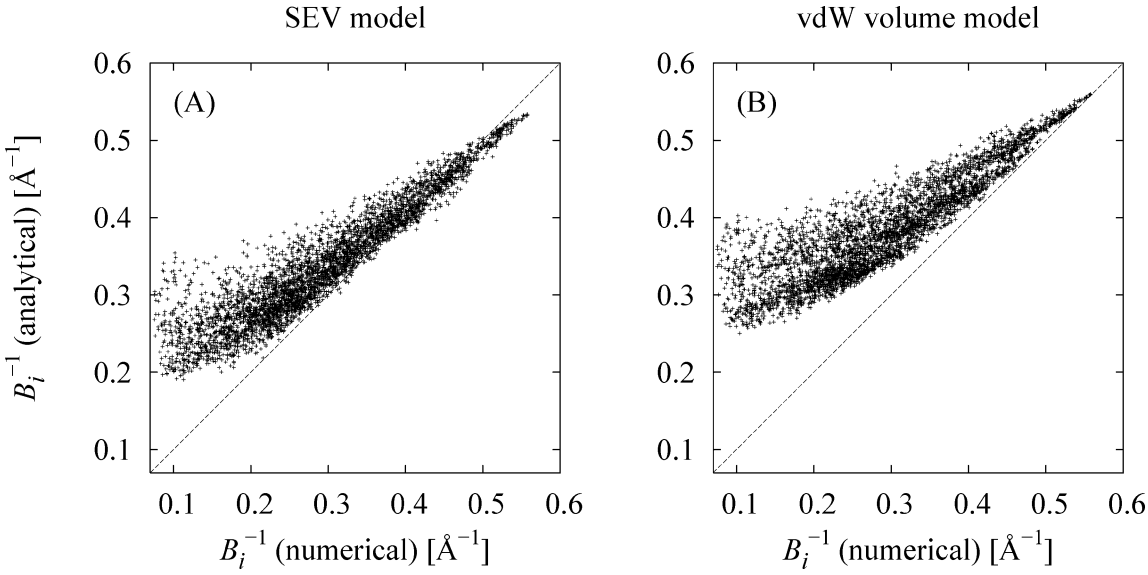


Figure 8. Comparisons between numerical and analytical inverse Born radii for the heavy atoms of the same protein conformations as in Figure 7. (A) Analytical Born radii computed using the present SEV model. (B) Analytical Born radii computed using the van der Waals volume model (ref 42).

previously reported,[42] except that the solvent excluded volume (SEV) of the solute was employed here rather than the van der Waals volume. The integration grid over the SEV was obtained by taking advantage of the particular way that the GEPOL algorithm describes the SEV of the solute; GEPOL iteratively places auxiliary spheres of various dimensions in the interstitial spaces between solute atoms in such a way that the van der Waals volume of the solute plus the auxiliary spheres accurately reproduces the SEV of the solute. Therefore in the present application a grid point was considered part of the SEV of the solute if it was contained within any solute atom or any one of the auxiliary spheres placed by GEPOL. The default 1.4 Å solvent probe radius was chosen for the numerical computation of the SEV with the GEPOL program to assess the accuracy of the model with respect to a full representation of the solute solvent excluded volume as in the GBMV series of models.[93,94] The results of this validation (Figure 8) show that the analytical SEV pairwise descreening model described above is able to yield Born radii which are not as affected by the spurious high dielectric interstitial spaces present in the van der Waals volume description of the solute. With the van der Waals volume model (Figure 8B) the Born radii of the majority of solute atoms start to significantly deviate from the reference values for Born radii larger than about 2.5 Å ($B^{-1} = 0.4$ Å^{-1}). Born radii computed with the analytical SEV model instead (Figure 8A) track the reference values reasonably well further up to about 4 Å ($B^{-1} = 0.25$ Å^{-1}). Despite this significant improvement most Born radii are still underestimated by the improved model (and, consequently, the inverse Born radii are overestimated—see Figure 8), particularly those of nonpolar atoms near the hydrophobic core of the larger proteins. These regions tend to be loosely packed and tend to contain interstitial spaces too large to be correctly handled by the present model. Because it mainly involves groups of low polarity, this limitation has a small effect on the GB solvation energies. It has however a more significant effect on the van der Waals solute−solvent interaction energy estimates through eq 23, which systematically overestimate the magnitude of the interaction for atoms buried in hydrophobic protein core. While the present model in general ameliorates in all respects the original AGBNP model, we are currently exploring ways to address this residual source of inaccuracy.

3.2. Small Molecule Hydration Free Energies. The validation and parametrization of the hydrogen bonding and cavity correction parameters have been performed based on the agreement between experimental and predicted AGBNP2 hydration free energies of a selected set of small molecules, listed in Table 2, containing the main functional groups present in proteins. This set of molecules includes only small and relatively rigid molecules whose hydration free energies can be reliably estimated using a single low energy representative conformation[115] as was done here. Table 2 lists for each molecule the experimental hydration free energy, the AGBNP2 hydration free energy computed without hydrogen bonding (HB) corrections and the default $\gamma = 117$ cal/mol/Å^2 surface tension parameter, denoted by AGBNP2/SEV, as well as the hydration free energy from the AGBNP2 model including the HB correction term and the parameters listed in Table 1. For comparison, the corresponding predictions with the original AGBNP1[42] model are reported in the Supporting Information.

Going down the results in Table 2, we notice a number of issues addressed by the new implementation. With the new surface area implementation and without corrections (third column in Table 2), the hydration free energies of the normal alkanes are too small compared to experiments; furthermore, in contrast with the experimental behavior, the predicted hydration free energies incorrectly become more favorable with increasing chain length. A similar behavior is observed for the aromatic hydrocarbons. Clearly this is due to the rate of increase of the positive cavity term with increasing alkane size which is insufficient to offset the solute−solvent van der Waals interaction energy term, which becomes more negative with increasing solute size. We have chosen to

Table 2. Experimental and Predicted Hydration Free Energies of a Set of Small Molecules

molecule	exptl[a,b]	AGBNP2/SEV[a,c]	AGBNP2[a,d]
n-ethane	1.83	0.98	1.80
n-propane	1.96	0.92	1.97
n-butane	2.08	0.88	2.14
n-pentane	2.33	0.78	2.26
n-hexane	2.50	0.70	2.40
cyclopentane	1.20	0.34	1.63
cyclohexane	1.23	0.05	1.50
benzene	−0.87	−1.50	−1.14
toluene	−0.89	−1.66	−0.94
acetone	−3.85	−1.09	−3.83
acetophenone	−4.58	−2.74	−5.07
ethanol	−5.01	−4.77	−5.30
phenol	−6.62	−4.51	−5.38
ethanediol	−9.60	−7.99	−9.87
acetic acid	−6.70	−2.73	−7.05
propionic acid	−6.48	−2.58	−6.38
methyl acetate	−3.32	−0.10	−3.92
ethyl acetate	−3.10	−0.02	−3.60
methyl amine	−4.56	−2.39	−4.37
ethyl amine	−4.50	−2.24	−3.95
dimethyl amine	−4.29	−1.95	−3.21
trimethyl amine	−3.24	−1.78	−2.39
acetamide	−9.71	−6.81	−10.45
N-methylacetamide	−10.08	−4.75	−7.51
pyridine	−4.70	−3.62	−5.30
2-methylpyridine	−4.63	−2.94	−4.22
3-methylpyridine	−4.77	−2.82	−4.13
methanethiol	−1.24	−0.61	−1.46
ethanethiol	−1.30	−0.57	−1.22
neutral AUE[a,e]		1.90	0.45
acetate ion	−79.90	−77.32	−87.70
propionate ion	−79.10	−76.29	−86.29
methylammonium ion	−71.30	−73.21	−73.54
ethylammonium ion	−68.40	−70.63	−70.75
methyl guanidinium	−62.02[f]	−57.30	−69.81
ions AUE[a,g]		2.85	5.47

[a] In kcal/mol. [b] Experimental hydration free energy from ref 116 except where indicated. [c] AGBNP predicted hydration free energies with the default γ parameter for all atoms types ($\gamma = 117$ cal/mol/Å^2) and without HB corrections. [d] AGBNP predicted hydration free energies with optimized parameters listed in Table 1. [e] Average unsigned error of the AGBNP predictions for the neutral compounds relative to the experiments. [f] From ref 117. [g] Average unsigned error of the AGBNP predictions for the ionic compounds relative to the experiments.

address this shortcoming by increasing by 10.2% and 2.5% respectively the effective surface tensions for aliphatic and aromatic carbon atoms rather than decreasing the corresponding α parameters of the van der Waals term since the latter had been previously validated against explicit solvent simulations. We have chosen to limit the increases of the surface tension parameters to aliphatic and aromatic carbon atoms since the results for polar functional groups did not indicate that this change was necessary for the remaining atom types. With this new parametrization we achieve (compare the second and fourth columns in Table 2) excellent agreement between the experimental and predicted hydration free energies of the alkanes and aromatic compounds. Note that the AGBNP2 model, regardless of the parametrization, correctly predicts the more favorable hydration free energies of the cyclic alkanes relative to their linear analogues. AGBNP2, thanks to its unique decomposition of the nonpolar solvation free energy into an unfavorable cavity term and an opposing favorable term, is, to our knowledge, the only analytic implicit solvent implementation capable of describing correctly this feature of the thermodynamics of hydration of hydrophobic solutes.

The AGBNP2 model without corrections markedly underpredicts the magnitudes of the experimental hydration free energies of the compounds containing carbonyl groups (ketones, organic acids, and esters). The hydration free energies of alcohols are also underpredicted but by smaller amounts. Much better agreement with the experimental hydration free energies of these oxygen-containing compounds (see Table 2) is achieved after applying hydrogen bonding corrections with $h = -1.25$ kcal/mol for the carbonyl oxygen and $h = -0.4$ kcal/mol for both the hydrogen and oxygen atoms of the hydroxy group (Table 1). Note that the same parameters employed individually for carbonyl and hydroxy groups in ketones and alcohols are applied to the more complex carboxylic groups of acids and esters as well as amides and carboxylate ions. The thiol groups of organic sulfides required similar corrections as the hydroxy groups (Tables 1 and 2). The AGBNP2 model without corrections also markedly underpredicted the magnitude of the experimental hydration free energies of amines and amides and, to a smaller extent, of compounds with nitrogen-containing heterocyclic aromatic rings. The addition of HB corrections of −0.25 kcal/mol for amine hydrogens and $h = -2.0$ kcal/mol for both amine and aromatic nitrogen atoms yields improved agreement (Table 2), although the effect of N-methylation is still overemphasized.

3.3. Miniprotein Results. As described in section 2.3, we have performed restricted MD simulations of a series of so-called miniproteins (trp-cage, cdp-1, and fsd-1) to study the extent of the agreement between the conformational ensembles generated with the original AGBNP implementation (AGBNP1) and the present implementation (AGBNP2) with respect to explicit solvent generated ensembles. The results of earlier studies[4,54,55] suggest that the AGBNP/OPLS-AA model correctly reproduces for the most part the backbone secondary structure features of protein and peptides. The tests in the present study are therefore focused on side chain conformations. The backbone atoms were harmonically restricted to remain within approximately 3 Å Cα root-mean-square deviation of the corresponding NMR experimental models. We structurally analyzed the ensembles in terms of the extent of occurrence of intramolecular interactions.

As shown in Table 3, we measured a significantly higher average number of intramolecular hydrogen bonds and ion pairing in the AGBNP1 ensembles relative to the explicit solvent ensembles for all miniproteins studied. The largest deviations are observed for cdp-1 and fsd-1, two miniproteins particularly rich in charged side chains, with on average nearly twice as many intramolecular hydrogen bonds compared to explicit solvent. Many of the excess intramolecular hydrogen bonds with AGBNP1 involve interactions between polar groups (polar side chains or the peptide backbone) and the side chains of charged residues. For example, for cdp-1 we observe (see Table 3) approximately eight hydrogen bonds between polar and charged groups on average compared to nearly none with explicit solvation.

Table 3. Average Number of Some Types of Intramolecular Electrostatic Interactions in the Explicit Solvent Conformational Ensembles, and the Ensembles Generated from Simulations Using the AGBNP1 and AGBNP2 Effective Potentials for the trp-cage, cdp-1, and fsd-1 Miniproteins

miniprotein	explicit	AGBNP1	AGBNP2
	Intramolecular Hydrogen Bonds		
trp-cage	13.5	18.3	15.3
cdp-1	12.6	24.5	15.4
fsd-1	14.1	24.6	14.3
all	40.2	67.4	45.0
	Polar−Polar Hydrogen Bonds		
trp-cage	12.9	17.1	13.9
cdp-1	12.5	16.4	14.1
fsd-1	12.0	15.0	12.9
all	37.4	48.5	40.9
	Polar−Charged Hydrogen Bonds		
trp-cage	0.6	1.2	1.4
cdp-1	0.1	8.1	1.3
fsd-1	2.1	9.6	1.4
all	2.8	18.9	4.1
	Ion Pairs		
trp-cage	0.3	1.0	1.0
cdp-1	2.5	2.9	2.7
fsd-1	1.4	4.6	4.0
all	4.2	8.5	7.7

Despite the introduction of empirical surface tension correction to penalize ion pairs,[55] AGBNP1 overpredicts ion pair formation. We found that ion pairing involving arginine was particularly overstabilized by AGBNP1 as we observed stable ion pairing between arginine and either glutamate or aspartate residues during almost the entire duration of the simulation in virtually all cases in which this was topologically feasible given the imposed backbone restrains. In contrast, with explicit solvation some of the same ion pairs were seen to form and break numerous times, indicating a balanced equilibrium between contact and solvent-separated conformations. This balance is not reproduced with implicit solvation, which instead strongly favors ion pairing. In any case, the relative stability of ion pairs appeared to depend in subtle ways on the protein environment as, for example, the two ion pairs between arginine and glutamate of cdp-1 were found to be stable with either explicit solvation or AGBNP1 implicit solvation whereas other Arg−Glu ion pairs in trp-cage and fsd-1 were found to be stable only with implicit solvation.

This analysis generally confirms quantitatively a series of past observations made in our laboratory indicating that the original AGBNP implementation tends to be biased toward conformations with excessive intramolecular electrostatic interactions, at the expense of more hydrated conformations in which polar groups are oriented so as to interact with the water solvent. During the process of development of the modifications to address these problems, we found it useful to rescore with varying AGBNP formulations and parametrizations the miniprotein conformational ensembles obtained with AGBNP1 and explicit solvation, rather than performing simulations with each new parametrizations. An example of this analysis is shown in the first row of plots of Figure 9, which compare the probability distributions of the

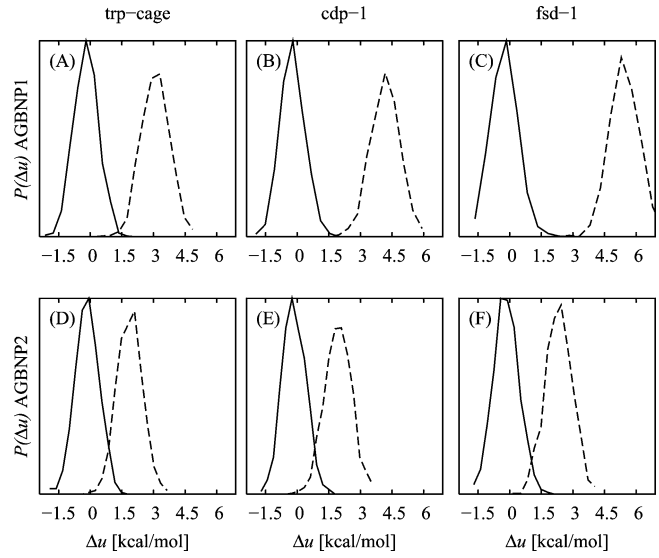


Figure 9. Potential energy distributions of the conformational ensembles for the trp-cage (first column, panels A, D), cdp-1 (second column, panels B, E), and fsd-1 (third column, panels C, F) miniproteins obtained using the AGBNP1/OPLS-AA (first row, panels A−C; full line) and AGBNP2/OPLS-AA (second row, panels D−F; full line) effective potentials and explicit solvation (dashed line). The distributions are shown as a function of the energy gap per residue (Δu) relative to the mean effective potential energy of the implicit solvent ensemble distribution.

AGBNP1 effective potential energies over the conformational ensembles generated with AGBNP1 implicit solvation and with explicit solvation. These results clearly show that the AGBNP1/OPLS-AA effective potential disfavors conformations from the explicit solvent ensemble relative to those generated with implicit solvation. The AGBNP1 energy scores of the explicit solvent ensembles of all miniproteins are shifted toward higher energies than those of the AGBNP1 ensemble, indicating that conformations present in the explicit solvent ensemble would be rarely visited when performing conformational sampling with the AGBNP1/OPLS-AA potential. AGBNP1/OPLS-AA assigns a substantial energetic penalty (see Figure 9A−C) to the explicit solvent ensemble relative to the AGBNP1 ensemble (on average 3.3, 4.4, and 5.7 kcal/mol per residue for, respectively, the trp-cage, cdp-1, and fsd-1 miniproteins). This energetic penalty, being significantly larger than thermal energy, rules out the possibility that conformational entropy effects could offset it to such an extent so as to equalize the relative free energies of the two ensembles. Detailed analysis of the energy scores shows that, as expected, the AGBNP1 implicit solvent ensemble is mainly favored by more favorable electrostatic Coulomb interaction energies due to its greater number of intramolecular electrostatic contacts relative to the explicit solvent ensemble (see above). Conversely, the AGBNP1 solvation model does not assign sufficiently favorable hydration free energy to the more solvent-exposed conformations obtained in explicit solvation so as to make them competitive with the more compact conformations of the AGBNP1 ensemble.

Similar energetic scoring analysis with the AGBNP2 model (see Figure 1 of the Supporting Information) with and

without hydrogen bonding to solvent corrections showed that the introduction of the SEV model for the solute volume significantly reduced the energetic gap between the explicit solvent and AGBNP1 conformational ensembles, and that the introduction of the hydrogen bonding corrections further favors the explicit solvent ensemble. We proceeded to vary the AGBNP2 parameters to achieve the best possible scoring of the explicit solvent ensembles relative to the AGBNP1 ensembles while maintaining an acceptable level of agreement with the small molecule experimental hydration free energies. This procedure eventually yielded the parameters listed in Table 1, which produce small molecule hydration free energies in good agreement with the experiments (Table 2), as well as energy distributions for the three miniproteins that, while still favoring the AGBNP1 ensembles, displayed energy gaps between the explicit and AGBNP1 implicit solvation ensembles comparable to thermal energy and smaller than the spread of the energy distributions.

The energy scoring experiments on the explicit solvent and AGBNP1 ensembles described above were very useful for tuning the formulation of the AGBNP2 model without requiring running lengthy MD simulations. They do not, however, guarantee that the conformational ensembles generated with the AGBNP2 solvation model will more closely match the explicit solvent ensembles than those generated with AGBNP1. This is because the new solvation model could introduce new energy minima not encountered with AGBNP1 or explicit solvation that would be visited only by performing conformational sampling with AGBNP2. To validate the model in this respect, we obtained MD trajectories with the AGBNP2 implicit solvent model and compared the corresponding probability distributions of the effective energy with those of the explicit solvent ensembles similarly as above. The results for the three miniproteins, shown in Figure 9D−F, indicate that the AGBNP2-generated ensembles display significantly smaller bias (mean energy gaps per residue of 2.0, 2.1, and 2.5 kcal/mol for, respectively, the trp-cage, cdp-1, and fsd-1 miniproteins) than AGBNP1 (Figure 9A−C), which yielded energy gaps of 3.3, 4.4, and 5.7 kcal/mol per residue, respectively. This observation shows that AGBNP2 produces conformational ensembles with energy distributions that more closely match on average that of the explicit solvent ensemble without producing unphysical minima that deviate significantly from it.

We have analyzed structural features of the conformational ensembles obtained with the AGBNP1 and AGBNP2 models to establish the degree of improvement achieved with the new model with respect to intramolecular interactions. The salient results of this analysis are shown in Table 3. This table reports for each miniprotein the average number of intramolecular hydrogen bonds and ion pairs. The number of hydrogen bonds is further decomposed into those involving only polar groups (including the backbone) and those involving a polar group and the side chain of a charged residue (arginine, lysine, aspartate, and glutamate). As noted above, it is apparent from these data that the AGBNP1 model produces conformations with too many hydrogen bonds and ion pairs. The majority of the excess hydrogen bonds with AGBNP1 involve residue side chains. Similarly, too many ion pairs are observed in the AGBNP1 ensemble particularly for the fsd-1 miniprotein (4.6 ion pairs on average with AGBNP1 compared to only 1.4 in explicit solvent). The AGBNP2 ensembles, in comparison, yield considerably fewer intramolecular hydrogen bonds. For instance, the average number of hydrogen bonds for cdp-1 is reduced from 24.5 with AGBNP1 to 15.4 with AGBNP2, which is to be compared with 12.6 in explicit solvent. With AGBNP2 the number of polar−polar hydrogen bonds is generally in good agreement with explicit solvation. However, the greatest improvement is observed with polar−charged interactions. For example, the number of polar−charged hydrogen bonds of fsd-1 is reduced by almost 10-fold in going from AGBNP1 to AGBNP2 to reach good agreement with explicit solvation. Importantly, a significant fraction of the excess polar−charged interactions observed with AGBNP1 corrected by AGBNP2 are interactions between the peptide backbone and charged side chains that would otherwise interfere with the formation of secondary structures.

With AGBNP2 we observe small but promising improvements in terms of ion pair formation. The average number of ion pairs of cdp-1 consistently agrees between all three solvation models, and the only possible ion pair in trp-cage is more stable in both implicit solvent formulations than in explicit solvent (it occurs in virtually all implicit solvent conformations compared to only 30% of the conformations in explicit solvent). However, the average number of ion pairs for fsd-1 is reduced from 4.6 with AGBNP1 to 4.0 with AGBNP2. We observe good agreement between the number of ion pairs involving lysine with either AGBNP1 or AGBNP2 and explicit solvation. However, ion pairs involving arginine are generally more stable with implicit solvation than with explicit solvation. The agreement in the number of ion pairs with cdp-1 is due to the fact that for this miniprotein the two possible ion pairs involving arginine result stable with explicit solvation as well as with implicit solvation. For the other two miniproteins, however, ion pairs involving arginine that are marginally stable with explicit solvation are found to be significantly more stable with implicit solvation, although less so with AGBNP2 solvation.

4. Discussion

Modern implicit solvent models for biomolecular simulations are generally based on the uniform dielectric continuum representation of the solvent which is accurately modeled by the Poisson−Boltzmann (PB) equation.[9] Generalized Born (GB) models,[10] which approximate the PB formalism, are applicable to molecular dynamics thanks to their low computational complexity. GB models have reached a high level of accuracy compared to PB following the introduction of more realistic solute volume descriptions[87,100] and of higher order corrections to the Coulomb field approximation.[118−120] However, at the molecular level water is sometimes poorly described by uniform continuum models. Even the best GB models have been found to deviate considerably from, for example, explicit solvent benchmarks.[121,127] The nonlinear and asymmetric dielectric response of water stems primarily from the finite extent and

internal structure of water molecules.[1] The modeling of effects due to water granularity is important for the proper description of molecular association equilibria. Integral equation methods[122] provide an accurate implicit solvation description from first principles; however, despite recent progress,[123] they are not yet applicable to molecular dynamics of biomolecules. The primary aim of the present study has been to formulate an analytical and computational efficient implicit solvent model incorporating solvation effects beyond those inherent in standard continuum dielectric models and, by so doing, achieving an improved description of solute conformational equilibria.

In this work we have developed the AGBNP2 implicit solvent model which is based on an empirical (but physically motivated) first solvation shell correction function parametrized against experimental hydration free energies of small molecules and the results of explicit solvent molecular dynamics simulations of a series of miniproteins. The correction function favors conformations of the solute in which polar groups are oriented so as to form hydrogen bonds with the surrounding water solvent, thereby achieving a more balanced equilibrium with respect to the competing intramolecular hydrogen bond interactions. A key ingredient of the model is an analytical prescription to identify and measure the volume of hydration sites on the solute surface. Hydration sites that are deemed too small to contain a water molecule do not contribute to the solute hydration free energy. Conversely, hydration sites of sufficient size form favorable interactions with nearby polar groups. This model thus incorporates the effects of both water granularity and nonlinear first shell hydration effects.

The GB and nonpolar models in the AGBNP2 implicit solvent model provide the linear continuum dielectric model basis of the model as well as a description of nonelectrostatic hydration effects.[42] In this work the GB and solute−solvent dispersion interaction energy models are further enhanced by replacing the original van der Waals solute volume model with a more realistic solvent excluded volume (SEV) model. The new volume description improves the quality of the Born radii of buried atoms and atoms participating in intramolecular interactions which would otherwise be underestimated due to high dielectric interstitial spaces present with the van der Waals volume description.[88] GB models with these characteristics have been previously proposed. The GBMV series of models[87,93,94] represent the SEV on a grid which, although accurate, is computationally costly and lacks frame of reference invariance. The pairwise descreening based GB^{OBC} model[120] introduced an empirical rectifying function to increase the Born radii of buried atoms in an averaged, geometry-independent manner. The GBn model[100] introduced the neck region between pairs of atoms as additional source of descreening, dampened by empirical parameters to account in an average way for overlaps between neck regions and between solute atoms and neck regions. The approach proposed here to represent the SEV consists of computing the atomic self-volumes, used in the pairwise descreening computation, using enlarged atomic radii so as to cover the interatomic interstitial spaces. The self-volume of each atom is then reduced proportionally to its solvent accessible surface area (see eq 31) to subtract the volume in van der Waals contact with the solvent. We show (Figure 8) that this model reproduces well Born radii computed from an accurate numerical representation of the SEV, noting that improvements for the Born radii of atoms in a loosely packed hydrophobic interior, while significant, are still not optimal. Although approximate, this representation of the SEV maintains the simplicity and computational efficiency of pairwise descreening schemes, while accounting for atomic overlaps in a consistent and parameter-free manner.

The new AGBNP2 model has been formulated to be employed in molecular dynamics conformational sampling applications, which require potential models of low computational complexity and favorable scaling characteristics, and with analytical gradients. These requirements have posed stringent constraints on the design of the model and the choice of the implementation algorithms. In the formulation of AGBNP2 we have reused as much as possible well-established and efficient algorithms developed earlier for the AGBNP1 model. For example, the key ingredient of the hydrogen bonding correction function is the free volume of a hydration site, which is computed using a methodology developed for AGBNP1 to compute atomic self-volumes. Similarly, the SEV-based pairwise descreening procedure employs atomic surface areas (see eq 31), computed as previously described.[42] AGBNP2 suffers additional computational cost associated with the SEV-based pairwise descreening procedure and the hydrogen bonding correction function. This handicap, however, is offset by having only one solute volume model in AGBNP2 rather than two in AGBNP1. AGBNP1 requires two separate invocations of the volume overlaps machinery (eq 2) for each of the two volume models it employs, corresponding to the van der Waals atomic radii for the pairwise descreening calculation and enlarged radii for the surface area calculation.[42] AGBNP2 instead employs a single volume model for both the pairwise descreening and surface area calculations. A direct CPU timing comparison between the two models cannot be reported at this time because the preliminary implementation of the AGBNP2 computer code used in this work lacks key data caching optimizations similar to those already employed in AGBNP1. Given the computational advantages of the new model discussed above, we expect to eventually obtain similar or better performance than with AGBNP1.

The AGBNP2 model has been parametrized against experimental hydration free energies of a series of small molecules and with respect to the ability of reproducing energetic and structural signatures of the conformational ensemble of three miniproteins generated with explicit solvation. These data sources are chosen so as to ensure that the resulting model would be applicable to both hydration free energy estimation and conformation equilibria, which are fundamental characteristics for models aimed at protein−ligand binding affinity estimation. On the other hand, experimental hydration free energies and explicit solvent conformational ensembles are to some extent incompatible with one another given the limitations of even the best fixed-charge force fields and explicit solvation models to reproduce experimental hydration free energies of small molecules with

high accuracy.[41,124,125] Mindful of this issue we did not seek a perfect correspondence with the experimental hydration free energy values. We first obtained parameters by fitting against the small molecule experimental hydration free energies and then adjusted the parameters to improve the agreement with the explicit solvent data, making sure that the predicted small molecule hydration free energies remained within a reasonable range relative to the experimental values. In practice this procedure yielded predicted hydration free energies in good agreement with the experimental values with the exception of the carboxylate and guanidinium ions (see Table 2), for which AGBNP2 predicts more favorable hydration free energies than the experiments, a consequence of the large hydrogen bonding corrections necessary to reduce the occurrence of intramolecular electrostatic interactions in the investigated proteins. As discussed further below, limitations in the description of hydration sites adopted for carboxylate and guanidinium ions may be partly the cause of the observed inconsistencies for these functional groups.

The parametrization and quantitative validation of the model, which is the primary focus of this work, has been based on comparing the effective potential energy distributions of implicit solvent conformational ensembles with those of explicit solvent ensembles. We observed that the AGBNP1 solvation model energetically ranked explicit solvent conformations significantly less favorably than implicit solvent conformations. The AGBNP2 model is characterized by smaller energetic bias relative to the explicit solvent ensembles, indicating that conformational sampling with the AGBNP2/OPLS-AA energy function produces conformations that more closely match those obtained with explicit solvation. This result is a direct consequence of employing the more realistic solvent excluded volume description of the solute, which yields larger Born radii for buried groups, as well as the hydrogen bonding to solvent corrections, which favor solvent exposed conformations of polar groups. Furthermore, comparison of the energy distributions of the AGBNP2 and explicit solvent ensembles for the three miniproteins (Figure 9D−F) shows, in contrast to the AGBNP1 results, that the AGBNP2 bias for the two more charge-rich miniproteins (cdp-1 and fsd-1) is similar to that of the least charged one (trp-cage). This suggests that the residual energetic bias of the AGBNP2 model is probably related to the nonpolar model rather than the electrostatic model. Future studies will address this particular aspect of the model.

The energy scoring studies conducted in this work indicate that AGBNP2 is a significant improvement over AGBNP1. They also show, however, that the new model falls short of consistently scoring explicit solvent conformations similarly to implicit solvent conformations. Although an optimal match between energy distributions is a necessary condition for complete agreement between implicit and explicit solvation results, it is unrealistic to expect to reach this ultimate goal at the present level of model simplification. Increasing the magnitude of the hydrogen bonding corrections can improve the agreement between the explicit and implicit solvation energy distributions, albeit at the expense of the quality of the predicted small molecule hydration free energies. It seems likely that the no parametrization of the current model would yield both good relative conformational free energies and hydration free energies. Future work will pursue the modeling of additional physical and geometrical features, such us the use of variable dielectric approaches to model polarization effects,[126] necessary to improve the agreement between implicit and explicit solvation energy distributions. The energy gap between the implicit solvent and explicit solvent energy distributions used here is, we believe, a meaningful measure of model quality and could serve as a useful general validation tool to compare the accuracy of implicit solvent models.

The excessive number of intramolecular electrostatic interactions involving charged groups has been one of the most noticeable shortcomings of GB-based implicit solvent models.[127] To correct this tendency, we have in the past adopted in the AGBNP1 model ad hoc strategies aimed at either destabilizing electrostatic intramolecular interactions[54] or, alternatively, stabilizing the competing solvent-separated conformations.[55] This work follows the latter approach using a more robust and physically motivated framework based on locating and scoring hydration sites on the solute surface as well as adopting a more realistic volume model. Structural characterization of the conformational ensembles has shown that AGBNP2 produces significantly fewer intramolecular interactions than AGBNP1, reaching good agreement with explicit solvent results. The reduction of intramolecular interactions has been the greatest for interactions between polar and charged groups. We believe the excessive tendency toward the formation of intramolecular interactions to be the root cause of the high melting temperatures of structured peptides[64] predicted with AGBNP1. Given the reduction of intramolecular interactions achieved with AGBNP2, we expect the new model to yield more reasonable peptide melting temperatures, a result which we hope to report in future publications.

Less-visible improvements have been obtained for ion pairs involving arginine side chains which remain more stable with implicit solvation than with explicit solvation. However, significantly, with AGBNP2 we observed a more dynamic equilibrium between ion-paired and solvent-separated conformations of arginine side chains that was not observed with AGBNP1. This result is promising because it indicates that the AGBNP2 solvation model, although still favoring ion-paired conformations, produces a more balanced equilibrium, which is instead almost completely shifted toward contact conformations with AGBNP1. Nevertheless it is apparent that the AGBNP treatment of the guanidinium group of arginine is not as good as for other groups. This limitation appears to be shared with other functional groups containing sp^2-hybridized nitrogen atoms as evidenced, for example, by the relatively lower quality of the hydration free energy predictions for amides and nitrogen-containing aromatic compounds (Table 2). Similar implicit solvent overstabilization solvation of arginine-containing ion pairs has been observed by Yu et al.[85] in their comparison of Surface Generalized Born (SGB) and SPC explicit solvation with the OPLS-AA force field. Despite quantitative differences, the explicit solvent studies (with the TIP3P water model) of Masunov and Lazaridis[128] and Hassan,[129] using the CHARMM force field, and that of Mandell at al.,[130] using the OPLS-AA force field, have confirmed that arginine forms the

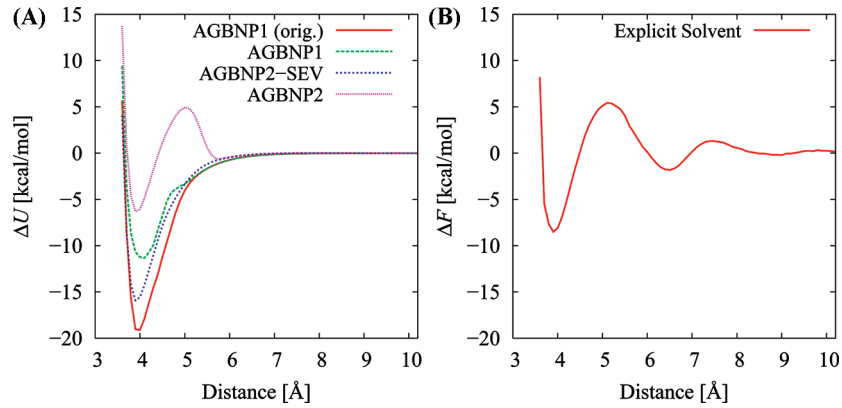


Figure 10. Potential of mean force of ion pair formation between propyl guanidinium and ethyl acetate in the coplanar orientation with AGBNP implicit solvation (A) and explicit solvation (B; ref 130). In (A) "AGBNP1 (orig.)" refers to the original AGBNP1 parametrization,[42] "AGBNP1" refers to the AGBNP1 model used in this work which includes a surface tension parameter correction for the carboxylate group aimed at reducing the occurrence of ion pairs,[55] "AGBNP2" refers to the current model, and "AGBNP2-SEV" refers to the current model without hydrogen bonding and surface tension corrections. The potentials of mean force are plotted with respect to the distance between the atoms of the protein side chain analogues equivalent to the Cζ of arginine and the Cγ of aspartate.

strongest ion pairing interactions, especially in the bidentate coplanar conformation. These observations are consistent with the present explicit solvent results using OPLS-AA and the SPC water model, where we find that most of the ion pairs of the miniproteins were found to involve arginine side chains. In contrast to our present implicit solvent results, however, the work of Masunov and Lazaridis[128] indicated that the GB-based implicit solvent model that they analyzed[14] produced potentials of mean force for arginine-containing ion pairs in general agreement with explicit solvation.

To rationalize the present implicit solvent results concerning ion pair formation, it has been instructive to analyze the potentials of mean force (PMFs) of ion pair association with the AGBNP model. As an example, Figure 10 shows the PMF for the association of propyl guanidinium (arginine side chain analogue) and ethyl acetate (aspartate and glutamate analogue) in a bidentate coplanar conformation (similar to the arrangement used previously)[85,128−130] for various AGBNP implementations. The corresponding explicit solvent PMF obtained by Mandell et al.[130] is also shown in Figure 10 for comparison. The original AGBNP1 parametrization[42] clearly leads to an overly stable salt bridge with the contact conformation scored at approximately −19 kcal/mol relative to the separated conformation, compared with −8.5 kcal/mol with explicit solvation. The AGBNP1 parametrization analyzed here, which includes an empirical surface area correction to reduce the occurrence of ion pairs,[55] yields a contact free energy (−11 kcal/mol) in much better agreement with explicit solvation, although the shape of the PMF is not properly reproduced. The present AGBNP2 model without hydrogen bonding corrections (labeled "AGBNP2-SEV" in Figure 10) yields a PMF intermediate between the original and corrected AGBNP1 parametrizations. The AGBNP2 model with hydrogen bonding corrections yields the PMF with the closest similarity to the one obtained in explicit solvent. Not only the contact free energy (−6.5 kcal/mol) is in good agreement with the explicit solvent result, but, importantly, it also reproduces the solvation barrier of the PMF at 5 Å separation, corresponding to the distance below which there is insufficient space for a water layer between the ions.

It is in this range of distances that the greatest discrepancies are observed between PMFs with explicit solvation and some GB-based implicit solvation models[85,128] that do not model effects due to the finite size of water molecules. Both the hydrogen bonding correction and the SEV volume description employed in AGBNP2, which are designed to take into account the granularity of the water solvent—the hydrogen bonding correction through the minimum free volume of water sites (eq 37) and the SEV model through the water radius offset (eq 28)—make it possible to reproduce the solvation barrier typical of molecular association processes in water.

It is notable in the PMF results shown in Figure 10 the lack of a free energy maximum with the AGBNP2/SEV model (AGBNP2 without HB corrections), which would be expected on the basis of results with the GBMV model, indicating that a SEV treatment of the GB model leads to a higher and much broader PMF maximum relative to explicit solvent.[88] There are two possible factors contributing to this discrepancy. The first is that the OPLS-AA force field used in this work seems to consistently produce stronger ionic interactions than the CHARMM force field (on which the GBMV model is based) as suggested by the relatively small free energies of salt bridge formation obtained with CHARMM-based implicit solvent models[88,128] relative to OPLS-AA-based ones (see for example ref 85 and the present results with AGBNP1). Because the shape of the PMF at intermediate separations is determined by a delicate balance between attractive electrostatic interactions and repulsive desolvation forces, stronger electrostatic interactions with OPLS-AA are potentially responsible in part for a missing or smaller PMF maximum. The lack of the PMF maximum with AGBNP2/SEV is most likely also due to the reduced radius offset used in AGBNP[42] used to construct the SEV. The small probe radius leads to a smaller reduction, compared to a full SEV treatment, of the high dielectric volume surrounding the ionic groups as they approach each other. The consequence is

a smaller rate of increase of the desolvation penalty which in turn leads to a smaller or absent PMF maximum.

Increasing the magnitude of the AGBNP radius offset is not feasible as we observed that the Gaussian overlap approximation for the overlap volumes (eq 10 of ref 42) breaks down for atomic radii much larger than the van der Waals radii. On the other hand, as the results in Figure 10 show, the added desolvation provided by the short-range HB function is able to properly correct this deficiency, yielding a PMF maximum in good correspondence with explicit solvation. This shows that the HB function as parametrized is likely taking into account not only short-range nonlinear hydration effects but also inaccuracies in the GB and nonpolar models, as well as approximations in the implementation such as the small probe radius discussed above.

The good correspondence between the AGBNP2 and explicit solvent PMFs for propyl guanidinium and ethyl acetate (Figure 10) stands in contrast with the residual AGBNP2 overprediction of arginine salt bridges compared to explicit solvation (Table 3). We observed that, in the majority of arginine salt bridges occurring with AGBNP2, the guanidinium and carboxylate groups interact at an angle rather than in the coplanar configuration discussed above. We have confirmed that the PMF of ion pair formation for an angled conformation (not shown) indeed shows a significantly more attractive contact free energy than the coplanar one. This result indicates that the in-plane placement of the hydration sites for the carboxylate groups (see the Appendix) does not sufficiently penalize angled ion pair arrangements. This observation is consistent with the need for introducing an isotropic surface area based hydration correction for carboxylate groups (the reduced γ parameter for the carboxylate oxygen atoms in Table 1), which showed some advantage in terms of reducing the occurrence of salt bridges. Future work will focus on developing a more general hydration shell description for carbonyl groups and related planar polar groups to address this issue.

5. Conclusions

We have presented the AGBNP2 implicit solvent model, an evolution of the AGBNP1 model we have previously reported, with the aim of incorporating hydration effects beyond the continuum dielectric representation. To this end a new hydration free energy component based on a procedure to locate and score hydration sites on the solute surface is used to model first solvation shell effects, such as hydrogen bonding, which are poorly described by continuum dielectric models. This new component is added to the generalized Born and nonpolar AGBNP models which have been improved with respect to the description of the solute volume description. We have introduced an analytical solvent excluded volume (SEV) model which reduces the effect of artifactual high dielectric interstitial spaces present in conventional van der Waals representations of the solute volume. The new model is parametrized and tested with respect to experimental hydration free energies and the results of explicit solvent simulations. The modeling of the granularity of water is one of the main principles employed in the design of the empirical first shell solvation function and the SEV model, by requiring that hydration sites have a minimum available volume based on the size of a water molecule. We show that the new volumetric model produces Born radii and surface areas in good agreement with accurate numerical evaluations. The results of molecular dynamics simulations of a series of miniproteins show that the new model produces conformational ensembles in much better agreement with reference explicit solvent ensembles than the AGBNP1 model with respect to both structural and energetics measures.

Future development work will focus on improving the modeling of some functional groups, particularly ionic groups involving sp^2 nitrogen, which we think are at the basis of the residual excess occurrence of salt bridges, and on the optimization of the AGBNP2 computer code implementation. Future work will also focus on further validation of the model on a wide variety of benchmarks including protein homology modeling and peptide folding.

Acknowledgment. This work was supported in part by National Institute of Health Grant GM30580. The calculations reported in this work have been performed at the BioMaPS High Performance Computing Center at Rutgers University funded in part by NIH shared instrumentation Grant 1 S10 RR022375.

Appendix A: Hydration Site Locations

Figure 11 shows the location of the hydration sites for the functional groups listed in Table 1. Each hydration site is represented by a sphere of 1.4 Å radius. The distance d_{HB} between the donor or acceptor heavy atom and the center of the hydration site sphere is set to 2.5 Å.

There is a single linear geometry for HB donor groups. The corresponding hydration site is placed at a distance d_{HB} from the heavy atom donor along the heavy atom−hydrogen bond.

Acceptor trigonal geometries have one or two hydration sites depending on whether the acceptor atom is bonded to, respectively, two or one other atom. In the former case the water site is placed along the direction given by the sum of the unit vectors corresponding to the sum of the NR_1 and NR_2 bonds (following the atom labels in Figure 11). In the latter case the W_1 site (see Figure 11) is placed in the R_1CO plane forming an angle of 120° with the CO bond. The W_2 site is placed similarly.

Acceptor tetrahedral geometries have one or two hydration sites depending on whether the acceptor atom is bonded, respectively, to three or two other atoms. In the former case the water site is placed along the direction given by the sum of the unit vectors corresponding to the sum of the NR_1, NR_2, and NR_3 bonds. In the latter case the positions of the W_1 and W_2 sites are given by

$$\mathbf{w}_1 = \mathbf{O} + d_{HB}(\cos\theta\mathbf{u}_1 + \sin\theta\mathbf{u}_2)$$

$$\mathbf{w}_2 = \mathbf{O} + d_{HB}(\cos\theta\mathbf{u}_1 - \sin\theta\mathbf{u}_2)$$

where $\mathbf{O}$ is the position of the acceptor atoms, $\theta = 104.4°$, and $\mathbf{u}_1$ and $\mathbf{u}_2$ are, respectively, the unit vectors corresponding to the OR_1 and OR_2 bonds.

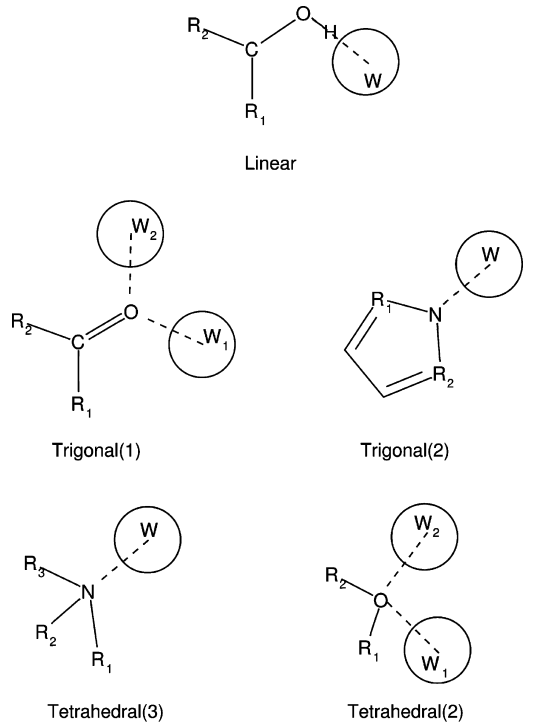


Figure 11. Diagram illustrating the hydration site locations for each of the functional group geometries used in this work. Linear, hydrogen bond donor; trigonal(1) and trigonal(2), trigonal planar geometries with, respectively, one and two covalent bonds on the acceptor atom; tetrahedral(2) and tetrahedral(3), tetrahedral geometries with, respectively, two and three covalent bonds on the acceptor atom. Representative molecular structures are shown for each geometry.

Appendix B: Gradients of GB and van der Waals Energies

The component of the gradient of the AGBNP2 van der Waals energy at constant self-volumes is the same as in the AGBNP1 model (see Appendix C of ref 42). In AGBNP2 the expression for the component of the gradient corresponding to variations in the atomic scaling factors, s_{ij}, includes pair corrections at all overlap levels because of the presence of multibody volumes in V''_{ij}. In addition, a new component corresponding to the change in surface areas appears:

$$\left(\frac{\partial \beta_j}{\partial \mathbf{r}_i}\right)_Q = -\frac{1}{4\pi}\sum_k \frac{\partial s_{kj}}{\partial \mathbf{r}_i} Q_{kj} = -\frac{1}{4\pi}\sum_k \frac{1}{V_k}\frac{\partial V'_k}{\partial \mathbf{r}_i} Q_{kj} \tag{39}$$

$$-\frac{1}{4\pi}\sum_k \frac{1}{V_k}\frac{\partial V'_{kj}}{\partial \mathbf{r}_i} Q_{kj} \tag{40}$$

$$+\frac{1}{4\pi}\sum_k \frac{1}{V_k} p_k \frac{\partial A_k}{\partial \mathbf{r}_i} Q_{kj} \tag{41}$$

Equation 39 leads to the same expression of the derivative component as in the AGBNP1 model (eq 72 in ref 42) (except for the extra elements in the two-body terms due to the inclusion of the $1/2V_{kj}$ correction term). Equation 40 corresponds to the component of the derivative due to variations in V'_{jk}, the volume to be added to the self-volumes of j and k to obtain the s_{jk} and s_{kj} scaling factors. In the AGBNP1 model this component included only two-body overlap volumes; in AGBNP2 this term instead includes all overlap volumes greater than zero. Finally, eq 41, where A_k is the surface area of atom k, leads to the component of the derivatives of the GB and vdW terms due to variations of the exposed surface area. The latter two terms are new for AGBNP2.

B.1. Component of Derivative from eq 40. From eq 63 in ref 42 and eq 40 we have

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \sum_{jk} W_{kj}\frac{\partial V'_{kj}}{\partial \mathbf{r}_i} \tag{42}$$

where W_{kj} has the same expression as in eq 69 in ref 42. In working with eq 42 it is important to note that, whereas V'_{kj} is symmetric with respect to swapping the j and k indices, W_{kj} and W_{jk} are different from each other. Substituting eq 30 into eq 42 and expanding over symmetric terms we obtain

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \frac{1}{2}\sum_{jk} W_{kj}\frac{\partial V_{kj}}{\partial \mathbf{r}_i} - \frac{1}{3}\sum_{jkl} W_{kj}\frac{\partial V_{jkl}}{\partial \mathbf{r}_i} + \frac{1}{2}\frac{1}{4}\sum_{jklp} W_{kj}\frac{\partial V_{jklp}}{\partial \mathbf{r}_i} - \ldots \tag{43}$$

Equation 43 is simplified by noting that

$$\frac{\partial V_{jk\ldots}}{\partial \mathbf{r}_i} = \delta_{ij}\frac{\partial V_{ik\ldots}}{\partial \mathbf{r}_i} + \delta_{ik}\frac{\partial V_{ji\ldots}}{\partial \mathbf{r}_i} + \ldots \tag{44}$$

Equation 44 is inserted in eq 43 and sums are reduced accordingly; then symmetric terms are collected into single sums by reindexing the summations, obtaining

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \frac{1}{2}\sum_j (W_{ij}+W_{ji})\frac{\partial V_{ij}}{\partial \mathbf{r}_i} - \frac{1}{3}\sum_{j<k}[(W_{ij}+W_{ji})+(W_{jk}+W_{kj})+(W_{ik}+W_{ki})]\frac{\partial V_{ijk}}{\partial \mathbf{r}_i} + \frac{1}{4}\sum_{j<k<l}[(W_{ij}+W_{ji})+(W_{ik}+W_{ki})+(W_{il}+W_{li})+(W_{jk}+W_{kj})+(W_{jl}+W_{lj})+(W_{kl}+W_{lk})]\frac{\partial V_{ijkl}}{\partial \mathbf{r}_i} - \ldots \tag{45}$$

The corresponding expression for the gradient of ΔG_{GB} is similar but employs the U_{ij} factors of eq 78 of ref 42 rather than W_{ij}.

B.2. Component of Derivative from eq 41. Inserting eq 41 in eq 63 of ref 42 gives

$$4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q3} = \sum_{jk} W_{kj}p_k\frac{\partial A_k}{\partial \mathbf{r}_i} = \sum_k W_k p_k \frac{\partial A_k}{\partial \mathbf{r}_i}$$

which is the same expression as that for the gradient of ΔG_{cav} (see Appendix A of ref 42) with the replacement

$$\gamma_k \rightarrow \frac{1}{4\pi}W_k p_k$$

The corresponding expression for the gradient of ΔG_{GB} follows from the substitution:

$$\gamma_k \rightarrow \frac{1}{4\pi} U_k p_k$$

B.3. Derivatives of HB Correction Energy. From eq 37 we have

$$\frac{\partial \Delta G_{\mathrm{hb}}}{\partial \mathbf{r}_i} = \sum_s h_s S'(w_s) \frac{\partial w_s}{\partial \mathbf{r}_i} \tag{46}$$

Inserting eqs 35 and 36 in eq 46 gives

$$\frac{\partial \Delta G_{\mathrm{hb}}}{\partial \mathbf{r}_i} = -\sum_{sj} \frac{h_s S'(w_s)}{V_s} \frac{\partial V_{sj}}{\partial \mathbf{r}_i} + \sum_{s,j<k} \frac{h_s S'(w_s)}{V_s} \frac{\partial V_{sjk}}{\partial \mathbf{r}_i} - \ldots \tag{47}$$

where

$$\frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_i} = \left(\frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_i}\right)_{\mathbf{r}_s} + \frac{\partial \mathbf{r}_s}{\partial \mathbf{r}_i} \frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_s} \tag{48}$$

where the first term on the right-hand side represents the derivative of the overlap volume with respect to the position of atom i keeping the position of the water site s fixed, and the second term reflects the change of overlap volume due to a variation of the position of the water site caused by a shift in position of atom i. The latter term is nonzero only if i is one of the parent atoms of the water site.

Supporting Information Available: Figure showing potential energy distributions of the AGBNP1 and explicit solvent conformational ensembles for the the trp-cage, cdp-1, and fsd-1 miniproteins scored with the AGBNP2-SEV/OPLS-AA and AGBNP2/OPLS-AA effective potentials; table listing experimental and AGBNP1 predicted hydration free energies of the set of small molecules in Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531–567.

(2) Feig, M.; Brooks, C. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(3) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(4) Felts, A. K.; Andrec, M.; Gallicchio, E.; Levy, R. Protein Folding and Binding: Effective Potentials, Replica Exchange Simulations, and Network Models. In *Water and Biomolecules—Physical Chemistry of Life Phenomena*; Springer Science: New York, 2008.

(5) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(6) Onufriev, A. *Annu. Rep. Comput. Chem.* **2008**, *4*, 125–137.

(7) Chen, J.; Brooks, C.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(8) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027–2094.

(9) Baker, N. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(10) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrikson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–129.

(11) Zhang, L.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591–607.

(12) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045–1066.

(13) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(14) Dominy, B. N.; Brooks, C. L. I. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(15) Banks, J.; et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(16) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(17) Ben-naim, A. *Hydrophobic Interactions*; Plenum Press: New York, 1980.

(18) Kauzmann, W. *Adv. Protein Chem.* **1959**, *14*, 1–63.

(19) Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

(20) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.

(21) Honig, B.; Yang, A.-S. *Adv. Protein Chem.* **1995**, *46*, 27–58.

(22) Sturtevant, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 2236–2240.

(23) Williams, D. H.; Searle, M. S.; Mackay, J. P.; Gerhard, U.; Maplestone, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1172–1178.

(24) Froloff, N.; Windemuth, A.; Honig, B. *Protein Sci.* **1997**, *6*, 1293–1301.

(25) Siebert, X.; Hummer, G. *Biochemistry* **2002**, *41*, 2965–2961.

(26) Ooi, T.; Oobatake, M.; Nemethy, G.; Sheraga, H. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.

(27) Lee, M. R.; Duan, Y.; Kollman, P. A. *Proteins* **2000**, *39*, 309–316.

(28) Hünenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. *Biochemistry* **1999**, *38*, 2358–2366.

(29) Simonson, T.; Brünger, A. T. *J. Phys. Chem.* **1994**, *98*, 4683–4694.

(30) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(31) Rapp, C. S.; Friesner, R. A. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 173–183.

(32) Fogolari, F.; Esposito, G.; Viglino, P.; Molinari, H. *J. Comput. Chem.* **2001**, *22*, 1830–1842.

(33) Pellegrini, E.; Field, M. J. *J. Phys. Chem. A* **2002**, *106*, 1316–1326.

(34) Curutchet, C.; Cramer, C. J.; Truhlar, D. G.; Ruiz-Lòpez, M. F.; Rinaldi, D.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2003**, *24*, 284–297.

(35) Jorgensen, W.; Ulmschneider, J.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264–16270.

(36) Wallqvist, A.; Covell, D. G. *J. Phys. Chem.* **1995**, *99*, 13118–13125.

(37) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(38) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523–9530.

(39) Wagoner, J.; Baker, N. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(40) Chen, J.; Brooks, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.

(41) Mobley, D.; Bayly, C.; Cooper, M.; Shirts, M.; Dill, K. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(42) Gallicchio, E.; Levy, R. *J. Comput. Chem.* **2004**, *25*, 479–499.

(43) Wallqvist, A.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2001**, *105*, 6745–6753.

(44) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047–2053.

(45) Zhou, R.; Huang, X.; Margulis, C.; Berne, B. *Science* **2004**, *305*, 1605–1609.

(46) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.

(47) Hummer, G.; Garde, S.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. *J. Phys. Chem. B* **1998**, *102*, 10469–10482.

(48) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570–4577.

(49) Pitarch, J.; Moliner, V.; Pascual-Ahuir, J.-L.; Silla, A.; Tuñón, I. *J. Phys. Chem.* **1996**, *100*, 9955–9959.

(50) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243–9244.

(51) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.

(52) Zacharias, M. *J. Phys. Chem. A* **2003**, *107*, 3000–3004.

(53) Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.

(54) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310–321.

(55) Felts, A.; Gallicchio, E.; Chekmarev, D.; Paris, K.; Friesner, R.; Levy, R. *J. Chem. Theory Comput.* **2008**, *4*, 855–858.

(56) Dong, F.; Wagoner, J.; Baker, N. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4889–4902.

(57) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

(58) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, C. W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(59) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(60) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comput. Chem.* **2001**, *22*, 1857–1879.

(61) Chekmarev, D.; Ishida, T.; Levy, R. *J. Phys. Chem. B* **2004**, *108*, 19487–19495.

(62) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801–6806.

(63) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.

(64) Weinstock, D.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R.; Wu, K.; Baum, J. *J. Am. Chem. Soc.* **2007**, *129*, 4858–4859.

(65) Weinstock, D.; Narayanan, C.; Baum, J.; Levy, R. *Protein Sci.* **2008**, *17*, 950–954.

(66) Ravindranathan, K.; Gallicchio, E.; Levy, R. *J. Mol. Biol.* **2005**, *353*, 196–210.

(67) Messina, T.; Talaga, D. *Biophys. J.* **2007**, *93*, 579–585.

(68) Ravindranathan, K.; Gallicchio, E.; Friesner, R. A.; McDermott, A. E.; Levy, R. M. *J. Am. Chem. Soc.* **2006**, *128*, 5786–5791.

(69) Ravindranathan, P.; Gallicchio, E.; McDermott, A.; Levy, R. *J. Am. Chem. Soc.* **2007**, *129*, 474–475.

(70) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. *J. Chem. Theory Comput.* **2007**, *3*, 256–277.

(71) Lapelosa, M.; Gallicchio, E.; Ferstandig Arnold, G.; Arnold, E.; Levy, R. M. *J. Mol. Biol.* **2009**, *385*, 675–691.

(72) Tjong, H.; Zhou, H. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(73) Tjong, H.; Zhou, H. *J. Chem. Phys.* **2007**, *126*, 195102.

(74) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008–3022.

(75) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6760–6764.

(76) Grant, J. A.; Pickup, B.; Sykes, M. J.; Kitchen, C.; Nicholls, A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913–4922.

(77) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693–1698.

(78) Levy, R.; Belhadj, M.; Kitchen, D. *J. Chem. Phys.* **1991**, *95*, 3627–3633.

(79) Alper, H.; Levy, R. M. *J. Phys. Chem.* **1990**, *94*, 8401–8403.

(80) Morozov, A.; Kortemme, T. *Adv. Protein Chem.* **2005**, *72*, 1–38.

(81) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.

(82) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(83) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133–152.

(84) Vitalis, A.; Pappu, R. *J. Comput. Chem.* **2009**, *30*, 673–699.

(85) Yu, Z.; Jacobson, M.; Josovitz, J.; Rapp, C.; Friesner, R. *J. Phys. Chem. B* **2004**, *108*, 6643–6654.

(86) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.

(87) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606.

(88) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769–14772.

(89) Lee, B.; Richards, F. *J. Mol. Biol.* **1971**, *55*, 379–400.

(90) Pascual-Ahuir, J. L.; Silla, E. *J. Comput. Chem.* **1990**, *11*, 1047–1060.

(91) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591–1608.

(92) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.

(93) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(94) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719–729.

(95) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(96) Kratky, K. W. *J. Stat. Phys.* **1981**, *25*, 619–634.

(97) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(98) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381–1392.

(99) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(100) Mongan, J.; Simmerling, C.; McCammon, J.; Case, D.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(101) Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J. *Biochemistry* **2004**, *43*, 7421–7431.

(102) Dahiyat, B.; Sarisky, C.; Mayo, S. *J. Mol. Biol.* **1997**, *273*, 789–796.

(103) Dahiyat, B.; Mayo, S. *Science* **1997**, *278*, 82–87.

(104) Snow, C.; Zagrovic, B.; Pande, V. *J. Am. Chem. Soc.* **2002**, *124*, 14548–14549.

(105) Pitera, J.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–7592.

(106) Zhou, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280–13285.

(107) Paschek, D.; Hempel, S.; García, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17754–17759.

(108) Nosè, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

(109) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695–1697.

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

(111) Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Gregersen, B.; Klepeis, J.; Kolossváry, I.; Moraes, M.; Sacerdoti, F.; Salmon, J.; Shan, Y.; Shaw, D. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*; IEEE: Tampa, FL, 2006.

(112) Martyna, G.; Tobias, D.; Klein, M. *J. Comput. Phys.* **1994**, *101*, 4177–4189.

(113) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(114) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(115) Mobley, D.; Chodera, J.; Dill, K. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(116) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563–595.

(117) Vorobyov, I.; Li, L.; Allen, T. *J. Phys. Chem. B* **2008**, *112*, 9588–9602.

(118) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(119) Im, W.; Lee, M.; Brooks, C. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(120) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383–394.

(121) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(122) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873–886.

(123) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871–882.

(124) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

(125) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.

(126) Zhu, K.; Shirts, M.; Friesner, R. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(127) Zhou, R.; Berne, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.

(128) Masunov, A.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 1722–1730.

(129) Hassan, S. *J. Phys. Chem. B* **2004**, *108*, 19501–19509.

(130) Mandell, D. J.; Chorny, I.; Groban, E.; Wong, S.; Levine, E.; Rapp, C.; Jacobson, M. *J. Am. Chem. Soc.* **2007**, *129*, 820–827.

CT900234U

JCTC Journal of Chemical Theory and Computation

Docking Ligands on Protein Surfaces: The Case Study of Prion Protein‡

Agata Kranjc,[†,#] Salvatore Bongarzone,[†,#] Giulia Rossetti,[†,#] Xevi Biarnés,[†,#,||] Andrea Cavalli,[§,⊥] Maria Laura Bolognesi,[§] Marinella Roberti,[§] Giuseppe Legname,[¶,#] and Paolo Carloni*[,†,#,||]

Statistical and Biological Physics Sector, Neurobiology Sector, International School for Advanced Studies (SISSA), SISSA-Unit, Italian Institute of Technology, 34014 Trieste, Italy, Department of Pharmaceutical Sciences, Alma Mater Studiorum, University of Bologna, 40126 Bologna, Italy, Department of Drug Discovery and Development, Italian Institute of Technology, 16163 Genova, Italy, and CNR-INFM-DEMOCRITOS Modeling Center for Research in Atomistic Simulation, 34014 Trieste, Italy

Received May 20, 2009

Abstract: Molecular docking of ligands targeting proteins undergoing fibrillization in neurodegenerative diseases is difficult because of the lack of deep binding sites. Here we extend standard docking methods with free energy simulations in explicit solvent to address this issue in the context of the prion protein surface. We focus on a specific ligand (2-pyrrolidin-1-yl-*N*-[4-[4-(2-pyrrolidin-1-yl-acetylamino)-benzyl]-phenyl]-acetamide), which binds to the structured part of the protein as shown by NMR (Kuwata, K. et al. *Proc Natl Acad Sci U.S.A.* **2007**, *104*, 11921−11926). The calculated free energy of dissociation (7.8 ± 0.9 kcal/mol) is in good agreement with the value derived by the experimental dissociation constant ($K_d = 3.9\ \mu$M, corresponding to $\Delta G^0 = -7.5$ kcal/mol). Several binding poses are predicted, including the one reported previously. Our prediction is fully consistent with the presence of multiple binding sites, emerging from NMR measurements. Our molecular simulation-based approach emerges, therefore, as a useful tool to predict poses and affinities of ligand binding to protein surfaces.

Introduction

Recent developments in molecular docking protocols (MDPs) allow one to predict accurately ligand poses in their target binding sites.[1] In several cases, the reason for their success lays in the coupling of traditional scoring function-based approaches with molecular simulation approaches[2] (such as soft harmonic modes,[3] molecular dynamics simulations,[4,5] and relaxed complex method[6,7]). The latter introduces conformational flexibility of the target, accounting for the fact that proteins are in constant motion between different conformational states.[8] These may be locally altered when a ligand is bound.[8]

In spite of these successes, there are still many important cases for which MDPs are challenged. These include the prediction of the following poses: transition metal and/or alkylating drugs, ligands causing large structural changes, and ligands *not* binding to specific pockets. The latter is common in proteins undergoing fibrillization in neurodegenerative diseases. Here we propose an enhanced molecular docking protocol (EMD, Scheme 1) that extends MDPs with

* Corresponding author. Telephone: +39 0403787407. Fax: +39 0403787528. E-mail: carloni@sissa.it.

‡ Agata Kranjc, Salvatore Bongarzone, Giulia Rossetti, and Xevi Biarnés contributed equally to this work.

† Statistical and Biological Physics Sector, International School for Advanced Studies (SISSA).

¶ Neurobiology Sector, International School for Advanced Studies (SISSA).

SISSA-unit, Italian Institute of Technology.

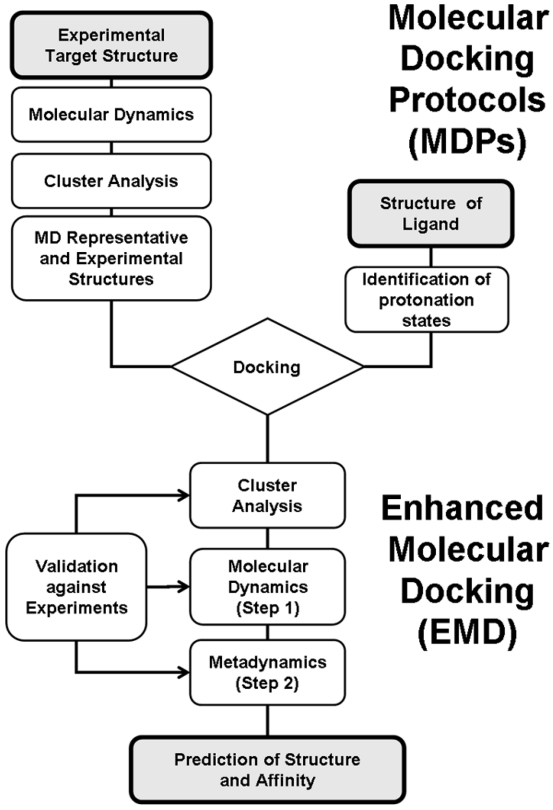
§ Department of Pharmaceutical Sciences, University of Bologna.

⊥ Department of Drug Discovery and Development, Italian Institute of Technology.

|| CNR-INFM-DEMOCRITOS Modeling Center for Research in Atomistic Simulation.

10.1021/ct900257t CCC: $40.75
Published on Web 08/14/2009

Scheme 1. MDPs are used to guess putative ligand binding regions on target surfaces based on structural information of the two separated moieties. Structural information of the target may come from the experiment and, in some cases, also the molecular simulation. Ligands may be docked on the entire structure (like in this work) or a putative binding site. Cluster analysis is used to group molecular dynamics (MD) conformers and/or ligand/target adducts into representative structures. In the EMD approach, MD simulations may be used to relax the structures and investigate the role of hydration. Enhanced sampling simulation techniques in explicit solvent (here metadynamics) allow the exploration of the ligand binding space and the prediction of the free energy of the binding. Comparison against experimental data, in this case structural information inferred by NMR chemical shift perturbations as well as with affinity measurements, allows the protocol to be validated[17]



free energy simulations in explicit solvent to predict the structure and the energetics of ligands binding to protein surfaces.

Protein misfolding, followed by self-association and subsequent deposition, has been observed in the brain tissues of patients affected by different neurodegenerative disorders. Diverse proteins have been shown to follow this process, including amyloid-β (in Alzheimer's disease), α-synuclein (in Parkinson's disease), huntingtin (in Huntington's disease), and prion protein (in prion diseases).[9–11] The protocol we propose here may be exploited in these cases for the design of ligands that, by stacking onto protein surfaces, may disrupt protein−protein interactions and, thus, inhibit protein self-assembly.

In this study, we apply our EMD protocol in the context of the cellular form of human prion protein ($HuPrP^C$).

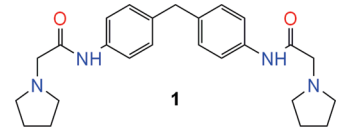


Figure 1. Chemical structure of 2-pyrrolidin-1-yl-*N*-[4-[4-(2-pyrrolidin-1-yl-acetylamino)-benzyl]-phenyl]-acetamide (GN8) considered in this work.

$HuPrP^C$ may convert into a pathogenic form (PrP^{Sc}, scrapie prion protein),[12] which is involved in the epidemics of the bovine spongiform encephalopathy (BSE) and the new variant Creutzfeldt-Jakob disease (nvCJD).[13,14] For these diseases, neither an early diagnosis nor a cure is currently available.[15] Therefore, there is great interest in designing ligands binding to $HuPrP^C$, which may interfere with its conversion and interaction with other self-aggregating proteins, such as amyloid-β.[16]

Recently, the ligand GN8 (2-pyrrolidin-1-yl-*N*-[4-[4-(2-pyrrolidin-1-yl-acetylamino)-benzyl]-phenyl]-acetamide, **1** in Figure 1)[17] has been shown to bind $MoPrP^C$ in the μM range. $MoPrP^C$ is highly similar to the $HuPrP^C$. The sequence similarity is as high as 98%, and the root-mean-square difference (rmsd) of the backbone between the molecular dynamics (MD) structures of $HuPrP^C$ (PDB code: 1HJM)[18] is the same (0.26 ± 0.02 nm) as that between the NMR structures of $HuPrP^C$ and $MoPrP^C$ (PDB code: 1AG2)[19] (0.27 nm). Therefore, significant changes of the structure on passing from the mouse to the human protein are not expected. For the **1**-$MoPrP^C$ adduct, NMR chemical shift perturbations of $MoPrP^C$ on protein residues induced by ligand binding have been reported.[17] These affect most significantly amino acid residues on one side of the protein surface (Arg156, Asn159 @ H1−S2 loop, Lys194 @ H2, Glu196, Thr199 @ H2−H3 loop, and Val210 @ H3). In addition, Val189 and Thr192, located on the other side of the PrP^C surface, are also perturbed (Figure 2A). All these residues are conserved on passing from $MoPrP^C$ to $HuPrP^C$. These perturbations have been ascribed to ligand binding, suggesting that multiple binding sites may be present. An ad hoc model of the **1**-$MoPrP^C$ adduct, constructed by docking and energy minimization, exhibited a single binding mode of GN8 connecting Asn159 and Glu196.[17] Subsequent quantum mechanical studies,[20] based on this model, pointed out that these two residues, along with Gln160 and Lys194, are important for the binding. However, such a single binding mode was not consistent with the presence of the contacts between the ligand and Val189, Thr192, and Thr199. A similar NMR study to map prion protein binding sites has been made only for one other ligand, quinacrine (Supporting Information, Figure S1).[21] The latter, however, has been later suggested not to bind to PrP^C but rather to PrP^{Sc} or other chemical chaperons involved in the prion propagation.[22]

The key ingredient of the protocol proposed in Scheme 1 is given by the type of free energy approach used. Several powerful methods are available for predicting ligand binding free energies by means of molecular simulation.[23,24] However, predictions have been made so far to targets with binding sites well characterized by X-ray crystallography or NMR experiments. Here we use the metadynamics[25] ap-

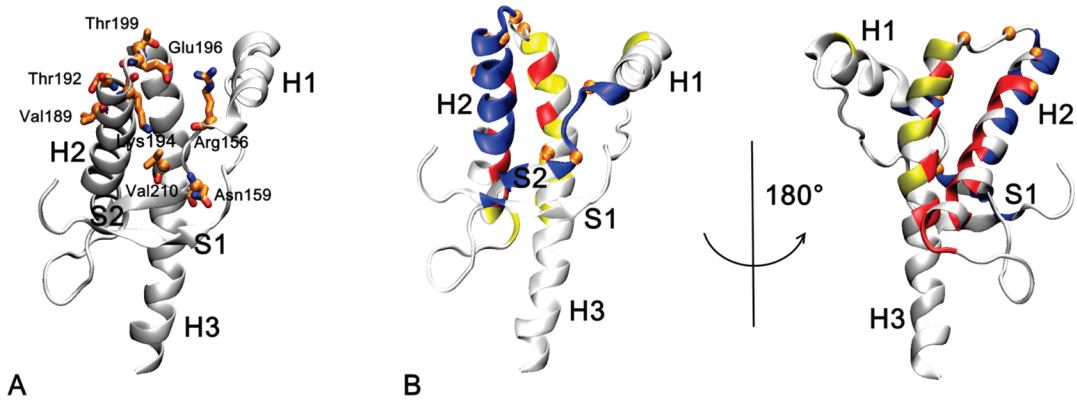


Figure 2. (A) Residues involved in GN8 binding to the prion protein (in licorice), as emerging from chemical shift changes.[17] (B) Three different binding regions (I, II, and III shown in blue, red, and yellow, respectively), as obtained after MDP procedures (see Scheme 1). Orange spheres represent compound **1** binding aminoacids defined by the NMR chemical shift study. The figure shows HuPrP^C. This is very similar to the MoPrP^C (sequence similarity = 98%) for which experiments have been carried out.

proach in its bias exchange variant.[26] This approach provides the free energy as a function of several reaction coordinates (such as geometrical distances, polar contacts, and water-mediated interactions), which characterize the ligand both binding to its target and dissociating from it.[27–30] Although GN8-PrP^C interaction energies have been provided by quantum chemical methods,[20] no calculation of free energy has been so far reported. As with several other techniques,[24] it may allow also simulating the whole molecular recognition process. This in turn may allow characterizing multiple binding sites of the ligand onto the proteins surface, such as those emerging from NMR in the **1**-MoPrP^C complex.

The proposed EMD protocol turns out to provide structural prediction consistently with the NMR data and affinity, which is in agreement with experimental data. The EMD protocol emerges, therefore, as a useful approach to investigate ligands sticking on protein surfaces.

Results and Discussion

In this study, we focus on the binding of the compound **1** to the surface of the HuPrP^C protein. Compound **1** is a symmetric molecule, composed of two pyrrolidine rings connected by acetamides to a diphenylmethane core (Figure 1). Two different conditions are considered: at neutral pH, where experimental affinity has been measured,[17] and at acidic pH, where NMR chemical shift perturbations[17] have been used to provide information on the amino acids involved in the binding.

We use the computational protocol summarized in Scheme 1: (i) Identification of the ligand protonation state at neutral and acidic pH. (ii) Use of MDPs to provide a first guess of the putative binding regions (iii). Use of MD simulations to relax the structure in an aqueous solution (step 1 in EMD). (iv) Use of metadynamics to predict the energetics of the binding of **1** to the protein. (v) Use of metadynamics to predict the binding poses of the compound (step 2 in EMD).

1. Protonation State of Compound 1. This compound can exist in different protonation states in which none, one, or both the tertiary nitrogen atoms of the pyrrolidine rings are protonated (Supporting Information, Figure S2). At pH = 7.4, at which the K_d has been measured, approximate pK_a calculations based on ref 31 (see Methods Section for details) allow us to suggest that, in water, **1** is present not only in the neutral form ($\mathbf{1}^0$) but also in the monoprotonated one ($\mathbf{1}^+$) (Supporting Information, Figure S2). In the latter, one of the two pyrrolidine nitrogen atoms is protonated.

At pH = 4.5, at which the NMR experiments were performed, the same calculations lead us to the conclusion that, in water, the ligand is mainly diprotonated ($\mathbf{1}^{2+}$), with both pyrrolidine nitrogen atoms protonated. Small amounts of $\mathbf{1}^+$ are also present. The calculated concentration of $\mathbf{1}^0$ in bulk water is very low (Supporting Information, Figure S2). However, one should keep in mind that the ligand−protein binding does not occur in pure water and the influence of the electrostatic field of the protein should be accounted for. Indeed, simple electrostatic potential calculations (see Methods Section for further details) show an increase of the positive charge density in the region of the protein defined by the NMR contacts (Supporting Information, Figure S6). This suggests that the protein environment will favor the accumulation of neutral $\mathbf{1}^0$. Therefore, binding poses involving the neutral form should be considered even at an acidic pH. Based on these results, we performed calculations on all of the three protomers.

2. Binding Regions of HuPrP^C Emerging from MDP. The three protomers were docked independently to the HuPrP^C NMR structure and to 20 different conformers obtained from a 20 ns MD simulations of the protein in aqueous solution. The putative binding regions I, II, and III were identified (Figure 2B). I is defined by the H2 helix and the loop connecting β-sheet S2 and helix H1. II consists of the H2−H3 helices. III is defined by the H3 helix, the N-term of H2 helix, and the loop between H1 helix and S1 β-sheet.

Binding region I is the only site which involves residues changing chemical shifts upon binding with compound **1**, and it is closer to all the other residues involved in the binding.[17] It was, therefore, the only one selected for subsequent free energy studies.

The adducts for each of the three protomers docked at the binding region I underwent 10 ns of MD calculations in an

aqueous solution. The ligands maintained completely ($\mathbf{1}^+$ and $\mathbf{1}^{2+}$) or partially ($\mathbf{1}^0$) the pose identified in the docking (see Supporting Information for details). Most importantly, the structural determinants of the three protomers turned out to be consistent with most ligand−protein contacts identified by NMR (Supporting Information, Table S1). However, the ligand−protein contacts with Val189, Thr192, Thr199, and Val210 could not be predicted. The simpler docking approach, combined with the energy minimization of protomer $\mathbf{1}^0$, provided similar results (Figure 3 in Kuwata et al.[17]).

Free energy calculations were used to explore the ligand binding space in an explicit solvent. These simulations identified alternative binding poses for each protomer of the ligand and predicted the dissociation free energy for $\mathbf{1}^0$ and $\mathbf{1}^+$. The free energy simulations were performed as a function of six collective variables that took into account rearrangements of the ligand and the protein, the hydrogen-bond contacts and the water bridges (see Methods Section). These variables have been already used to characterize ligand-target molecular recognition processes using the metadynamics approach.[27–30]

3. The HuPrP^C−$\mathbf{1}^0$ Complex. In the lowest free energy cluster identified by the metadynamics calculations, $\mathbf{1}^0$ is located in the wide cleft formed by helices H1, H2, and H3 (1°.B1 in Figure 3A), similar to the model proposed by Kuwata et al. for MoPrP^C.[17] The contacts $\mathbf{1}^0$ forms with the HuPrP^C are consistent with the reported chemical shift changes on Glu196, upon GN8 binding (Supporting Information, Table S1), as well as with a recent quantum chemical study.[20] They are also consistent with the chemical shift changes on Arg156, Thr199, and Val210 upon GN8 binding (Supporting Information, Table S1). The phenyl groups of $\mathbf{1}^0$ form a π-cation interaction with Arg156, and a water-mediated hydrogen-bond is present between Thr199 and the pyrrolidine nitrogen (N1; Figure 1). The pyrrolidine ring forms hydrophiobic interactions with Val210 (as well as with Pro158 and Thr183). The HuPrP^C−$\mathbf{1}^0$ complex is further stabilized with a direct hydrogen-bond between Thr190 and the carbonyl group of $\mathbf{1}^0$ (O2; Figure 1). The unbound state of HuPrP^C−$\mathbf{1}^0$ system corresponds to a conformation in which the ligand has no contact with the protein. The conformation of Lys194 changes upon ligand dissociation (Supporting Information, Figure S3). This is consistent with the significant chemical shift change reported for this amino acid.[17] This contrasts with a recent quantum chemical investigation, which points to the role of Lys194 for the bridging conformation of the GN8/PrP^C complex.[20] This discrepancy may be due to the fact that here we consider free energies in solution, while ref 20 presents interaction energies *in vacuo*. Smaller conformational changes were also observed for other residues present in the H2−H3 loop (res. 195−199, Supporting Information, Figure S3). These rearrangements were not observed with the MD calculations (see Supporting Information), possibly because they are induced during the ligand binding process simulated here.

The unbound state of HuPrP^C−$\mathbf{1}^0$ is 5.5 kcal/mol ± 0.9 higher in energy with respect to the bound state described previously. The ligand is not completely detached from the protein, although it is already separated by five layers of water molecules between the two moieties. The remaining free energy for the complete dissociation was estimated as the mean electrostatic interaction between the two molecules in implicit solvent (see Methods Section). This turned out to be −0.7 kcal/mol. Thus, we estimated the dissociation free energy to be 4.7 kcal/mol in our simulation conditions. Considering also the concentration of the species in the simulation box (see Methods Section), the standard free energy of dissociation is estimated to be 7.8 kcal/mol. This is in very good agreement with the experimental value of 7.5 kcal/mol (corresponding to $K_d = 3.9$ μM) reported by Kuwata et al.[17]

4. The HuPrP^C−$\mathbf{1}^+$ Complex. Four different stable conformations of $\mathbf{1}^+$ were identified on HuPrP^C surface (Figure 3B). In the lowest free energy state, $\mathbf{1}^+$ lays along the loop connecting helices H2 and H3 (1^+.B1 in Figure 3B). It forms a remarkable hydrophobic interaction with Thr199, consistently with the chemical shift changes reported for this residue.[17] The amidic nitrogen atoms of $\mathbf{1}^+$ (N3 and N4; Figure 1) are hydrogen-bonded to Thr201 and Asn197, respectively. This induces a subtle conformational change of the Glu196 and Asn197 backbone upon ligand binding, which may be the reason for the chemical shift displacement reported experimentally for Glu196 (Supporting Information, Table S1). Additionally, the neutral pyrrolidine ring of $\mathbf{1}^+$ is kept by the hydrophobic cleft formed by Ile184, Thr188, Phe198, Val203, and Met206 further stabilizing the complex. No water-mediated interactions were observed between $\mathbf{1}^+$ and HuPrP^C.

The free energy difference between the bound state of HuPrP^C−$\mathbf{1}^+$ (1^{1+}.B1 in Figure 3B) and the corresponding unbound state, with the corrections described above, turns out to be 8.6 kcal/mol. This is similar to that predicted for $\mathbf{1}^0$ and is in good agreement with the experimental data.

5. The HuPrP^C−$\mathbf{1}^{2+}$ Complex. Five different stable conformations of $\mathbf{1}^{2+}$ were identified on HuPrP^C surface (Figure 3C). In the most stable conformation, $\mathbf{1}^{2+}$ binds yet in another position of HuPrP^C, laying along helix H2 (1^{2+}.B1 in Figure 3C). Half of a part of $\mathbf{1}^{2+}$ is in close contact with the HuPrP^C surface in the cleft formed by Val189, Thr192, and Thr193. Indeed, these positions were reported to interact directly with the ligand according to NMR experiments (Supporting Information, Table S1). Two layers of water molecules are present between the protein surface and the rest of the molecule, presumably due to the presence of Lys185. In the other accessible conformations, $\mathbf{1}^{2+}$ covers different regions of the protein surface (Supporting Information, Figure S4). The interaction with Lys194 is conserved in the majority of them. This result is consistent with the chemical-shift changes of this residue upon ligand binding.[17] The dissociation free energy of $\mathbf{1}^{2+}$ was not calculated as, according to our calculations based on pK_a estimations, this protomer is not present at the conditions in which the K_d was measured.[17]

In summary, the EMD protocol enables to identify binding poses of the **1** protomers to the HuPrP^C surface

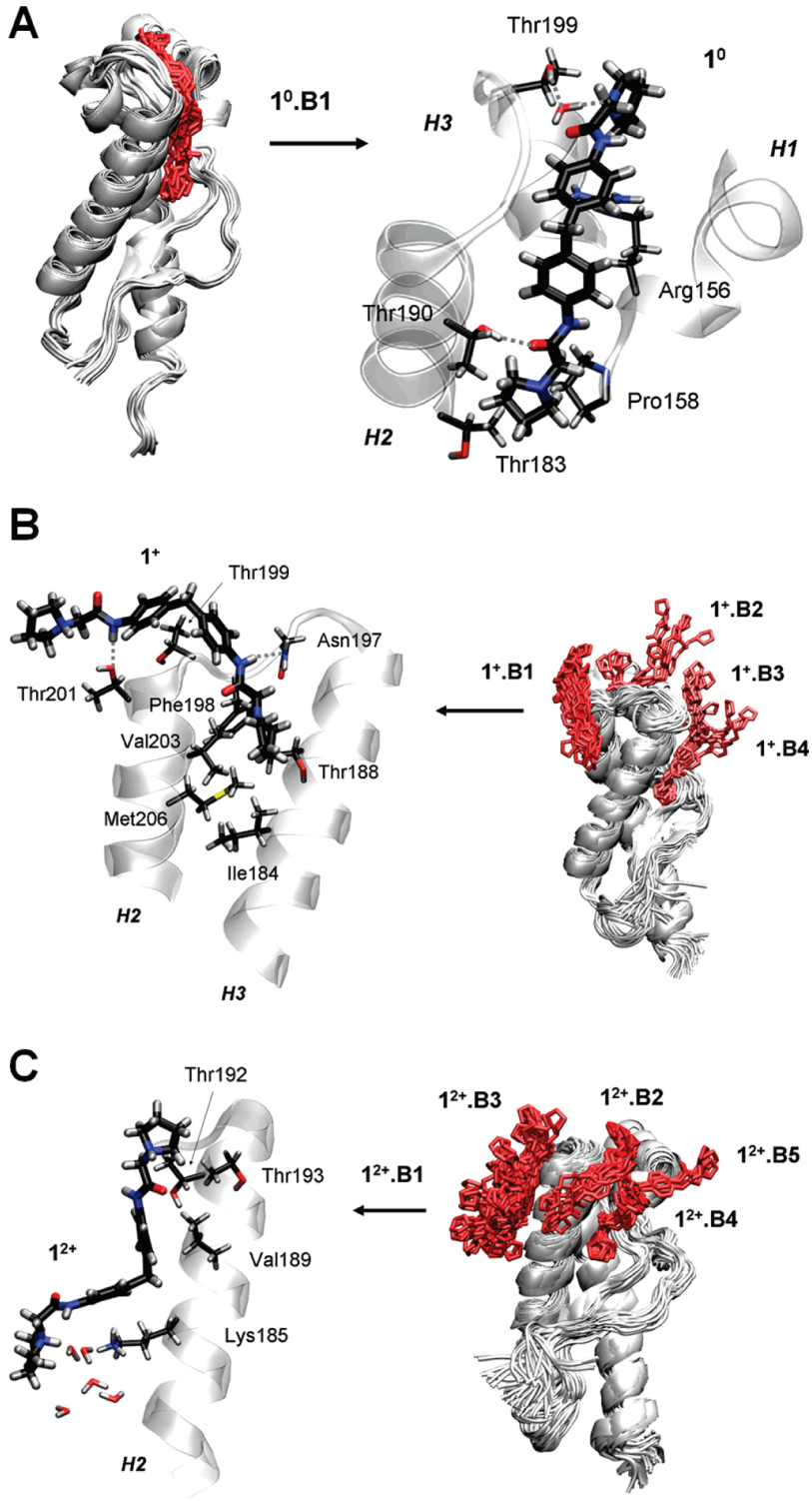


Figure 3. Three dimensional structures of HuPrP^C in complex with **1** protomers: (A) $\mathbf{1}^0$, (B) $\mathbf{1}^+$, and (C) $\mathbf{1}^{2+}$. These structures correspond to the bound-state free energy minima (1°.B1, 1^+.B1−B4, 1^{2+}.B1−B5), as calculated with the metadynamics method (see Methods Section). Close-ups on the ligands and the binding sites are also shown.

other than that proposed by Kuwata et al.[17] The multiple binding sites pattern that we observed from our simulations (Figure 4A) provides a structural basis for the NMR contacts (Figure 4B). The NMR contacts are observed in distant positions of HuPrP^C; in fact, they are located in opposite parts of the protein surface.[17] This result could only be obtained when we extended MDPs with enhanced sampling simulations. The predicted value of the K_d was in good agreement with experiments.[17]

Conclusions

The integration of docking algorithms with MD based simulations has been shown to be convenient in computer-assisted drug design.[2] We suggest here that its extension with

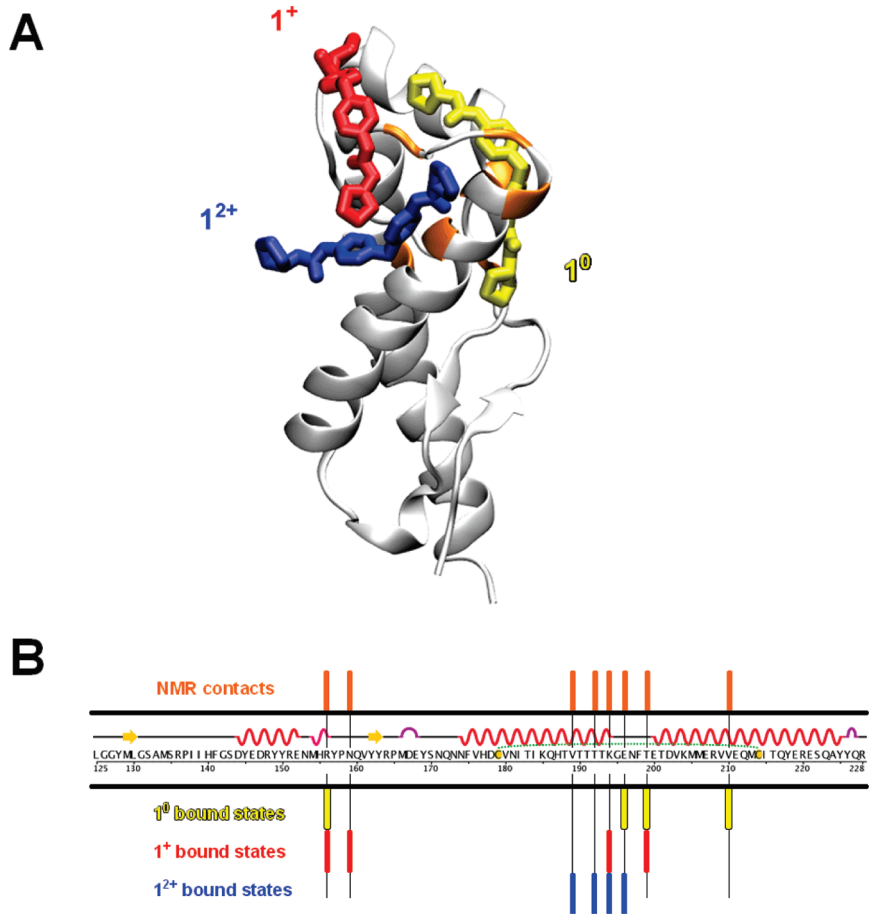


Figure 4. (A) Superimposition of the most populated binding poses of the three GN8 protomers: $\mathbf{1}^0$ (yellow), $\mathbf{1}^+$ (red), and $\mathbf{1}^{2+}$ (blue). (B) HuPrP^C sequence. The residues experimentally found to be involved in binding are highlighted in orange bars.[17] Those emerging from the calculations are shown with the same color code as (A).

free energy simulations, based on metadynamics, is useful to predict the adducts and the energetics of small ligands binding to cavity-less proteins. This approach enables the identification of the preferential ligand binding poses on protein surfaces with their corresponding binding affinity. This information is helpful for the subsequent improvement of the lead compounds in drug design. Docking simulations at protein surface sites are still in their infancy, and therefore, our study could provide a step towards the development of a computational protocol able to identify small organic molecules that interfere with protein−protein interactions occurring in the fibrillization process.

Here, we have focused on the ligand **1** (Figure 1) that targets the cellular form of the prion protein (HuPrP^C), the main agent involved in prion diseases.[17] Given the lack of deep binding pockets along the protein structure, it is reasonable to assume that a small molecule will not bind specifically to a single site. NMR data indicates that this is indeed the case.[17] This data shows that, for compound **1**, there are few hot spots far away from each other on the prion protein surface. This is a clear indication that **1** cannot bind to a unique position. Our study shows that compound **1** can adopt different protonation states at physiological and acid pH, at which experiments have been carried out.

Several poses of the species most present at an acidic pH, the diprotonated state ($\mathbf{1}^{2+}$), are found at the protein surface (Figure 3C). Similar binding sites are observed for $\mathbf{1}^+$ (Figure 3B), which is present in smaller content at an acidic pH. The neutral form ($\mathbf{1}^0$), which also may be present at an acidic pH, binds to a specific region of HuPrP^C surface along the shallow cleft formed by helices H1, H2, and H3 (Figure 3A).

Taken altogether, these multiple binding modes (Figure 4A) are consistent with all of the experimentally predicted contacts between the compound **1** and the different parts of the protein (Figure 4B).[17] The lack of a unique binding site is also coherent with the fact that different PrP^C binders, such as PrP^C antibodies,[32,33] molecular chaperones,[15,34] and DNA aptamers,[35] may interact with different regions of PrP^C surface. The multiple-site binding pattern that arises from our simulations might be an important facet for the anti-fibrillization potency of compound **1**, as it shows different points for disrupting the protein−protein interactions among HuPrP. This result could not be obtained by applying only standard MDP protocols (Scheme 1). In fact, it was necessary to simulate the whole binding process of compound **1** from the solution to the protein surface by means of enhanced sampling MD techniques like metadynamics. Most importantly, the predicted dissociation free energy turned out to be in very good agreement with the experimental data. Based on these encouraging results, the EMD protocol may now be used to predict the potency of ligands interacting with protein surfaces or target proteins without a unique binding

site. These include oligomers,[65] β-amyloid, as well as α-sinucleyn[66] and other unstructured proteins.

Methods

Identification of Binding Sites. The residues interacting with **1** are located in HuPrP^C C-term, for which the NMR structure is available (residues 125−228, PDB ID: 1HJM).[18] Protonation states were assigned by the web server H++[36] assuming pH 7.4. Putative binding sites were identified by (i) molecular simulations (using the GROMACS package[37] and (ii) docking procedure (using the GOLD[38,39] and the Autodock programs[40]).

1. Molecular Simulations. The protein was inserted into a cubic box of water molecules, ensuring that the solvent shell would extend for at least 0.8 nm around the system. Three sodium counterions were added. The AMBER99 force field[41,42] was used for the protein. Sodium ions were modeled with the AMBER-adapted Aqvist potential.[43] The water molecules were described by the TIP3P model.[44] The system was minimized imposing harmonic position restraints of 1000 kJ·mol^{-1}·nm^{-2} on solute atoms, allowing the equilibration of the solvent without distorting the solute structure. After an energy minimization of the solvent and the solute without harmonic restraints, the temperature was gradually increased from 0 to 298 K. This was performed by increasing the temperature from 0 to 298 K in 12 steps in which the temperature was increased by 25 K in 100 ps of MD.

Constant temperature−pressure (T = 298 K, P = 1 bar) 20-ns dynamics was then performed through the Nosé-Hoover[45,46] and Andersen-Parrinello-Rahman[47,48] coupling schemes. Periodic boundary conditions were applied. The final simulation box equilibrated at around 6.69 × 6.69 × 6.69 nm. Long-range electrostatic interactions were treated with the particle mesh Ewald (PME)[49,50] method, using a grid with a spacing of 0.12 nm combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points. The cutoff radius for the Lenard-Jones interactions as well as for the real part of PME calculations was set to 0.9 nm. The pair list was updated every 2 steps, and the LINCS algorithm[51] was used to constrain all bond lengths involving hydrogen atoms allowing us to use a time step of 2 fs.

The MD trajectory of prion protein alone was clustered with the gromos method[52] and as result 20 different conformations were obtained, which were used along with the NMR structure for docking of compound **1**.

2. Docking. Titration curves for compound **1** in bulk solution were calculated by the ChemAxon software[31] showing that this molecule is present in two protonation states at pH = 7.4: neutral (**1**^{0}) and monoprotonated (**1**$^+$); while at pH = 4.5 it exists mostly in the diprotonated form (**1**$^{2+}$) (Figure 1 and S2). This method has been used because it appears to be rather reliable: in a calculation of pK_a of 1000 molecules, less than 0.5% calculations turned out to differ by more than 0.5 pH unit from the experimental value.[31] Since the acidity of compound **1** is expected to change in the proximity of the protein, electrostatic potential calculations in implicit solvent[64] were addressed for the prion protein in two different conditions: pH = 4.5 for NMR measurements conditions and pH = 7 for affinity measurements conditions. The electrostatic potential surfaces were calculated using the APBS package[64] and visualized with VMD[68]. The calculation parameters were 0.3 Å grid spacing, 129^3 meshes, solvent and protein dielectrics of 78.54 and 40, respectively.

All three protomers underwent geometry optimization at the B3LYP/6-31G** level of theory by means of the Gaussian03 software (g03).[53]

The optimized structures, **1**^{0}, **1**$^+$ and **1**$^{2+}$, were docked to the NMR structure of HuPrP^C and to its 20 different conformations, as obtained after the cluster analysis of the MD trajectory.

The GOLD 3.1[38,39] and Autodock 3.0.5[40] programs were used. In GOLD, the docking area was defined as a sphere of 3.5 nm radius around the His187, so that the whole protein was screened. The ChemScore (CS)[54] and GoldScore (GS)[38] scoring functions were used for ranking. For each protomer and scoring function, 100 docking runs were performed.

In Autodock,[40] a Lamarckian genetic search algorithm was used to identify low energy binding sites and orientations of **1** protomers. Binding modes were ranked by a scoring function implemented in the Autodock. A point grid with a spacing 0.0475 nm was used. A point grid was centered to the center of mass of the protein, its dimensions were 12.6 × 12.6 × 12.6 nm. Gasteiger atom charges were assigned to the protein atoms using AutoDock tools. Water molecules were excluded from the protein before docking. One hundred randomly seeded runs were performed. The binding poses were identified by the AClAP 1.0 clustering procedure.[55]

3. **Hydration and Thermal Stability of 1−HuPrP^C Adducts.** A 10 ns MD simulation of the adducts (HuPrP^C−**1**^{0}, HuPrP^C−**1**$^+$ and HuPrP^C−**1**$^{2+}$) allowed for proper hydration of the system and identification of the collective motions that may be essential for PrP^C−ligand interactions. The protomers were bound to the binding region I (Figure 2B). The simulation protocol was the same as for the free protein. For the three ligands, the gaff force field[42,56] was used. The atomic restrained electrostatic potential (RESP) charges[57,58] were calculated by using the respectively module of AMBER after geometry optimization and electrostatic potential calculations of each protomer at the B3LYP/6-31G** level of theory by means of the g03 software.[53]

4. **Dissociation Free Energy Calculations.** The dissociation free energies of **1**^{0}, **1**$^+$, and **1**$^{2+}$ were calculated using metadynamics[25] in its bias-exchange variant[26] as a function of collective variables (CVs), which should be relevant for describing the dissociation process. CVs used in this work are: (i) the distance between the center of mass of the ligand and the protein binding region; (ii) the number of polar contacts between the ligand and one portion of the protein binding region I; (iii) the number of polar contacts between the ligand and the other portion of the protein binding region I; (iv) the number of water bridge contacts between the ligand and the protein binding region I; (v) the rmsd difference of the system with respect to an equilibrated MD structure taken from the previous section; and (vi) the rmsd fluctuation of the residues defining the protein binding region I. The choice of these CVs was based on previous ligand−target interaction metadyamics studies[27–30] as well as by observations based on the former MD simulations (see Supporting Information for more details). The calculations

do not require in principle the previous knowledge of the protein−ligand adduct structure. However, for computational efficiency we exploit the fact that all the target regions detected from NMR are in the close proximity of region I. Therefore, here we explored only this region.

Six independent metadynamics simulations were run in parallel. Each replica was biased by different one-dimensional time-dependent potentials, which were built as a function of each of the collective variables defined above. Exchanges among replicas were attempted every 10 ps using a metropolis acceptance criterion.[26] Similar setup was shown to improve the sampling of the configurational space and the convergence of the results.[26,59–63] At the end of the different replica simulations, the explored phase space, in terms of the six collective variables used in this study, was clustered using the gromos method.[52] The clustering radius for each collective variable was set to 0.1, 0.2, 0.4, 2.5, 0.05, and 0.02 nm, respectively. The free energy corresponding to each cluster was then reconstructed from the populations of clusters observed during the simulations. The free energy value was corrected by the corresponding bias potentials acting on that cluster as in a usual weighted histogram analysis.[62] Details on this procedure can be found in references,[59–63] and are summarized in Supporting Information together with the converged free energy profiles.

Two reference states of the ligand-protein system, bound and unbound, needed to be defined to provide the corresponding dissociation free energy value. The bound state was considered as the lowest free energy cluster. The unbound state was considered to be a cluster showing no contacts with the binding regions I (lowest values of CVs *ii* and *iii*) and at the same time with a higher rmsd with respect to the initial docked structure (highest value of CV *v*). Given the size of the simulation box, the ligand is not fully detached from the protein in its unbound state. Therefore, the residual dissociation energy of the unbound state was roughly estimated in implicit solvent using an adaptive Poisson−Boltzmann solver. The APBS package[64] was used with the same parameters described previously. It was estimated as the difference in solvation energy of the complex minus the solvation energy of each component plus the intermolecular Coulomb interaction. The standard free energy of dissociation was obtained by applying the following relationship: $\Delta G^0 = \Delta G - RT \ln([L])$, where ΔG is the total dissociation free energy as a result of our simulation, R is the molar constant, and $[L]$ is the concentration of the ligand in our simulation box (i.e., 5.5 mM, corresponding to one molecule in 6.69^3 nm^3). The standard free energy is related to the dissociation equilibrium constant (K_d) by $\Delta G^0 = -RT \ln(K_d)$.

Acknowledgment. The authors acknowledge A. Laio, F. Marinelli, and F. Pietrucci for fruitful discussions and for providing the necessary information to perform the analysis of the bias-exchange metadynamics simulations. X.B. acknowledges the financial support from the Government of Catalonia through a Beatriu de Pinós fellowship (BP-A 2007 http://ww.gencat.cat/agaur). P.C. acknowledges financial support from IIT (http://www.iit.it).

Supporting Information Available: Details of the MD simulations of compound **1** protomers in complex with HuPrP^C and additional information of the bias-exchange metadynamics procedure. The relevant experimental NMR data referenced in this work are summarized. Parameters of the biasing potential in the metadynamics calculations are provided. Furthermore, six supplementary figures, as referenced in the main text, are reported. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug. Discov.* **2004**, *3*, 935–49.

(2) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. *Med. Res. Rev.* **2006**, *26*, 531–68.

(3) May, A.; Zacharias, M. *Biochim. Biophys. Acta* **2005**, *1754*, 225–31.

(4) Mangoni, M.; Roccatano, D.; Di Nola, A. *Proteins* **1999**, *35*, 153–62.

(5) Pak, Y.; Wang, S. *J. Phys. Chem. B* **2000**, *104*, 354–9.

(6) Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. *J. Am. Chem. Soc.* **2002**, *124*, 5632–3.

(7) McCammon, J. A. *Biochim. Biophys. Acta* **2005**, *1754*, 221–4.

(8) Carlson, H. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–52.

(9) Bucciantini, M.; Giannoni, E.; Chiti, F.; Baroni, F.; Formigli, L.; Zurdo, J.; Taddei, N.; Ramponi, G.; Dobson, C. M.; Stefani, M. *Nature* **2002**, *416*, 507–11.

(10) Cavalli, A.; Bolognesi, M. L.; Minarini, A.; Rosini, M.; Tumiatti, V.; Recanatini, M.; Melchiorre, C. *J. Med. Chem.* **2008**, *51*, 347–72.

(11) Soto, C. *Nat. Rev. Neurosci.* **2003**, *4*, 49–60.

(12) Prusiner, S. B. *Science* **1982**, *216*, 136–44.

(13) Caughey, B.; Baron, G. S. *Nature* **2006**, *443*, 803–10.

(14) Prusiner, S. B. *Trends Biochem. Sci.* **1996**, *21*, 482–7.

(15) Trevitt, C. R.; Collinge, J. *Brain* **2006**, *129*, 2241–65.

(16) Laurén, J.; Gimbel, D. A.; Nygaard, H. B.; Gilbert, J. W.; Strittmatter, S. M. *Nature* **2009**, *457*, 1128–1132.

(17) Kuwata, K.; Nishida, N.; Matsumoto, T.; Kamatari, Y. O.; Hosokawa-Muto, J.; Kodama, K.; Nakamura, H. K.; Kimura, K.; Kawasaki, M.; Takakura, Y.; Shirabe, S.; Takata, J.; Kataoka, Y.; Katamine, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11921–6.

(18) Calzolai, L.; Zahn, R. *J. Biol. Chem.* **2003**, *278*, 35592–6.

(19) Riek, R.; Hornemann, S.; Wider, G.; Billeter, M.; Glockshuber, R.; Wuthrich, K. *Nature* **1996**, *382*, 180–2.

(20) Ishikawa, T.; Ishikura, T.; Kuwata, K. *J Comput Chem* 2009 [Online early access]. DOI: 10.1002/jcc.21265. Published Online: Apr 30, 2009. http://www3.interscience.wiley.com/cgi-bin/fulltext/122370880/HTMLSTART. Accessed June 28, 2009.

(21) Vogtherr, M.; Grimme, S.; Elshorst, B.; Jacobs, D. M.; Fiebig, K.; Griesinger, C.; Zahn, R. *J. Med. Chem.* **2003**, *46*, 3563–4.

(22) Kirby, L.; Birkett, C. R.; Rudyk, H.; Gilbert, I. H.; Hope, J. *J. Gen. Virol.* **2003**, *84*, 1013–20.

(23) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(24) Rodinger, T.; Pomes, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–70.

(25) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–6.

(26) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–9.

(27) Branduardi, D.; Gervasio, F. L.; Cavalli, A.; Recanatini, M.; Parrinello, M. *J. Am. Chem. Soc.* **2005**, *127*, 9147–55.

(28) Fiorin, G.; Pastore, A.; Carloni, P.; Parrinello, M. *Biophys. J.* **2006**, *91*, 2768–2777.

(29) Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2005**, *127*, 2600–2607.

(30) Vargiu, A. V.; Ruggerone, P.; Magistrato, A.; Carloni, P. *Nucleic Acids Res.* **2008**, *36*, 5910–21.

(31) Calculator Plugins were used for structure property prediction and calculation. *Marvin 5.0.0*; ChemAxon: Budapest, Hungary, 2008; http://www.chemaxon.com; (accessed May 29, 2008).

(32) Campana, V.; Zentilin, L.; Mirabile, I.; Kranjc, A.; Casanova, P.; Giacca, M.; Prusiner, S. B.; Legname, G.; Zurzolo, C. *Biochem. J.* **2009**, *418*, 507–15.

(33) Williamson, R. A.; Peretz, D.; Pinilla, C.; Ball, H.; Bastidas, R. B.; Rozenshteyn, R.; Houghten, R. A.; Prusiner, S. B.; Burton, D. R. *J. Virol.* **1998**, *72*, 9413–8.

(34) Kaneko, K.; Zulianello, L.; Scott, M.; Cooper, C. M.; Wallace, A. C.; James, T. L.; Cohen, F. E.; Prusiner, S. B. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10069–74.

(35) Takemura, K.; Wang, P.; Vorberg, I.; Surewicz, W.; Priola, S. A.; Kanthasamy, A.; Pottathil, R.; Chen, S. G.; Sreevatsan, S. *Exp. Biol. Med.* **2006**, *231*, 204–214.

(36) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Res.* **2005**, *33*, W368–71.

(37) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(38) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–48.

(39) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–23.

(40) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–62.

(41) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–88.

(42) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2005**, *26*, 114–114.

(43) Aaqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

(44) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(45) Hoover, W. G. *Phys. Rev. A:At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.

(46) Nose, S. *Mol. Phys.* **1984**, *52*, 255–268.

(47) Nosé, S.; Klein, M. L. *Mol. Phys.* **1983**, *50*, 1055–76.

(48) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–92.

(50) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(51) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(52) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236–40.

(53) Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; 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.; and Pople, J. A.; Gaussian, Inc.: Wallingford, CT, 2004.

(54) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. *Proteins* **2002**, *49*, 457–71.

(55) Bottegoni, G.; Rocchia, W.; Recanatini, M.; Cavalli, A. *Bioinformatics* **2006**, *22*, e58−65.

(56) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–74.

(57) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(58) Cornell, W.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.

(59) Leone, V.; Lattanzi, G.; Molteni, C.; Carloni, P. *PLoS Comput. Biol.* **2009**, *5*, e1000309.

(60) Piana, S.; Laio, A.; Marinelli, F.; Van Troys, M.; Bourry, D.; Ampe, C.; Martins, J. C. *J. Mol. Biol.* **2008**, *375*, 460–70.

(61) Todorova, N.; Marinelli, F.; Piana, S.; Yarovsky, I. *J. Phys. Chem. B* **2009**, *113*, 3556–3564.

(62) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. *PloS Comp. Biol.* **2009**, *5*, e1000452.

(63) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. *J. Am. Chem. Soc.* **2009**; in press. Epub ahead of print, available at http://pubs.acs.org/doi/abs/10.1021/ja903045y.

(64) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–41.

(65) Esteras-Chopo, A.; Morra, G.; Moroni, E.; Serrano, L.; Lopez de la Paz, M.; Colombo, G. *J. Mol. Biol.* **2008**, *383*, 266–280.

(66) Herrera, F. E.; Chesi, A.; Paleologou, K. E.; Schmid, A.; Munoz, A.; Vendruscolo, M.; Gustincich, S.; Lashuel, H. A.; Carloni, P. *PLoS One* **2008**, *3*, e3394.

(67) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14.1*, 33–38.

CT900257T

JCTC Journal of Chemical Theory and Computation

Intramolecular Basis Set Superposition Error Effects on the Planarity of DNA and RNA Nucleobases

David Asturiol,[†] Miquel Duran,[‡,‡] and Pedro Salvador*[,‡,‡]

Institut de Química Computacional, Parc Científic i Tecnològic de la Universitat de Girona, Edifici Jaume Casademont, Pic de Peguera 15 (la Creueta), 17003 Girona, Spain, and Institut de Química Computacional and Departament de Química, Universitat de Girona, Campus de Montilivi, 17071 Girona, Spain

Received January 30, 2009

Abstract: Molecules of utmost importance like DNA and RNA nucleobases are predicted to be nonplanar by a typical ab initio method, such as second order Møller−Plesset perturbation theory (MP2) combined with standard Pople's basis sets. Similarly to the case of other planar aromatic systems, these pitfalls can be explained in terms of intramolecular basis set superposition error (BSSE) effects, induced by local basis set deficiencies. We demonstrate that conventional BSSE correction techniques such as the Counterpoise method can account for this wrong behavior and provide proper correction whenever spurious results occur, mainly in case of thymine, uracil and guanine but also to lower extent for adenine and cytosine. We also show that special care must be taken when assessing the BSSE by means of ghost-orbital calculations for strongly overlapping fragments. Often molecular orbitals in the isolated fragment calculation have a different orientation as in the ghost-orbital calculation. This can lead to bogus derivatives of the CP-correction term, essential to account for geometry and vibrational BSSE effects.

Introduction

In the last years a number of studies[1−10] have reported conventional ab initio calculations at the correlated level producing nonplanar minima for systems such as benzene, polycyclic aromatic hydrocarbons, and other nonrigid cyclic systems, such as the hexacoordinated carbon anion $(B_6C)^{2-}$. The planar stationary structures corresponded to saddle-points on the potential energy surface (PES) with one or more imaginary frequencies, associated to low-lying out-of-plane vibrational modes.

Until very recently the origin of such anomalies had not yet been clearly determined. Martin et al.[4] studied in detail the benzene molecule and suggested a basis set superposition error (BSSE) origin. Moran et al.[6] reported a wide number of correlated calculations combined with Pople's basis sets for aromatic systems that produced nonplanar minima. In these cases, their analysis revealed a strongly geometry-dependent two-electron basis set incompleteness error (BSIE) which increased for nonplanar geometries. However, no alternatives but the use of a different (more balanced) basis set or the careful extrapolation to the complete basis set limit were provided as a solution for these problematic cases.

Recently,[11] we have offered a solution to the problem based on the use of conventional BSSE correction techniques which have been successfully applied to correct for BSSE in ab initio descriptions of intermolecular complexes. We have confirmed that the origin of the reported pitfalls arise from local basis set deficiencies. This clearly indicates the fact that intra- and intermolecular BSSE have common origin and can be managed in a similar fashion. The same philosophy has been applied very recently by Balabin et al.[12] to obtain accurate energy differences for several conformations of normal alkanes.

The origin of the intermolecular BSSE in ab initio calculations is rooted on the use of truncated basis sets centered on the atomic positions. The interaction energy

* To whom correspondence should be addressed. E-mail: pedro.salvador@udg.edu.

[†] Institut de Química Computacional, Parc Científic i Tecnològic de la Universitat de Girona.

[‡] Institut de Química Computacional and Departament de Química, Universitat de Girona.

10.1021/ct900056u CCC: $40.75
Published on Web 07/31/2009

between two molecules is usually calculated subtracting the energy of the fragments from the energy of the complex. In such a calculation, the energy of the fragments is obtained using the basis set of each fragment, whereas the energy of the complex is calculated using the basis set of all fragments.

It is well-known that BSSE affects not only the interaction energy of the complex but also its potential energy surface. This translates into geometrical and vibrational effects associated with the change of the position of the stationary point and nearby curvature of the PES. Several examples of how BSSE can dramatically affect the geometry of intermolecular complexes can be found in the literature.[13−15]

For a single molecule, there is no a priori problem with the fact that atoms or groups of atoms make use of basis functions centered on other parts of the molecule, as it takes into account polarization and charge transfer effects. However, if the use of these external basis sets is the result of a lack of flexibility of the fragment's own basis set, these local basis set deficiencies may result into spurious stationary points and vibrational frequencies associated. One can refer to such phenomena as intramolecular BSSE.[16−20]

There are several strategies to correct for BSSE[21−24] but the Counterpoise method[21] (CP) is the most widely used due to its simplicity. The CP-correction to the energy for a system formally composed of N interacting fragments reads as

$$\delta^{CP}(\vec{R}) = \sum_{i}^{N} \varepsilon_i^i(\vec{R}) - \varepsilon_i^{\text{full}}(\vec{R}) \quad (1)$$

where $\varepsilon_i^i(\vec{R})$ and $\varepsilon_i^{\text{full}}(\vec{R})$ represent the energy of the i-th fragment of the system calculated with its own basis set and with the full basis set of the system (ghost-orbital calculation), respectively. Explicit dependence on the atomic positions has been included to stress that the CP correction is geometry-dependent. Also, it is worth to note that the electronic state of each fragment must be specified in terms of its charge and spin multiplicity. The CP-correction, applied as an additive correction term to the total energy,[25]

$$E^{CP}(\vec{R}) = E(\vec{R}) + \delta^{CP}(\vec{R}) \quad (2)$$

provides BSSE corrected energies, as well as any property that can be obtained from the total energy of the system or its derivatives, namely, stationary points, vibrational frequencies, dipole moments, etc.

In the case of intermolecular complexes, the definition of the fragments is usually trivial; each interacting molecule is considered as a fragment forming the so-called supermolecule. This permits to obtain BSSE-corrected interaction or stabilization energies with respect to the corresponding fragments. Some ambiguities in the specification of the fragments may arise in the specific case of charged intermolecular complexes or interactions involving open-shell species, namely, which fragment bears the charge and which multiplicity is to be specified. Nevertheless, such cases have already been considered and satisfactory results have been obtained.[19]

The situation is rather different in the case of intramolecular BSSE correction. One must face the ambiguity of subdividing the molecular unit into subunits from which a BSSE-correction will be determined. This may involve the rupture of chemical bonds, and therefore, the specification of the electronic state of each individual fragment becomes another source of arbitrariness. An alternative to avoid such unsaturated fragments is to estimate BSSE by modeling the system with a proper intermolecular complex with the same geometry.[26−28] Such approach seems to provide reasonable results for single-point calculations but it is certainly nontrivial how it could be applied for CP-corrected geometry optimizations and frequency calculations.

Another possible general approach for a molecule could be to use atomic fragments. However, the CP philosophy might be difficult to accomplish in this case (vide infra). Ideally, atoms ought to be promoted to the hybridization state they appear in the molecule when determining the basis set extension effects, and this is not a trivial task. The use of atomic fragments also has dramatic implications with the computational cost associated to the CP-correction, which is roughly $N + 1$ times ($N =$ the number of fragments) that of the uncorrected calculation. For any midsize molecule such computational scheme can easily become unfeasible.

It is worth noting that in the case of a more involved methodology that has proven to be very successful in correcting for BSSE in the intermolecular case, namely the Chemical Hamiltonian Approach (CHA), the fragments are essentially defined by their associated basis functions so that charge and multiplicity do not need to be specified for each fragment. In fact, the CHA method does not rely on any extra fragment calculations since the total energy obtained in the calculations is already free from BSSE. The reason for not using it in this context is that the method does not seem to behave properly for strongly overlapping fragments, which is the case of intramolecular BSSE (a variant[16] of the method at the Hartree−Fock level was developed some years ago with promising results for intramolecular BSSE correction, too). Nevertheless, what one learns from an a priori BSSE correction method such as the CHA is that the CP method is essentially an ingenious balance error technique that, as pointed out by Mayer,[29] assumes that the energy lowering induced by the use of external basis functions for a given fragment within the supermolecule is the same as the one obtained by considering the single fragment together with the whole set of basis functions (ghost orbitals). The more similar the electronic state of the isolated fragment is to the *local state* of that fragment within the molecule, the more appropriate the CP-correction will be.

CP-corrected calculations are nowadays carried out quite routinely without paying much attention to the electronic state of the fragment's calculations. A notably exception are Alexander and co-workers, who for a simple system like $B\cdots H_2$ brought up the use of diabatic states carefully chosen in order to obtain accurate CP-corrections for different electronic states.[30] Fortunately, numerical experience so far has shown that the CP correction is, in this aspect, quite robust. For instance, in the case of charged intermolecular complexes such as the protonated water dimer, quite reasonable results were obtained using either two or three fragments.[19] For a genuine open-shell complex such as HF···NO, we described[31] difficulties in the selection of proper ghost-

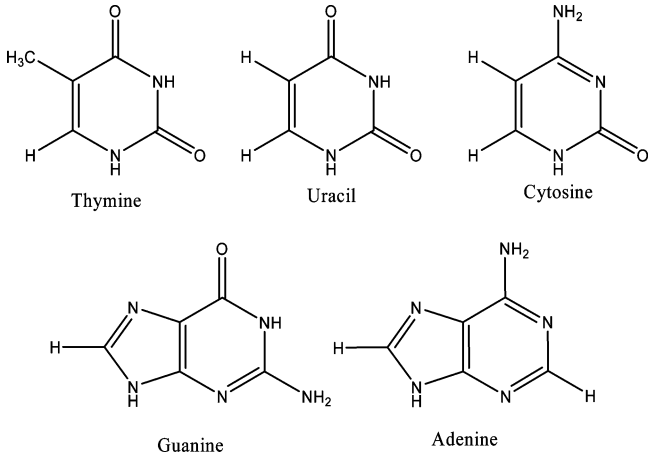


Figure 1. Nucleobases considered in this study.

orbital states for obtaining CP-correction for different electronic states, but the energetic differences between them were very small.

As mentioned above, in a recent work[11] we have tackled the correction of intramolecular BSSE effects in the case of benzene and several arenes, including charged systems such as cyclopentadienyl and indenyl anions at the MP2 and Configuration Interaction with Singlet and Doublet excitations (CISD) levels of theory combined with standard Pople's basis sets. In all cases, the problems were associated with out-of-plane bending low-lying modes for which one or more imaginary frequencies can be found (as large as 1181*i* at the MP2/6-311++G** level of theory). Simple inspection showed that the intramolecular BSSE did not affect bonding distances or angles (otherwise stretching and other bending modes would have been affected). Taking this in consideration, we showed that such specific intramolecular basis set deficiencies could be solved by taking as fragments diatomic C−H units constituting the arenes. To maintain the molecule's symmetry all C−H fragments must be equivalent. In the case of charged systems, careful choice of the CP correction combining charged and neutral C−H moieties had to be designed.

The aim of this work is to show that the anomalous behavior observed for benzene and other arenes seems to be in fact quite common for cyclic planar molecules with π-systems. Here we describe a series of spurious imaginary frequencies associated with out-of-plane bending modes for planar stationary points of adenine, cytosine, thymine, guanine, and uracil nucleobases (see Figure 1) obtained at the MP2 level of theory and Pople's standard basis sets. Luckily, what they seem to be the most widely used basis sets in the literature for these systems, namely 6-31G* and 6-311G*, do not present any spurious imaginary frequency in any case. However, if a better description is needed and diffuse functions are added to those bases, thymine, uracil and guanine optimized planar structures can present one or more imaginary frequencies.

These molecules, apart from being essential building blocks of life, are especially important in the photophysics field as they are some of the hotter molecules of the last and present decade. The fact that these molecules can present such pitfalls at the MP2 level is indeed more relevant as the use of MP2 is very common in photophysics studies.[13,32−35] A normal procedure in such works[36−38] is to optimize ground state structures including dynamic correlation, therefore, the use of MP2 is rather general. In addition, Franck−Condon vertical excitations are carried out at the ground state minimum structure, and all subsequent studies on the excited states are started from that point. Thus, it is essential to get a proper starting point (ground state minimum) to perform an accurate study as minor geometrical changes can imply a change in the order of the states.

Another point of interest is to determine to what extent such basis set deficiencies are localized in a region of the molecule and whether it would be sufficient to correct for intramolecular BSSE only locally, that is, using a specific Counterpoise function that would take into account only a subset of atoms of the system. This would likely be the case of an intermolecular hydrogen bond formation or for instance the interaction between the two ends of a long chain-like molecule. In such cases, a local treatment may be of use not just because BSSE correction would be irrelevant in most parts of the molecule but also to avoid spurious effects of the CP-correction itself. In this respect, with this paper we also aim to show that when the overlap between fragments is strong (i.e., when breaking a chemical bond) the electronic state of the fragment and that of the ghost-orbital calculation might differ, causing spurious CP corrections. Our results indicate that in the intramolecular case it is not of utmost importance which is the electronic state of each fragment, but that the isolated fragment calculation and the corresponding ghost orbital calculation must correspond to the *same* state to obtain a proper BSSE removal.

Computational Details

All ab initio calculations have been carried out with Gaussian 03[39] program. Standard CP-corrected geometry optimizations and vibrational frequency calculations have been performed using the automatic procedure as implemented in Gaussian 03 at the MP2 level of theory (frozen core). For special Counterpoise function definitions we have also used our own code, which allows us to exploit symmetry if any and also permits the use of different specific Gaussian keywords for each fragment calculation (with the Counterpoise keyword the process is automatized in Gaussian 03 but all fragment calculations share the same options).

Thymine and uracil nucleobases were optimized within *Cs* symmetry. No symmetry constraint other than planar ring was used for cytosine, guanine and adenine. An active space including all π orbitals (10 electrons in 8 orbitals for thymine) was used for the Complete Active Space Self Consistent Field (CASSCF) calculations.

Results and Discussion

Let us focus first in the particular case of thymine. We have performed geometry optimizations and frequency calculations at the MP2 and CASSCF levels of theory for the same group of basis sets used by Moran et al. in the benzene case (over 24 basis sets featuring Pople's 3-21G, 6-31G, and 6-311G families and Dunning's cc-pVXZ basis). At the MP2 level,

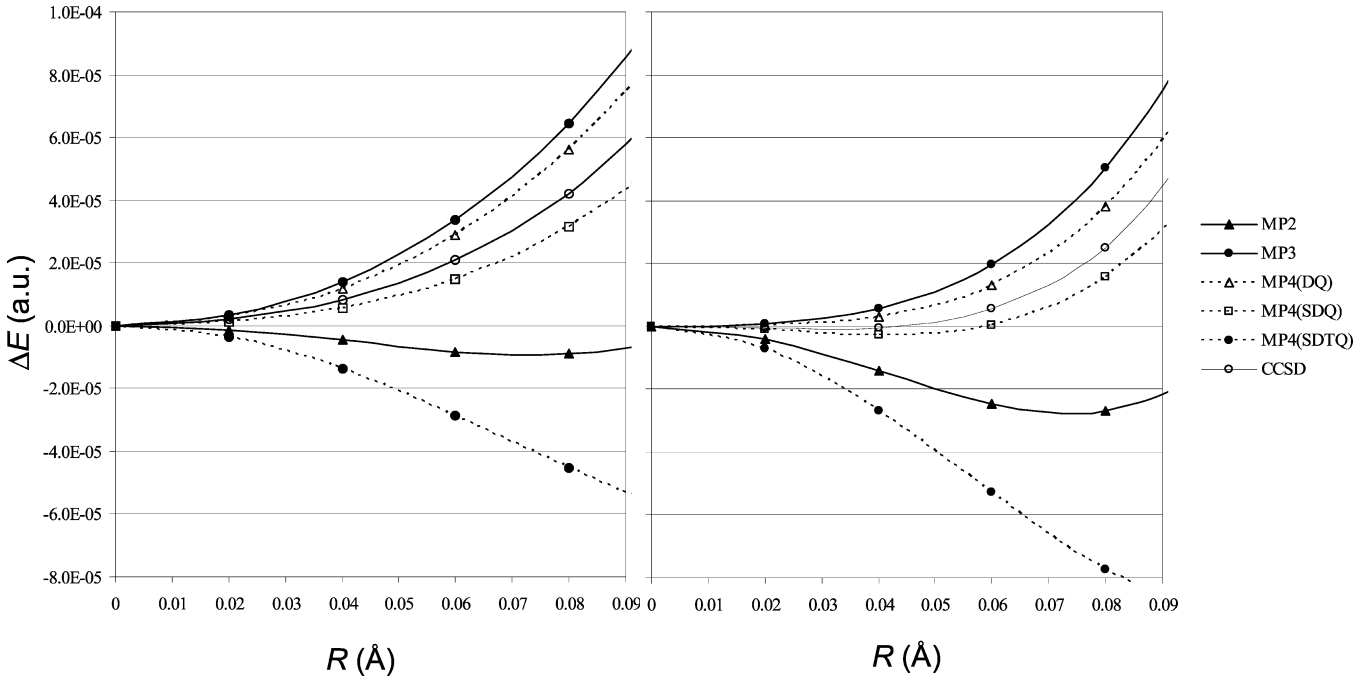


Figure 2. CCSD and MPn energies along the vibrational mode associated to the imaginary frequency for thymine at the MP2 level with the 6-31+G* (left) and 6-311+G*(right) basis sets.

we have obtained imaginary frequencies for the planar optimized structures for 12 of the basis sets used (see Table S1 in the Supporting Information). For particularly unbalanced basis sets such as 6-311++G and 6-311+G, up to three imaginary frequencies have been found. The results are slightly worse than in the case of benzene, as for thymine also the MP2/6-31+G* and MP2/6-31+G** lead to spurious results. Again, these spurious imaginary frequencies were found in correspondence to out of plane bending low-lying modes.

At the CASSCF level, no imaginary frequencies have been found in any case showing that the problems could be associated to two-electron excitations to high energy virtual orbitals with large diffuse character. We have also performed single-point calculations at higher levels of theory with the 6-31+G* and 6-311+G* basis sets along the (uncorrected) vibrational mode associated to the imaginary frequency at the MP2 level of theory (see Figure 2). In both cases, MP3 and MP4(DQ) energies produce the correct profile. The inclusion of the triples in the MP4 energy expression leads to a wrong profile and the inclusion of singles leads to wrong description only for the large basis set. This is also the case for the CCSD method, for which the energy profile is extremely flat in the case of the 6-311+G* basis. This simple analysis shows that these problems are not restricted to the MP2 level and may occur even at the CCSD level. It is also difficult to infer that the problem at the MP2 level might be due to a convergence problem of the MPn series. To answer these questions a much more systematic study would be required, which is beyond the scope of the present work.

Motivated by these similarities with the benzene case, we have attempted an analogous approach for the BSSE removal based upon (mainly) diatomic fragments. Due to the heteroatomic character of the nucleobases (See Scheme 1) we encounter, aside of C−H moieties, other units such as C═O, N−H or CCH_3 fragments. Lewis structures suggested the use of doublet and singlet multiplicities for C−H and C═O fragments, and numerical evidence recommended to use triplet and quadruplet for N−H and CCH_3 fragments, respectively. The reasons behind this choice will be made clear later on.

Table 1. CP-Corrected and Uncorrected Frequencies of Optimized Planar Structures of Pyrimidine Nucleobases[a]

	thymine		uracil		cytosine	
	MP2	CP-correct.	MP2	CP-correct.	MP2	CP-correct.
6-31G*	107	106	134	135	128	128
	138	139	159	156	203	202
	148	153	371	369	357	358
6-31+G*	*80i*	71	*20i*	86	96	105
	100	109	132	138	190	193
	140	153	315	335	337	349
6-311G*	103	104	137	136	123	124
	139	141	151	152	203	201
	147	151	370	370	360	359
6-311+G*	*160i*	75	*113i*	93	67	108
	84	108	103	137	189	193
	140	151	280	342	309	359

[a] Imaginary frequencies are displayed in italics.

The structures were reoptimized according to the total CP-corrected energy and CP-corrected frequency calculations were performed on the CP-optimized planar stationary structures. In Table 1, we present the three lowest vibrational frequencies for thymine, uracil and guanine obtained for four selected basis set cases. With the above-mentioned fragment definition the CP procedure provided excellent results in all cases. The imaginary frequencies were removed in the problematic cases and no significant effect was observed for those which showed proper behavior.

Nucleobases present less symmetry constraints than benzene and other arenes considered in our previous work. Thus, in the present case, one has more freedom to choose proper

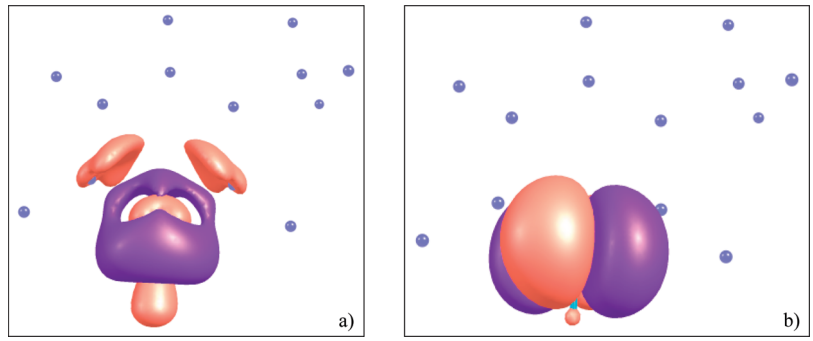


Figure 3. Density difference plot between ghost-orbital and isolated calculation for a N−H fragment in thymine for (a) triplet and (b) singlet electronic states. The position of the ghost-atoms is shown with semitransparent blue spheres. See text for isosurface values.

fragments. In fact, if the basis set deficiencies would be rather localized, one could use a Counterpoise function including only those fragments that would be needed to correct for such deficiencies. Accordingly, we have explored several fragment definitions and Counterpoise functions for this system. Some of our findings are described next.

First of all, the use of different multiplicity on the N−H fragments has a dramatic effect on the low lying out-of-plane mode. The reason is that, in the singlet case, conventional ghost orbital calculations lead to a qualitatively different state than that of the isolated fragment calculation. The explanation is simple in terms of molecular orbital occupations. In the singlet case the HOMO corresponds to one of the p_x,p_y degenerate orbitals in the isolated fragment calculation. However, in the ghost-orbital calculation, this degeneracy is broken and the in-plane p orbital is stabilized by the presence of ghost-orbitals (mainly of s symmetry) of the neighboring atoms. Energetically speaking there is no apparent problem in the energy difference between the isolated and ghost-orbital calculations. However, if the HOMO in the isolated fragment calculation does not happen to have the same orientation as in the ghost-orbital calculation artificial effects appear beyond energy correction, namely, first- and second derivatives of the energy. This can be visualized by comparing the difference between the two densities obtained with and without ghost orbitals at the Hartree−Fock level, as shown in Figure 3. The position of the ghost-atoms is represented by semitransparent blue spheres. The isosurface value in the triplet case is set to 0.0005. Thus, the differences are very small and partially localized in the closest carbon ghost-atoms. Because of BSSE-like basis set extensions, the density is redistributed in the ghost-orbital calculation, increasing in the vicinity of the closest atoms from which the basis functions are used and slightly decreasing in the region close to the nuclei. However, in the case of the singlet calculation, the density difference between the ghost-orbital calculation and the isolated fragment is much larger (isosurface value is set to 0.005 for clarity) and localized in the N−H unit. The typical polarized picture suggests that the two densities correspond to two rotated electron distributions. The inclusion of such energy (and specially energy derivative) differences in eq 1 leads to an essentially wrong CP-correction, which has no correcting effect on the out-of-plane molecular distortions and introduces spurious effects on the stretching modes associated to the N−H moieties. A similar effect has been observed for the rather unchemical C−CH_3 fragment arising from the methyl substituent in the heterocycle. The fragment in the doublet state exhibits a double bond between the carbon atoms whose orientation is again strongly affected by the presence of ghost-orbitals. Such a fragment definition leads to meaningless CP-corrected frequencies. Of course, these problems could be solved simply by proper rotation of the orbitals of isolated fragment calculation, but this might not be easily achieved in automatized procedures such as the Counterpoise keyword in Gaussian 03, for instance. This just shows that one must be very careful in these cases when carrying out routine ghost-orbital calculations to quantify basis set extension effects.

We also explored the effect of using multiplicity specification for the C−H and C═O fragments and no noticeable differences were observed. In the case of the C−H fragment one might foresee similar problems associated with the partial occupation of degenerate p orbitals in the low spin case. It seems rather fortunate that the conventional ghost-orbital calculation lead to a similar orientation of the SOMO orbital.

Another point of interest is to determine to which extent the intramolecular BSSE exhibited by this system is a local effect or not. For this we have considered separately each fragment contribution to the Counterpoise correction and obtained the corresponding CP-corrected frequencies. It is worth to mention that the CP-optimization does not lead to meaningful geometry changes with respect to the conventional MP2 calculation. For instance, the largest deviation on the internal coordinates of thymine at the MP2/6-31+G* level induced upon CP-optimization were just 0.009 Å and 0.3° in bond distance and angles, respectively. That means that one can reasonably obtain frequency corrections with partial Counterpoise functions on the same CP-optimized geometry obtained with the full Counterpoise correction, which largely simplifies the following analysis.

At the MP2/6-31+G* level, the lowest lying out-of-plane vibration for planar thymine shows an imaginary frequency of 80*i*. The use of a CP-correction including contributions from just one of the six fragments depicted in Figure 4 does not lead in any case to a change in the topology of the stationary point. The value of the imaginary frequency decreases in all cases, reaching a highest value of 40*i* in the best case, for the N−H fragment in ortho position with respect to the C−H group (number 4 in Figure 4). Already

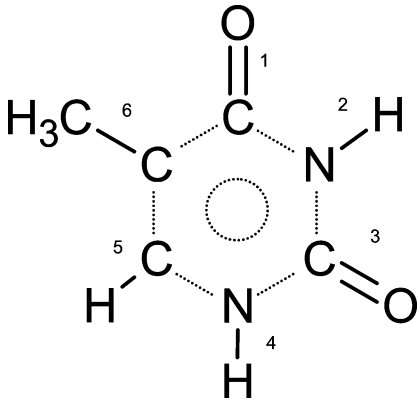


Figure 4. Intramolecular fragments used for CP-correction in thymine.

when considering two adjacent fragment's contributions at a time one can observe a change in the topology. Using fragments 4−5 and 3−4 in the Counterpoise function the lowest frequency assumes values of 39 and 37 cm^{-1}, respectively. Nevertheless, a similar value (35 cm^{-1}) is obtained including distant fragments in para position like 1 and 4. Other combinations involving the C−H fragment also provide corrections in the proper direction. With the progressive inclusion of more fragment contributions to the CP-correction the value of the lowest frequency increases monotonically up to the reported value of 71 cm^{-1} when using all six fragments. For instance, using the three adjacent fragments 3,4,5 the value is 51 cm^{-1} and including all contributions except that of the $C-CH_3$ fragment, an almost converged value of 63 cm^{-1} is obtained. With this analysis we can conclude that the intramolecular BSSE effects were to some extent localized around the N−H moiety in ortho position with respect to the C−H group. However, its removal is not enough to produce a change of topology and contributions from adjacent C−H and C═O groups must also be taken into account. We have also seen that contributions from distant fragments with little chemical significance such as the $C-CH_3$ could be safely ignored from the CP function if necessary. It arises from our results that BSSE effects seem to be quite delocalized on the heterocycle and accordingly to get a proper BSSE removal all fragment contributions should be taken into account. Nevertheless, we do not expect this to be a general trend for intramolecular BSSE problems. Further studies on the determination of the local character of BSSE effects in intramolecular hydrogen bonding situations are currently under work.

After considering the thymine case in deep detail, in the following, we will describe the results obtained for the rest of nucleobases we have considered.

The three lowest vibrational frequencies obtained for all DNA and RNA nucleobases at the MP2 level with four representative basis sets are given in Tables 1 and 2. The CP-corrected values obtained after proper intramolecular BSSE removal are also provided. Thymine and uracil structures were optimized with *Cs* symmetry, whereas guanine, adenine, and cytosine present a planar ring with a NH_2 group slightly out of plane due to pyramidalization that breaks the symmetry. CP-corrected results obtained suggest that geometrical reoptimizations might not be necessary as only very minor changes are observed between the uncorrected and CP-corrected geometries. Nevertheless, all CP-corrected frequency calculations were carried out upon CP-optimized geometries.

Table 2. CP-Corrected and Uncorrected Frequencies of Optimized Planar Structures of Purine Nucleobases[a]

	adenine		guanine	
	MP2	CP-correct.	MP2	CP-correct.
6-31G*	159	158	129	127
	207	205	151	152
	272	272	192	194
6-31+G*	126	151	*55i*	116
	185	192	128	133
	275	262	165	178
6-311G*	160	161	131	133
	213	208	156	155
	273	274	195	194
6-311+G*	139	150	8	113
	192	195	127	131
	274	276	164	175

[a] Imaginary frequencies are displayed in italics.

Fortunately, all optimized structures using the common 6-31G* and 6-311G* basis sets were characterized as true minima. The CP-correction (see below) did not change this situation and in fact, the values of the three lowest frequencies were obtained within a deviation of 3% with respect to the corresponding conventional calculation.

When using diffuse functions the intramolecular BSSE effects can be very important. Concerning the CP-correction, pyrimidine derivatives present no special difficulties for a proper fragment definition. They are characterized by a six member ring and six substituents, (except for cytosine, which has one unsubstituted position in the ring). Uracil is very similar to thymine as they only differ by a methyl group, which in the case of uracil is a C−H unit. Accordingly, the results obtained for uracil follow the same tendency as those observed for thymine. Imaginary frequencies associated to the low lying out-of-plane mode are found for the 6-31+G* and 6-311+G* basis sets. When no diffuse functions are included in the basis set, the planar structures correspond to true minima. Again, the CP-correction using the analogous fragment definitions as in the case of thymine is able to account for this pitfall in both cases, with very little effect on the already correct descriptions.

Interestingly, no imaginary frequencies have been found in the case of cytosine, despite his similarity with thymine and especially with uracil (see Figure 1). Yet, again for the 6-31+G* and 6-311+G* basis sets the values of the lowest vibrational frequencies are somewhat too small compared with the results obtained with more balanced basis sets. Hence, there are some intramolecular BSSE effects but not to the extent of changing the topology of the planar stationary points. In fact, cytosine shares with the other two nucleobases the same three substituents that were observed to contribute more to the BSSE effects in thymine. This shows again that for such heterocyclic systems the BSSE effects are rather subtle and delocalized. For completeness we have performed also CP-corrected optimizations and frequency calculations for this system. The absence of substituent in ortho position with respect to the C═O group leads to the difficulty of

dealing with a single-atom fragment; highly symmetric in its isolated state but not in the presence of ghost-orbitals. To avoid the problems described before we have considered the N atom as a fragment in high spin state. As alternative, we have explored also the use of a larger fragment involving two adjacent positions of the ring, namely NC-NH_2. Both fragment definitions lead to very similar results (maximum deviation of 8 cm^{-1} in the third frequency at the MP2/6-31+G* level), and only the results from the first option are reported in Table 1. Even though no imaginary frequencies were observed, the CP-corrected frequencies are more similar among the four basis sets used than the uncorrected ones. For instance, the somewhat too low values of 67 and 309 cm^{-1} obtained with the 6-311+G* basis set are blue-shifted by 40 and 50 cm^{-1} to a value in much better agreement with the one obtained with the 6-311G* basis set.

Finally, the results obtained for adenine and guanine basis sets are collected in Table 2. The molecule of adenine seems to be the less prone to intramolecular BSSE effects. No imaginary frequencies have been observed for the planar optimized structures and only a slight drop of about 20−25 cm^{-1} in the value of the lowest vibrational frequency when including diffuse functions has been observed. The situation in the case of guanine is different as again difficulties are observed specially with the 6-31+G* basis set, for which an imaginary frequency of 55*i* is obtained. The value of 8 cm^{-1} obtained with the 6-311+G* basis set can also be considered as spurious.

Purine bases are characterized by a heterocyclic six-membered ring fused to an imidazole ring. Therefore, the definition of fragments to account for intramolecular BSSE must be somewhat different to that of the pyrimidine bases discussed above. The Lewis structures for each molecule show the presence of a double bond involving a C−C pair in the edge of the two fused rings which probably should not be broken. Following this premise, we ended up with C═O, N−H, and C═C diatomic fragments and larger N═C-NH_2 and N═CH fragments involving the unsaturated N atoms, which were considered in high spin state. Once again, the results obtained are very satisfactory. For guanine, the two wrong vibrational frequencies obtained with the 6-31+G* and 6-311+G* basis are efficiently removed with the CP-correction. Even in the case of adenine, where the BSSE effects were less pronounced, the CP-correction induced a slight blue-shift in the lowest vibrational frequencies to final values much closer to those obtained with more balanced basis sets. Several other fragment definitions were also tested, for instance involving a central NC═CN fragment. The CP-corrected results were proved to be very similar, provided situations like those described in detail in the case of thymine were not present.

Conclusions

We have shown that MP2 calculations combined with conventional basis sets including diffuse functions such as the 6-31+G* or 6-311+G* can incorrectly predict imaginary frequencies associated to out-of-plane vibrational modes of planar optimized structures of molecules of utmost importance such as the DNA and RNA nucleobases. Other basis sets like cc-pVXZ and aug-cc-pVXZ seem to be less prone to basis set deficiency problems and are more recommended for vibrational frequencies analysis.

The origin of such pitfalls has been demonstrated to be rooted in intramolecular basis set deficiencies, which eventually lead to intramolecular BSSE effects, similarly to the case of benzene and other planar arenes for which such problems have already been detected and analyzed in detail.

The application of conventional BSSE-correction techniques, such as the Counterpoise method, provide once again proper assessment and correction whenever spurious results occur and do not produce meaningful effects in those cases already correctly described. However, special care must be taken when dealing with strongly overlapping fragments (i.e., when breaking a chemical bond). Even though our results indicate that it is not of utmost importance which is the electronic state of each fragment, it is very important to make sure that isolated fragment and the associated ghost orbital calculations must correspond to the same state with the same orientation of singly occupied degenerate orbitals to obtain a proper BSSE removal.

Acknowledgment. Support for this work from the Spanish Ministerio de Ciencia y Tecnología (project No. CTQ2005-02698/BQU) and from the FPU program (Grant No. AP2004-4774). We are also grateful to Dr. Lluís Blancafort for useful discussions.

Supporting Information Available: Table showing the lowest harmonic vibrational frequencies. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dkhissi, A.; Adamowicz, L.; Maes, G. *J. Phys. Chem. A* **2000**, *104*, 2112–2119.

(2) Goodman, L.; Ozkabak, A. G.; Thakur, S. N. *J. Phys. Chem.* **1991**, *95*, 9044–9058.

(3) Lampert, H.; Mikenda, W.; Karpfen, A. *J. Phys. Chem. A* **1997**, *101*, 2254–2263.

(4) Martin, J. M. L.; Taylor, P. R.; Lee, T. J. *Chem. Phys. Lett.* **1997**, *275*, 414–422.

(5) Michalska, D.; Zierkiewicz, W.; Bienko, D. C.; Wojciechowski, W.; Zeegers-Huyskens, T. *J. Phys. Chem. A* **2001**, *105*, 8734–8739.

(6) Moran, D.; Simmonett, A. C.; Leach, F. E.; Allen, W. D.; Schleyer, P. V.; Schaefer, H. F. *J. Am. Chem. Soc.* **2006**, *128*, 9342–9343.

(7) Saeki, M.; Akagi, H.; Fujii, M. *J. Chem. Theory Comput.* **2006**, *2*, 1176–1183.

(8) Shahbazian, S. *Chem. Phys. Lett.* **2007**, *443*, 147–151.

(9) Simandiras, E. D.; Rice, J. E.; Lee, T. J.; Amos, R. D.; Handy, N. C. *J. Chem. Phys.* **1988**, *88*, 3187–3195.

(10) Torii, H.; Ishikawa, A.; Takashima, R.; Tasumi, M. *J. Mol. Struct. (TheoChem)* **2000**, *500*, 311–321.

(11) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Phys.* **2008**, *128*, 144108.

(12) Balabin, R. M. *J. Chem. Phys.* **2008**, *129*, 164101.

(13) Hobza, P.; Havlas, Z. *Theor. Chem. Acc.* **1998**, *99*, 372–377.

(14) Salvador, P.; Paizs, B.; Duran, M.; Suhai, S. *J. Comput. Chem.* **2001**, *22*, 765–786.

(15) Salvador, P.; Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **2000**, *113*, 5666–5674.

(16) Halasz, G. J.; Vibok, A.; Suhai, S.; Mayer, I. *Int. J. Quantum Chem.* **2002**, *89*, 190–197.

(17) Jensen, F. *Chem. Phys. Lett.* **1996**, *261*, 633–636.

(18) Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2001**, *105*, 1944–1950.

(19) Salvador, P.; Duran, M.; Dannenberg, J. J. *J. Phys. Chem. A* **2002**, *106*, 6883–6889.

(20) Sellers, H.; Almlof, J. *J. Phys. Chem.* **1989**, *93*, 5136–5139.

(21) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(22) Iwata, S.; Nagata, T. *Theor. Chem. Acc.* **2007**, *117*, 137–144.

(23) Mayer, I. *Int. J. Quantum Chem.* **1983**, *23*, 341–363.

(24) Nagata, T.; Iwata, S. *J. Chem. Phys.* **2004**, *120*, 3555–3562.

(25) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024–11031.

(26) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8–12.

(27) Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442*, 42–46.

(28) Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.

(29) Mayer, I.; Hamza, A. *Int. J. Quantum Chem.* **2003**, *92*, 174–180.

(30) Alexander, M. H. *J. Chem. Phys.* **1993**, *99*, 6014–6026.

(31) Salvador, P.; Mayer, I. *J. Chem. Phys.* **2004**, *120*, 5882–5889.

(32) Broo, A.; Holmen, A. *J. Phys. Chem. A* **1997**, *101*, 3589–3600.

(33) Cai, Z. L.; Sha, G. H.; Zhang, C. H.; Huang, M. B. *Chem. Phys. Lett.* **1991**, *178*, 273–278.

(34) Colominas, C.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, *118*, 6811–6821.

(35) Leszczynski, J. *Int. J. Quantum Chem.* **1992**, *19*, 43–55.

(36) Hudock, H. R.; Levine, B. G.; Thompson, A. L.; Satzger, H.; Townsend, D.; Gador, N.; Ullrich, S.; Stolow, A.; Martinez, T. J. *J. Phys. Chem. A* **2007**, *111*, 8500–8508.

(37) Kobayashi, R. *J. Phys. Chem. A* **1998**, *102*, 10813–10817.

(38) Perun, S.; Sobolewski, A. L.; Domcke, W. *J. Phys. Chem. A* **2006**, *110*, 13238–13244.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery Jr., J. A.; 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, 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*, revison C.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

CT900056U