JCTC Journal of Chemical Theory and Computation

Protein−Ligand Complexes: Computation of the Relative Free Energy of Different Scaffolds and Binding Modes

Julien Michel,[†] Marcel L. Verdonk,[‡] and Jonathan W. Essex*,[†]

School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom, and Astex Therapeutics Ltd., 436 Cambridge Science Park, Cambridge, CB4 0QA, United Kingdom

Received April 2, 2007

Abstract: A methodology for the calculation of the free energy difference between a pair of molecules of arbitrary topology is proposed. The protocol relies on a dual-topology paradigm, a softening of the intermolecular interactions, and a constraint that prevents the perturbed molecules from drifting away from each other at the end states. The equivalence and the performance of the methodology against a single-topology approach are demonstrated on a pair of harmonic oscillators, the calculation of the relative solvation free energy of ethane and methanol, and the relative binding free energy of two congeneric inhibitors of cyclooxygenase 2. The stability of two alternative binding modes of an inhibitor of cyclin-dependent kinase 2 is then investigated. Finally, the relative binding free energy of two structurally different inhibitors of cyclin-dependent kinase 2 is calculated. The proposed methodology allows the study of a range of problems that are beyond the reach of traditional relative free energy calculation protocols and should prove useful in drug design studies.

Introduction

Free energy is an important thermodynamic property, and its knowledge permits the prediction of a wide variety of chemical phenomena ranging from binding to phase transitions.[1] Ever since the first applications of the free energy perturbation methodology (FEP) to the calculation of relative free energies were reported,[2] considerable methodological efforts have been devoted to improving the simulation protocols. These efforts are partly justified by the reports of several studies, confirming that relative binding free energies of protein−ligand complexes can be predicted with good precision and accuracy, making the technology an attractive tool for drug design.[3−7] Together with a tremendous increase in computational power, these refinements have made binding free energy calculations sufficiently rapid such that it becomes feasible in certain circumstances to consider their routine application in a drug design environment.[8] The free energy difference between two molecules A and B in a given medium can be calculated for example by thermodynamic integration:

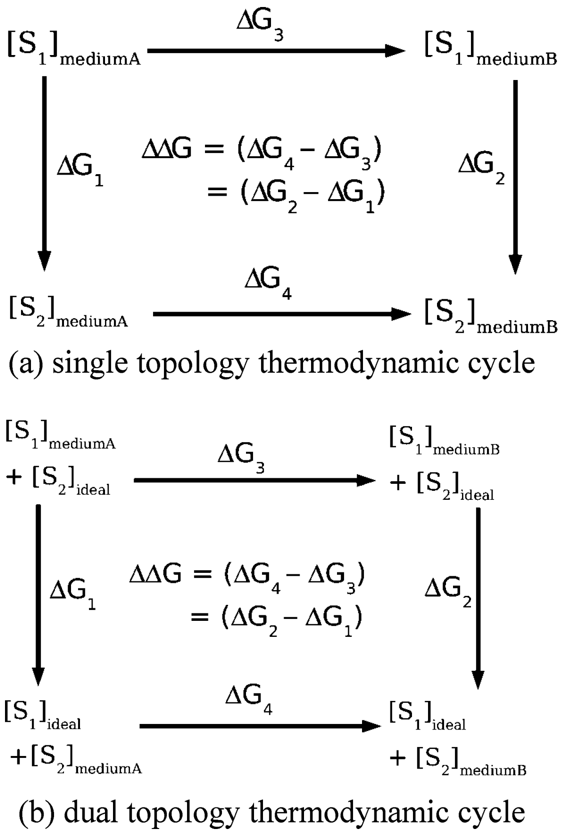
$$\Delta G_{\mathrm{medium,A}\rightarrow\mathrm{B}} = \int_0^1 \frac{\partial G(\lambda)}{\partial \lambda}\,\mathrm{d}\lambda = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \qquad (1)$$

λ is a coupling parameter that allows the smooth transformation of the potential energy function $U(\lambda=0)$, appropriate for molecule A, into a potential energy function appropriate for molecule B, $U(\lambda=1)$. The brackets in eq 1 denote an ensemble average corresponding to the derivative of the potential energy function $U(\lambda)$ with respect to λ (free energy gradients).

Most protein−ligand binding free energy studies have considered series of congeneric inhibitors of a protein. Two main reasons dictated these choices. First, the free energy perturbation or thermodynamic integration equations converge more easily for similar compounds. This difficulty can be circumvented by running longer simulations. As increasing computer power becomes more and more affordable, this solution becomes increasingly feasible. Second, a practical scheme for the smooth transformation of the potential energy function A $U(\lambda=0)$ into the potential energy function B

[*] Corresponding author e-mail: J.W.Essex@soton.ac.uk.
[†] School of Chemistry, University of Southampton.
[‡] Astex Therapeutics Ltd.

10.1021/ct700081t CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/04/2007


$[S_1]_{mediumA}$
ΔG_3
$[S_1]_{mediumB}$
ΔG_1
$\Delta\Delta G = (\Delta G_4 - \Delta G_3)$
$= (\Delta G_2 - \Delta G_1)$
ΔG_2
$[S_2]_{mediumA}$
ΔG_4
$[S_2]_{mediumB}$
$[S_1]_{mediumA} + [S_2]_{ideal}$
ΔG_3
$[S_1]_{mediumB} + [S_2]_{ideal}$
ΔG_1
$\Delta\Delta G = (\Delta G_4 - \Delta G_3)$
$= (\Delta G_2 - \Delta G_1)$
ΔG_2
$[S_1]_{ideal} + [S_2]_{mediumA}$
ΔG_4
$[S_1]_{ideal} + [S_2]_{mediumB}$

(a) single topology thermodynamic cycle

(b) dual topology thermodynamic cycle

Figure 1. Thermodynamic cycles that relate the difference in free energy between S_1 and S_2 in two media A and B. S_1 and S_2 could be two small molecules and media A and B, water and vacuum, in which case, the double free energy difference will correspond to the relative hydration free energy of S_2 with respect to S_1. If the media A and B represent a solvated protein and pure water, then the double free energy difference will correspond to the relative binding free energy of S_2 with respect to S_1. While the horizontal processes corresponding to ΔG_3 or ΔG_4 are often measured experimentally, the vertical processes corresponding to ΔG_1 or ΔG_2 are usually easier to calculate in a computer simulation. The first cycle implements the single-topology approach where S_1 is converted into S_2 by smooth variation of its force field parameters and geometry. The second cycle implements the dual-topology approach where the interaction energy of S_1 with its medium is gradually turned off while the interaction energy of S_2 is gradually turned on.

$U(\lambda{=}1)$ has to be proposed. This is often done by a single-topology paradigm shown in Figure 1a. In this approach, the free energy difference between two molecules A and B is calculated by interpolating the force field parameters of molecules A and B. This often requires the coupling of internal coordinate changes to the coupling parameter λ, and the presence of dummy atoms if the topologies of molecules A and B differ. This scheme makes it difficult to devise internal coordinate changes that would convert molecule A into a structurally different molecule B. As a result, many ligands that do not share a similar topology are not considered for a free energy simulation because they would require a complex system setup. This severely limits the applicability of the technique in a rational drug design context. A methodology that readily allows the consideration of sets of structurally diverse ligands is therefore desirable. In the dual-topology paradigm described in Figure 1b, the force field parameters and internal coordinates of molecules A and B are no longer coupled with λ. Rather, a free energy change occurs because the interaction energy of the molecules A and B are coupled to λ.[9] In this approach, compounds A and B can have arbitrary geometries as no scheme to convert the internal coordinates of molecules A into B is necessary, and it seems therefore that the problem of ligand diversity can be solved by this technique. However, practical applications of the dual-topology paradigm suffer from two main difficulties which have limited usage of this technique.

First, noisy free energy gradients can be recorded at the beginning or end of the perturbation ($\lambda = 0.0$ or 1.0). This problem is related to the functional form of the Lennard-Jones (LJ) nonbonded equation, which makes it difficult to turn off completely the LJ interaction at one site. This difficulty can be avoided by not running simulations at the end states and extrapolating the free energy gradients.[10] A more satisfactory solution is to use softened intermolecular energy functions that ensure that the LJ interaction can be turned off smoothly.[11,12] Second, at either end of the simulation, one compound is completely decoupled from its environment. As a result, it could drift away from its initial position, leading to serious convergence issues. A related problem occurs if the intramolecular energy of the ligand is turned off, as in this situation bonds between atoms are broken. If the bond energy is described by a harmonic term, this invariably leads to a divergence of the free energy change.[13] These problems can be avoided by enforcing a restraint that keeps the ligands in the binding site and by not turning off the intramolecular energy terms.[13−16]

Other more original approaches that allow the calculation of binding free energies for diverse compounds have been proposed. Schafer et al. have suggested the use of a nonphysical reference compound that is designed to maximize phase space overlap with a series of compounds of interest. A single simulation is then carried out with this reference compound, and numerous configurations are stored for analysis.[17,18] Each compound in the series can then be mapped onto the reference compound. The free energy difference is then calculated by traditional FEP. The approach is, in principle, very efficient, as a single simulation of a protein−ligand complex is required. However, it can be difficult to devise a reference compound that has a good phase space overlap with all the compounds to be studied, making the calculations potentially imprecise.[19] In addition, ligand flexibility has yet to be addressed.[20]

Another approach that has recently gained popularity involves the calculation of absolute rather than relative binding free energies by either the double decoupling method,[21] potential of mean force approaches,[22] or a combination of these.[23] Absolute binding free energy calculations can be demanding, as the complete annihilation of a ligand can require extensive simulation of several intermediate states to yield precise answers. In addition, it can be difficult to deal with the large structural changes that can occur in the binding site if the ligand is removed. For example, a large

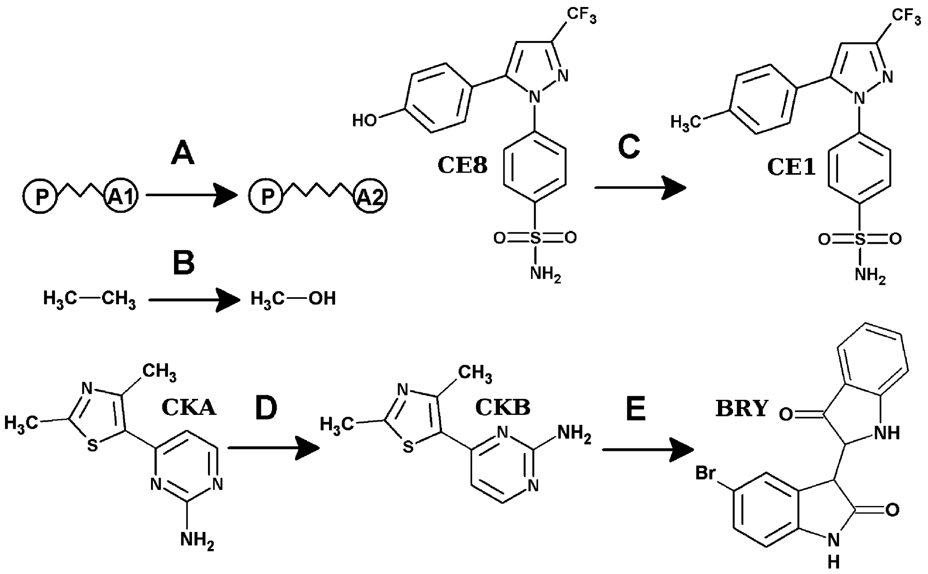


Figure 2. The systems considered in this study. System A is the perturbation of harmonic oscillator P-A1 into harmonic oscillator P-A2. System B is the perturbation of ethane into methanol in a box of TIP4P water. System C is the perturbation between CE8 and CE1, two congeneric inhibitors of the protein COX2. System D is the perturbation of one binding mode of an inhibitor of CDK2, denoted CKA, into an alternative binding mode to the same protein, denoted CKB. System E is the perturbation of the previous inhibitor in the binding mode CKB, into another CDK2 inhibitor, 5-bromoindirubin, and denoted BRY.

change in hydration pattern and conformation of the binding site of the protein OppA is observed between unliganded and liganded structures.[24,25] HIV protease is also known to undergo substantial conformational changes upon inhibitor binding.[26] Such large conformational changes are not sampled easily and rapidly by conventional Monte Carlo or molecular dynamics simulations, leading to potentially imprecise absolute binding free energies.

Methods that calculate relative binding free energies suffer less from these issues, as the binding site is always occupied by a ligand, and it can be expected that relative binding free energies will converge more rapidly than absolute binding free energies. In addition, if the ligands share some common structural features, they often benefit from cancellation of systematic errors in the force field parameters. Finally, the experimental binding affinity data are often available in the form of IC50s, and their conversion to an absolute free energy scale is not always straightforward, making direct comparison of the calculated absolute binding free energies more difficult than their relative counterparts.

In this article, we show that it is possible to efficiently compute the relative binding free energy of substantially different ligands. This is achieved by combining Monte Carlo sampling with a dual-topology approach and a constraint. The formulation is general and demonstrated to give results identical to a single-topology protocol. We then apply the methodology to classes of problems that are not easily handled by a single-topology approach. The range of perturbations covered in this study is shown in Figure 2.

Many existing dual-topology implementations used in the context of relative free energy calculations are in fact best described as hybrid single-/dual-topology methodologies because a portion of the ligand is invariant during the perturbation.[14,16,20] This is necessary to avoid the ligand drifting when it is fully decoupled from its environment, as is commonly experienced in absolute binding free energy calculations.[15] For such a scheme to be practical, the two ligands of interest must share common structural features and occupy the same regions of space. The present work builds on previous efforts from diverse groups[13−16,19−22,27] and strives to overcome these limitations to propose an efficient binding free energy calculation scheme generally applicable to ligand binding studies.

Methods

The free energy calculations were performed with the program ProtoMS2.1.[28] The replica exchange thermodynamic integration (RETI) method[29] was used to construct the free energy profiles, and the necessary ensemble of states were formed using Metropolis Monte Carlo sampling.[30] In the simulations of ethane and methanol, the solvent was represented by a periodic box of 533 TIP4P molecules.[31] In the other systems, the solvent was either modeled by a ball of TIP4P water of 22 Å radius centered on the ligands or by a generalized Born surface area (GBSA) model.[32] The GBSA model employed here is an implementation of the pairwise descreening approximation of Hawkins et al.[33] which has been parametrized to be used in conjunction with AM1/BCC atomic partial charges.[34] Because the generalized Born interaction energy term cannot be broken down into pairwise terms, it does not integrate well with typical Monte Carlo simulations. We have shown, however, that with the adoption of specialized Monte Carlo moves it is possible to sample rigorously the equilibrium distribution of a biomolecular system solvated by an accurate GBSA potential with a crude, more efficient potential and thus recover most of the efficiency loss with minimal approximations.[35]

Models and parameters for the solutes were obtained from a previous study[36] or were derived using the GAFF force field[37] and the AM1/BCC method[38] to obtain atomic partial charges. Models for the proteins were also taken from a previous study or created according to similar guidelines.[36] The PDB codes of the crystallographic structures used to construct the protein models were 1CX2 (COX2),[39] 1PXJ (CDK2 inactivated),[40] and 2C5O (CDK2 activated).[41]

The bond angles and torsions for the side chains of protein residues within 10 Å of any heavy atom of the ligand and all the bond angles and torsions of the ligand were sampled during the simulation, with the exception of rings. The bond lengths of the protein and ligand were constrained. A 10 Å residue-based cutoff feathered over the last 0.5 Å was employed in all simulations. In the generalized Born simulations, a cutoff of 20 Å for the calculation of the Born radii was applied.

For the explicit solvent simulations of the inhibitors in the bound state, solvent moves were attempted with a probability of 85.7%, protein side-chain moves with a probability of 12.8%, and solute moves with a probability of 1.4%. In the unbound state, solvent moves were attempted 98.4% of the time. In the simulation of ethane and methanol, solvent moves, solute moves, and volume moves were attempted 99%, 0.9%, and 0.1% of the time, respectively. The move probabilities were selected on the basis of the number of protein residues, solvent molecules, and according to previous studies.[35,36] All of the simulations were conducted at 25 °C.

For the perturbation of ethane into methanol, the free energy gradients were accumulated at 11 equally spaced values of the coupling parameter λ. The system was first well pre-equilibrated at one value of the coupling parameter, and each simulation was further equilibrated for 5 million (M) moves before collecting statistics for 25M moves. For the congeneric COX2 inhibitors, 12 values of the coupling parameter λ were employed (0.00, 0.10, ..., 0.90, 0.95, and 1.00) in the single-topology calculations to be consistent with a previous study,[36] and the equilibration and data collection phases consisted of 10M and 30M moves, respectively. For the implicit solvent simulations, these quantities were reduced to 100 000 (100K) and 1M moves, respectively. For the perturbations in CDK2, 21 equally spaced λ values were used; each window was equilibrated for 30 M moves, and data were collected for 50M moves (750K and 1.8M moves, respectively, in the implicit solvent simulations). For a given set of conditions, the free energy change was taken as the mean of five independent simulations and the error estimate as one standard error from the mean. Depending on the system and the simulation conditions, each simulation required 12−36 h on 11−21 2.2 GHz Opteron machines. The calculation of the free energy differences by the dual-topology method was achieved by coupling the two solutes of interest through eq 2:

$$U(\lambda) = U_0 + \lambda U(S_2) + (1-\lambda)U(S_1) \qquad (2)$$

where $U(S_x)$ represents the energy terms associated with the solute X being turned off or on, and U_0 represents the energy terms for the rest of the system. Note that only the intermolecular energy of the solutes is turned on or off. The intramolecular nonbonded energy and the bonded terms are not modified. A fully decoupled solute is thus transferred to a gas-phase environment. As a result, in the dual-topology method, only intermolecular energy terms contribute to the free energy gradients. By contrast, in the single-topology method, intramolecular as well as intermolecular terms typically contribute to the free energy gradients. To avoid numerical instabilities when one solute is fully decoupled from its surrounding medium, a separation-shifted soft-core functional form (eq 3) for the solute i nonbonded energy was implemented:[12]

$$U_{\text{nonbonded},\lambda} = (1-\lambda)4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}{}^{12}}{(\lambda\delta\sigma_{ij}+r_{ij}{}^2)^6}\right) - \left(\frac{\sigma_{ij}{}^{6}}{(\lambda\delta\sigma_{ij}+r_{ij}{}^2)^3}\right)\right] + \frac{(1-\lambda)^n q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda + r_{ij}{}^2)}} \qquad (3)$$

where the parameters n and δ were introduced to control the softness of the Coulombic and Lennard-Jones interactions, respectively. These parameters were adjusted for individual perturbations to ensure optimum softening. For the implicit solvent simulations, the generalized Born energy was scaled by the same exponent n, while linear scaling was used for the surface area term. The internal degrees of freedom of each solute were sampled independently. However, to avoid solute drift at either end of the perturbation, the translations and rotations of the pair of ligands were coupled; that is, each individual Monte Carlo move translates and rotates both ligands by the same amount. Internal energy terms are not turned on or off, and a fully decoupled solute is thus in an ideal state. In this state, the free energy of the solute is invariant to rigid body translations and rotations, and the present constraint does not therefore contribute to the free energy change. Such a constraint also has the advantage of being easily implemented in a Monte Carlo simulation package.

In essence, the present dual-topology technique can be considered as a large single-topology perturbation where all the atoms of the first ligand are gradually converted from fully interacting atoms to dummy atoms, while all the atoms of the second ligand are gradually converted from dummy atoms to fully interacting atoms with the additional presence of one dummy bond between the center of geometry of the two ligands. Because the intramolecular interactions (bonded and nonbonded terms) are retained, this corresponds to exchanging a ligand from an ideal gas molecule state to a condensed phase and vice versa. This approach differs from absolute binding free energy schemes that use harmonic restraints or anchors to keep one ligand in place in the binding site[15,16,23] as the interactions of the fully interacting ligands at $\lambda = 0$ or $\lambda = 1$ with their surroundings are not biased by the present constraint. Where applicable, the simulations were also performed by a standard single-topology approach.

Results

Relative Free Energy of a Pair of Harmonic Oscillators. Before applying the above-described methodology in studies of biomolecular systems, it is important to verify that the simulation results are in agreement with other existing methods. In particular, we wish to establish whether or not the calculated relative free energies need to be corrected because of the nature of the constraint employed in the dual-topology scheme described above. Because the free energies calculated in biomolecular systems can often exhibit a significant statistical uncertainty which could hide systematic differences, we first investigated a simple model of protein−ligand binding for which a relative binding free energy can be determined analytically.

Consider the following system:

$$A_1,P + A_2,0 \leftrightarrow A_1,0 + A_2,P \tag{4}$$

A_1 and A_2 are single atoms. They interact with environment P through a harmonic potential of identical equilibrium length $r_0 = 2.0$ length units and different force constants K_1 and K_2 of values 20 and 10 kcal/mol/length unit, respectively. Thus, A_1 and A_2 "bind" differently to P. The symbol A_1,0 means that A_1 is a gas-phase-like environment where it does not experience any intermolecular potential. When classical statistical mechanics is used and for this strictly one-dimensional problem, it is possible to derive an analytical expression for the relative free energy of A_2 and A_1.[42]

$$A_2 - A_1 = \Delta A = \frac{kT}{2}\ln\left(\frac{K_2}{K_1}\right) \tag{5}$$

At a temperature T of 298 K, this quantity amounts to −0.2053 kcal/mol. This relative free energy was estimated by three different free energy calculation techniques:

- *A single-topology scheme* where the force constant of the harmonic oscillator A_1 is changed by linear combination into the force constant of harmonic oscillator A_2.
- *The present dual-topology scheme* where both atoms are simulated simultaneously, but their interaction energy with P is scaled. In addition, both atoms are constrained to occupy the same position at every Monte Carlo move.
- *A double-decoupling approach* where the absolute binding free energies of A_1 and A_2 are first determined by application of the double-decoupling methodology.[21] The difference in their absolute binding free energies should give the correct relative binding free energy. Unlike the double-annihilation method,[43] in the double-decoupling approach, the decoupled ligand is restrained to occupy a well-defined region of space. The advantages of the restraint are two-fold. First, it makes the calculation reversible by avoiding the convergence issues associated with a ligand drifting out of a binding site. Second, because the volume the decoupled ligand occupies is defined by the restraint, the calculated absolute binding free energy can be related to a standard absolute binding free energy. Here, the ligands were restrained by a hardwall potential. When this form of restraint is used and according to Gilson et al.,[21] the absolute binding free energy can be calculated as

$$\Delta A_{\mathrm{bind}}{}^{0} = \Delta A_{\mathrm{bind}}^{\mathrm{sim}} + kT\ln\frac{V_{\mathrm{hardwall}}}{V_{\mathrm{standard}}} \tag{6}$$

Table 1. Free Energy of Two Harmonic Oscillators[a]

experiment	forward	backward
single topology	−0.2055 ± 0.0003	0.2052 ± 0.0001
dual topology	−0.2055 ± 0.0001	0.2054 ± 0.0001
decoupling A_1[b]	−1.8387 ± 0.0000	1.8387 ± 0.0000
decoupling A_2[b]	−1.8076 ± 0.0000	1.8076 ± 0.0000
ΔΔA[b]	0.0311 ± 0.0000	−0.0311 ± 0.0000
decoupling A_1[c]	−1.5255 ± 0.0028	1.5252 ± 0.0003
decoupling A_2[c]	−1.3217 ± 0.0002	1.3220 ± 0.0002
ΔΔA[c]	0.2038 ± 0.0028	−0.2032 ± 0.0003
decoupling A_1[d]	−1.3268 ± 0.0454	1.5248 ± 0.0008
decoupling A_2[d]	−1.2377 ± 0.0118	1.3197 ± 0.0010
ΔΔA [d]	0.0892 ± 0.0469	−0.2051 ± 0.0013
decoupling A_1[e]	−1.0204 ± 0.1285	1.5205 ± 0.0013
decoupling A_2[e]	−0.9004 ± 0.0484	1.3153 ± 0.0018
ΔΔA[e]	0.1200 ± 0.1373	−0.2052 ± 0.0022

[a] The figures are in kilocalories per mole and are the average of five independent single-window FEP simulations of 10M moves. One standard error is plotted as an estimate of the precision. Forward is for the perturbation of A_1 into A_2. Backward is for the perturbation of A_2 into A_1. For the absolute binding free energy calculations, a standard volume of 2.0 length units was arbitrarily defined. The hardwall was centered at the equilibrium bond length value of each oscillator. [b] A hardwall of width 0.1 length unit was applied. [c] A hardwall of width 0.5 length unit was applied. [d] A hardwall of width 1.0 length unit was applied. [e] A hardwall of width 2.0 length unit was applied.

where $\Delta A_{\mathrm{bind}}^{\mathrm{sim}}$ is the free energy of turning off the interactions of the ligand with its environment in the presence of a hardwall and V_{hardwall} and V_{standard} are the hardwall and standard state volumes, respectively. For this simple, one-dimensional example, here, we arbitrarily define a standard "volume" of 2.0 length units. The simulation results are listed in Table 1. It can be seen that the single- and dual-topology schemes agree with the analytical solution. The backward runs are more precise than the forward run. This behavior was well interpreted by Kofke:[44] because A_1 has a larger force constant than A_2, the low-energy configurations of oscillator A_1 are a subset of the low-energy configurations of oscillator A_2, and sampling with the second oscillator gives a more thorough coverage of the configuration space.

In Table 1, the free energy changes for turning off one oscillator (decoupling A_1 or decoupling A_2) are also listed for different hardwalls. The difference between the free energy changes calculated for A_1 and A_2 and with the same hardwall should equal the relative binding free energy of both atoms. It can be seen that the hysteresis between the forward and backward free energy change is a function of the hardwall size. As expected, the backward runs are generally much more precise. From the absolute binding free energies obtained with the backward run, the relative free energy of A_1 and A_2 is not reproduced with the first hardwall, reproduced fairly with the second hardwall, and very well reproduced by the last two hardwalls. The behavior of the simulations is well understood in terms of the explanations put forward by Gilson and co-workers: the hardwall potential must not be too small so as to exclude the low-energy states of A1 or A2, but if it is too large, the simulations become less reversible and precise.[21]

This simple example demonstrates that the relative free energy calculated by a standard single-topology approach, the present dual-topology scheme, and a double-decoupling methodology (provided the hardwall is suitably chosen) all agree with the analytical result. There is therefore no

Table 2. Relative Free Energies Calculated for a Series of Molecular Systems[a]

experiment	solvent	coupling	soft core	free energy[b]
$\Delta\Delta G_{solv}$ ethane → methanol	TIP4P	single	NA	-5.95 ± 0.06[c]
$\Delta\Delta G_{solv}$ ethane → methanol	TIP4P	dual	$n = 0, \delta = 1.0$	-6.00 ± 0.05
$\Delta\Delta G_{solv}$ ethane → methanol	TIP4P	dual	$n = 1, \delta = 1.0$	-6.10 ± 0.06
$\Delta\Delta G_{solv}$ ethane → methanol	TIP4P	dual	$n = 2, \delta = 1.0$	-6.19 ± 0.10
$\Delta\Delta G_{solv}$ ethane → methanol	TIP4P	dual	$n = 1, \delta = 2.0$	-6.10 ± 0.16
$\Delta\Delta G_{solv}$ CE8 → CE1	TIP4P	single	NA	4.54 ± 0.04[c]
$\Delta\Delta G_{solv}$ CE8 → CE1	TIP4P	dual	$n = 0, \delta = 1.25$	4.61 ± 0.38
$\Delta\Delta G_{bind}$ CE8 → CE1	TIP4P	single	NA	-2.99 ± 0.07
$\Delta\Delta G_{bind}$ CE8 → CE1	TIP4P	dual	$n = 0, \delta = 1.25$	-2.72 ± 0.58
$\Delta\Delta G_{solv}$ CE8 → CE1	GBSA	single	NA	6.56 ± 0.01[c]
$\Delta\Delta G_{solv}$ CE8 → CE1	GBSA	dual	NA	6.53 ± 0.03
$\Delta\Delta G_{bind}$ CE8 → CE1	GBSA	single	NA	-2.52 ± 0.03
$\Delta\Delta G_{solv}$ CE8 → CE1	GBSA	dual	$n = 0, \delta = 1.25$	-2.82 ± 0.29
ΔG_{prot} CKA → CKB, inactivated	TIP4P	dual	$n = 0, \delta = 1.50$	7.97 ± 0.49
ΔG_{prot} CKA → CKB, inactivated	GBSA	dual	$n = 0, \delta = 1.50$	4.71 ± 0.29
ΔG_{prot} CKA → CKB, activated	TIP4P	dual	$n = 0, \delta = 1.50$	-5.80 ± 0.41
ΔG_{prot} CKA → CKB, activated	GBSA	dual	$n = 0, \delta = 1.50$	-4.56 ± 0.11
$\Delta\Delta G_{solv}$ CKB → BRY, activated	TIP4P	dual	$n = 0, \delta = 1.50$	3.07 ± 0.30
$\Delta\Delta G_{solv}$ CKB → BRY	GBSA	dual	NA	7.14 ± 0.01
$\Delta\Delta G_{bind}$ CKB → BRY	TIP4P	dual	$n = 0, \delta = 1.50$	-0.48 ± 0.50
$\Delta\Delta G_{bind}$ CKB → BRY	GBSA	dual	$n = 0, \delta = 1.50$	-5.62 ± 0.09

[a] The figures are in kilocalories per mole. $\Delta\Delta G_{solv}$ is a solvation free energy; $\Delta\Delta G_{bind}$ is a binding free energy, and ΔG_{prot} is the free energy difference between two ligands bound to a protein. [b] The error estimate is taken as one standard error from five independent simulations. [c] To obtain a relative solvation free energy with the single-topology method, the perturbation must also be carried out in vacuum. For the perturbation of ethane to methanol and CE8 to CE1, the free energy changes in the gas phase were respectively 2.69 ± 0.01 and 13.54 ± 0.01 kcal mol^{-1}

evidence from this system, or from the other systems that will be discussed later, that the present dual-topology scheme, in a Monte Carlo framework, requires a correction term arising from the applied constraints.

Relative Solvation Free Energy of Ethane and Methanol. The calculated relative solvation free energies of ethane and methanol for a series of single- and dual-topology calculations are listed in Table 2. It is apparent that, within statistical sampling error, the single- and dual-topology coupling schemes give the same free energy change. This suggests that the implementation of the dual-topology scheme is correct. It can also be seen that the parameters of the soft core influence the precision of the calculations. With δ set to 1.0, increasing n increases the spread of the individual simulation results. This can be understood by inspecting Figure 3a. The solute is perturbed from an apolar (ethane) to a polar (methanol) molecule, and as λ increases, it experiences stronger Coulombic interactions with the solvent. Increasing the parameter n results in the solute−solvent Coulombic interactions being restored later in the perturbation. This causes the free energy gradients to vary more rapidly in the second half of the perturbation. Because the free energy change is estimated by trapezium integration, smooth variations of the free energy gradients with λ should give more precise free energies, than profiles that change more rapidly. Also, as seen in Figure 3b and for this system at least, rapid variations of the free energy gradients are associated with larger error estimates and hence greater imprecision. When δ is increased from 1.0 to 2.0 with a constant value of n, the free energy gradients vary fairly smoothly, but the standard error on the estimate of the same free energy gradients obtained from five independent simulations fluctuates more, leading to increased imprecision. If δ is set to too high a value, then the solute−solvent Lennard-Jones energy is softened "too much" and the volume of space the solute can occupy varies more during the simulation, effectively making the perturbation more difficult. On the other hand, we have run a single simulation with δ set to 0.25 and n set to 1. The free energy change was -7.63 ± 0.57 kcal mol^{-1}, a figure that does not agree within error bounds with the previous free energy changes. Inspection of the free energy gradients reveals that most of the imprecision arises from widely fluctuating free energy gradients at the end of the perturbation. The difficulty here is that the soft core is not soft enough to allow a smooth decoupling of the molecule of ethane from the solvent, a problem typically observed in dual-topology simulations without a soft core.[12] Note that no such difficulty is observed at the beginning of the perturbation because the molecule of methanol fits inside the volume of ethane.

The present observations suggest that the soft-core parameters can be tuned to increase the precision of the calculated free energy change. In this process, analysis of the smoothness of the free energy gradient profiles and their statistical errors can be of valuable assistance.

Relative Binding Free Energy of Congeneric Inhibitors. The relative solvation free energy of two congeneric inhibitors of COX2[39,45] calculated by single- and dual-topology approaches is reported in Table 2. While the two methods give similar answers, it is clear that the single-topology calculations are much more precise. The same trend is observed for the relative binding free energy calculations: similar answers are obtained, but the single-topology calculations are 8−10 times more precise.

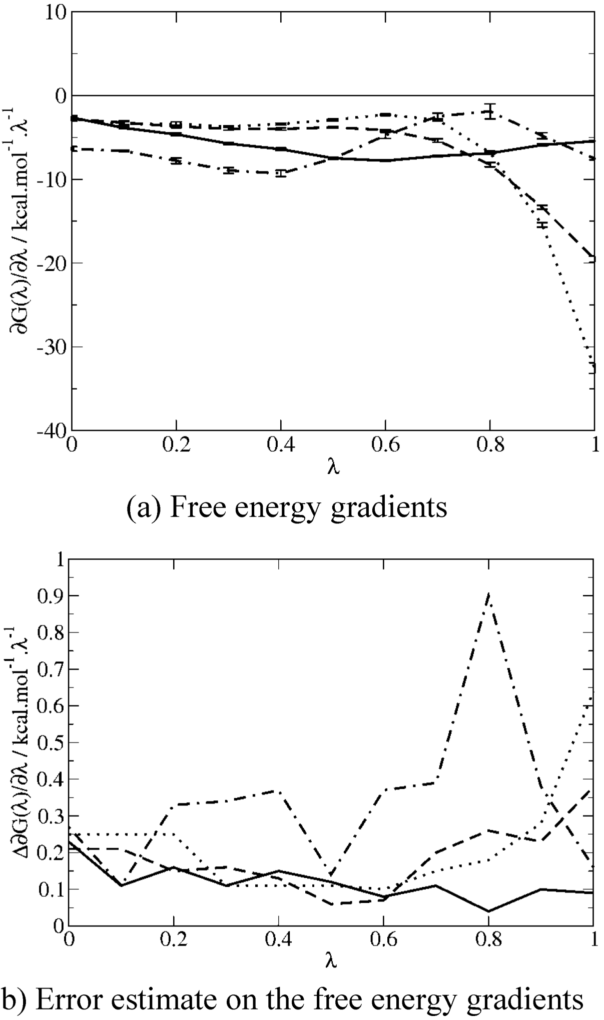


Figure 3. Ethane to methanol. (a) The free energy gradients for the perturbations carried out with various soft-core parameter sets. For the solid line, $n = 0$ and $\delta = 1.0$; for the dashed line, $n = 1$ and $\delta = 1.0$; for the dotted line, $n = 2$ and $\delta = 1.0$; for the dashed-dotted line, $n = 1$ and $\delta = 2.0$. (b) The standard error of the free energy gradients for the different parameter sets of the soft core.

The percentage of RETI moves that are accepted at each value of the coupling parameter λ is plotted in Figure 4. Because, in a RETI move, swaps of the configurations generated at neighboring values of λ are periodically attempted, the acceptance rate measures how much the equilibrium distributions of the neighboring replicas differ. As the free energy difference will converge more readily if the replicas are similar, a plot of this acceptance rate should give indications as to the difficulty in obtaining precise free energy differences along a given pathway. In the single-topology simulations, the replicas exchange more readily. In addition, the exchange rate is lower for the perturbation in the unbound state. This analysis corroborates the larger standard deviation of the dual-topology results. There remains the need to establish why the dual-topology simulations are much less precise. A plausible explanation can be put forward. In the single-topology simulation, the perturbation (mutation of a methyl group into a hydroxyl group) is well-localized on the scaffold. Other atoms in the molecule make a minor contribution to the free energy change (by small variations of their atomic partial charges). In the dual-topology simulations, there are twice as many internal degrees of freedom to sample independently (those of each solute molecule) and the intermolecular interactions of every atom contribute directly to the free energy gradients. Thus, the dual-topology simulations should be intrinsically more difficult to carry out with sufficient precision. In Figure 5, the free energy gradients recorded by the single- or dual-topology technique at one value of the coupling parameter λ are plotted. It is evident that the free energy gradients recorded with the dual-topology technique fluctuate much more than those recorded with the single-topology technique. The data extracted from the simulation thus support the present explanation.

In Figure 5, it is apparent that the free energy gradients recorded with the single- and dual-topology techniques differ. This is because both techniques convert one ligand into

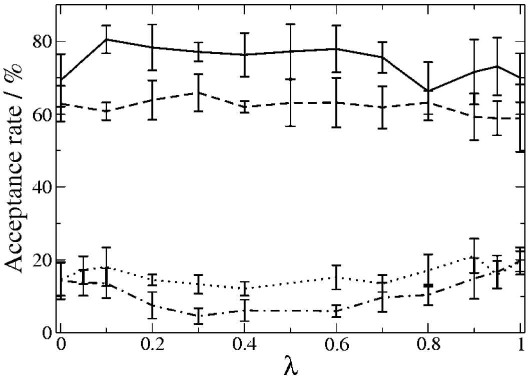


Figure 4. The acceptance rate of the RETI moves as a function of the coupling parameter λ for two congeneric inhibitors of COX2. The solid line is for the single-topology simulation of the bound state. The dashed line is for the single-topology simulation of the unbound state. The dotted line is for the dual-topology simulation of the bound state. The dashed-dotted line is for the dual-topology simulation of the unbound state. Each point is the average of five independent simulations, and the error bars show one standard deviation.

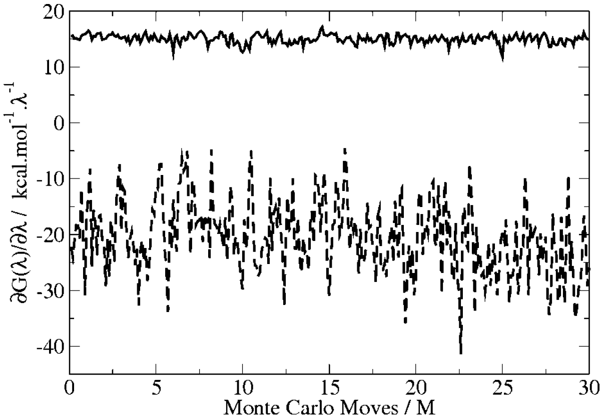


Figure 5. Fluctuations in the free energy gradients recorded during a Monte Carlo simulation. The data was extracted from an explicit solvent simulation perturbing the two congeneric inhibitors of COX2 bound to the protein at $\lambda = 0.50$. The solid line shows the gradients recorded with the single-topology protocol. The dashed line shows the gradient recorded with the dual-topology protocol.

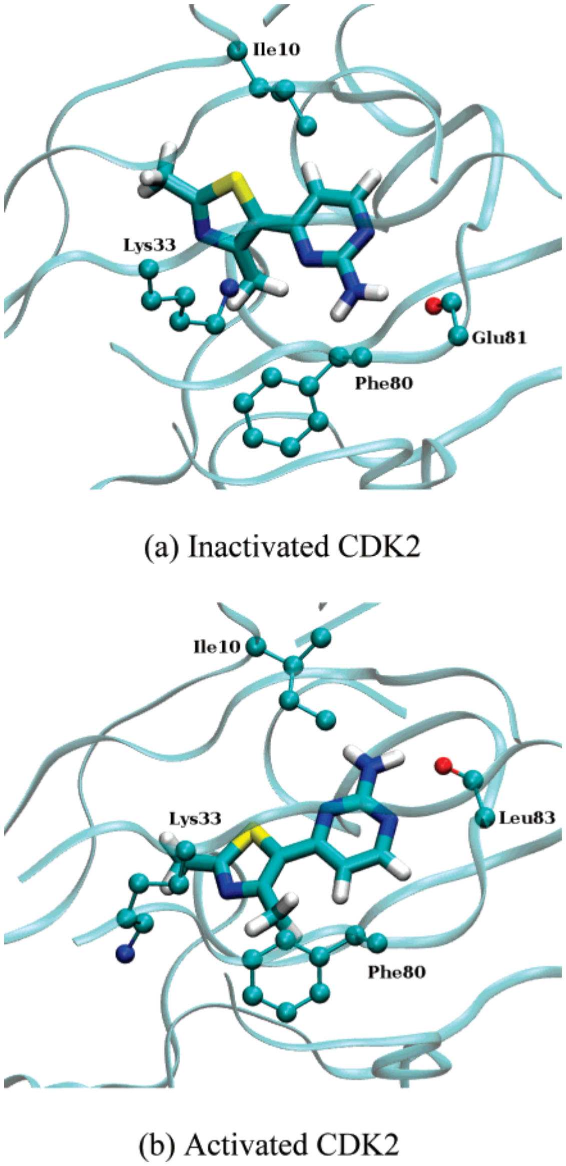
another by a different pathway, and in general, one does not expect similar gradients. The thermodynamic cycles shown in Figure 1 require only that the relative binding or solvation free energies be identical. Table 2 shows that, to within statistical error, this is indeed the case. Finally, in the previous system, ethane to methanol, no large difference in precision was observed between single- and dual-topology calculations. This is presumably because the numbers of atoms and internal degrees of freedom were much smaller.

Overall, these observations are in line with the findings of Pearlman that compared dual- and single-topology methods in a simple "ethane to ethane" perturbation[46] as well as comments from Shobana et al. who reported a hybrid single-/dual-topology technique.[27]

We have recently reported that binding free energy calculations can be carried out in an implicit solvent with good accuracy.[36] An implicit treatment of the solvent has the advantage of removing any sampling difficulties for the simulation in the unbound state, and the relative solvation free energies by the single- and dual-topology approaches are found once again to yield identical results to within a very narrow error interval (Table 2). There is still, however, a substantial difference in the precision of the relative binding free energy between the single- and dual-topology methods. In this system, the binding site is shielded from the solvent, and there is little precision to be gained for dual topology by adopting an implicit solvent model.

Binding Mode of a CDK2 Inhibitor in an Inactivated and Activated Complex. Cyclin-dependent kinase 2 (CDK2) plays an important role in the control of the cell cycle, is believed to be an important target for the development of cancer treatments, and is the focus of intense effort in drug development. CDK2 exists in an inactivated form, but the binding of cyclins A or E and subsequent phosphorylation of Thr160 causes significant conformational changes which greatly increase its phosphorylation activity.[47] Recently, Kontopidis et al. reported that a number of CDK2 inhibitors adopt different binding modes when complexed to an inactivated or activated CDK2.[41] Figure 6 shows the binding mode of 4-(2,4-dimethyl-1,3-thiazol-5-yl)pyrimidin-2-amine in inactivated and activated CDK2.

In inactive CDK2, the amino group of the ligand forms a hydrogen bond with the backbone carbonyl of Glu81. In the active complex, the amino group interacts with the backbone carbonyl of Leu83. This is achieved by a 180° flip of the pyrimidine ring. In addition, the thiazole ring moves substantially. Previous studies suggest that, in the inactivated form, favorable protein−ligand interactions are formed with Ile10 and Lys33.[40] In the activated form, the thiazole ring packs favorably against Phe80. The binding mode in inactivated and activated CDK2 will be referred to in the text as CKA and CKB, respectively. The flip of the pyrimidine ring, coupled with the translation and rotation of the thiazole ring, would be difficult to simulate with a single-topology method.



(a) Inactivated CDK2

(b) Activated CDK2

Figure 6. The binding mode of 4-(2,4-dimethyl-1,3-thiazol-5-yl)pyrimidin-2-amine in inactivated (top panel, PDB code 1PXJ) and activated (bottom panel, PDB code 2C5O) CDK2. The backbone carbonyl group of Glu81 and Leu83 is shown in CPK representation in the inactivated and activated CDK2, respectively. The side chains of Ile10, Lys33, and Phe80 are shown in CPK representation. The backbone of CDK2 is shown in ribbon representation. The ligand is shown in licorice representation. Hydrogen atoms on the protein are not shown, for clarity.

By contrast, setting up a dual-topology simulation is no more difficult than in the previous example. Here, we investigate with what accuracy the relative stability of each binding mode in both CDK2 forms can be predicted, and whether or not the results support the crystallographic evidence. The predicted relative free energies are listed in Table 2. The calculations were performed with explicit and implicit models of waters. Both types of calculation correctly identify the crystallographic binding mode as more stable, that is, that the CKA binding mode is favored in inactivated CDK2, while the CKB binding mode is favored in activated CDK2. The implicit solvent simulations are more precise than the explicit solvent simulations. This is because the water molecules in the vicinity of the ligand have to reorganize substantially to accommodate the change of binding mode. The difficulty of the explicit solvent simula-

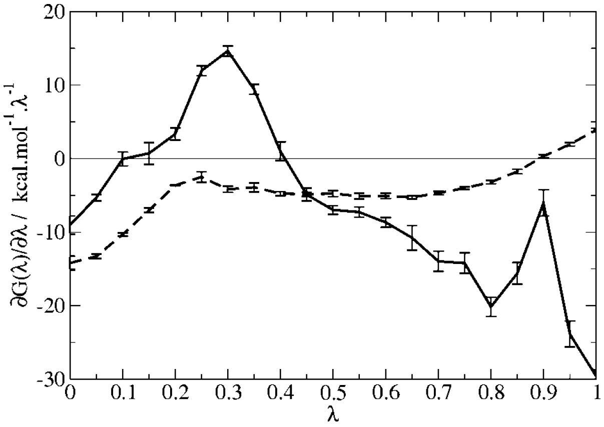


Figure 7. The free energy gradients recorded in the perturbation of CKA into CKB, bound to activated CDK2 (PDB code 2C5O). The solid line is for the explicit solvent simulation, and the dashed line is for the implicit solvent simulation. Error bars are shown for the free energy gradients obtained at each value of the coupling parameter λ.

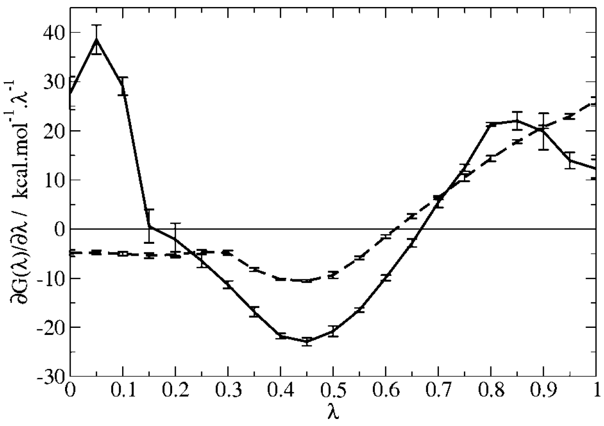


Figure 8. The free energy gradients recorded in the perturbation of CKB into BRY, bound to activated CDK2 (PDB code 2C5O). The solid line is for the explicit solvent simulation, and the dashed line is for the implicit solvent simulation. Error bars are shown for the free energy gradients obtained at each value of the coupling parameter λ.

tions is apparent when inspecting a plot of the free energy gradients in Figure 7. The implicit solvent simulations exhibit a smoother plot, with lower error bars on the individual free energy gradients. In addition, the acceptance rate of the RETI moves is on average twice as large as in the explicit solvent simulations (data not shown). However, the actual ΔG values differ, particularly in the inactivated CDK2. As no experimental measure of the relative stability of the two binding modes is available, it is not possible to conclude which protocol gives the more accurate answer. Certainly, one would not expect the behavior of the water molecules in the binding site to be well approximated by the GBSA model. However, the dual-topology method with both explicit and implicit solvation is able to assign the correct binding mode of the inhibitor to each kinase structure. This should be a simple test of the methodology as each protein structure is preorganized to stabilize a single binding mode. It is nevertheless important and comforting that the calculations are able to reproduce this trend clearly.

Relative Binding Free Energy of Two Different Scaffolds. The main objective of this research is to propose a method to calculate the relative binding free energy of structurally diverse molecules. To demonstrate this, we selected the inhibitor CKB described in the previous paragraph and attempted to perturb it into the CDK2 inhibitor 5-bromoindirubin (BRY) shown in Figure 2.[48] This system was selected because the two inhibitors share no common structural features and yet occupy the same position in the binding site. Relative solvation and binding free energies were calculated with implicit and explicit solvent models. The free energy gradients recorded for the perturbation in the bound state are shown in Figure 8. Observations similar to the previous system can be made. The free energy gradients in the explicit solvent simulations fluctuate more readily, and precise free energy estimates are harder to obtain than with the implicit solvent simulations. This is likely to be because transformation of CKB into the larger inhibitor BRY requires extensive solvent reorganization of the partially solvated binding pocket. Such a difficulty is of course not observed in the implicit solvent simulations. However, the binding free energies listed in Table 2 for each solvation model differ qualitatively. In the explicit solvent simulations, BRY and CKB are deemed inhibitors of similar potency to within the precision of the calculation, while BRY is predicted to be more potent by over 5 kcal mol^{-1} in the implicit solvent simulations. Such a large discrepancy might be related to the large difference in predicted relative solvation free energy. In implicit water, BRY is predicted 4 kcal mol^{-1} less stable than in explicit water; this would make the binding of BRY (which involves partial desolvation) more favorable. The reported IC50s for BRY and CKB are 1 μM and 6.5 μM, respectively, suggesting that BRY is a moderately more potent inhibitor.[48,49] However, the assay conditions differ, and not all the necessary parameters to permit conversion of the IC50s into inhibition constants were reported. Thus, rather than focusing on the accuracy of the results (which depends on the quality of the force field), we wish to emphasize the reproducibility of the calculations (which depends on the nature of the coupling scheme and the extent of sampling).

Discussion

In drug design, the identification of a promising scaffold (“hit”) is often judged more difficult than the structure-based optimization of a given scaffold into a high-affinity compound (“lead”). Because of the computational expense required to obtain precise predictions and limitations in standard protocols, successful applications of free energy calculations have often been limited to lead optimization scenarios. The present methodology allows in principle for free energy differences between compounds of arbitrary shape to be calculated. This feature was highlighted by performing two different perturbations of CDK2 inhibitors that would have been difficult to set up in a single-topology paradigm. Because of ever increasing computational resources, sufficient precision in the predicted free energy changes to allow practical applications is now obtained at an affordable cost. Obvious applications include the investigation of the relative stability of different ligand binding modes predicted by docking programs, and significantly expanding the scope of free energy calculations in drug design by allowing consideration of structurally diverse compounds.

The setup of the perturbation is easier than in a single-topology approach, because it is no longer necessary to devise a scheme to convert gradually the topology of one ligand into the other. This makes it easier to set up rapidly a large number of compounds. The automation of such a task will be important to allow free energy calculations in a virtual screening context. However, before such an ambitious goal will be accomplished, a number of methodological challenges will have to be solved. The more the ligands differ in structure, the more likely it is that induced fit effects differ. Reliable free energy difference will only be obtained from a simulation if the sampling algorithms can relax the protein binding site sufficiently. In addition, structurally different ligands are more likely to have different low-energy conformers, when bound to the protein and in the aqueous phase. An accurate prediction of the free energy change will require thorough sampling of the energy minima of these conformers. This difficulty can be lessened by adopting an implicit solvent model, but with likely increased inaccuracy of the force field. Clearly, the force fields will have to accurately reproduce the energy difference between these minima to yield meaningful free energy changes. It has been suggested that this would be a major difficulty in absolute binding free energy calculations.[50] For flexible ligands, the complete sampling of their numerous energy minima will, in addition, require improved sampling algorithms. At the present, it is doubtful that Monte Carlo (or molecular dynamics) sampling methods can accomplish this generally. In these contexts, the proposed methodology could be used to test advanced sampling methods and calibrate force fields. Despite these limitations, a viable strategy for a free energy calculation drug design technology emerges from the results presented in this paper. When the present methodology is used, it is possible to calculate with reasonable precision the free energy change of small inhibitors of diverse structure. It could be therefore used to screen a number of low-molecular-weight scaffolds (e.g., heterocyclic rings). As these compounds are likely to have only a few rotatable bonds, existing sampling techniques are more likely to sample sufficiently the low-energy rotamers of the ligands in the unbound and bound states and yield precise, converged, results. In addition, if the screened compounds bind in the same part of the binding site, differences in induced fit effects should be lessened. Because this approach is compatible with popular developments in fragment-based screening technologies, it could be used to assist in the setup of fragment libraries. Once promising heterocycles are identified, substituent optimization should be accomplished by a single-topology method as more precise free energy estimates can be obtained for the same amount of computational resources. Alternative coupling schemes that combine dual- and single-topology features could also be envisaged, but such schemes would probably involve a complex system setup which would restrict their applicability.

Conclusion

We have described a methodology that allows the calculation of free energy differences between molecules of arbitrary shape and position. The methodology makes use of a dual-topology coupling scheme, a soft-core nonbonded energy function, and a constraint on the translation/rotation of the pair of solutes that can be easily implemented in a Monte Carlo simulation. Results identical to single-topology, double-decoupling, and analytical approaches are obtained for the relative free energy of a pair of harmonic oscillators. The method is as precise as a standard single-topology approach on the simple calculation of the relative solvation free energy of ethane and methanol. It proves less precise when applied to the calculation of the relative binding free energy of two inhibitors of COX2, although the two methods give identical results to within statistical error. However, as illustrated by two examples involving inhibitors of CDK2, the dual-topology method is readily applied to classes of problems that are beyond the reach of single-topology approaches. Precision can also be improved by adopting an implicit solvent approach, albeit at the expense of some accuracy. The computational expense is similar to standard single-topology approaches. This study highlights the strengths and weaknesses of single- and dual-topology methods for the calculation of relative free energies and suggests when one approach should be considered over the other. The present methodology demonstrates that relative binding free energy calculations between structurally diverse ligands can be computed with good precision and a reasonable amount of computational expense. As such, it should prove attractive to calculate relative binding free energies between sets of ligands that would have been previously only considered feasible by more challenging absolute binding free energy calculation methodologies. It is hoped that the present method will extend the scope of free energy simulations and find useful applications in drug design studies.

Acknowledgment. We thank the University of Southampton for funding this work and EPSRC (GR/R06137/01) for providing computational resources.

References

(1) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.

(2) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050−3054.

(3) Miyamoto, S.; Kollman, P. A. *Proteins: Struct. Funct. Bioinf.* **1993**, *16*, 226−245.

(4) Essex, J. W.; Severance, D. L.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **1997**, *101*, 9663−9669.

(5) Fox, T.; Scanlan, T. S.; Kollman, P. A. *J. Am. Chem. Soc.* **1997**, *119*, 11571−11577.

(6) Price, M. L. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 9455−9466.

(7) Udier-Blagovic, M.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2004**, *47*, 2389−2392.

(8) Jorgensen, W. L. *Science* **2004**, *303*, 1813−1818.

(9) Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. *Science* **1989**, *244*, 1069−1072.

(10) Simonson, T. Chapter 9. In *Computational Biochemistry and Biophysics*, 1st ed.; Marcel Dekker: New York, 2001; p 169.

(11) Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529−539.

(12) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025−9031.

(13) Boresch, S.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 103−118.

(14) Boresch, S.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 119−136.

(15) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.

(16) Roux, B.; Nina, M.; Pomès, R.; Smith, J. C. *Biophys. J.* **1996**, *71*, 670−681.

(17) Schafer, H.; van Gunsteren, W. F.; Mark, A. E. *J. Comput. Chem.* **1999**, *20*, 1604−1617.

(18) Pitera, J. W.; van Gunsteren, W. F. *J. Phys. Chem. B* **2001**, *105*, 11264−11274.

(19) Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem.* **2003**, *24*, 1730−1739.

(20) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750−6754.

(21) Gilson, M. K.; Given, J. A.; Bush, B.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047−1069.

(22) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825−6830.

(23) Deng, Y. Q.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255−1273.

(24) Sleigh, S. H.; Tame, J. R. H.; Dodson, E. J.; Wilkinson, A. J. *Biochemistry* **1997**, *36*, 9747−9758.

(25) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.

(26) Hornak, V.; Okur, A.; Rizzo, R.; Simmerling, C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 915−920.

(27) Shobana, S.; Roux, B.; Andersen, O. S. *J. Phys. Chem. B* **2000**, *104*, 5179−5190.

(28) Woods, C. J.; Michel, J. *ProtoMS2.1*; in-house Monte Carlo software, 2006.

(29) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703−13710.

(30) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

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

(32) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.

(33) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(34) Michel, J.; Taylor, R. D.; Essex, J. W. *J. Comput. Chem.* **2004**, *25*, 1760−1770.

(35) Michel, J.; Taylor, R. D.; Essex, J. W. *J. Chem. Theory Comput.* **2006**, *2*, 732−739.

(36) Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Med. Chem.* **2006**, *49*, 7427−7439.

(37) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(38) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132−146.

(39) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. *Nature* **1996**, *384*, 644−648.

(40) Wang, S. D.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; McNae, I.; Wu, S. Y.; McInnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. *J. Med. Chem.* **2004**, *47*, 1662−1675.

(41) Kontopidis, G.; McInnes, C.; Pandalaneni, S.; McNae, L.; Gibson, D.; Mezna, M.; Thomas, M.; Wood, G.; Wang, S.; Walkinshaw, M.; Fischer, P. *Chem. Biol.* **2006**, *13*, 201−211.

(42) McQuarrie, D. A. *Statistical Mechanics*, 1st ed.; Harper and Row: New York, 1976.

(43) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. *J. Chem. Phys.* **1988**, *89*, 3742−3746.

(44) Kofke, D. A. *Mol. Phys.* **2004**, *102*, 405−420.

(45) Marnett, L. J.; Kalgutkar, A. S. *Curr. Opin. Chem. Biol.* **1998**, *2*, 482−490.

(46) Pearlman, D. A. *J. Phys. Chem.* **1994**, *98*, 1487−1493.

(47) Collins, I.; Garrett, M. D. *Curr. Opin. Pharmacol.* **2005**, *5*, 366−373.

(48) Jautelat, R.; Brumby, T.; Schafer, M.; Briem, H.; Eisenbrand, G.; Schwahn, S.; Kruger, M.; Lucking, U.; Prien, O.; Siemeister, G. *ChemBioChem* **2005**, *6*, 531−540.

(49) Wu, S. Y.; McNae, I.; Kontopidis, G.; McClue, S. J.; McInnes, C.; Stewart, K. J.; Wang, S. D.; Zheleva, D. I.; Marriage, H.; Lane, D. P.; Taylor, P.; Fischer, P. M.; Walkinshaw, M. D. *Structure* **2003**, *11*, 399−410.

(50) Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2006**, *49*, 5880−5884.

CT700081T

JCTC Journal of Chemical Theory and Computation

Density Functional and Semiempirical Molecular Orbital Methods Including Dispersion Corrections for the Accurate Description of Noncovalent Interactions Involving Sulfur-Containing Molecules

Claudio A. Morgado, Jonathan P. McNamara, Ian H. Hillier,* Neil A. Burton, and Mark A. Vincent

School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

Received March 22, 2007

Abstract: We describe the use of density functional theory (DFT-D) and semiempirical (AM1-D and PM3-D) methods having an added empirical dispersion correction, to treat noncovalent interactions between molecules involving sulfur atoms. The DFT-D method, with the BLYP and B3LYP functionals, was judged against a small-molecule database involving sulfur-π, S−H···S, and C−H···S interactions for which high-level MP2 or CCSD(T) estimates of the structures and binding or interaction energies are available. This database was also used to develop appropriate AM1-D and PM3-D parameters for sulfur. The DFT-D, AM1-D, and PM3-D methods were further assessed by calculating the structures and binding energies for a set of eight sulfur-containing base pairs, for which high-level ab initio data are available. The mean absolute deviations (MAD) for both sets of structures shown by the DFT-D methods are 0.04 Å for the intermolecular distances and less than 0.7 kcal mol^{-1} for the binding and interaction energies. The corresponding values are 0.3 Å and 1.5 kcal mol^{-1} for the semiempirical methods. For the complexes studied, the dispersion contributions to the overall binding and interaction energies are shown to be important, particularly for the complexes involving sulfur-π interactions.

Introduction

It is now recognized that noncovalent interactions involving aromatic side chains play an important role in determining protein dynamics during folding and also protein−ligand recognition. A variety of interactions involving π-systems has been studied including $\pi-\pi$, cation-π, alkyl-π, amino-π, oxygen-π, and sulfur-π. Of these interactions, sulfur-π contacts have been the subject of only limited theoretical study in spite of the recognition of their importance in biological systems.[1−7] In the late 1970s frequent and close contacts between side chains of sulfur-containing amino acids (Met and Cys) and aromatic amino acids (Tyr, Trp, and Phe) were first recognized in crystal structures of globular proteins. Morgan et al.[8] identified proteins that contained one or more chains of alternating “sulfur and π-bonded atoms” and established a minimum distance of 5 Å for S···C(sp^2) van der Waals contacts. Sulfur-π interactions were found to occur more frequently than originally thought as exemplified by database searches carried out by Morgan et al.[9] and Reid et al.[10] In the statistical analysis of PDB data performed by Reid et al. a preference for the placement of divalent sulfur at the edge and slightly above the plane of aromatic rings was found. More recent structural analyses have considered the interaction of Met, Cys, and disulfide bridges separately with aromatic residues, and this work showed that although only a small number of examples of Met···aromatic interactions were found, Cys residues had a much stronger preference for facial contacts with aromatic residues.[11] A statistical analysis[12] of crystal structures[13] also provided evidence for interactions between disulfide units and aromatic residues.

* Corresponding author phone: +44 (0)161 275 4686; fax: +44 (0)161 275 4734; e-mail: Ian.Hillier@manchester.ac.uk.

10.1021/ct700072a CCC: $37.00
Published on Web 06/23/2007

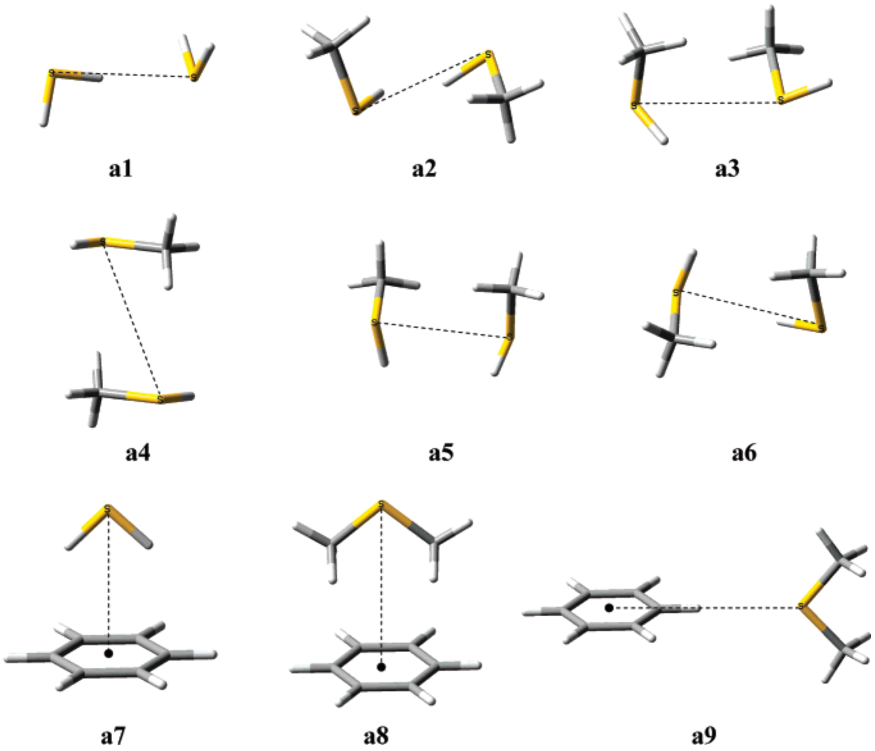


Figure 1. Small-molecule database structures showing interaction distances (dotted lines). Complexes **a1**: hydrogen sulfide dimer (C_s); **a2**−**a6**: methanethiol dimers (C_1, except for **a4** which possesses C_i symmetry); **a7** benzene with hydrogen sulfide (C_{2v}); **a8**, **a9**: benzene with dimethylsulfide (both C_{2v}). For **a7** and **a8** the intermolecular distance corresponds to the distance between the sulfur atom and the center of mass of the benzene ring.

On the experimental side, Viguera and Serrano[14] investigated the contribution of sulfur-π interactions to the stability of α-helices employing spectroscopic techniques along with the AGADIR[15] algorithm. Phe···Cys and Phe···Met interactions were found to provide −2.0 kcal mol^{-1} and −0.65 kcal mol^{-1}, respectively, to helix stability. Tatko and Waters[16,17] studied Phe···Met and Trp···Met interactions and found that they were similar in magnitude to Phe···Lys or Trp···Lys interactions, with a value of −0.30 kcal mol^{-1}, and also found that the interaction of the Met side chain with the face of the aromatic residue was quite modest.

In view of the importance of π-stacking interactions in biological and other systems,[18] there have been an increasing number of studies to calculate these interactions which are often dominated by dispersive contributions. MP2 calculations with quite large basis sets are generally considered to be the minimum acceptable level for an ab initio study, although a higher level of electron correlation such as CCSD(T) is to be preferred.[18] In this respect, a number of studies at the post-Hartree−Fock level of the interaction of benzene with small sulfur containing molecules such as methanethiol,[3,4] dimethyl sulfide,[5] and H_2S have been reported,[6] including a recent CCSD(T) study of the C_6H_6···H_2S complex.[7]

For the efficient study of large biomolecules the use of density functional theory (DFT) or perhaps a semiempirical method would be desirable, but here the accuracy of the predictions is of particular concern. There have been a number of investigations to determine how appropriate are different functionals for the study of π-stacking interactions, and the new functionals of Truhlar,[19] along with the older half and half functional of Becke,[20] have shown promise. An alternative strategy is to add an empirical correction of the form R^{-6} to a density functional scheme to yield DFT-D models, rather than to tackle the difficult task of computing the dispersive term quantum mechanically.[21−23] Self-consistent-charge density functional tight-binding methods including such a dispersive correction (SCC DFTB-D) have also been developed.[24] DFT-D methods have been shown to be remarkably successful in predicting the binding energies of the *JSCH-2005* database[18] of 156 noncovalent biological complexes compiled by Hobza and co-workers.[25,26] For this same database, semiempirical models also including an empirical dispersive correction (AM1-D, PM3-D) have been developed which on average yield interaction energies to within 1−1.5 kcal mol^{-1} of the high-level ab initio values [MP2 or CCSD(T)].[27]

In this paper we investigate the use of the DFT-D method[22,23,25,26] and semiempirical methods (AM1-D, PM3-D)[27] to describe a range of noncovalent sulfur interactions in a number of model systems and compare these with the results of high level ab initio calculations [MP2 or CCSD(T)]. We have chosen to follow the approach of Jurečka et al.[18] in developing a small-molecule database to evaluate these more approximate methods. This database (Figure 1) contains different sulfur-π, S−H···S, and C−H···S interactions (**a1**−**a9**) for which high-level ab initio data (e.g., MP2 or CCSD(T)] have been reported (**a1**−**a7**)[7,23,28] or are calculated in this work (**a8**, **a9**). In our database hydrogen-bonded interactions have been considered in the H_2S dimer (**a1**) and in some of the CH_3SH dimers (**a2**, **a6**),

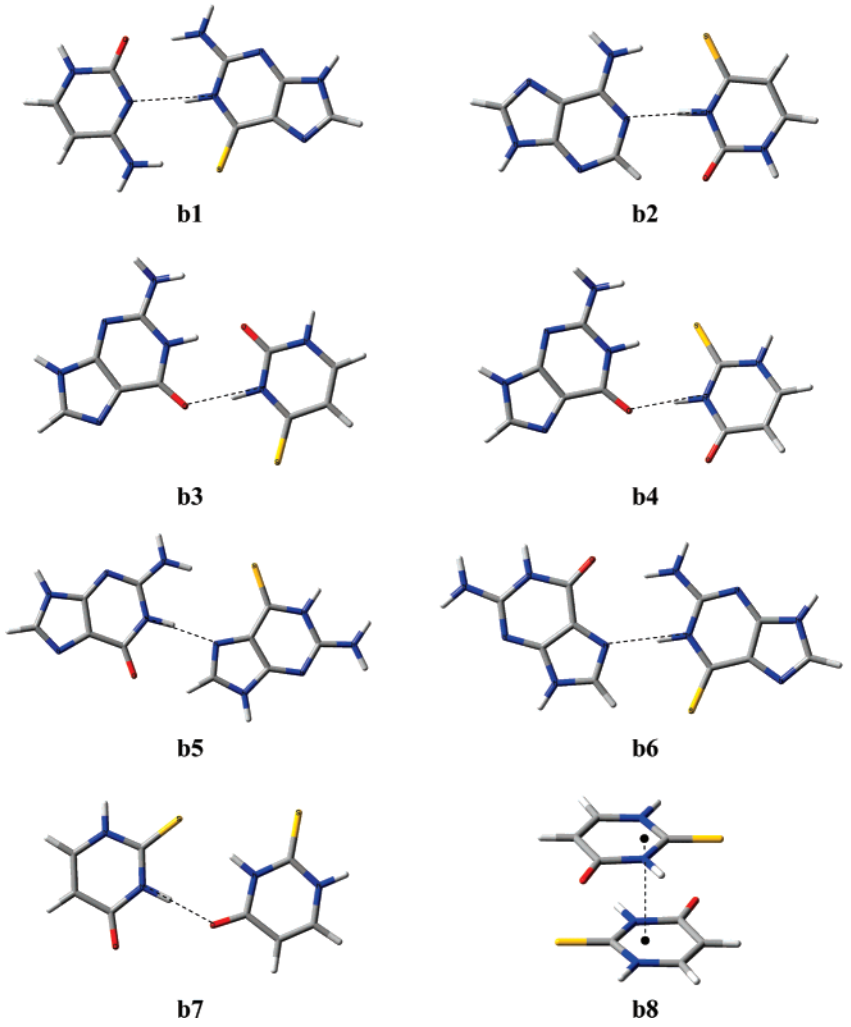


Figure 2. Biomolecule database structures showing interaction distances (dotted lines). Complexes **b1**: 6-thioG···C WC (C_s); **b2**: A···4-thioU WC (C_s); **b3**: G···4-thioU wobble (C_1); **b4**: G···2-thioU wobble (C_1); **b5**: G···6-thioG (C_s); **b6**: 6-thioG···G (C_s); **b7**: (2-thioU)$_2$ (C_s); **b8**: (2-thioU)$_2$ (C_1). In the **b8** complex the intermolecular distance is measured between the centers of mass of each monomer, and the monomers are parallel to each other.

whereas the other CH_3SH dimers (**a3**−**a5**) serve as models for C−H···S contacts, as does the $C_6H_6 \cdots CH_3SCH_3$ complex (**a9**). Sulfur-π interactions are considered in the complexes of benzene with H_2S (**a7**) and CH_3SCH_3 (**a8**). We use this same small molecule database to develop sulfur parameters for use in the AM1-D and PM3-D schemes and examine the strengths and weaknesses of these and the DFT-D method for treating important noncovalent interactions involving sulfur atoms, which are relevant for studying large biomolecules. We also examine the use of some alternative functionals for describing these interactions.

Computational Details

Empirical Dispersive Correction. In the approach developed by Grimme,[22,23] a pairwise additive potential of the form C_6/R^6 is used to account for long-range dispersion effects that can be particularly poorly described with some density functionals and with current semiempirical methods (AM1, PM3).[29,30] For each model, the dispersion corrected total energy is given by

$$E_{\mathrm{Total}} = E + E_{\mathrm{disp}} \tag{1}$$

where E is the normal self-consistent DFT or semiempirical energy (AM1, PM3) and E_{disp} is an empirical term containing the dispersion correction

$$E_{\mathrm{disp}} = -s_6 \sum_i \sum_j \frac{C_6^{ij}}{\mathrm{R}_{ij}^6} f_{\mathrm{dmp}}(R_{ij}) \tag{2}$$

Here, the summation is over all atom pairs, C_6^{ij} is the dispersion coefficient for the pair of atoms i and j (calculated from the atomic C_6 coefficients), s_6 is a scaling factor that depends on the density functional or semiempirical method used, and R_{ij} is the interatomic distance between atoms i and j.[22] A damping function is used in order to avoid near singularities for small distances. This function is given by

$$f_{\mathrm{dmp}}(R_{ij}) = \frac{1}{1 + e^{-\alpha(R_{ij}/R_0 - 1)}} \tag{3}$$

where R_0 is the sum of atomic van der Waals radii, and α is a parameter determining the steepness of the damping function. We note that initially the DFT-D method used the combination rule[22]

$$C_6^{ij} = 2C_6^iC_6^j/(C_6^i + C_6^j) \quad (4)$$

whereas more recently the geometric mean was employed[23]

$$C_6^{ij} = \sqrt{C_6^iC_6^j} \quad (5)$$

DFT Calculations. In this work we have used both combination rules in the DFT-D calculations.[22,23] The values for the C_6, R_0, s_6, and α parameters were taken from the respective parametrizations, with one exception (Table 1). In the original DFT-D formalism, atomic C_6 coefficients were only quoted for the elements H, C, N, O, F, and Ne.[22] A corresponding C_6 coefficient for sulfur was determined following an algorithm proposed by Halgren[31] giving a C_6 value of 10.3 J nm^6 mol^{-1}, substantially larger than the value used for the geometric mean (5.57 J nm^6 mol^{-1}).[23] We also use a van der Waals radius of 1.87 Å for sulfur (Table 1).

The DFT-D calculations reported herein have been performed using a locally modified version of GAUSSIAN 03.[32] Although other functionals may be used within the DFT-D formalism, here all calculations have been performed with the dispersion corrected BLYP[33] and B3LYP[34] methods and the TZV(2d,2p) basis set.[35] Following Grimme we do not consider basis set superposition errors (BSSE) in view of the quite large basis sets employed. In this work the DFT-D method using the BLYP functional with the combination rule given in eq 4[22] is referred to as BLYP-D*. The BLYP-D and B3LYP-D notation refers to the use of the geometric mean, eq 5.[23] We have also compared some of the results of our DFT-D calculations with those using a number of alternative density functionals suggested to be appropriate for the description of π-stacking interactions. Thus, DFT calculations with the MPW1B95,[36] MPWB1K,[36] PW6B95,[19] and PWB6K[19] functionals were carried out using GAUSSIAN 03, with a local implementation of the PW6B95 and PWB6K functionals. Geometry optimizations using these four functionals were performed with the 6-31+G(d,p) basis set,[37] and the structures were optimized in the corresponding point group. We have not computed counterpoise-corrections to the binding energies, as the purpose of these calculations was to evaluate the performance of these functionals to describe sulfur-π interactions in the absence of such corrections, looking into their applicability to large molecular systems. As in the case of the DFT-D calculations, the numerical integration was performed employing the weighting scheme of Becke along with grids of ultrafine quality. Reported binding and interaction energies refer to electronic energies; zero-point vibrational as well as thermal corrections are not included.

Computational Results and Discussion

Small-Molecule Database. *DFT Calculations*. In Table 2 we report our calculations of the small-molecule database using the DFT-D (BLYP-D*, BLYP-D, B3LYP-D) method. We note that for the $(H_2S)_2$ (**a1**) and $(CH_3SH)_2$ complexes (**a2**−**a6**) we report binding energies (energy of the complex with respect to relaxed monomers), whereas for the complexes involving benzene (**a7**−**a9**) we report interaction energies (energy of the complex with respect to unrelaxed monomers). For the BLYP-D* method we also report the dispersion contribution (ΔE_{disp}) to the binding or interaction energy (ΔE).

Table 1. Atomic C_6 Coefficients (J nm^6 mol^{-1}) and van der Waals Radii, R_0 (Å)[a,b]

	C_6	R_0
H	0.16 (0.14)	1.11 (1.001)
C	1.65 (1.75)	1.61 (1.452)
N	1.11 (1.23)	1.55 (1.397)
O	0.70 (0.70)	1.49 (1.342)
S	10.30[c] (5.57)	1.87 (1.683)

[a] BLYP-D*, AM1-D, and PM3-D values and (in parentheses) BLYP-D and B3LYP-D values. [b] α and s_6 values are respectively as follows: 23, 1.4 (BLYP-D*, AM1-D, and PM3-D); 20, 1.2 (BLYP-D); and 20, 1.05 (B3LYP-D). [c] Determined using algorithm of Halgren.[31]

We first consider our results for the hydrogen sulfide (**a1**) and methanethiol dimers (**a2**−**a6**). It is now generally accepted that strong hydrogen bonds can be accurately described using most standard density functionals.[38] In the case of a DFT-D approach the inclusion of an empirical C_6/R^6 correction may lead to some “double counting” of correlation effects,[22] which may be detrimental to the description of more strongly bound systems. However, in the case of the H_2S and the CH_3SH dimers, where the S−H···S interactions are weaker (and C−H···S interactions are even weaker), dispersion effects are expected to be important.

The binding energies and intermolecular separations obtained for the H_2S dimer agree well with the reference values,[23] showing a tendency of the DFT-D method to slightly overestimate the binding. In the case of the CH_3SH dimer, all three DFT-D schemes successfully identify the five stationary structures reported in the work of Cabalerio-Lago and Rodríguez-Otero,[28] with intermolecular orientations quite close to those obtained using the MP2 method. At this level, the binding energies of the different CH_3SH dimers span a narrow energy range of only 0.68 kcal mol^{-1}, and, as a result, it is not surprising that none of the dispersion corrected DFT methods reproduce the MP2 ordering of these energies. In fact, the DFT-D methods predict complex **a6** to have the largest binding, whereas at the MP2 level **a3** has the largest (Table 2). We also find that for the CH_3SH dimers, at the BLYP-D* level the dispersion contributions although relatively small (−1.60 to −2.68 kcal mol^{-1}) are important since they contribute between 60 and 92% to the overall binding energies (Table 2).

All the $(CH_3SH)_2$ structures show interactions between the sulfur atom in one molecule and a hydrogen atom in the methyl group of the second molecule. For the C−H···S contacts there is good agreement between the DFT-D calculations and the reference structures; the DFT-D calculations yield H···S distances close to 3.0 Å and C−H···S angles ranging from 130 to 150°. Regarding hydrogen-bonding, only structures **a2** and **a6** present interactions between a sulfur atom and a SH group, and only **a2** clearly exhibits this interaction, with an H···S distance and an S−H···S angle of 2.68 Å and 164.8° (BLYP-D*), respectively. At the BLYP-D and B3LYP-D levels these geometrical parameters are 2.69 Å, 164.7° and 2.69 Å, 164.4°, respectively. Overall, for the five $(CH_3SH)_2$ dimers alone, the

Table 2. Dispersion Corrected DFT Intermolecular Distances (Å) and Binding Energies and Interaction Energies (kcal mol^{-1}) for the Small-Molecule Database Complexes[a,b]

		BLYP-D*			BLYP-D		B3LYP-D		reference[e]	
complex[c]		R	ΔE_{disp}	ΔE	R	ΔE	R	ΔE	R	ΔE
$(H_2S)_2$ (C_s)	**a1**	4.13	−0.85	−1.91	4.12	−1.95	4.11	−1.97	4.10	−1.70
$(CH_3SH)_2$ (C_1)	**a2**	4.01	−1.63	−2.70	4.01	−2.86	4.00	−2.88	4.01	−2.28
$(CH_3SH)_2$ (C_1)	**a3**	3.87	−2.30	−2.52	3.78	−2.95	3.76	−3.27	3.76	−2.68
$(CH_3SH)_2$ (C_i)	**a4**	4.65	−1.60	−2.28	4.64	−2.68	4.63	−2.81	4.65	−2.00
$(CH_3SH)_2$ (C_1)	**a5**	4.17	−2.41	−2.66	4.10	−3.06	4.09	−3.30	4.11	−2.50
$(CH_3SH)_2$ (C_1)	**a6**	4.00	−2.68	−2.91	3.93	−3.33	3.93	−3.44	3.94	−2.46
$C_6H_6\cdots H_2S$ (C_{2v})	**a7**	3.83	−2.46	−2.14	3.69	−2.66	3.67	−2.93	3.80	−2.74
$C_6H_6\cdots CH_3SCH_3$ (C_{2v})	**a8**	4.92	−3.49	−1.96	4.81	−1.95	4.78	−2.27	4.94	−3.00
$C_6H_6\cdots CH_3SCH_3$ (C_{2v})	**a9**	5.53	−0.66	−0.83	5.40	−0.46	5.37	−0.61	5.46	−1.21
MAD[d]		0.04		0.41	0.04	0.57	0.05	0.62		

[a] TZV(2d,2p) basis set. [b] Binding energies for complexes **a1**−**a6**, interaction energies for complexes **a7**−**a9**. [c] Refer to Figure 1 for definition of interaction distance (R). [d] Mean absolute deviation. [e] Reference data: **a1**: RI-MP2/aug-cc-pVTZ;[23] **a2**−**a6**: MP2/aug-cc-pVDZ/cc-pVDZ (diffuse functions on hydrogen excluded);[28] **a7**: CCSD(T)/aug-cc-pVQZ;[7] **a8**, **a9**: MP2/aug-cc-pVDZ, this work.

BLYP-D*, BLYP-D, and B3LYP-D methods yield S⋯S distances with MAD (mean absolute deviation) values of 0.05, 0.01, and 0.01 Å, respectively, and give binding energies with corresponding MAD values of 0.29, 0.59, and 0.76 kcal mol^{-1}.

We next consider the $C_6H_6\cdots H_2S$ complex (**a7**), where we have chosen a C_{2v} structure with the sulfur atom located above the center ring (hydrogens-down configuration) since CCSD(T) data are available for this complex.[7] The BLYP-D* method predicts an intermolecular separation of 3.83 Å (refer to Figure 1) which differs by only 0.03 Å from the CCSD(T)/aug-cc-pVQZ value. At the BLYP-D and B3LYP-D levels the sulfur−benzene distances are a little short being 3.69 and 3.67 Å, respectively (Table 2). For this complex, compared to the CCSD(T) calculation, the interaction energy is underestimated by the BLYP-D* method by 0.60 kcal mol^{-1}, although this difference is still within the expected accuracy of the DFT-D model. On the other hand, the BLYP-D or B3LYP-D methods yield interaction energies (−2.66 and −2.93 kcal mol^{-1}) somewhat closer to the CCSD(T) value (−2.74 kcal mol^{-1}) even though the interaction distances differ from the reference values by more than 0.1 Å (Table 2). Importantly, for the $C_6H_6\cdots H_2S$ complex the BLYP-D* calculations indicate that in the absence of the dispersion correction the interaction between the respective monomers is repulsive (Table 2). Overall, for the $C_6H_6\cdots H_2S$ complex, the BLYP-D* method best describes the geometry of sulfur-π interactions with an interaction energy close to the ab initio value.[7]

We have used the MP2/aug-cc-pVDZ level of theory to obtain reference data for the $C_6H_6\cdots CH_3SCH_3$ complexes (**a8** and **a9**, Table 2). At this level, the interaction distance for the $C_6H_6\cdots H_2S$ complex (**a7**, hydrogens-down C_{2v} configuration) is calculated to be the same as the value obtained from the CCSD(T)/aug-cc-pVQZ calculation,[7] although the MP2 interaction energy is a little larger [MP2, −3.06 kcal mol^{-1}; CCSD(T), −2.74 kcal mol^{-1}]. In view of the excellent agreement between the MP2 and CCSD(T) calculations, we have also calculated two different C_{2v} configurations of the $C_6H_6\cdots CH_3SCH_3$ complex (**a8**, **a9**) at the MP2/aug-cc-pVDZ level of theory (Table 2).

Comparing to data obtained at the MP2/aug-cc-pVDZ level for the $C_6H_6\cdots CH_3SCH_3$ complex (**a8**, **a9**) only, BLYP-D* exhibits intermonomer distances closest to the reference structures with a MAD value of 0.05 Å, whereas BLYP-D and B3LYP-D predict intermolecular distances with slightly larger deviations, underestimating these quantities in both cases (MADs: 0.10 and 0.13 Å, Table 2). Most importantly, for the orientation involving a sulfur-π interaction (**a8**), the dispersion energy (−3.49 kcal mol^{-1}, Table 2) contributes significantly to the overall interaction energy (−1.96 kcal mol^{-1}), in the absence of dispersion, the interaction being repulsive.

The calculations reported herein indicate that the BLYP-D* method is a little better than BLYP-D and B3LYP-D models for describing the intermolecular geometries of van der Waals complexes involving sulfur-π interactions especially when the sulfur atom is located above the π-surface, as indicated by the interaction distances and interaction energy MADs for the complexes **a7** and **a8** alone: 0.03 and 0.82 (BLYP-D*), 0.12 and 0.57 (BLYP-D), and 0.15 Å and 0.46 kcal mol^{-1} (B3LYP-D).

Finally, to compare the BLYP-D* method to other density functionals we have also carried out calculations with the MPW1B95,[36] MPWB1K,[36] PW6B95,[19] and PWB6K[19] functionals on those complexes involving benzene from our small-molecule database. The results are summarized in Table 3. It can be seen that all of the functionals describe the complex **a7** well but give intermolecular distances that are too great for complexes **a8** and **a9**. The best results are obtained with the PWB6K functional, particularly for complex **a7**, where the intermolecular distance and interaction energies differ from the reference CCSD(T) data[7] by only 0.02 Å and 0.23 kcal mol^{-1}, respectively. Compared to the reference MP2 calculations, PWB6K overestimates the equilibrium distances in complex **a8** by 0.13 Å but yields the same intermolecular separation as does BLYP-D* for complex **a9** (Table 2).

We turn now to calculations of these complexes using the dispersion corrected semiempirical methods.

Parametrization of Semiempirical Methods. The semiempirical calculations reported herein were performed using our own local semiempirical program.[39] As in our previous work[27] these calculations (AM1-D, PM3-D) use the combi-

Table 3. DFT Intermolecular Distances (Å) and Interaction Energies (kcal mol^{-1}) for Selected Small-Molecule Database Complexes^a

		MPW1B95		MPWB1K		PW6B95		PWB6K		reference^d	
complex^b		R	ΔE	R	ΔE	R	ΔE	R	ΔE	R	ΔE
$C_6H_6 \cdots H_2S$	**a7**	3.85	−2.30	3.83	−2.49	3.87	−2.46	3.82	−2.97	3.80	−2.74
$C_6H_6 \cdots CH_3SCH_3$	**a8**	5.33	−1.02	5.29	−1.08	5.31	−1.32	5.07	−1.78	4.94	−3.00
$C_6H_6 \cdots CH_3SCH_3$	**a9**	5.59	−0.73	5.58	−0.74	5.57	−0.88	5.53	−1.03	5.46	−1.21
MAD^c		0.19	0.97	0.17	0.88	0.18	0.76	0.07	0.54		

^{a} 6-31+G(d,p) basis set. ^{b} Refer to Figure 1 for definition of interaction distance (R). ^{c} Mean absolute deviation. Corresponding MAD values for these complexes only, at the various DFT-D levels are (R,ΔE): BLYP-D*: 0.04, 0.67; BLYP-D: 0.10, 0.63; B3LYP-D: 0.13 Å, 0.51 kcal mol^{-1}. ^{d} Reference data: **a7**: CCSD(T)/aug-cc-pVQZ;[7] **a8**, **a9**: MP2/aug-cc-pVDZ, this work.

Table 4. Sulfur Parameters^a

parameter^b	AM1-D		PM3-D	
U_{ss} (eV)	−57.235044	(−56.694056)	−50.249536	(−49.895371)
U_{pp} (eV)	−48.307513	(−48.717049)	−43.968965	(−44.392583)
β_s (eV)	−3.311308	(−3.920566)	−8.397415	(−8.827465)
β_p (eV)	−7.256468	(−7.905278)	−7.594232	(−8.091415)
α (Å^{-1})	2.309315	(2.461648)	2.234331	(2.269706)

^{a} (In parentheses) published AM1 and PM3 parameters.[29,30] All other parameters remain unchanged. ^{b} (In parentheses) units.

nation rule given in eq 4 and the associated C_6, R_0, s_6, and α parameters (Table 1), along with our modified semiempirical parameters for H, C, N, and O.[27]

In their development of the PM3$_{BP}$ Hamiltonian (for the treatment of hydrogen-bonding in nucleobase pairs), Giese et al.[40] found that the most appropriate parametrization scheme involved allowing the PM3$_{BP}$ parameters to vary only a little from the default values so that the final parameter set would be transferable to other chemical systems. This is the approach adopted herein, and we chose to modify only the U_{ss}, U_{pp}, β_s, β_p, and α parameters of sulfur (for both AM1[29] and PM3[30] methods); all the remaining parameters for sulfur are unchanged. Given the lack of complete structural data from the high level ab initio studies and the very good agreement between the BLYP-D* calculations and the ab initio reference data for the interaction distances in the small-molecule database (Table 2), we chose as our reference data (for the parametrization), the BLYP-D* structures (Table 2), in conjunction with the interaction energies taken from the ab initio data (Table 2).[7,23,28] The sulfur parameters were optimized using a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, details of which are available elsewhere.[41] Since we require the new semiempirical methods to be capable of accurately predicting interaction energies for structures extracted from experiment, we chose quite large weighting factors for the energy contributions to our fitting function, in line with our previous work.[27] The initial values of the parameters were set equal to the standard AM1[29] or PM3[30] values.

In Table 4 we report our modified AM1-D and PM3-D parameters for sulfur. We see that the final parameters only deviate a little from the standard AM1 or PM3 values. For the AM1-D model, the largest change is found for the β_s parameter which decreases in magnitude by 15.5%, while for the PM3-D method the largest change is found for the β_p parameter which also decreases in magnitude by 6.1%. These relatively small changes are not unexpected given that suitable AM1 SRP values for the calculation of thermal rate constants have been reported by allowing a variation in the standard values of only ±5%.[42]

In Table 5 we report the structures and binding (**a1**−**a6**) or interaction energies (**a7**−**a9**) for the complexes in the small-molecule database (Figure 1) calculated using the corrected semiempirical methods (AM1-D, PM3-D). In line with the DFT-D calculations for the $(H_2S)_2$ (**a1**) and $(CH_3SH)_2$ complexes (**a2**−**a6**) we report binding energies, whereas for the complexes involving benzene (**a7**−**a9**) we report interaction energies. In addition for the AM1-D and PM3-D methods we report the dispersion contribution (ΔE_{disp}) to the overall binding or interaction energies (ΔE). For comparison we have also calculated these complexes using the current semiempirical methods, AM1[29] and PM3.[30]

We see that compared to the AM1 and PM3 methods, the inclusion of a dispersive correction (AM1-D, PM3-D) leads to essentially a halving of the average error in the binding and interaction energies [MADs: 1.24 (AM1); 1.28 (PM3); 0.85 (AM1-D); 0.66 (PM3-D) kcal mol^{-1}, Table 5]. Interestingly, both the AM1 and PM3 methods predict all the complexes to be bound including those with sulfur-π interactions (**a7**, **a8**). We see that as far as the binding and interaction energies are concerned, our new semiempirical methods perform almost as well as the BLYP-D* method (MAD: 0.41 kcal mol^{-1}, Table 2), although the excellent agreement is not surprising given that these methods were parametrized using the BLYP-D* structures.

For the dispersion corrected methods, the largest difference between the semiempirical and reference interaction energies occurs for the $C_6H_6 \cdots CH_3SCH_3$ complex (**a8**) with deviations of 2.83 (AM1-D) and 3.11 (PM3-D) kcal mol^{-1}, respectively. In spite of this large discrepancy, all other binding and interaction energies only differ from the reference value by less than 1 kcal mol^{-1} at the PM3-D level. Excluding complex **a8**, the AM1-D and PM3-D MADs for the binding/interaction energies (0.60 and 0.32 kcal mol^{-1}, respectively) are essentially comparable to the BLYP-D* values (Table 2). For the five $(CH_3SH)_2$ dimers, the relative ordering of the reference ab initio binding energies (**a3** > **a5** > **a6** > **a2** > **a4**) is not well reproduced at the AM1-D and PM3-D levels, in line with our DFT-D results (Table 2). However, unlike the DFT-D calculations (Table 2), the PM3-D method does in fact correctly predict **a3** to be the most stable arrangement of $(CH_3SH)_2$.[28] We note that at the AM1 and PM3 levels we were unable to locate all of the methanethiol stationary structures (Table 5). For the complexes involving benzene, the interaction energies at the AM1 and PM3 levels

Table 5. Semiempirical Intermolecular Distances (Å) and Binding Energies and Interaction Energies (kcal mol^{-1}) for the Small-Molecule Database Complexes[a]

		AM1		PM3		AM1-D			PM3-D			reference[d]	
complex[b]		*R*	Δ*E*	*R*	Δ*E*	*R*	Δ*E*$_{disp}$	Δ*E*	*R*	Δ*E*$_{disp}$	Δ*E*	*R*	Δ*E*
$(H_2S)_2$ (C_s)	**a1**	3.05	−3.84	3.23	−5.56	3.85	−0.93	−1.62	3.87	−0.87	−1.99	4.10	−1.70
$(CH_3SH)_2$ (C_1)	**a2**	3.67	−2.00	3.72	−2.12	3.92	−1.95	−2.34	3.97	−1.69	−2.31	4.01	−2.28
$(CH_3SH)_2$ (C_1)	**a3**	4.09[e]	−1.69[e]	4.27	−1.90	4.11	−3.07	−3.38	4.35	−2.56	−2.96	3.76	−2.68
$(CH_3SH)_2$ (C_i)	**a4**	4.30	−1.28	4.43	−1.55	4.61	−2.75	−3.11	4.58	−2.35	−2.91	4.65	−2.00
$(CH_3SH)_2$ (C_1)	**a5**	4.26	−1.21	4.38	−1.56	4.83	−2.87	−3.59	4.46	−2.33	−2.94	4.11	−2.50
$(CH_3SH)_2$ (C_1)	**a6**	3.66[f]	−2.06[f]	3.66[f]	−2.13[f]	4.17	−4.07	−3.89	4.51	−2.78	−2.67	3.94	−2.46
$C_6H_6\cdots H_2S$ (C_{2v})	**a7**	4.61	−0.55	4.57	−0.41	3.54	−3.26	−2.66	3.74	−2.73	−2.16	3.80	−2.74
$C_6H_6\cdots CH_3SCH_3$ (C_{2v})	**a8**	6.12	−0.18	4.86	−0.55	4.37	−6.20	−5.83	4.45	−5.83	−6.11	4.94	−3.00
$C_6H_6\cdots CH_3SCH_3$ (C_{2v})	**a9**	4.95	−0.85	5.02	−1.02	5.04	−0.96	−0.95	5.08	−0.95	−1.32	5.46	−1.21
MAD[c]		0.55	1.24	0.41	1.28	0.31		0.85	0.30		0.66		

[a] Binding energies for complexes **a1**−**a6**, interaction energies for complexes **a7**−**a9**. [b] Refer to Figure 1 for definition of interaction distance (*R*). [c] Mean absolute deviation. [d] Reference data: **a1**: RI-MP2/aug-cc-pVTZ;[23] **a2**−**a6**: MP2/aug-cc-pVDZ/cc-pVDZ (diffuse functions on hydrogen excluded);[28] **a7**: CCSD(T)/aug-cc-pVQZ;[7] **a8**, **a9**: MP2/aug-cc-pVDZ, this work. [e] Close to the **a5** structure. [f] Close to the **a2** structure.

Table 6. Dispersion Corrected DFT Intermolecular Distances (Å) and Binding Energies (kcal mol^{-1}) for the Sulfur-Containing Base Pairs in the Biomolecule Database[a]

		BLYP-D*			BLYP-D		B3LYP-D		reference[c]	
complex[b]		*R*	Δ*E*$_{disp}$	Δ*E*	*R*	Δ*E*	*R*	Δ*E*	*R*	Δ*E*
6-thioG···C WC pl (C_s)	**b1**	3.15	−4.03	−25.78	3.12	−27.35	3.11	−28.10	3.11	−25.50
A···4-thioU WC (C_s)	**b2**	2.94	−3.84	−13.78	2.87	−15.18	2.87	−15.36	3.04	−13.20
G···4-thioU (C_1)	**b3**	2.84	−3.83	−14.79	2.82	−15.92	2.80	−16.74	2.81	−15.90
G···2-thioU (C_1)	**b4**	2.80	−3.21	−13.27	2.79	−14.39	2.77	−14.99	2.77	−14.60
G···6-thioG pl (C_s)	**b5**	2.94	−3.68	−18.18	2.91	−19.71	2.90	−20.82	2.89	−19.00
6-thioG···G pl (C_s)	**b6**	2.97	−3.96	−18.84	2.93	−20.40	2.91	−21.49	2.92	−19.60
(2-thioU)$_2$ pl (C_s)	**b7**	2.86	−3.20	−10.47	2.84	−10.95	2.82	−11.37	2.80	−11.60
(2-thioU)$_2$ (C_1)	**b8**	3.6	−6.90	−4.23	3.5	−4.59	3.5	−4.99	3.6	−5.85
MAD[d]		0.05		0.95	0.05	0.94	0.04	1.35		

[a] TZV(2d,2p) basis set. [b] Refer to Figure 2 for definition of interaction distance (*R*). Abbreviations used: G (guanine), C (cytosine), A (adenine), U (uracil), WC (Watson−Crick), pl (planar). [c] Reference data: **b1**−**b7**: RI-MP2/CBS//RI-MP2/cc-pVTZ;[18] **b8**: MP2/6-31G*.[43] [d] Mean absolute deviation.

are all underestimated such that the inclusion of the dispersive correction leads to an improvement in the interaction energies. For example interaction energies for the complex involving H_2S are now −2.66 (AM1-D) and −2.16 (PM3-D) kcal mol^{-1}, close to the CCSD(T) value (−2.74 kcal mol^{-1}).[7] We see that for each of the complexes involving benzene (**a7**−**a9**) the dispersion contribution to the interaction energy at both the AM1-D and PM3-D levels is significant, in some cases being greater than the interaction energy (Δ*E*) itself (Table 5). However, despite the good agreement for the $C_6H_6\cdots H_2S$ complex, as previously noted, the interaction energy for the complex of benzene with dimethylsulfide (**a8**) is predicted to be almost twice that given by ab initio calculation at both the AM1-D and PM3-D levels [−5.83 (AM1-D); −6.11 (PM3-D); −3.00 kcal mol^{-1} (MP2)].

As far as the structures of the various complexes are concerned, the inclusion of the dispersive correction leads to an improvement in the intermolecular distances (Figure 1, Table 5), the MADs being reduced from 0.71 (AM1) and 0.41 (PM3) to 0.31 (AM1-D) and 0.30 Å (PM3-D). We note that although the AM1-D and PM3-D methods yield binding and interaction energies of comparable accuracy to the DFT-D values, the DFT-D geometries for these complexes are in fact much closer to the reference ones [MAD: 0.04 (BLYP-D*); 0.04 (BLYP-D); 0.05 Å (B3LYP-D)]. The MADs for the interaction distances at both the AM1-D and PM3-D levels are still quite large (Table 5), even though these methods have been parametrized using the structures from our BLYP-D* calculations (Table 2).

Biomolecule Database. We have calculated the structures and binding energies of a number of sulfur-containing base pairs taken from the *JSCH-2005* database of Jurečka et al.[18] and also from the work of Šponer et al.[43] to further test the DFT-D and the semiempirical AM1-D and PM3-D methods. For the semiempirical methods, this represents an important step in examining the transferability of our new modeling schemes to the study of a series of complexes *not* included in the reference parametrization data. The hydrogen-bonded sulfur-containing complexes were taken from the *JSCH-2005* database (complexes **b1**−**b7**);[18] the stacked base pair (2-thioU)$_2$, orientation **b8** is taken from the work of Šponer et al.[43] The results are presented in Tables 6 and 7.

DFT Calculations. At the BLYP-D*/TZV(2d,2p) level, the binding energies of all the thiobase complexes are in very good agreement with the reference data (Table 6). The largest difference for a hydrogen-bonded base pair is just 1.33 kcal mol^{-1} (**b4**), while in the case of the stacked (2-thioU)$_2$ complex (**b8**) the difference amounts to only 1.62 kcal mol^{-1}. Intermolecular distances are within the expected accuracy

Table 7. Semiempirical Intermolecular Distances (Å) and Binding Energies (kcal mol^{-1}) for the Sulfur-Containing Base Pairs in the Biomolecule Database

		AM1		PM3		AM1-D			PM3-D			reference[b]	
complex[a]		R	ΔE	R	ΔE	R	ΔE_{disp}	ΔE	R	ΔE_{disp}	ΔE	R	ΔE
6-thioG···C WC pl (C_s)	**b1**	3.15	−16.10	2.83	−17.50	3.06	−4.33	−21.38	2.76	−5.32	−27.12	3.11	−25.50
A···4-thioU WC (C_s)	**b2**	2.95	−5.76	2.81	−8.56	2.78	−4.10	−10.54	2.69	−4.63	−16.48	3.04	−13.20
G···4-thioU (C_1)	**b3**	3.02	−7.40	2.81	−5.52	2.96	−3.69	−15.24	2.74	−3.59	−13.85	2.81	−15.90
G···2-thioU (C_1)	**b4**	3.07	−8.24	2.80	−7.96	2.87	−3.44	−10.82	2.74	−3.36	−11.87	2.77	−14.60
G···6-thioG pl (C_s)	**b5**	2.99	−10.12	2.86	−10.96	2.84	−3.74	−15.73	2.71	−4.23	−21.17	2.89	−19.00
6-thioG···G pl (C_s)	**b6**	2.99	−10.12	2.79	−11.23	2.85	−4.26	−17.34	2.71	−4.71	−23.00	2.92	−19.60
(2-thioU)$_2$ pl (C_s)	**b7**	3.08	−6.08	2.81	−6.13	2.88	−3.13	−8.30	2.75	−3.42	−9.16	2.80	−11.60
(2-thioU)$_2$ (C_1)	**b8**	7.42[d]	−4.81	6.11[e]	−1.64	3.84	−9.70	−6.85	4.25	−7.86	−6.88	3.6	−5.85
MAD[c]		0.57	7.05	0.42	6.94	0.12		2.61	0.26		2.36		

[a] Refer to Figure 2 for definition of interaction distance (R). Abbreviations used: G (guanine), C (cytosine), A (adenine), U (uracil), WC (Watson−Crick), pl (planar). [b] Reference data: **b1**−**b7**: RI-MP2/CBS//RI-MP2/cc-pVTZ;[18] **b8**: MP2/6-31G*.[43] [c] Mean absolute deviation. [d] Corresponds to a planar hydrogen-bonded arrangement. [e] Rings are parallel but significantly displaced compared to the reference structure.

of the DFT-D method, showing a tendency to overestimation.[22,23,25,26] Overall, the BLYP-D* method exhibits MAD values of 0.05 Å and 0.95 kcal mol^{-1} for intermolecular distances and binding energies, respectively. On average, the binding energies calculated with the BLYP-D and B3LYP-D methods show absolute deviations from the reference data of just 0.94 and 1.35 kcal mol^{-1}, respectively, and with the exception of complexes **b2** and **b8**, intermolecular distances obtained with these methods are in very good agreement with the reference data. Finally, in comparison with BLYP-D*, both BLYP-D and B3LYP-D give shorter stacking and hydrogen-bonding distances and also larger binding energies.

Semiempirical Calculations. The AM1-D and PM3-D methods have previously been used to calculate the entire *JSCH-2005* database of Jurečka et al.[18] (excluding those complexes involving sulfur) and gave MADs of 1.1 and 1.3 kcal mol^{-1}, respectively, compared to 9.0 and 8.6 kcal mol^{-1} for the AM1 and PM3 methods.[27] For the complexes in the biomolecule database herein, we find a similar improvement in the MADs for both binding energies and geometries (Table 7). Overall the MADs are 2.61 kcal mol^{-1} (AM1-D) and 2.36 kcal mol^{-1} (PM3-D), considerably smaller than the corresponding values for the AM1 and PM3 methods, respectively (7.05 and 6.94 kcal mol^{-1}). As with the DFT-D calculations, the presence of sulfur in these systems does not seem to affect the overall performance of the AM1-D and PM3-D methods. We find that optimization of the stacked (2-thioU)$_2$ complex (**b8**) at the AM1-D and PM3-D levels yields the desired stacked configuration. However, at the AM1 level this structure collapses to a planar hydrogen-bonded arrangement of the two rings, and at the PM3 level the two rings remain essentially parallel but are significantly displaced compared to the MP2 structure, as indicated by the interaction distance of 6.11 Å (Table 7).

Conclusions

These calculations have reinforced our previous finding that the DFT-D method is a very accurate and efficient scheme for calculating a wide range of intermolecular interactions where dispersive contributions are particularly important. The MAD values we find for sulfur containing molecules are similar to those found for molecules having only first row atoms. The extension of this scheme to semiempirical wavefunctions is found to yield MAD values a little larger than the DFT-D ones but still within 2 kcal mol^{-1} of the high level ab initio values. Thus, as we found for molecules containing first row atoms, the PM3-D and AM1-D methods can be surprisingly accurate and computationally economic. In view of this, even though these DFT and semiempirical methods with a R^{-6} correction might be deemed to be more 'approximate' than DFT methods requiring a specific functional, they do represent a good compromise between high level ab initio methods and empirical force field schemes.

Acknowledgment. We thank the North West Development Agency (NWDA) and the NWGrid Consortium for the provision of computing resources. We thank Dr. Y. Zhao for data using the PWB6K functional.

Supporting Information Available: Cartesian coordinates of the BLYP-D* optimized complexes **a1**−**a9** and **b1**−**b8**. This material is available free of charge via the Internet at http://pubs.acs.org.

References

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

(2) Némethy, G.; Scheraga, H. A. *Biochem. Biophys. Res. Commun.* **1981**, *98*, 482.

(3) Cheney, B. V.; Schulz, M. W.; Cheney, J. *Biochim. Biophys. Acta* **1989**, *996*, 116.

(4) Duan, G. L.; Smith, V. H., Jr.; Weaver, D. F. *Mol. Phys.* **2001**, *99*, 1689.

(5) Pranata, J. *Bioorg. Chem.* **1997**, *25*, 213.

(6) Hermida-Ramón, J. M.; Cabalerio-Lago, E. M.; Rodríguez-Otero, J. *J. Chem. Phys.* **2005**, *122*, 204315.

(7) Tauer, T. P.; Derrick, M. E.; Sherrill, C. D. *J. Phys. Chem. A* **2005**, *109*, 191.

(8) Morgan, R. S.; Tatsch, C. E.; Gushard, R. H.; McAdon, J. M.; Warme, P. K. *Int. J. Pept. Protein Res.* **1978**, *11*, 209.

(9) Morgan, R. S.; McAdon, J. M. *Int. J. Pept. Prot. Res.* **1980**, *15*, 177.

(10) Reid, K. S. C.; Lindley, P. F.; Thornton, J. M. *FEBS Lett.* **1985**, *190*, 209.

(11) Samanta, U.; Pal, D.; Chakrabarti, P. *Proteins: Struct., Funct., Genet.* **2000**, *38*, 288. Pal, D.; Chakrabarti, P. *J. Biomol. Struct. Dyn.* **2001**, *19*, 115. Pal, D.; Chakrabarti, P. *J. Biomol. Struct. Dyn.* **1998**, *15*, 1059.

(12) Zauhar, R. J.; Colbert, C. L.; Morgan, R. S.; Welsh, W. J. *Biopolymers* **2000**, *53*, 233.

(13) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380.

(14) Viguera, A. R.; Serrano, L. *Biochemistry* **1995**, *34*, 8771.

(15) Muñoz, V.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 399.

(16) Tatko, C. D.; Waters, M. L. *Protein Sci.* **2004**, *13*, 2515.

(17) Waters, M. L. *Biopolymers* **2004**, *76*, 435.

(18) Jureèka, P.; Šponer, J.; Èerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(19) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

(20) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

(21) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748.

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

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

(24) Valdés, H.; Øeha, D.; Hobza, P. *J. Phys. Chem. B* **2006**, *110*, 6385. Dobeš, P.; Oteypka, M.; Strnad, M.; Hobza, P. *Chem. Eur. J.* **2006**, *12*, 4297.

(25) Anthony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.

(26) Morgado, C.; Vincent, M. A.; Hillier, I. H.; Shan, X. *Phys. Chem. Chem. Phys.* **2007**, *9*, 448.

(27) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.

(28) Cabaleiro-Lago, E. M.; Rodríguez-Otero, J. *J. Phys. Chem. A* **2002**, *106*, 7440.

(29) Dewar, M. J. S.; Zoebisch, E.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1993**, *115*, 5348.

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

(31) Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827.

(32) 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.; Hratchain, 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.; 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, 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.

(33) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(34) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. Stephens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(35) Schäfer, A.; Huber, C.; Ahlrichs, R. J. *J. Chem. Phys.* **1994**, *100*, 5829. Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(36) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

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

(38) Tsuzuki, S.; Leuthi, H. P. *J. Chem. Phys.* **2001**, *114*, 3949.

(39) Mohr, M.; McNamara, J. P.; Wang, H.; Rajeev, S. A.; Ge, J.; Morgado, C.; Hillier, I. H. *Faraday Discuss.* **2003**, *124*, 413.

(40) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275.

(41) McNamara, J. P.; Sundararajan, M.; Hillier, I. H.; Ge, J.; Campbell, A.; Morgado, C. *J. Comput. Chem.* **2006**, *27*, 1307. McNamara, J. P.; Sundararajan, M.; Hillier, I. H. *J. Mol. Graphics Modell.* **2005**, *24*, 128.

(42) Jitariu, L.; Wang, H.; Hillier, I. H.; Pilling, M. J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2459.

(43) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem. A* **1997**, *101*, 9489.

CT700072A

JCTC Journal of Chemical Theory and Computation

Empirical Correction to Molecular Interaction Energies in Density Functional Theory (DFT) for Methane Hydrate Simulation

Qi-Shi Du,*,[†,‡] Peng-Jun Liu,[†] and Jun Deng[§]

Department of Chemistry, Hainan Normal University, Haikou, Hainan 571158, China, Key Laboratory of Subtropical Bioresource Conservation and Utilization, Guangxi University, Nanning, Guangxi 530004, China, and Nanning New and High-Tech Incubator, 68 Keyuan Avenue, Nanning, Guangxi 530007, China

Received January 30, 2007

Abstract: A general and empirical method is proposed for correction of London dispersion and other deficiencies in density functional theory (DFT). This method is based on the existing Lennard-Jones (L-J) equation and van der Waals parameters. The benchmark of energy correction is set as the energy difference between DFT and more accurate methods, for example CCSD(T). The energy correction includes all differences between CCSD(T) and DFT, dispersion energy, configuration interaction, induction interaction, residual correlation, and other effects. The energy correction is expressed as a linear combination of van der Waals potentials of nonbonded atomic pairs. The combination coefficients are determined using a least-squares approach in a training set of molecular pairs. The coefficients then can be used for the energy corrections in DFT calculations in a molecular family. Three correction equations of molecular pair interaction energy, water−water, water−methane, and methane−methane, are derived for methane hydrate simulation. The correction equation of the water−water pair is applied in the DFT calculation of water pentamer, yielding good intermolecular potential energy surfaces (PES), very close to the results of CCSD(T) over the active interaction range from 2.1 Å to 8.0 Å.

1. Introduction

The London dispersion energy[1] plays an important role in molecular interactions in molecular clusters, solutions, solids, and biological macromolecules. Density functional theory (DFT) within the Kohn−Sham formulation and with presently available exchange-correlation functionals does not describe the dispersion interaction correctly.[2] The true correlation energy functional should include the van der Waals interaction,[3] and future generations of optimized effective potentials[4−7] may correct this deficiency of DFT. Since the pioneer work of Lein et al.[2] in 1999, considerable progress has been achieved toward a better description for weak molecular interactions within DFT.[8,9] The efforts for molecular dispersion correction in DFT approaches can be classified into three categories. (1) In the theoretical approaches, or nonempirical approaches, the molecular system is divided into two subsystems, and the dispersion interaction between them is calculated from intermolecular perturbation theory,[10,11] or from the dynamic polarizabilities,[12,13] or from ground state densities only.[14−17] (2) Semiempirical approaches reparametrize the existing density functionals so that they describe dispersion properly. The mainstream in this category is represented by modification of the exchange functionals.[18−21] (3) In empirical dispersion approaches[22] empirical van der Waals potentials, in most cases using the $-c_6/R^6$ formula, are added to the DFT energy. The a posteriori empirical dispersion energy E_{dis} is calculated separately from the DFT calculation.

* Corresponding author phone: 086-771-327-0730/086-771-323-8107; e-mail: duqishi@yahoo.com.

[†] Hainan Normal University.

[‡] Guangxi University.

[§] Nanning New and High-Tech Incubator.

10.1021/ct700026d CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/13/2007

Great efforts have been made for the correct description of weak molecular interaction energy in DFT. Several dispersion interaction correction methods are suggested.[23−25] Some of them are effective in part interaction distance and may fail at other distances and are not proper to all types of molecular interactions. For example, the generalized-gradient approximation (GGA) DFT fails to handle the long-range part of the potential and is not appropriate in the simulation of physical sorption processes.[23,24] There is still a long way to go for a comprehensive solution to this problem.[17]

Because of the simplicity and high efficiency of empirical approaches, some progress has been made in recent years.[22,26] Direct inclusion of London forces with a posteriori functions into the nonlocal part of the correlation potential[27] and into pseudopotentials[28] has been suggested recently by several authors.[27,29,30] Bordner et al. presented a method, which exploits the virtually unlimited number of ab initio calculations, as compared with experimental data to directly derive van der Waals parameters.[26] Zhechkov et al.[22] use modified van der Waals potential, U_0-U_1r^n-U_2r^{2n}, for a better description at short distances. The results of DFT are basis set dependent. Therefore, it is difficult to find general a posteriori potential equations and van der Waals parameters, which are appropriate for all basis sets. In this study we suggest a general empirical method for a better description of the molecular dispersion interaction in DFT, using existing van der Waals parameters and L-J potential equations. It is effective over the active molecular interaction range, short distance, van der Waals equilibrium region, and long distance. This method is also efficient for other differences between DFT and higher level approaches, including dispersion energy, configuration interaction, induction interaction, residual correlation, and other effects.

2. Method and Scheme

In the derivation of the molecular pair interaction energy correction equation for DFT, the benchmark of energy correction is defined as the energy difference between DFT and the more accurate method. In this study we use CCSD(T) as the benchmark calculation, which stands for coupled-cluster (CC) theory with single, double, and part or full triple excitations

$$E_{\mathrm{bench}}(r_i) = \Delta E(r_i) = E_{\mathrm{CCSD-T}}(r_i) - E_{\mathrm{DFT}}(r_i) \qquad (i = 1, 2, ..., n) \quad (1)$$

where r_i is the ith distance between two molecules, and n is the total number of distances. The energy correction is assumed to be a summation of van der Waals potentials over all nonbonded atomic pairs in the molecular cluster

$$E_{\mathrm{corr}}(r_i) = \sum_{l=1}^{m} c_l U_l^{\mathrm{vdw}}(r_i;d_l,\sigma_l) \quad (i = 1, 2, ..., n) \qquad (2)$$

where $U_l^{\mathrm{vdw}}(r_i)$ is the van der Waals potential of the lth atomic pair at distance r_i, m is the number of van der Waals pairs, c_l is the coefficient for the van der Waals potential of the lth atomic pair, and d_l and σ_l are atomic van der Waals parameters. Because water and methane are small molecules, we use the distance r_i of the heavy atomic pair as the distances of all atomic van der Waals pairs. The $U_l^{\mathrm{vdw}}(r_i)$ could be any type of van der Waals potential, Lennard-Jones, Exp-6, or Morese potential. If the L-J (6-12) potential is used, then the van der Waals potential takes the form

$$U_l^{\mathrm{vdw}}(r_i) = -2d_l\left(\frac{\sigma_l}{r_i}\right)^6 + d_l\left(\frac{\sigma_l}{r_i}\right)^{12} \qquad (i = 1,2,..., n;\ l = 1,2,..., m) \quad (3)$$

The squared residues between benchmarks $E_{\mathrm{bench}}(r_i)$ and corrected values $E_{\mathrm{corr}}(r_i)$ by eq 2 are summed over the whole interaction range

$$Q = \sum_{i=1}^{n}\{[E_{\mathrm{bench}}(r_i) - E_{\mathrm{corr}}(r_i)]^2\} = \sum_{i=1}^{n}\{[E_{\mathrm{bench}}(r_i) - \sum_{l=1}^{m} c_l U_l^{\mathrm{vdw}}(r_i)]^2\} \quad (4)$$

A least-squares approach is used to find the best combination coefficients $\{c_l, l=1, ..., m\}$. When the sum of squared residues takes its minimum value, the derivatives of Q to combination coefficients c_k are zero, $\partial Q/\partial c_k = 0$, leading to simultaneous linear equations

$$\text{for } \frac{\partial Q}{\partial c_k} = 0: \quad \sum_{l=1}^{m} c_l \sum_{i=1}^{n} U_l^{\mathrm{vdw}}(r_i)U_k^{\mathrm{vdw}}(r_i) = \sum_{i=1}^{n} E_{\mathrm{bench}}(r_i)U_k^{\mathrm{vdw}}(r_i) \qquad (k = 1, ..., m) \quad (5)$$

Equation 5 is $m\times m$ simultaneous linear equations and can be solved using a general inverse matrix to get the least-squares solution. The combination coefficients $\{c_l, l=1, ..., m\}$ and eq 2 then can be used to correct other DFT calculations of molecular interaction energies in a molecular family. For the L-J (6-12) potentials the combination coefficients c_l also can be combined into the van der Waals parameters d_l. The molecular interaction energy correction method can be performed independently outside of the DFT software package.

A big difference in our interaction energy correction equation with other empirical methods is that we use both the attractive branch and the repulsive branch of the van der Waals potential. In other empirical dispersion correction equations[22] based on the van der Waals potential only the attractive branch $-c_6/R^6$ is used, because the dispersion interaction is always attractive. The combination of two branches makes the correction equation easy to fit different molecular pair interaction energy, polar−polar, polar−nonpolar, and nonpolar−nonpolar.

3. Calculation Examples

In this section we derive the interaction energy correction equations of three molecular pairs: water−water, water−methane, and methane−methane, which will be used in methane hydrate simulation in our continuing study.

3.1. Benchmark for Corrections. In the benchmark calculations Xi'An-CI is used to compute the molecular interaction energies of three molecular pairs for future

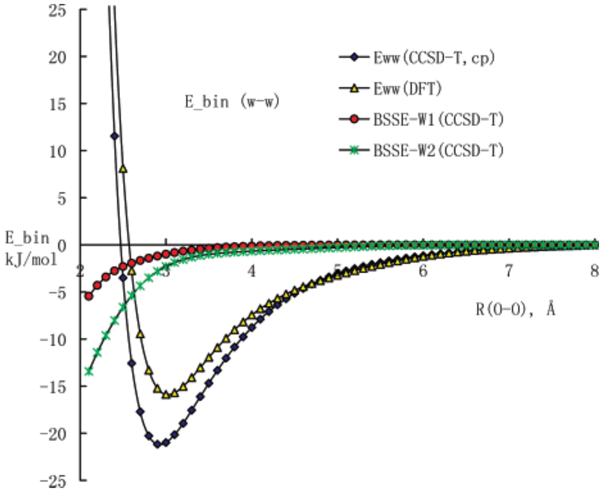


Figure 1. The molecular interaction energies of the water dimer calculated by CCSD(T)/TZVP of Xi'An-CI and BLYP/TZVP of DFT without dispersion correction. The CP-corrections for BSSE of two water molecules are performed in CCSD(T)/TZVP calculations.

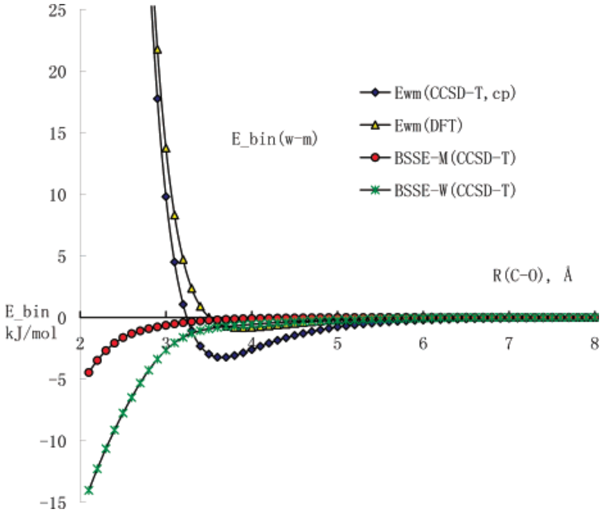


Figure 2. The molecular interaction energies of the water−methane calculated by CCSD(T)/TZVP of Xi'An-CI and BLYP/TZVP of DFT without dispersion correction. The CP-corrections for BSSE of water and methane molecules are performed in CCSD(T)/TZVP calculations.

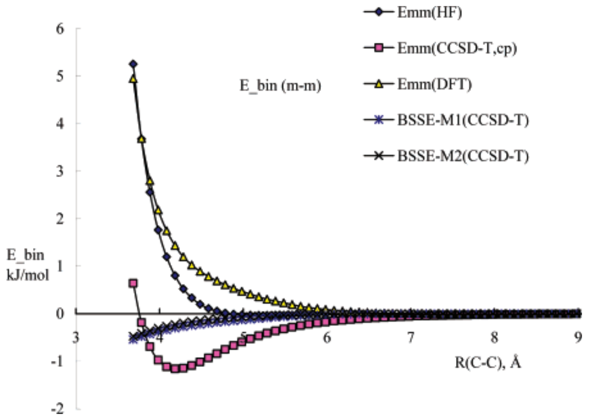


Figure 3. The molecular interaction energies of the methane−methane calculated by HF/cc-pVTZ and BLYP/TZVP of DFT without dispersion correction and CCSD(T)/cc-pVTZ of Xi'An-CI after CP-correction. According to the HF and DFT calculations, there is no attractive interaction between two methane molecules.

applications in methane hydrate study. Xi'An-CI is a configuration interaction (CI) software package based on the graphical unitary group approach (GUGA) developed by Prof. Wen's group.[31−34] The molecular interaction range is from 2.1 Å to 8.0 Å with an increment of 0.1 Å. The CCSD(T) method and the basis sets TZVP and cc-pVTZ are used in the benchmark calculations. Figure 1 shows the molecular interaction energy of the water dimer obtained by using CCSD(T)/TZVP of Xi'An-CI and BLYP/TZVP of Gauss-03 DFT (no dispersion correction). Figure 2 shows the results of water−methane using the same methods. The geometries of water and methane molecules were optimized at the CCSD(T)/TZVP level and kept constant in all molecular configurations. The rigid monomer approximation may produce an error of -0.2 ± 0.1 kJ mol^{-1} according to ref 38, which is tolerant in this study. In order to correct the basis set superposition error (BSSE) caused by the basis set TZVP, we performed the counterpoise (CP) corrections for water and methane at all distances, and the BSSE CP-correction curves are shown in the figures. The CP-correction curves for BSSE of two water molecules in Figure 1 are quite different. The reason is that they have different orientations. The CP-corrected molecular interaction energy of CCSD(T)/TZVP is computed according to the following equation

$$E^{cp}_{W-W}(r) = E_{W-W}(r;s_{W1},s_{W2}) - E_{W1}(r;s_{W1},s_{gW2}) - E_{W2}(r;s_{gW1},s_{W2}) \quad (6)$$

where s_{w1} and s_{w2} are basis functions of real water molecules 1 and 2, and s_{gw1} and s_{gw2} are basis functions of "ghost" water molecules 1 and 2, respectively.

As shown in Figure 1, after CP-correction the hydrogen bond energy of the water dimer (E_{hb}=−21.17 kJ/mol and r_e=2.90 Å) calculated with CCSD(T)/TZVP of Xi'An CI is very close to the results of very accurate CCSD(T)(FULL)/IO249[39] calculations (E_{hb}=−21.15 kJ/mol and r_e=2.912 Å). The notation "CCSD(T)(FULL) /IO249" stands for coupled-cluster (CC) with single, double, and full triple excitation configurations, and IO249 is a basis set for water with 249 basis functions. In Figure 1 the hydrogen bond energy calculated by DFT is much smaller than the results of CCSD(T). No CP-correction is made for DFT calculation. If CP-correction is performed for DFT, then the H-b energy will be even smaller.

In Figure 3 we show the calculation results of the molecular interaction energy between two methane molecules using CCSD(T)/cc-pVTZ of Xi'An-CI, BLYP/TZVP of DFT, and HF/cc-pVTZ. According to the calculation results of HF/cc-pVTZ and BLYP/TZVP without dispersion correction, there is no attractive interaction between two methane molecules. If CP-correction is performed for DFT and HF calculations, the interaction energies of the methane pair will be even positive. Because both HF and DFT cannot describe dispersion interaction.

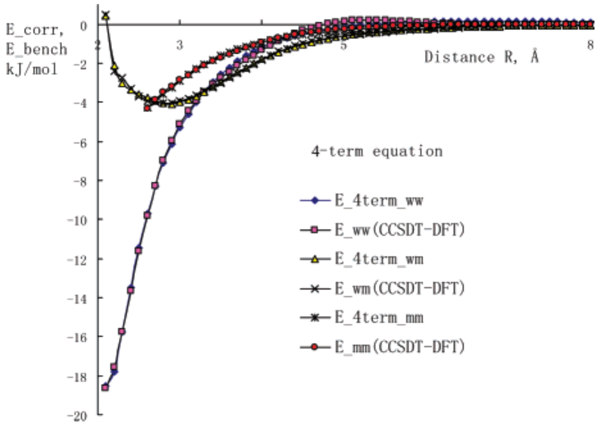


Figure 4. Comparison of the correction benchmark of interaction energy $E_{\mathrm{bench}}(r) = E_{\mathrm{CCSDT}}(r) - E_{\mathrm{DFT}}(r)$ and the empirical correction of interaction energy $E_{\mathrm{corr}}(r)$ of water−water, water−methane, and methane−methane pairs. The energy corrections $E_{\mathrm{corr}}(r)$ are obtained by the curve fitting technique to the benchmark interaction energy $E_{\mathrm{bench}}(r)$ using the L-J (4, 6-8,12) 4-term potential equation.

Table 1. Calculation Results of Water Dimer Using 4 Types of van der Waals Potential Equations

method	6-term	4-term	2-term	2-term+d
SEE	0.3913	0.3843	0.3776	0.2427
R	0.9967	0.9967	0.9967	0.9986
C_1 (OO_atr)	4.8236	2.0412	4.1205	6.0515
C_2 (OH_atr)	2.6249	−1.4331		
C_3 (HH_atr)	−1.5629	1.7016		
C_4 (OO_rpl)	−0.0537	0.2211	0.2211	−0.1110
C_5 (OH_rpl)	−2.6815			
C_6 (HH_rpl)	0.9131			

Figure 4 shows the energy differences of water−water, water−methane, and methane−methane interactions between CCSD(T) and DFT calculations

$$E_{\mathrm{bench}}(r) = \Delta E(r) = E_{\mathrm{CCSD-T}}(r) - E_{\mathrm{DFT}}(r) \qquad (7)$$

This energy difference is defined as the correction benchmark of the molecular interaction energy for DFT. The interactions of water−water, water−methane, and methane−methane represent the polar−polar, polar−nonpolar, and nonpolar−nonpolar molecular interactions, respectively. As shown in Figure 4, the energy correction benchmarks $E_{\mathrm{bench}}(r)$ of three types of molecular interactions take very different forms. In the water dimer the energy correction is negative at all distances, which can be corrected by using only attractive branch $-2d(\sigma/r)^6$ of the L-J (6-12) potential. However, the energy correction benchmark $E_{\mathrm{bench}}(r)$ of water−methane has a shallow and broad well, and it can be redressed by using both attractive and repulsive branches of the L-J (6-12) potential. The energy correction benchmark of methane−methane interaction for DFT is a negative upward flat slope. From the theoretical viewpoint, dispersion is attractive in all molecular interaction range. In this study the energy correction benchmark $E_{\mathrm{bench}}(r)$ for DFT, defined by eq 1, actually includes all differences between advanced CCSD(T) and DFT. Along with the dispersion, it may include configuration interaction, induction interaction, residual correlation, and other effects.

3.2. Correction Equation for Water Dimer. The correction method described in section 2 is applied to the water dimer. Four types of van der Waals potential equations are used in the energy corrections, and the results are summarized in Table 1. In the water dimer there are 9 atomic van der Waals pairs classified into 3 types, O−O, 4O−H, and 4H−H. The first energy correction is a 6-term equation, using L-J (6-12) potentials of three atomic pairs

$$E_{\mathrm{corr}}(r_l) = c_1\left[-2d_{\mathrm{OO}}\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^6\right] + c_2 4\left[-2d_{\mathrm{HH}}\left(\frac{\sigma_{\mathrm{HH}}}{r_l}\right)^6\right] + c_3 4\left[-2d_{\mathrm{OH}}\left(\frac{\sigma_{\mathrm{OH}}}{r_l}\right)^6\right] + c_4\left[d_{\mathrm{OO}}\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^{12}\right] + c_5 4\left[d_{\mathrm{HH}}\left(\frac{\sigma_{\mathrm{HH}}}{r_l}\right)^{12}\right] + c_6 4\left[d_{\mathrm{OH}}\left(\frac{\sigma_{\mathrm{OH}}}{r_l}\right)^{12}\right] \qquad (8)$$

In eq 8 we separate the attractive branch and the repulsive branch in the L-J equation of each atomic pair and assign them different combination coefficients, because they may have different contributions to $E_{\mathrm{corr}}(r)$. The standard estimated error (SEE) and correlation coefficient (*R*) are 0.3913 and 0.9967, respectively, as shown in Table 1. However, we find that some atomic van der Waals pairs have negative contributions to the molecular interaction energy. It is difficult to give a theoretical explanation for the “negative contributions”. In eq 8 the three attractive terms play the same role, and the role of the three repulsive terms is the same, too. In the next trial we use a 4-term equation, three attractive terms and one repulsive term

$$E_{\mathrm{corr}}(r_l) = c_1\left[-2d_{\mathrm{OO}}\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^6\right] + c_2 4\left[-2d_{\mathrm{HH}}\left(\frac{\sigma_{\mathrm{HH}}}{r_l}\right)^6\right] + c_3 4\left[-2d_{\mathrm{OH}}\left(\frac{\sigma_{\mathrm{OH}}}{r_l}\right)^6\right] + c_4 d_{\mathrm{OO}}\left[\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^{12}\right] \qquad (9)$$

As shown in Table 1, the 4-term eq 9 gives the results as good as the results of the 6-term eq 8, and the standard estimated error (SEE) of the 4-term eq 9 is even smaller than the SEE of the 6-term eq 8. Actually, the 4-term equation can be reduced to a 2-term equation, because the terms with the same exponent are one term after combination. In the next trial we use a 2-term equation, in which only the attractive term and the repulsive term of the heavy atomic pair O−O are used

$$E_{\mathrm{corr}}(r_l) = c_1\left[-2d_{\mathrm{OO}}\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^6\right] + c_4 d_{\mathrm{OO}}\left[\left(\frac{\sigma_{\mathrm{OO}}}{r_l}\right)^{12}\right] \qquad (10)$$

Surprisingly, the results of the 2-term eq 10 are better than that of both the 6-term and 4-term equations, as shown in Table 1.

Careful observation of Figure 4 reveals that the energy correction $E_{\mathrm{corr}}(r)$ of the water dimer turns soft at short distances ∼2.2 Å, and the largest errors are in the short

Table 2. Calculation Results of Water−Methane Using 4 Types of van der Waals Potential Equations

method	8_term	2_term	2_term+d	2-term+d*
SEE	1.0583	1.0020	0.6404	0.2405
R	0.8221	0.8221	0.9214	0.9836
C_1 (CO_atr)	−0.4232	1.1849	2.5340	4.8243
C_2 (OH_atr)	−0.5870			
C_3 (CH_atr)	1.2659			
C_4 (HH_atr)	0.8645			
C_5 (CO_rpl)	−0.0007	0.0897	0.1982	0.9850
C_6 (OH_rpl)	0.0191			
C_7 (CH_rpl)	0.1422			
C_8 (HH_rpl)	0.0513			

distance. For a better energy correction, we introduce a damping function[40] on the van der Waals potential

$$E_{\text{corr}}(r_l) = c_1\ f_{\text{dmp}}(r_l;d_{\text{OO}})\left[-2d_{\text{OO}}\left(\frac{\sigma_{\text{OO}}}{r_l}\right)^6\right] + c_4\ f_{\text{dmp}}(r_l;d_{\text{OO}})d_{\text{OO}}\left[\left(\frac{\sigma_{\text{OO}}}{r_l}\right)^{12}\right] \quad (11)$$

The form of damping function $f_{\text{dmp}}(r;\ d_{\text{OO}})$ is as follows

$$f_{\text{dmp}}(r;d_{\text{OO}}) = 1.0 - e^{-(r-d_{\text{OO}}-r_0)} \quad (12)$$

In eq 12 r_0 is a constant, and for the water dimer we use $r_0 = 1.75$. As shown in Table 1 the calculation results of eq 11 are much better than the results of the other three equations. The damping function is a purely empirical one, which makes the van der Waals potential softer in the short distance and not much change in the other distances.

3.3. Correction Equation for Water−Methane. Following the same procedure used for the water dimer, we calculate the interaction energy correction functions for the water−methane. In the molecular pair H_2O-CH_4 there are 15 atomic vdW pairs (C−O, 4O−H, 2C−H, and 8H−H), which are classified into four types of van der Waals potentials. For the energy correction of the water−methane interaction we use four types of van der Waals potential equations. The first equation is 8-term L-J (6-12) potentials, including 4 attractive terms and 4 repulsive terms of the four types of atomic pairs. The second equation is a 2-term equation, only the L-J (6-12) potential of the heavy atomic pair (C−O) is used. In Table 2 we find that the results of the 2-term equation are better than the results of the 8-term equation. The third equation (2-term+d) is built by applying damping functions on the 2-term equation. The fourth equation (2-term+d*) uses the same equation as (2-term+d), but the two exponents in the L-J potential are optimized, 4 for the attractive branch and 8 for the repulsive branch.

In Table 2 we find that the damping functions improve the energy correction remarkably for the 2-term equation. The L-J (4-8) potential gives much better results (*R*=0.9936 and SEE=0.2405) than the L-J (6-12) potential does for the water−methane interaction. The optimization of exponents of the L-J potential in the 2-term equation is necessary for this polar−nonpolar molecular interaction pair.

Table 3. Calculation Results of Methane−Methane Using 4 Types of van der Waals Potential Equations

method	6_term	2_term	2_term+d	2-term+d[a]
SEE	0.0662	0.0667	0.0639	0.1442
R	0.9982	0.9981	0.9982	0.9906
C_1 (CC_atr)	0.1814	6.0759	6.3352	2.6400
C_2 (HH_atr)	1.4069			
C_3 (CH_atr)	1.6612			
C_4 (CC_rpl)	4.2606	3.3713	2.5648	−4.8178
C_5 (HH_rpl)	5.4061			
C_6 (CH_rpl)	−1.7799	--	--	--

[a] Use the same potential equation as column 4 (2_term+d), the two exponents in L-J are 4 for the attractive branch and 8 for the repulsive branch.

Table 4. Four-Term L-J (4,6-8,12) Potential Equation for Water−Water, Water−Methane, and Methane−Methane Interactions

mol	R	SEE	C_1(atr, 4)	C_2(atr, 6)	C_3(rpl, 8)	C_4(rpl, 12)
w−w	0.9989	0.2244	−3.4933	11.9112	5.0585	−0.3427
w−m	0.9888	0.2127	2.7248	3.0574	0.0232	−0.0616
m−m	0.9997	0.0274	−4.1609	20.1582	24.5779	−2.6374

3.4. Correction Equation for Methane−Methane. The molecular pair methane−methane represents the nonpolar−nonpolar molecular interaction, in which dispersion makes the main contribution. The same correction procedure is applied to the methane−methane interaction energy correction for DFT calculations. Because the dispersion interaction is very weak in the nonpolar molecular pair, we use a higher basis set in the benchmark calculation, CCSD(T)/cc-pVTZ, and a lower basis set in the DFT calculation, BLYP/TZVP. Therefore, the correction for molecular interaction energy may reach the results of the larger basis set using only the smaller basis set in DFT calculations. The calculation results are summarized in Table 3. In Table 3 the 2-term+d L-J (6-12) potential equation gives the best correction for the methane−methane interaction. And the L-J (4-8) potential equation 2-term+d* does not give better results than the L-J (6-12) equation 2-term+d does for this molecular pair. It is best to use a 4-term L-J (4,6-8,12) potential equation for the methane hydrate simulation, two attractive terms with exponents 4 and 6 and two repulsive terms with exponents 8 and 12

$$E_{\text{corr}}(r_l) = c_1\ f_{\text{dmp}}(r_l;d)\left[-d2\left(\frac{\sigma}{r_l}\right)^4\right] + c_2\ f_{\text{dmp}}(r_l;d)\left[-d2\left(\frac{\sigma}{r_l}\right)^6\right] + c_3\ f_{\text{dmp}}(r_l;d)\left[d\left(\frac{\sigma}{r_l}\right)^8\right] + c_3\ f_{\text{dmp}}(r_l;d)\left[d\left(\frac{\sigma}{r_l}\right)^{12}\right] \quad (13)$$

The 4-term equations give the best energy corrections for all three molecular pairs, water−water, water−methane, and methane−methane, as shown in Table 4 and Figure 4.

3.5. Check Correction Effect in Water Pentamer. The basics structural unit of methane hydrate is a dodecahedron cell consisting of 20 water molecules and a methane molecule in the center. Each face of the dodecahedron cell is a water

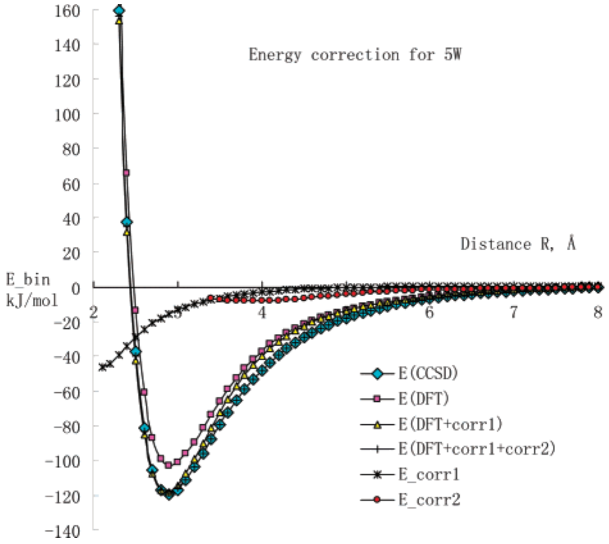


Figure 5. The dispersion-corrected water−water binding energy in the planar water pentamer for BLYP/TZVP calculations and comparison with CCSD(T)/TZVP. The 4-term dispersion-correction equation is used derived from the water dimer. E_corr1 is the energy correction for hydrogen-bonded water pairs in pentamer, and E_corr2 is the energy correction for non-hydrogen-bonded water pairs in pentamer, respectively.

pentamer, which is a regular pentagon of hydrogen-bonded five water molecules. In this example the molecular interaction energy correction equation $E_{\mathrm{corr}}(r)$ obtained from the water dimer is applied to correct the DFT BLYP/TZVP calculation of the molecular interaction energy in water pentamer and compared with the CCSD(T)/TZVP calculation. The dispersion-corrected water−water binding energy in the water pentamer is expressed by the following two formulas

$$E^{5\mathrm{w}}_{\mathrm{bin1}}(r) = E^{\mathrm{DFT}}_{\mathrm{bin}}(r) + \frac{5}{2}E^{\mathrm{ww}}_{\mathrm{corr1}}(r) \qquad (14)$$

$$E^{5\mathrm{w}}_{\mathrm{bin2}}(r) = E^{\mathrm{DFT}}_{\mathrm{bin}}(r) + \frac{5}{2}[E^{\mathrm{ww}}_{\mathrm{corr1}}(r) + E^{\mathrm{ww}}_{\mathrm{corr2}}(r)] \qquad (15)$$

where $E^{\mathrm{ww}}_{\mathrm{corr1}}(r)$ is the energy correction for 5 hydrogen-bonded water pairs (w1−w2, w2−w3, w3−w4, w4−w5, and w5−w1) in the pentamer, and $E^{\mathrm{ww}}_{\mathrm{corr2}}(r)$ is the energy correction for 5 non-hydrogen-bonded water pairs (w1−w3, w1−w4, w2−w4, w2−w5, and w3−w5) in the pentamer, respectively. The energy corrections are calculated using the 4-term eq 13, and the coefficients listed in Table 4 are derived from the water dimer. In eqs 14 and 15 $E^{\mathrm{DFT}}_{\mathrm{bin}}(r)$ is the total binding energy in the water pentamer obtained by using the DFT calculation without dispersion

$$E^{\mathrm{DFT}}_{\mathrm{bin}}(r) = E^{\mathrm{DFT}}_{\mathrm{total}}(r) - E^{\mathrm{DFT}}_{\mathrm{total}}(\infty) \qquad (16)$$

The dispersion-corrected total water binding energies in the water pentamer are shown in Figure 5. If we only use the energy correction $E^{\mathrm{ww}}_{\mathrm{corr1}}(r)$, which is the dispersion correction for hydrogen-bonded water pairs in the pentamer, there is a ∼7 kJ mol^{-1} gap from 3.8 Å to 4.5 Å between the dispersion-corrected curve and the curve of CCSD(T)/TZVP, because the energy correction $E^{\mathrm{ww}}_{\mathrm{corr1}}(r)$ does not include the dispersion compensation for cointeractions from non-hydrogen-bonded water molecules (w1−w3, w1−w4, w2−w4, w2−w5, and w3−w5). After further correction using $E^{\mathrm{ww}}_{\mathrm{corr2}}(r)$, which is the dispersion correction for non-hydrogen-bonded water pairs in the pentamer, the dispersion-corrected water−water binding energy curve of BLYP/TZVP fits the CCSD(T)/TZVP curve very well over the full interaction range.

4. Discussion and Conclusion

Because of the theoretical limitations, DFT in its usual local and gradient approximations fails to describe the molecular dispersion interaction correctly, and dispersion-correction is necessary. The empirical correction methods developed in the literature usually only uses the attractive branch of van der Waals equations and fail to correct the dispersion interaction at all interaction distances and for different molecular pairs. The empirical correction method proposed in this study uses both the attractive and repulsive branches of existing van der Waals potential equations and atomic parameters, which gives good energy corrections at all interaction distances. The molecular pairs water−water, water−methane, and methane−methane represent three types of molecular interactions: polar−polar, polar−nonpolar, and nonpolar−nonpolar interaction, respectively. Based on the calculation results of these three examples and the application to the water pentamer, we offer the following conclusions: (1) Because the dispersion behavior of the molecular interaction energy between polar−polar, polar−nonpolar, and nonpolar−nonpolar molecules is very different, it is not appropriate to use only the attractive branch of the L-J potential to redress the molecular dispersion interaction in DFT. Both attractive and repulsive branches are needed. (2) More potential terms of atomic vdW pairs do not certainly give better corrections for dispersion interactions. Use of only the vdW potential of a heavy atomic pair, for water−water it is the O−O pair, for water−methane it is the C−O pair, and for methane−methane it is the C−C pair, gives better results than the use of all the atomic vdW pairs. (3) For the polar−polar and nonpolar−nonpolar molecular interactions, the L-J (6-12) potential equation gives a good description for the molecular interaction energy correction; however, for the polar−nonpolar molecular interaction, the L-J (4-8) potential equation works better than the L-J (6-12) equation does. The 4-term L-J (4,6-8,12) potential equation, two attractive terms with the exponents 4 and 6 and two repulsive terms with the exponents 8 and 12, gives the best descriptions for all three types of molecular interactions. (4) A damping function is essential for a good correction to molecular dispersion interactions at a short distance.

The combination coefficients of vdW potential terms, optimized in a training set, can be used for the molecular interaction energy correction in a molecular family. This empirical and general correction method is efficient for dispersion interactions as well as configuration interaction, induction interaction, residual correlation, and other effects. It also can be used to reach the results of a larger basis set using only a smaller basis set. It can be performed independently outside the DFT software package.

Acknowledgment. This work is supported by the Chinese National Science Foundation (NSFC) under contract number 20373048 and by the Hainan Natural Science Foundation (HNSF) under contract number 60505. The authors thank Prof. Zhen-Yi Wen and graduate student An-Yang Li (Northwest University, China) for the use of their advanced software package Xi'An-CI.

References

(1) London, F. Theory and systematic of molecular forces. *Phys. Z.* **1930**, *63*, 245−267.

(2) Lein, M.; Dobson, J. F.; Gross, E. K. U. Towards the description of van der Waals interactions within density-functional theory. *J. Comput. Chem.* **1999**, *20*, 12−22.

(3) Dobson, J. F.; Das, M. P. *Topics in Condensed Matter Physics*; Nova: New York, 1994; pp 121−142.

(4) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463−1473.

(5) Chakarova, S. D.; Schröder, E. van der Waals interactions of polycyclic aromatic hydrocarbon dimmers. *J. Chem. Phys.* **2005**, *122*, 054102.

(6) Cybulski, S. M.; Seversen, C. E. Critical examination of the supermolecule density functional theory calculations of intermolecular interactions. *J. Chem. Phys.* **2005**, *122*, 014117.

(7) Xu, X.; Goddard, W. A. The extended Perdew-Burke-Ernzerhof functional with improved accuracy for thermodynamic and electronic properties of molecular systems. *J. Chem. Phys.* **2004**, *121* 4068−4082.

(8) Tao, J.; Perdew, P.; Staroverov, S.; Scuseria, G. Climbing the Density Functional Ladder: Nonempirical Meta−Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(9) Tao, J.; Perdew, P. Test of a nonempirical density functional: Short-range part of the van der Waals interaction in rare-gas dimers. *J. Chem. Phys.* **2005**, *122*, 114102.

(10) Hesselmann, A.; Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **2003**, *367*, 778−784.

(11) Misqitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion Energy from Density-Functional Theory Description of Monomers. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(12) Osinga, V. P.; van Gisbergen, S. J. A.; Snijders, J. G.; Baerends, E. J. Density functional results for isotropic and anisotropic multipole polarizabilities and C_6, C_7, and C_8 Van der Waals dispersion coefficients for molecules. *J. Chem. Phys.* **1997**, *106*, 5091−5101.

(13) Adamovic, I.; Gordon, M. S. Dynamic polarizability, dispersion coefficient C6, and dispersion energy in the effective fragment potential method. *Mol. Phys.* **2005**, *103*, 379−387.

(14) Andersson, Y.; Langreth, D. C.; Lundqvist, B. I. van der Waals Interactions in Density-Functional Theory. *Phys. Rev. Lett.* **1996**, *76*, 102−105.

(15) Dobson, J. F. Prospects for a van der Waals density functional. *Int. J. Quantum Chem.* **1998**, *69*, 615−618.

(16) Sato, T.; Tsuneda, T.; Hirao, K. A density-functional study on π-aromatic interaction: Benzene dimer and naphthalene dimer. *J. Chem. Phys.* **2005**, *123*, 104307.

(17) Ortmann, F. F.; Bechstedt, F. Semiempirical van der Waals correction to the density functional description of solids and molecular structures. *Phys. Rev. B* **2006**, *73*, 205101.

(18) Lacks, D. J.; Gordon, R. G. Pair interactions of rare-gas atoms as a test of exchange-energy-density functionals in regions of large density gradients. *Phys. Rev. A* **1993**, *47*, 4681−4690.

(19) Adamo, C.; Barone, V. Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: The *m*PW and *m*PW1PW models. *J. Chem Phys.* **1998**, *108*, 664−675.

(20) Xin, X.; Goddard, W. A., III The X3LYP extended density functional for accurate descriptions of nonbond interactions, spin states, and thermochemical properties. *PNAS* **2004**, *101*, 2673−2677.

(21) Kurita, N.; Inoue, H.; Sekino, H. Adjustment of Perdew-Wang exchange functional for describing van der Waals and DNA base-stacking interactions. *Chem. Phys. Lett.* **2003**, *370*, 161−169.

(22) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duarte, H. A. An efficient *a posteriori* treatment for dispersion interaction in density-functional-based tight binding. *J. Chem. Theory Comput.* **2005**, *1*, 841−848.

(23) Valdes, H.; Sordo, J. A. Ab initio and DFT studies on van der Waals trimers: The OCS·(CO2)2 complexes. *J. Comput. Chem.* **2002**, *23*, 444−455.

(24) Jensen, F. *Introduction to Computational Chemistry*; Odense University: Odense, Denmark, 1999.

(25) Tran, F.; Weber, J.; Wesolowski, T. A.; Cheikh, F.; Ellinger, Y.; Pauzat, F. Physisorption of molecular hydrogen on polycyclic aromatic hydrocarbons: a theoretical study. *J. Phys. Chem. B* **2002**, *106*, 8689−8696.

(26) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. Direct derivation of van der Waals force field parameters from quantum mechanical interaction energies. *J. Phys. Chem. B* **2003**, *107*, 9601−9609.

(27) Dion, M.; Rydberg, H.; Schröder, E.; Landgreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(28) von Lilienfeld, O. A.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. Optimization of Effective Atom Centered Potentials for London Dispersion Forces in Density Functional Theory. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(29) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. *J. Chem. Phys.* **2001**, *114*, 5149−5155.

(30) Rappé, A. K.; Goddard, W. A., III A new model for charge distributions in molecular systems. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(31) Wang, Y. B.; Zhai, G. H.; Suo, B.; Gan, Z. T.; Wen, Z. Y. Hole−particle correspondence in CI calculations. *Chem. Phys. Lett.* **2003**, *375*, 134−140.

(32) Wang, Y. B.; Wen, Z. Y.; Zhang, Z. Y.; Du, Q. S. New realization of loop driving CI. *J. Comput. Chem.* **1992**, *13*, 187−198.

(33) Gan, Z. T.; Su, K. H.; Wang, Y. B.; Wen, Z. Y. CI Benchmark Calculations on PC. *J. Comput. Chem.* **2001**, *22*, 560−563.

(34) Suo, B.; Zhai, G. H.; Wang, Y. B.; Wen, Z. Y.; Hu, X. Q.; Li, L. M. Parallelization of MRCI based on hole-particle symmetry. *J. Comput. Chem.* **2005**, *26*, 88−96.

(35) Kvenvolden, K. A. Gas hydrates-geological perspective and global change. *Rev. Geophys.* **1993**, *31*, 173−188.

(36) Cao, Z.; Tester, J. W.; Trout, B. L. Computation of the methane−water potential energy hypersurface via *ab initio* methods. *J. Chem. Phys.* **2001**, *115*, 2550−2559.

(37) Sloan, E. D. Introductory overview: hydrate knowledge development. *Am. Mineral.* **2004**, *89*, 1155−1161.

(38) Klopper, W.; vanDuijneveldt-van de Rijdt, J. G. C. M.; vanDuijneveldt, F. B. Computational determination of equilibrium geometry and dissociation energy of the water dimer. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227−2234.

(39) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(40) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. Dispersion corrections to density functionals for water aromatic interactions. *J. Chem. Phys.* **2004**, *120*, 2693−2699.

CT700026D

JCTC Journal of Chemical Theory and Computation

Weakly Bonded Complexes of Aliphatic and Aromatic Carbon Compounds Described with Dispersion Corrected Density Functional Theory

Enrico Tapavicza, I-Chun Lin, O. Anatole von Lilienfeld,[‡] Ivano Tavernelli,
Maurício D. Coutinho-Neto,[§] and Ursula Rothlisberger*

Laboratoire de Chimie et Biochimie Computationelle, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

Received March 1, 2007

Abstract: Interaction energies and structural properties of van der Waals complexes of aliphatic hydrocarbons molecules and crystals of aromatic hydrocarbon compounds are studied using density functional theory augmented with dispersion corrected atom centered potentials (DCACPs). We compare the performance of two sets of DCACPs, (a) DCACP-MP2, a correction for carbon only, generated using MP2 reference data and a penalty functional that includes only equilibrium properties and (b) DCACP-CCSD(T), a set that has been calibrated against CCSD(T) reference data using a more elaborate penalty functional that explicitly takes into account some long-range properties and uses DCACP corrections for hydrogen and carbon atoms. The agreement between our results and high level ab initio or experimental data illustrates the transferability of the DCACP scheme for the gas and condensed phase as well as for different hybridization states of carbon. The typical error of binding energies for gas-phase dimers amounts to 0.3 kcal/mol. This work demonstrates that only one DCACP per element is sufficient to correct for weak interactions in a large variety of systems, irrespective of the hybridization state.

1. Introduction

London dispersion forces are fundamental for the proper description of chemical and biological systems such as molecular liquids and crystals, proteins, and nucleic acids. Since these forces are purely due to electron correlation effects, they are difficult and computationally expensive to capture using conventional quantum chemical approaches. Kohn−Sham density functional theory (DFT)[1,2] is a very popular first principles electronic structure method due to its relatively high accuracy and low computational cost. DFT is in principle exact and should correctly describe the London dispersion forces if the true exchange-correlation (xc) functional were known. However, most of the conventional local approximations to xc-functionals are unable to describe dispersion reliably.[3−6] To account for these forces in DFT, several remedies have been proposed and implemented, for example, prepartitioning of the electron density,[7] solving the adiabatic connection formula for the long-range part of the interaction energy,[8] using symmetry adapted perturbation theory,[9,10] developing more sophisticated approximations to the xc-potential,[11,12] or adding an explicit dispersion term with a C_6 coefficient determined either empirically[5,13,14] or generated by the instantaneous dipole moment of the exchange hole.[15−17]

Recently, dispersion corrected atom centered potentials (DCACPs) have been introduced to account for London dispersion forces within self-consistent DFT calculations and have been shown to perform remarkably well for several cases.[18−21] Here, we systematically probe the transferability of DCACPs to various hybridization states other than the one used in the calibration and test their performance for

* Corresponding author e-mail: ursula.roethlisberger@epfl.ch, http://lcbcpc21.epfl.ch.

[‡] Current address: Department of Chemistry, New York University, New York, NY 10003.

[§] Current address: Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, Rua Santa Adélia, 166 Santo André, São Paulo, Brazil BR-09.210-170.

10.1021/ct700049s CCC: $37.00
Published on Web 08/08/2007

some solid-state systems. In addition we compare the performance of two generations of DCACPs, which differ in their reference method and in their calibration procedure. The study focuses on carbon which, in the case of biomolecular classical force fields, is the element that requires the largest number of different atom types, i.e., it requires very different Lennard-Jones coefficients for different hybridization states.[22] In stark contrast, in the DCACP approach only one carbon DCACP is employed for all hybridization states. The question we pose in this study is with which accuracy this large variety of systems can be described in spite of the limitation to a single effective potential. We study several weakly bonded complexes of carbon compounds using two sets of DCACPs previously calibrated against MP2 (DCACP-MP2[18]) and CCSD(T) (DCACP-CCSD(T)[23]) references, respectively. The performance of DCACPs for different hybridization states is investigated by studying a series of aliphatic hydrocarbon C_2H_n ($n = 2,4,6$) homo dimers as well as the methane−ethene hetero dimer in the gas phase. Moreover, we assess the transferability of DCACPs from the gas to the condensed phase for crystals of benzene and graphite. Our results suggest that DCACPs offer a cheap, pragmatic way to include the effect of London dispersion forces in DFT calculations that is strongly transferable, i.e., once calibrated, the same DCACP for each element can be employed in different hybridization states without additional tuning.

2. Computational Details

All DFT calculations were carried out with the plane wave code CPMD,[24,25] the xc-functional BLYP,[26,27] and pseudopotentials of Goedecker et al.[28] Two generations of DCACPs, DCACP-MP2[18] and DCACP-CCSD(T),[23] were employed in this study. These two sets differ in the following three respects: (a) the chosen reference during the calibration stage (MP2 for DCACP-MP2 and CCSD(T) for DCACP-CCSD(T)); (b) a ‘united atom’ approach for DCACP-MP2, i.e., only carbon atoms are corrected by a DCACP but not hydrogens, whereas DCACPs for both hydrogen and carbon have been used in DCACP-CCSD(T); and (c) for DCACP-MP2, the penalty functional included only the energy and forces at equilibrium distance,[18] whereas a new penalty functional was introduced to improve the description of the intermolecular midrange to long-range behavior in the calibration of the DCACP-CCSD(T) set.[29] The DCACPs were calibrated to correct the BLYP xc functional. Further details of the calibration of the two generations of DCACPs can be found in refs 18 and 29. For comparison purposes, standard BLYP calculations without dispersion correction were also carried out.

In order to calculate the interaction energy of the gas-phase dimers, the monomer geometry was first optimized using the pure BLYP functional. The BLYP optimized monomer geometries were then used to construct the dimers in all calculations. The dimer geometries of ethane, ethene, and ethyne shown in Figure 1 were placed in an isolated cell with dimensions 25 × 10 × 10 Å^3. A plane wave cutoff of 100 Ry was applied. For the methane−ethene complex, a cell measuring 26 × 13 × 12 Å ^{3} and a plane wave cutoff of 150 Ry was used. In all cases the intermolecular distance was varied along the *x*-axis. The orientation of the methane−ethene complex is indicated by the coordinate system in Figure 1. The interaction energies, defined as $E_{\mathrm{AB}}^{\mathrm{int}} = E_{\mathrm{AB}}^{\mathrm{total}} - E_{\mathrm{A}}^{\mathrm{total}} - E_{\mathrm{B}}^{\mathrm{total}}$, were calculated at various distances. In order to gauge the accuracy of the dispersion corrected DFT calculations, the corresponding MP2 and CCSD(T) calculations have been carried out using the ab initio program packages GAUSSIAN03[30] or MOLPRO.[31]

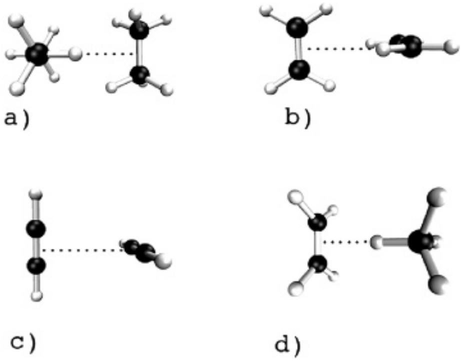


Figure 1. Geometries of the hydrocarbon dimers: (a) $(C_2H_6)_2$, (b) $(C_2H_4)_2$, (c) $(C_2H\ _2)_2$, and (d) $C_2H_4\cdots CH_4$. The intermolecular distance refers to the distance between the midpoints of the C−C bonds of the two molecules in the case of C_2H_n. In case of the $CH_4-C_2H_4$ complex, it is defined as the distance between the midpoint of the C−C bond in ethene and the hydrogen atom of methane pointing toward the ethene molecule.

The MP2 and CCSD(T) calculations were done at the same intermolecular distances with the monomers fixed at the BLYP optimized geometry. In both sets of calculations the aug-cc-pVTZ basis set and counterpoise corrections[32,33] were employed.

Calculations for the benzene crystal were based on the experimentally determined space group P_{bca}.[34,35] Total energy calculations were performed by isotropically varying the volume of a unit cell, whereas the internal coordinates of the four monomers were allowed to relax. The total energy of the monomer ($E_{\mathrm{mono}}^{\mathrm{total}}$) was calculated in an isolated cubic supercell with an edge of 12 Å. The cohesive energy of the benzene crystal was then calculated using $E^{\mathrm{cohesive}} = [E_{\mathrm{cryst}}^{\mathrm{total}} - 4{*}E_{\mathrm{mono}}^{\mathrm{total}}]/4$. A plane wave cutoff of 200 Ry was enough to ensure converged results with respect to the number of plane waves caused by the variation of the unit cell volume.[36]

For the determination of the interlayer binding (IB) energy of graphite, a monomeric graphene sheet containing 32 carbon atoms was geometry optimized under two-dimensional periodic boundary conditions (PBC) along the graphene plane (*xy*). Thereafter, the total energy of 3 sheets (96 atoms) in ABA and ABC packing was calculated with various interlayer distances while keeping the internal geometry of the layers rigid. A cell with dimensions 9.84 × 8.52 × 24.0 Å^3 (two-dimensional PBC along *x* and *y*) and a plane wave cutoff of 100 Ry were used. The IB energy per atom was calculated as E_{IB} (ABC)/atom = [E_{total} (ABC) − 3*E_{total} (A)]/64, where atoms in the first and third layer were counted as ‘half’ to allow for direct comparison to the corresponding result for the crystal.

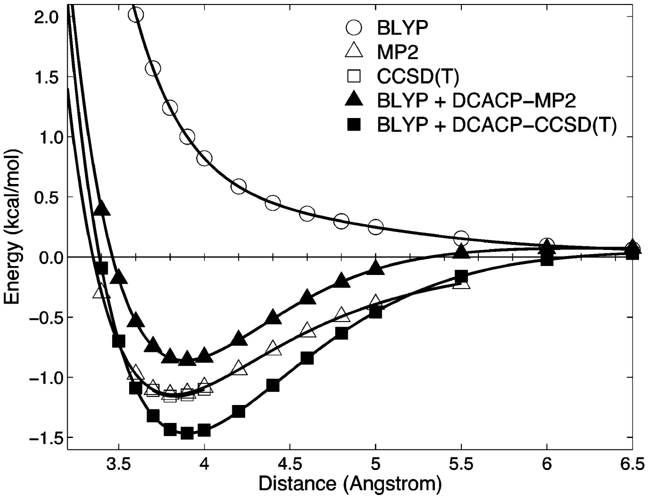


Figure 2. Interaction energy [kcal/mol] of the ethane dimer as a function of the intermolecular distance [Å]. For the definition of the intermolecular distance see the caption of Figure 1.

Table 1. Summary of the Calculations for the Gas-Phase Dimers[a]

	DCACP-MP2		MP2		Δ	
	E^{int}	r_{eq}	E^{int}	r_{eq}	E^{int}	r_{eq}
$(C_2H_6)_2$	−0.859	3.88	−1.145	3.82	0.286	0.06
$(C_2H_4)_2$	−1.287	3.79	−1.476	3.72	0.189	0.07
$(C_2H_2)_2$	−0.523	3.83	−0.295	3.98	−0.228	−0.15
$C_2H_4{\cdots}CH_4$	−0.385	3.08	−0.520	3.12	0.135	−0.04
MD					0.096	−0.02
MAD					0.209	0.08

	DCACP-CCSD(T)		CCSD(T)		Δ	
	E^{int}	r_{eq}	E^{int}	r_{eq}	E^{int}	r_{eq}
$(C_2H_6)_2$	−1.464	3.90	−1.158	3.83	−0.306	0.07
$(C_2H_4)_2$	−1.301	3.94	−1.390	3.75	0.089	0.19
$(C_2H_2)_2$	−0.135	4.15	−0.154	4.17	0.019	−0.02
$C_2H_4{\cdots}CH_4$	−0.547	3.09	−0.50[b]	3.12[b]	−0.047	−0.02
MD					0.061	0.06
MAD					0.115	0.08

[a] Equilibrium distances (r_{eq} [Å]) and binding energies at equilibrium distance (E^{int} [kcal/mol]) are shown together with the deviations (Δ) from the reference of the corresponding level of theory. The definitions of the labels 'MD' and 'MAD' are given in the text. [b] Values are taken from ref 38.

3. Results

The resulting potential energy curves of the DFT calculations for the hydrocarbon dimers are shown in Figures 2−5. The reference values of the equilibrium distances of the reference methods are also shown. As observed in refs 5 and 37, uncorrected BLYP results in a purely repulsive interaction energy for all vdW dimers studied and converges to zero in the dissociative limit (Figures 2−5). The equilibrium distances and binding energies for the dimers obtained with the two sets of DCACPs are summarized in Table 1. Mean deviations (MD), defined as $(1/N)(x^{DCACP} - x^{ref})$, and mean absolute deviations (MAD), defined as $(1/N)|x^{DCACP} - x^{ref}|$, for characteristic properties are also given. Here, N is the number of test systems, and x^{DCACP} and x^{ref} are the equilibrium distance or the binding energy at the equilibrium distance of the DCACP calculations and the reference calculations. Always the same level of theory which was used in the calibration for the DCACPs was chosen as reference. It should be mentioned that except for the ethyne dimer, MP2/aug-cc-pVTZ and CCSD(T)/aug-cc-pVTZ reference values differ by less than 0.1 kcal/mol for the binding energy at equilibrium distance and predict the same equilibrium distances within 0.05 Å (Table 1, Figures 2, 3, and 5).

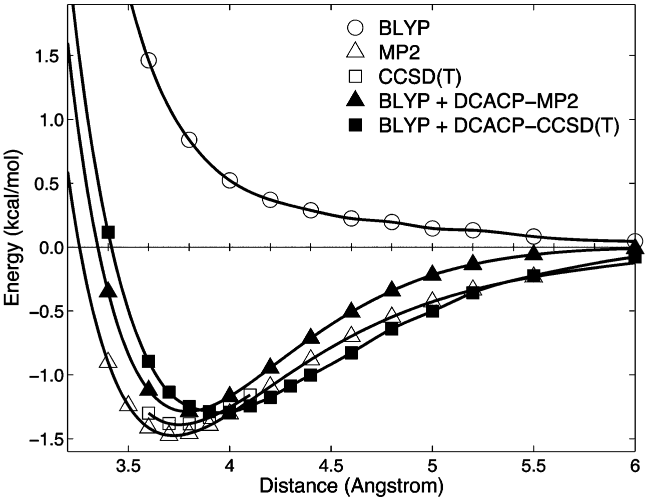


Figure 3. Interaction energy [kcal/mol] of the ethene dimer as a function of the intermolecular distance [Å]. For the definition of the intermolecular distance see the caption of Figure 1.

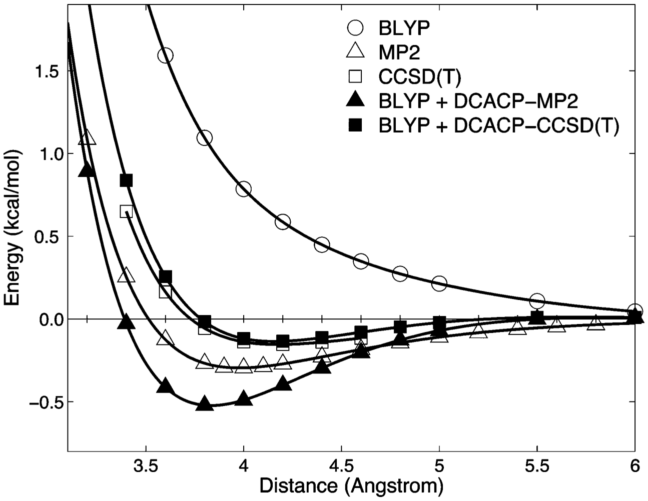


Figure 4. Interaction energy [kcal/mol] of the ethyne dimer as a function of the intermolecular distance [Å]. For the definition of the intermolecular distance see the caption of Figure 1.

The DCACP-MP2 corrected DFT calculations predict equilibrium distances within 0.08 Å of the corresponding MP2 calculations for all dimers except for $(C_2H_2)_2$ where a maximal deviation of 0.15 Å is found. However, ethyne dimer exhibits a very flat potential energy surface (PES) around the minimum (Figure 4), and the deviation of 0.15 Å corresponds to a maximal variation in the energy of only

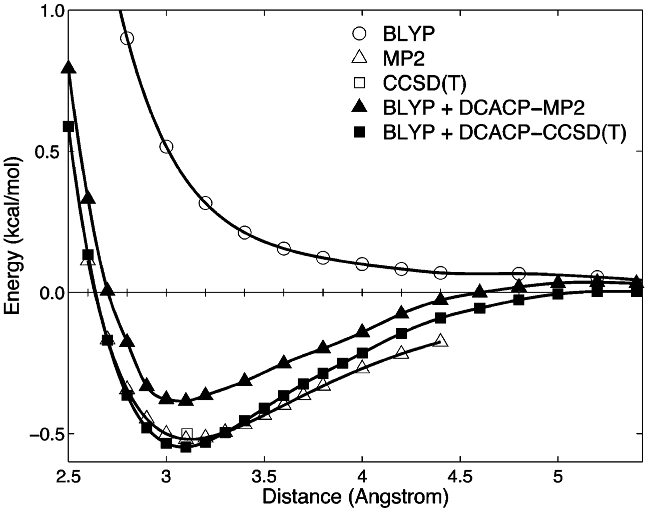


Figure 5. Interaction energy [kcal/mol] of the methane−ethene complex as a function of the intermolecular distance [Å]. The reference value was obtained with basis set extrapolated CCSD(T) calculations.[38] For the definition of the intermolecular distance see the caption of Figure 1.

0.05 kcal/mol, therefore, a higher tolerance in the prediction of the equilibrium distances is acceptable.

The DCACP-MP2 MAD for interaction energies amounts to 0.209 kcal/mol at the equilibrium distance, where the largest deviation of 0.286 kcal/mol is found in the case of $(C_2H_6)_2$.

Using the DCACP-CCSD(T) set, intermolecular equilibrium distances are predicted with similar accuracy with a MAD of 0.08 Å. The largest deviation of 0.19 Å is exhibited by the ethene dimer, which is larger than in the corresponding DCACP-MP2 result. The larger deviation in the prediction of the equilibrium distance in the case of ethene can be explained by the use of the second parameter in the penalty functional in the calibration of the DCACP-CCSD(T), which accounts for a better description of the slope of the PES around the minimum. The DCACP-MP2 predict the minimum distance of the ethene dimer with smaller deviation from the reference, but, compared to MP2, the resulting PES is too steep around the minimum and approaches the asymptotic limit too fast. In contrast the shape of DCACP-CCSD(T) PES is flatter, and the slope is similar to the CCSD(T) reference curve. The binding energy of the ethene dimer is very well reproduced with an error of only 0.089 kcal/mol.

In the case of the hetero dimer (methane−ethene) MP2 and CCSD(T) both predict similar interaction energies and equilibrium distances (Figure 5). Both DCACP sets reproduce very well the equilibrium distance (Table 1), but the interaction energy is more accurately predicted by the DCACP-CCSD(T). This is probably due to the use of an extra DCACP for hydrogen in the case of DCACP-CCSD(T), which is especially important for the case of the methane−ethene dimer, where a hydrogen of methane is pointing toward the ethylene molecule (Figure 1).

On average, the same accuracy is found between DCACP-MP2 and DCACP-CCSD(T) calculations with respect to the reference equilibrium distances which are predicted with an MAD of 0.08 Å, while the MAD of the DCACP-CCSD(T) interaction energies of 0.115 kcal/mol is slightly smaller than in the case of DCACP-MP2.

The better correlation between dispersion corrected DFT and the reference in the case of DCACP-CCSD(T) is presumably due to the separate calibration of hydrogen and carbon and the improved midrange and long-range behavior.[29] However, by construction in both cases, the DCACP interaction energies approach the dispersion uncorrected BLYP values because at large distances the effect of the DCACPs vanishes.[29] For the systems studied here, for distances typically larger than 5.5 Å the DCACPs correction approaches zero, and the corrected and uncorrected potentials converge to the same value. In general in all the discussed dimer interaction curves it can be seen that the DCACP-CCSD(T) PES approach the asymptotic limit (BLYP PES) more slowly than the first generation of DCACPs, which are based solely on a equilibrium penalty functional.

Comparing the deviations of the two DCACPs from the corresponding reference values for the four different dimers (Table 1), it can be seen that they differ in magnitude and that also the sign of the deviation can be different (e.g., Figure 2). For example, in case of the ethane dimer the reference values of MP2 and CCSD(T) are almost equal, but the DCACP-MP2 underestimates the binding energy while the DCACP-CCSD(T) overestimates it. This can be explained because of the different calibration procedures and because of the fact that in the case of DCACP-MP2 only the DCACP for carbon was used, while in DCACP-CCSD(T) simultaneous DCACPs for carbon and hydrogen atoms were included. The latter is especially important for cases in which the carbon−hydrogen ratio differs from the one of the calibration system. Therefore it cannot be expected that the deviation from the reference curve has the same sign and magnitude for the two generations of DCACPs.

To further probe the transferability of DCACPs to different hybridization states and to assess their long-range behavior, we also have studied the molecular crystal of benzene. With DCACP-MP2, a very shallow minimum (6.2 kcal/mol) is predicted at a ratio of 0.91 between the theoretical and the experimental density (ρ/ρ_{exp}) (Figure 6). Experimental cohesive energies for the benzene crystal range from values of 9.0−12.5 kcal/mol.[39−42] A better agreement with experimental results is obtained when DCACP-CCSD(T) are used. In this case the cohesive energy is evaluated as 12.1 kcal/mol, within the spread of experimental values. In addition, the predicted equilibrium density of the unit cell is at 0.95 times the experimental density, slightly closer to the experiment than the unit cell density predicted by DCACP-MP2. The underestimation of the binding energy of DCACP-MP2 compared to DCACP-CCSD(T) is probably due to its worse description in the midrange to long-range regime. In the case of DCACP-CCSD(T) where an additional term was introduced in the penalty functional, the description of this range is improved.[29]

Another test for the transferability to the solid state was carried out by investigating the relative stability of different graphite configurations. Graphite exists in a thermodynamic stable hexagonal (AB) and a metastable rhombohedral

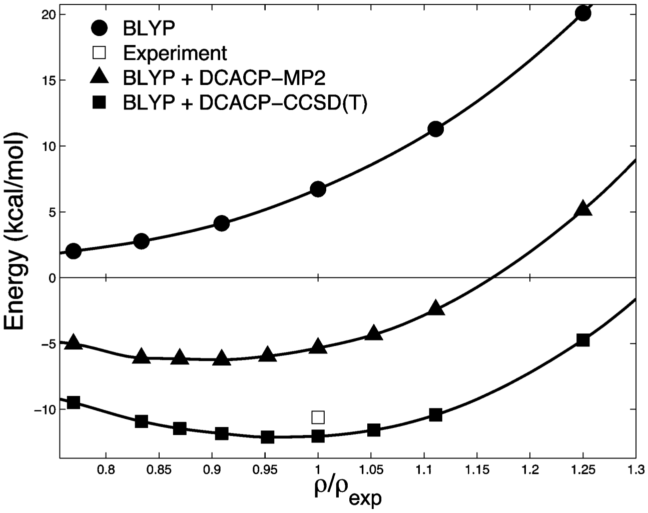


Figure 6. Cohesive binding energy of the benzene crystal [kcal/mol] as a function of the ratio between the calculated density ρ and the experimental density ρ_{exp}. The experimental value (10.6 kcal/mol) recommended for comparison in ref 42 is plotted for comparison.

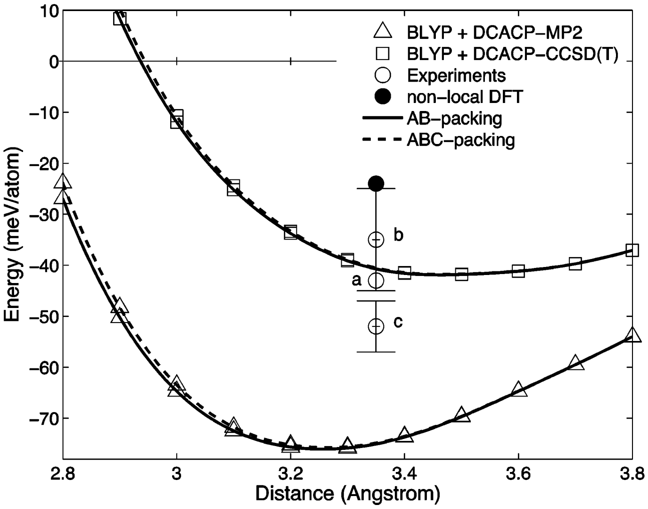


Figure 7. Interlayer binding energy [meV/atom] of three graphene sheets as a function of the interlayer distance [Å]. The experimental values (circle) of Girifalco et al. (a),[43] Benedict et al. (b),[44] and Zacharia et al. (c)[45] are plotted at the experimental interlayer distance of 3.35 Å according to ref 48. The experimental uncertainties, if available, are indicated by the error bars. A theoretical result from a recent nonlocal DFT calculation[49] is also shown as a filled circle.

configuration (ABC). In Figure 7, the IB energies for AB and ABC packing, obtained with the two sets of DCACPs, are shown. While both sets of DCACPs predict the ABC packing to be less favorable than the AB packing, consistent with the experimental observations, the difference in the IB energies between ABC and AB packing is at the limit of the numerical precision of this study. It is worth noting that the energy difference between the two configurations increases at short range and approaches zero in the dissociation limit.

Concerning absolute binding energies of hexagonal graphite, three different experimental IB energies exist in the literature for comparison: in the earliest publication, an IB energy of 43 meV/atom was measured.[43] More recent studies estimated IB energies to be 35 ± 10[44] and 52 ± 5[45] meV/atom. DCACP-MP2 corrected DFT predicts an IB energy of ≅76 meV/atom,[47] a clear overestimation with respect to all available experimental data, which is consistent with the fact that MP2 overestimates the binding of the benzene dimer.[46] DCACP-CCSD(T) predicts a binding energy of approximately 42 meV/atom agreeing within less than 10 meV/atom with all experimental values. However, the equilibrium interlayer distance of 3.30 Å predicted by the DCACP-MP2 calculations is slightly closer to the experimental value of 3.35 Å than the distance of 3.45 Å predicted by the DCACP-CCSD(T) calculations. As shown in Figure 7 DCACP-CCSD(T) predicts a flatter PES around the equilibrium distance.

4. Conclusion

Using dispersion corrected and uncorrected DFT, structures and interaction energies of dispersion dominated systems such as the dimers of ethane, ethene, ethyne, and the methane−ethene dimer were computed. For the hydrocarbon dimers, both DCACP-MP2 and DCACP-CCSD(T) predict the equilibrium intermolecular separations within a MAD of 0.08 Å to the corresponding reference values, and interaction energies of the dimers are predicted within a MAD of 0.209 and 0.115 kcal/mol, respectively. DCACPs exhibit good transferability with respect to all hybridization states (sp^3 to sp), where the typical maximum error for interaction energies and equilibrium distances amounts to 0.3 kcal/mol and 0.2 Å, respectively. For the gas-phase dimers, DCACP-CCSD(T) results correlate on average better with the corresponding post-Hartree−Fock results than the DCACP-MP2 set.

Furthermore, the transferability of DCACPs to the condensed phase has been investigated. Both sets of DCACPs drastically improve the description of crystals compared to the uncorrected BLYP results, which predict the crystals of benzene and graphite to be unstable.

For the benzene crystal, a qualitatively good description is obtained using the DCACP-MP2 set, and a highly accurate description is obtained with DCACP-CCSD(T). The latter predicts a cohesive energy which lies in the range of experimental values, and the density of the crystal is predicted within 5% of the experimental density.

Considering graphite, the DCACP-CCSD(T) IB energy is in very good agreement with the experiments, whereas DCACP-MP2 overestimates the IB energy. As in the case of benzene, the use of DCACPs generally leads to a drastic improvement of the calculated IB energies and geometries compared to conventional BLYP.

In summary, in the studied cases DCACPs lead to a clear improvement in the description of dispersion effects with respect to pure BLYP DFT and in most cases provide excellent results compared to high level ab initio data or experiments. A clear advantage of this method over empirical atom−atom corrections is its high transferabilty, i.e., a single DCACP per element is sufficient to account for weak interactions in various chemical environments.

The DCACP-CCSD(T) set, in general, leads to a better agreement with the corresponding reference data than DCACP-MP2, especially in the solid phase where long-range interactions are of high importance. The better performance of the DCACP-CCSD(T) can be attributed to an improved calibration procedure for a better description of the mid- to long-range interactions and to a separate calibration of hydrogen and carbon atoms. In addition, CCSD(T) is of higher accuracy than MP2. We therefore suggest the use of DCACP-CCSD(T) in DFT applications where dispersion forces are of importance.

Acknowledgment. The authors acknowledge discussions with Michele Parrinello and Ari P. Seitsonen. This work was supported by the Swiss National Science Foundation (Grant No. 200020-108063/1) and BIOMACH, Priority 3 (NMP-2002-3.4.1.1-3: Contract No. 505487-1).

Supporting Information Available: Geometries and parameters of the DCACP-CCSD(T). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(2) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(3) Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175.

(4) Pérez-Jordá, J. M.; Becke, A. D. *Chem. Phys. Lett.* **1995**, *233*, 134.

(5) Meijer, E. J.; Sprik, M. *J. Chem. Phys.* **1996**, *105*, 8684.

(6) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.

(7) Wesolowski, T. A.; Tran, F. *J. Chem. Phys.* **2003**, *118*, 2072.

(8) Kohn, W.; Meir, Y.; Makarov, D. E. *Phys. Rev. Lett.* **1998**, *80*, 4153.

(9) Patkowski, K.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2004**, *120*, 6849.

(10) Misquitta, A.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

(11) Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2005**, *122*, 114102.

(12) Langreth, D.; Dion, M.; Rydberg, H.; Schroder, E.; Hyldgaard, P.; Lundqvist, B. *Int. J. Quantum Chem.* **2005**, *101*, 599.

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

(14) Williams, R.; Malhotra, D. *Chem. Phys.* **2006**, *327*, 54.

(15) Becke, A.; Johnson, E. *J. Chem. Phys.* **2005**, *122*, 154104.

(16) Becke, A.; Johnson, E. *J. Chem. Phys.* **2005**, *123*, 154101.

(17) Johnson, E.; Becke, A. *J. Chem. Phys.* **2006**, *124*, 174104.

(18) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(19) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 195119.

(20) Tkatchenko, A.; von Lilienfeld, O. A. *Phys. Rev. B* **2006**, *73*, 153406.

(21) von Lilienfeld, O. A.; Andrienko, D. *J. Chem. Phys.* **2006**, *124*, 054307.

(22) 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. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(23) Parameters for DCACPs were tabulated in the Supporting Information. The same procedure as in ref 29 was used to generate the DCACPs; however, the parameters for hydrogen were calibrated against the CCSD(T) reference of the H2 dimer in the parallel configuration instead of the full configuration interaction method.

(24) *CPMD;* Copyright IBM Corp 1990−2001, Copyright MPI fuer Festkoerperforschung Stuttgart 1997−2001.

(25) CPMD consortium page. http://www.cpmd.org (accessed May 5, 2007).

(26) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(27) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(28) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703.

(29) Lin, I.-C.,; Coutinho-Neto, M. D.; Felsenheimer, C.; von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U. submitted to *Phys. Rev. B.* **2007**, *75*, 205131.

(30) 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.; Gomperst, 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.; 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 A.1*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(31) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, version 2006.1*; 2006.

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

(33) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.

(34) Cox, E.; Smith, J. *Nature* **1954**, *173*, 75.

(35) Cox, E.; Cruickshank, D.; Smith, J. *Proc. R. Soc. London., Ser. A* **1958**, *247*, 1.

(36) Dacosta, P.; Nielsen, O.; Kunc, K. *J. Phys. C: Solid. State. Phys.* **1986**, *19*, 3163.

(37) Tsuzuki, S.; Luthi, H. *J. Chem. Phys.* **2001**, *114*, 3949.

(38) Johnson, E.; Becke, A. *J. Chem. Phys.* **2005**, *123*, 024101.

(39) Chickos, J.; Acree, W. *J. Phys. Chem. Ref. Data* **2002**, *31*, 537.

(40) Oliver, G. D.; Eaton, M.; Huffman, H. M. *J. Am. Chem. Soc.* **1948**, *70*, 1502.

(41) Nakamura, M.; Miyazawci, T. *J. Chem. Phys.* **1969**, *51*, 3146.

(42) Schweizer, W.; Dunitz, J. *J. Chem. Theory Comput.* **2006**, *2*, 288.

(43) Girifalco, L.; Lad, R. *J. Chem. Phys.* **1956**, *25*, 693.

(44) Benedict, L. X.; Chopra, N. G.; Cohen, M. L.; Zettl, A.; Louie, S. G.; Crespi, V. H. *Chem. Phys. Lett.* **1998**, *286*, 490.

(45) Zacharia, R.; Ulbricht, H.; Hertel, T. *Phys. Rev. B* **2004**, *69*, 155406.

(46) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887.

(47) The value of 38 meV/atom derived by the DCACP-MP2 approach in ref 18 refers to the interaction of two graphene sheets. For a comparison with the interlayer binding of the graphite crystal, this value has to be multiplied by a factor of 2.

(48) Baskin, Y.; Mayer, L. *Phys. Rev.* **1955**, *100*, 544.

(49) Rydberg, H.; Dion, M.; Jacobsen, N.; Schröder, E.; Hyldgaard, P.; Simak, S. I.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2003**, *91*, 126402.

CT700049S

JCTC Journal of Chemical Theory and Computation

Simple Methods To Reduce Charge-Transfer Contamination in Time-Dependent Density-Functional Calculations of Clusters and Liquids

Adrian Lange and John M. Herbert*

Department of Chemistry, The Ohio State University, Columbus, Ohio 43210

Received May 23, 2007

Abstract: Using as benchmarks a series of increasingly large hydrated uracil clusters, we examine the nature and extent of charge-transfer (CT) contamination in condensed-phase, time-dependent density-functional theory. These calculations are plagued by a large number of spurious CT excitations at energies comparable to (and sometimes below) the valence excitation energies, even when hybrid density functionals are used. Spurious states below the first $n\pi^*$ and $\pi\pi^*$ states of uracil are observed in clusters as small as uracil−$(H_2O)_4$. Reasonable electronic absorption spectra can still be obtained, upon configurational averaging, despite pervasive CT contamination, but the spurious states add significantly to the cost of the calculations and severely complicate attempts to locate optically dark $n\pi^*$ states. The extent of CT contamination is reduced substantially by introducing an electrostatic (point charge) description of an extended solvent network, even in cases where the region of solvent described by density functional theory is large (>120 atoms). Alternatively, CT contamination may be reduced by eliminating certain excitation amplitudes from the linear response equations, with minimal loss of accuracy (<0.1 eV) in the valence excitation energies.

I. Introduction

Time-dependent density functional theory (TD-DFT) is currently the most popular method for calculating excited electronic states of gas-phase molecules with ∼10−200 atoms, owing to its favorable computational scaling (cubic or better with respect to system size)[1,2] and reasonable accuracy (0.2−0.3 eV for the lowest few valence excitations).[2−4] Condensed-phase TD-DFT calculations, on the other hand, are beset by serious contamination from spurious, low-energy charge-transfer (CT) excited states,[5−11] the proximate cause of which is TD-DFT's tendency to underestimate long-range CT excitation energies.[2,3,12−15] Although this problem is present already in the gas phase (and will manifest itself in TD-DFT calculations of well-separated molecules[14] or even sufficiently large single molecules),[15−17] it is much more pervasive in liquids and clusters.

Underestimation of long-range CT energetics is a consequence of incorrect asymptotic behavior on the part of the exchange-correlation potential,[14] and several long-range correction schemes have been developed recently in an attempt to alleviate this problem.[18−23] These corrections appear to mitigate CT problems for well-separated molecules in the gas phase, though only one of them has been tested in a cluster environment.[11] Furthermore, these corrections do not rectify all of the problems associated with the long-range behavior of existing density functionals,[24] and moreover the improved asymptotic behavior sometimes comes at the expense of diminished accuracy for ground-state properties.[25] In the present work, we explore some alternative methods for reducing CT contamination that are different from (though fully compatible with) these long-range correction schemes.

Several previous assessments of the performance of TD-DFT in liquids and clusters have focused exclusively on weakly allowed $n\rightarrow\pi^*$ excitations in systems such as aqueous acetone[5,6,10,11] and aqueous formamide.[9] In the case of acetone in liquid water[5,6] or water clusters,[10] spurious CT bands overlap the lowest $n\rightarrow\pi^*$ band at 4.5 eV when nonhybrid

* Corresponding author e-mail: herbert@chemistry.ohio-state.edu.

10.1021/ct700125v CCC: $37.00
Published on Web 08/02/2007

(but gradient-corrected) density functionals are employed. Hartree−Fock exchange *does* have the correct long-range behavior for CT states,[14] and hybrid functionals with 20−25% Hartree−Fock exchange are found to remove CT contamination from the lowest valence band, by pushing the offending CT states to ~1 eV higher in energy.[6]

In the present work, we use uracil as a typical example of a molecule possessing both bright states ($^1\pi\pi^*$) and dark states ($^1n\pi^*$). Our results for uracil−water clusters demonstrate that hybrid functionals alone do not guarantee that the lowest valence band will be free of CT contamination; clusters as small as uracil−$(H_2O)_4$ exhibit spurious CT states at energies comparable to or below the lowest $n\rightarrow\pi^*$ and $\pi\rightarrow\pi^*$ excitation energies. These extra states significantly increase the cost of the calculations, in both time and memory, and for a large cluster like uracil−$(H_2O)_{37}$, the memory bottleneck precludes us from calculating any states at all above 6 eV.

Two simple procedures to reduce CT contamination are examined here. First, we demonstrate that a mixed quantum mechanics/molecular mechanics (QM/MM) formalism significantly reduces the number of spurious CT states, as compared to calculations performed on the gas-phase QM region. This is true even for large QM regions and allows us to calculate a full electronic absorption spectrum for a QM region consisting of uracil−$(H_2O)_{37}$. In conjunction with liquid-phase QM/MM calculations, or on its own in the gas phase, spurious CT states can also be removed by omitting TD-DFT excitation amplitudes that correspond to long-range CT. For the present systems, this typically increases the valence excitation energies by $\lesssim$0.1 eV.

II. Computational Details

As the only long-range component of contemporary density functionals, Hartree−Fock exchange is known to reduce contamination from long-range CT excited states by pushing these states to higher excitation energies.[6,13,14,17] As such, our study will focus primarily on the hybrid functionals B3LYP[26,27] and PBE0,[28−30] though for comparison we present a few results obtained with the nonhybrid functional BLYP.[27,31] The PBE0 functional (also known as PBE1PBE)[29] consists of PBE correlation in conjunction with 25% Hartree−Fock exchange and 75% PBE exchange and has been specifically recommended for excited-state calculations.[32,33] While a larger fraction of Hartree−Fock exchange—for example, Becke's "half and half" mixture of Hartree−Fock and Slater exchange,[34] in conjunction with LYP[27] correlation—can reduce the overall number of CT states even further, this functional is less accurate for valence excitation energies[17] as well as for ground-state thermochemistry.[35] Newer, highly parametrized functionals that include full Hartree−Fock exchange may be superior in these respects,[36] but such functionals are not yet widely available, nor have they been widely tested. We shall restrict our attention to the popular hybrids B3LYP and PBE0.

All TD-DFT calculations reported here employ the Tamm−Dancoff approximation[37] and were performed using Q-Chem.[38] Only singlet excitations are considered. Density plots were rendered with the Visual Molecular Dynamics program[39] using a contour value of 0.001 au in all cases.

Table 1. Lowest Valence TD-DFT Excitation Energies ω for a Gas-Phase Isomer of Uracil−$(H_2O)_4$, at Its PBE0 /6-31+G* Geometry

functional	basis set	ω/eV $n\rightarrow\pi^*$	 $\pi\rightarrow\pi^*$
PBE0	6-31+G*	5.06	5.54
PBE0	6-311+(2d,2p)	5.02	5.46
PBE0	aug-cc-pVDZ	5.00	5.44
PBE0	aug-cc-pVTZ	5.00	5.45
B3LYP	6-31+G*	4.93	5.43
B3LYP	6-311+(2d,2p)	4.89	5.34
B3LYP	aug-cc-pVDZ	4.87	5.33
B3LYP	aug-cc-pVTZ	4.87	5.34

The basis-set dependence of the lowest $n\rightarrow\pi^*$ and $\pi\rightarrow\pi^*$ excitation energies in uracil−water clusters appears to be very mild, as demonstrated by benchmark calculations for uracil−$(H_2O)_4$ that are listed in Table 1. For both B3LYP and PBE0, excitation energies obtained with the 6-31+G* basis set differ by no more than 0.1 eV from those obtained with much larger basis sets. As such, all TD-DFT calculations will employ 6-31+G*, along with the SG-0 quadrature grid.[40]

Our interest lies in liquid-phase environments, and thus we wish to employ uracil−water geometries representative of aqueous uracil rather than a gas-phase cluster. We obtain such geometries from a molecular dynamics (MD) simulations of aqueous uracil at constant temperature (298 K) and density (0.9989 g/cm^3). Uracil was added to a pre-equilibrated, 25 Å × 25 Å × 25 Å periodic box of flexible water molecules, which was then re-equilibrated using 300 ps of MD. The AMBER99[41] and TIP3P[42] force fields (as implemented in the Tinker[43] software package) were used for uracil and for water, respectively. Following equilibration, uracil−water clusters were extracted from the simulation based on distance criteria that are described in section III. Water molecules near the uracil (according to these criteria) are included explicitly in the TD-DFT calculations, while additional water molecules up to 20.0 Å away (about 2300 molecules) are incorporated, in some cases, as TIP3P point charges.

III. Results and Discussion

A. CT Contamination in Uracil−Water Clusters. In an effort to understand just how "long range" the long-range CT problem in TD-DFT really is, we performed TD-DFT calculations on a sequence of increasingly large uracil−water clusters extracted from the MD simulation described in section II, by selecting all water molecules having at least one atom within a specified distance d of any uracil atom. All other water molecules were discarded. All clusters were generated from the same MD snapshot, so that each successively larger cluster contains the smaller clusters as its core, and these clusters range in size from bare uracil (when d = 1.5 Å) to uracil−$(H_2O)_{37}$ (when d = 4.5 Å).

For each cluster in this sequence, we calculated the first 40 TD-PBE0/6-31+G* excited states. Table 2 summarizes

Table 2. Summary of TD-PBE0/6-31+G* Calculations on Uracil−Water Clusters Extracted from a Single Snapshot of an Aqueous-Phase MD Simulation

				first $^1\pi\pi^*$ state		
d/Å[a]	no. water molecules	no. states below 6 eV	ω_{40}/eV[b]	state no.[c]	ω/eV	oscillator strength
1.5	0	5	9.05	2	5.33	0.1234
2.0	4	6	8.06	3	5.22	0.1394
2.5	7	11	7.46	5	5.28	0.1280
3.0	15	19	6.60	3	5.08	0.0807
3.5	18	20	6.50	4	5.18	0.0709
4.0	25	29	6.22	9	5.08	0.0190
4.5	37	59	5.65	18	5.06	0.1353

[a] Distance threshold for selecting water molecules. [b] Excitation energy of the 40th state above the ground state. [c] Indicates where the state appears in the TD-DFT excitation manifold.

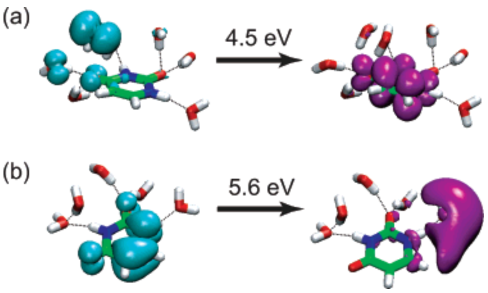


Figure 1. Typical examples of spurious CT excitations in small uracil−water clusters: (a) water-to-uracil CT and (b) uracil-to-water CT. Each excitation may be conceptualized as a rearrangement of the electron detachment density on the left into an attachment density on the right.

the results, including two simple measures of the extent of CT contamination: the excitation energy ω_{40} of the 40th state above the ground state and the number of excited states within 6 eV of the ground state. (In these clusters, the second electronic absorption band typically consists of a few states in the 6.0−6.5 eV range, so 6 eV provides a lower bound to the number of TD-DFT excited states that must be calculated in order to reach this second band.)

At the TD-PBE0/6-31+G* level, bare uracil possesses five excited states below 6 eV, the lowest two of which are an $n\pi^*$ dark state (at 4.56 eV) and a $\pi\pi^*$ bright state (at 5.33 eV). There are also two more dark states of mixed $n\pi^*$/Rydberg character, plus one optically allowed $n\pi^*$ state whose oscillator strength is 30% of that associated with the $\pi\pi^*$ state. (Uracil is slightly nonplanar in the geometries extracted from the MD simulation, so we use “bright” and “dark” as qualitative descriptions of transition intensities. The “optically allowed” $n\pi^*$ state, for example, correlates in a planar chromophore to an excitation out of an a'' lone pair orbital.)

The 40 excitations calculated for bare uracil reach 9 eV above the ground state, but due to the appearance of spurious CT states, ω_{40} drops as cluster size increases, while at the same time the number of states below 6 eV increases. By the time the cluster size reaches $d = 4.5$ Å [uracil−$(H_2O)_{37}$], the first 40 excited states reach only 5.65 eV, well below the energy of the second absorption band. At these energies, the density of excited states is ~60 states/eV, and using Q-Chem on a machine with 4 Gb of memory, we are unable to calculate enough states to reach 6 eV. Excluding core orbitals from the TD-DFT excitation space (which changes the excitation energies by $< 10^{-4}$ eV) reduces the required memory for the Davidson iterations[44] by a factor of $N_{\text{core}}/N_{\text{occupied}} \approx 0.21$ and (just barely) allows us to calculate the 59 states that are required to reach 6 eV, by which point the density of states has reached ~80 states/eV. (For comparison, multireference calculations of gas-phase uracil find a total of eight $n\pi^*$ and $\pi\pi^*$ states in the 5.0−7.0 eV range.)[45]

The results in Table 2 are for PBE0, but B3LYP paints a similar picture (with even a slightly larger number of spurious CT states, consistent with its slightly smaller fraction of Hartree−Fock exchange). We conclude that, despite their success for acetone in liquid water,[6] in certain systems the popular hybrid functionals B3LYP and PBE0 may still suffer from considerable CT contamination at or below the lowest valence excitation energies. Whereas a $\pi\pi^*$ bright state ought to be either the first or second excited state (depending on the order of the $n\pi^*$ and $\pi\pi^*$ states, which changes as a function of cluster size and geometry), we see from Table 2 that clusters as small as uracil−$(H_2O)_4$ exhibit spurious states below the first bright state. Apparently, the “long range” CT problem in TD-DFT can manifest even at hydrogen-bond distances, and even when using hybrid functionals with up to 25% Hartree−Fock exchange. That said, it should be emphasized that the problem is dramatically worse for nonhybrids—a BLYP calculation on the $d = 2.5$ Å cluster, for example, yields more than 40 states below 6 eV, even though there are only seven water molecules, while at $d = 3.0$ Å, CT states appear starting at 2.85 eV and the first $\pi\pi^*$ state is not even among the first 40 excited states!

In further contrast to the case of acetone in water, where no significant hybridization is observed between the water molecules and the acetone lone pairs,[5] we do observe hybridization between water and the carbonyl lone pairs of uracil. Consequently, the real $n\pi^*$ states (and sometimes even $\pi\pi^*$ states with some $n\pi^*$ character) are sometimes difficult to discern from the spurious CT states simply on the basis of the TD-DFT excitation amplitudes and Kohn−Sham molecular orbitals (MOs). Such ambiguity is avoided by instead examining electron attachment and detachment densities obtained from the eigenvectors of the difference density matrix between the ground and excited states.[46] The detachment density represents the part of the density that is removed from the ground state and rearranged in the excited-state to form the attachment density.[2] We make exclusive use of these densities in identifying the qualitative character of the excited states.

Typical examples of low-energy CT states appearing in small uracil−water clusters are illustrated in Figure 1, while Figure 2 depicts some typical CT states in a larger cluster. In small clusters, the CT states below about 5.5 eV are almost exclusively water-to-uracil CT states of the type depicted in Figure 1(a), where the detachment density is dominated by the out-of-plane lone pair on a single water molecule. Such states appear in larger clusters as well [Figure 2(a)], where the water molecule in question tends to be located at the surface of the cluster.

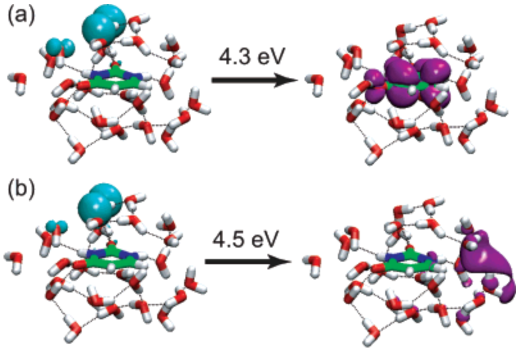


Figure 2. Typical examples of spurious CT excitations in a uracil−$(H_2O)_{25}$ cluster: (a) water-to-uracil CT and (b) water-to-water CT.

The appearance of these states is easy to understand. First note that the out-of-plane lone pairs on the water molecules (except possibly those at the center of a large cluster) are the highest occupied MOs (HOMOs) in the system, while the lowest unoccupied MO (LUMO) is always a uracil π^* orbital. In the limit of large separation between an occupied and a virtual MO, and absent any component of Hartree−Fock exchange, TD-DFT will predict CT between these orbitals at an excitation energy equal to the difference in their Kohn−Sham eigenvalues.[14,21] Thus, if any frontier occupied MOs are spatially separated from low-lying virtual MOs, then one *will* obtain spurious, low-energy CT excitations, unless a large component of Hartree−Fock exchange (greater than 25%, evidently) is employed. Such states should be anticipated in most condensed-phase systems.

In addition to the uracil π^* LUMO, larger uracil−water clusters also possess low-lying virtual MOs localized on the solvent that are not present in small clusters. This opens up another avenue to spurious CT in large clusters and affords water-to-water CT excitations such as that depicted in Figure 2(b). Occasionally these states have some uracil-to-water CT character as well, but mostly the uracil-to-water CT states [e.g., Figure 1(b)] appear at energies above 5.5 eV, though they proliferate rapidly at higher excitation energies.

The small-cluster CT states are intriguing, because a cluster like uracil−$(H_2O)_7$ might not immediately come to mind upon mention of "long-range" CT in TD-DFT. To emphasize that small clusters are indeed susceptible to CT contamination, we examine in detail the excited states of the $d = 2.5$ Å cluster, of which there are six within 5.5 eV of the ground state. Detachment densities for these six states are depicted in Figure 3, along with excitation energies (ω_i) and oscillator strengths (f_i). Attachment densities are not shown, as each one is dominated by the LUMO and resembles the attachment density shown in Figure 1(a). The detachment densities identify states 3 and 5 as the first $n\pi^*$ and $\pi\pi^*$ states, respectively, whereas the remaining states below 5.5 eV involve water-to-uracil CT of the type discussed above.

With regard to the oscillator strengths, we note that the $n\pi^*$ state borrows sufficient intensity to achieve an oscillator strength 25% as large as that of the nominal bright state, whereas the CT excitations are mostly dark, consistent with nearly nonoverlapping attachment and detachment densities. In system configurations where $n\pi^*/\pi\pi^*$ intensity borrowing is less significant, however, oscillator strengths for the low-energy CT states sometimes exceed that of the $n\pi^*$ state. Thus the real dark states cannot be identified simply from a list of excitation energies and oscillator strengths, but only by careful analysis of the MOs or (better yet) attachment/detachment densities.

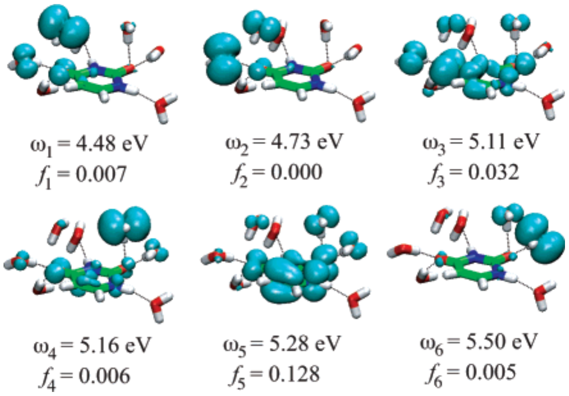


Figure 3. Excitation energies, oscillator strengths, and detachment densities for the lowest six TD-PBE0/6-31+G* excited states of a uracil−$(H_2O)_7$ cluster.

In larger clusters, however, CT states can undergo a type of ersatz intensity borrowing that greatly complicates interpretation of the vertical excitation spectrum. A hint as to this behavior is the overall decrease in the oscillator strength of the first $\pi\pi^*$ state as a function of cluster size (see Table 2), though the trend is not monotonic—the $\pi\pi^*$ intensity recovers at $d = 4.5$ Å, at least for this one particular cluster geometry. The reason for this diminished intensity is that, as the density of spurious CT states increases, there appear CT states with energies comparable to that of the $\pi\pi^*$ state, and these spurious excitations borrow intensity from the real bright state. Since oscillator strengths out of the ground state are positive and sum to a constant (the Thomas−Reiche−Kuhn sum rule),[47] this decreases the oscillator strength of the real bright state. (This explanation is only qualitative, since the sum rule is not exactly fulfilled within the Tamm−Dancoff approximation that we employ here.)[2]

In larger clusters, this form of intensity borrowing actually makes it difficult to determine which excitation is the real bright state. The $d = 4.0$ Å cluster, for example, exhibits five excited states between 5.05 and 5.20 eV that have significant intensity (states 7−11 in the excitation manifold), which are depicted in Figure 4. With the exception of state 11 (which has the smallest oscillator strength of the five), each of the detachment densities has a significant uracil π component, but in all cases there is a significant contribution from a water lone pair as well. All five of the attachment densities are dominated by the uracil π^* LUMO. State 9 is selected as the $\pi\pi^*$ state in Table 2 because its TD-DFT eigenvector contains a larger component of the uracil $\pi \rightarrow \pi^*$ excitation than any of the other four states, but note that this is *not* the strongest transition of the five, as in this case the spurious CT states have borrowed the majority of the oscillator strength of the $\pi\pi^*$ bright state. In reporting a vertical excitation spectrum, then, it is not appropriate simply

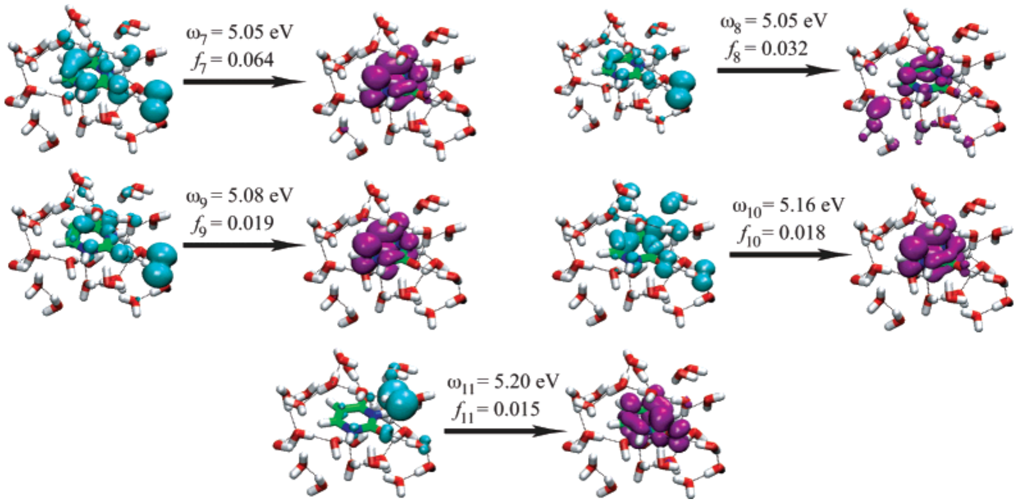


Figure 4. TD-PBE0/6-31+G* excitations for the $d = 4.0$ Å cluster, illustrating intensity borrowing by spurious CT states.

to report the transition with the largest oscillator strength as “the” bright state.

B. QM/MM Simulations of Aqueous Uracil. In a recent TD-PBE0 study of hydrated uracil,[48] it was found that a uracil−$(H_2O)_4$ complex embedded in a polarizable continuum induces a redshift of only 0.1 eV in the first uracil $\pi \rightarrow \pi^*$ excitation, whereas the experimentally measured solvatochromatic shift is about 0.5 eV. In fact, the polarizable continuum accounts for the entirety of the calculated shift; the four explicit water molecules do not modify the gas-phase excitation energy at all.[48] (A recent TD-BLYP study of *s*-tetrazine in aqueous solution also found that those water molecules that are directly hydrogen-bonded to the chromophore do not suffice to explain the observed solvatochromatic shift.)[8] As a next step, it seems natural to consider QM/MM simulations of aqueous uracil, using a QM region substantially larger than uracil−$(H_2O)_4$. Such calculations are discussed in the present section.

To make comparison with results in the previous section, we first consider a sequence of calculations whose QM regions are precisely the same series of increasingly large uracil−water clusters described in Table 2 of section IIIA. The MM region in these new calculations consists of all additional water molecules extracted from our MD simulation, out to a distance of 20.0 Å away from uracil. These MM water molecules (about 2300 in all) are incorporated as TIP3P point charges. Table 3, which is analogous to Table 2 in the previous section, summarizes the results of TD-PBE0/6-31+G* calculations on these QM/MM systems.

Addition of the MM solvent region has a very small effect on the excitation energy for the first $\pi\pi^*$ state, inducing a shift of no more than 0.07 eV, even in cases where the QM region consists only of uracil, or of uracil plus only a few water molecules. There is also no clear trend in the direction of this shift.

The MM solvent region does have one tremendously important effect, however: it dramatically reduces the number of spurious CT states at all values of *d*, the distance threshold for selecting QM water molecules. With the addition of point charges, even a large QM region like $d = 4.5$ Å (which is 13−14 Å across, and contains 123 atoms) affords only 10 excited states within 6 eV of the ground state. Absent the TIP3P charges, the same QM region affords an estimated 60 states below 6 eV. Figure 5(a) plots the growth in the number of low-energy excited states as a function of *d*, for TD-PBE0 calculations with and without MM point charges. For gas-phase clusters the number of states rises rapidly with cluster size, but this growth is very sluggish in the presence of MM point charges.

Table 3. Summary of TD-PBE0/6-31+G* QM/MM Calculations on Aqueous Uracil, as a Function of the Size of the QM Region

				first $^1\pi\pi^*$ state		
d/Å[a]	no. QM water molecules	no. states below 6 eV	ω_{40}/eV	state no.	ω/eV	oscillator strength
1.5	0	3	9.73	2	5.31	0.1396
2.0	4	5	8.66	2	5.25	0.1168
2.5	7	5	8.16	2	5.22	0.1707
3.0	15	8	7.59	2	5.15	0.1616
3.5	18	7	7.44	1	5.13	0.1041
4.0	25	8	7.15	1	5.09	0.1422
4.5	37	10	6.91	1	5.10	0.1624

[a] Distance criterion for selecting the QM region.

Although MM point charges eliminate many low-energy CT states (for reasons explained below), Figure 5(b) reveals a steady decrease in ω_{40} as a function of *d*, even for the QM/MM calculations, though the falloff is sharper in the absence of point charges. The decrease in ω_{40} indicates that the MM charges do not remove all spurious CT states, especially at higher excitation energies. While the QM/MM calculation at $d = 4.5$ Å yields only 10 states below 6 eV, there are another 30 states (mostly spurious) between 6.00 and 6.91 eV. On the other hand, a far greater number of spurious states appear in this energy régime when the MM charges are removed, and this is an important practical consideration, given that the number of excited states requested in a TD-DFT calculation determines the memory required for the Davidson iterations.[44] In fact, we are unable to locate the second electronic absorption band in gas-phase uracil−$(H_2O)_{37}$ ($d = 4.5$ Å) due to the large number of states required. The total memory requirement for such a calcula-

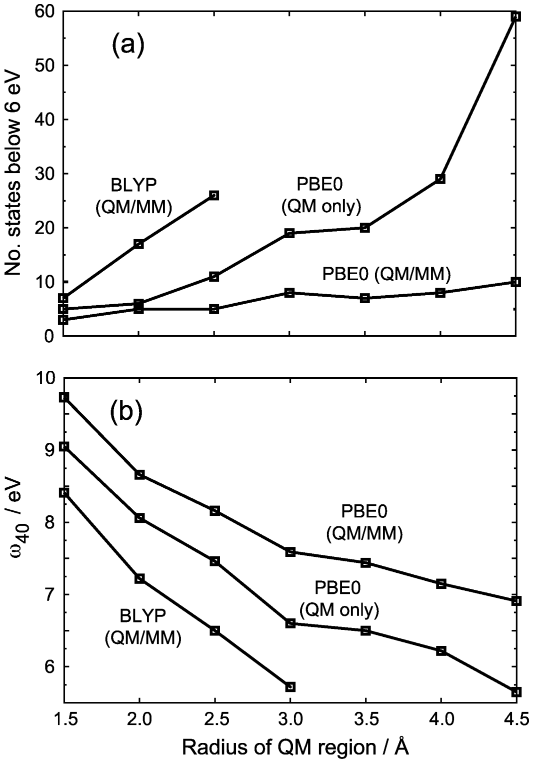


Figure 5. (a) Number of excited states within 6 eV of the ground state and (b) excitation energy of the 40th excited state, each as a function of the radius *d* of the QM region.

tion exceeds 4 Gb, even when core orbitals are excluded from the TD-DFT excitation space.

Examination of the low-energy excited states for the $d = 2.5$ Å cluster—which may be compared to the $d = 2.5$ Å results of the previous section that are depicted in Figure 3—provides a clue to the origin of this reduction in CT states. For the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ excitations (states 3 and 5 in Figure 3), we find that the attachment densities, detachment densities, and excitation energies are nearly unchanged by addition of the point charges. Each of the CT excitations (states 1, 2, 4, and 6 in Figure 3) is also unchanged in its qualitative character but is shifted to ~1 eV higher in energy. What was $\omega_1 = 4.48$ eV in the absence of point charges becomes $\omega_3 = 5.48$ eV in the QM/MM calculation.

In fact, for the QM/MM calculations we find that the $n\pi^*$ and $\pi\pi^*$ states are *always* the first and second excited states, irrespective of the size of the QM region. (Interestingly, the order of these two states changes as a function of solvent configuration, something that could not have been deduced from polarizable continuum models.)[48] Because there are few low-energy CT states, there is also no intensity-borrowing problem of the sort discussed in section IIIA. (The dip in the $\pi\pi^*$ oscillator strength that is observed at $d = 3.5$ Å results from substantial intensity borrowing on the part of the $n\pi^*$ state.)

To understand why the point charges wield such an influence on CT excitation energies, recall that in the large clusters of section IIIA, only the water molecules on the surface of the cluster contribute to the lowest-energy CT states (see Figure 2). Addition of the MM point charges has the effect of stabilizing the lone pair orbitals on these water molecules, lowering their Kohn—Sham eigenvalues and thereby increasing the excitation energy associated with water-to-uracil CT. Importantly, this stabilization is sufficient to remove low-energy CT states *only* in conjunction with a hybrid functional; TD-BLYP calculations are still beset by numerous CT states at low energies, even within a QM/MM framework.

Table 4. Summary of TD-BLYP/6-31+G* Calculations on Uracil—Water Clusters

d/Å[a]	includes TIP3P charges?[b]	no. states below 6 eV	ω_{40}/eV
1.5	no	12	7.89
2.0	no	28	6.48
2.5	no	>40	5.68
3.0	no	>40	4.67
1.5	yes	7	8.41
2.0	yes	17	7.22
2.5	yes	26	6.50
3.0	yes	>40	5.72

[a] Denotes the size of the cluster or the size of the QM region. [b] Point charges were used in some calculations to represent additional water molecules out to $d = 20.0$ Å.

To emphasize this point, Table 4 summarizes TD-BLYP calculations on our sequence of uracil—water clusters, both with and without point charges. (These data are plotted alongside TD-PBE0 results in Figure 5.) Although the MM solvent does reduce the number of states below 6 eV, the number of such states remains large, even in the QM/MM calculations. Using BLYP, attempts to locate the second absorption band quickly become intractable as cluster size increases.

These observations clarify the results of Bernasconi, Sprik, and Hutter,[5,6] who simulated electronic absorption spectra of aqueous acetone using plane-wave Carr—Parrinello MD. Using BLYP, these authors find that the lowest valence ($n\pi^*$) band is buried beneath a much broader and more intense CT band, comprised of several spurious CT states. In contrast, the B3LYP and PBE0 functionals shift the CT band upward by ~ 1 eV, well above the valence $n\pi^*$ band.[6] Our results show that this is only partially attributable to the use of hybrid functionals. Equally important is the fact that Bernasconi et al. use plane-wave DFT (hence periodic boundary conditions), which means that there were no "surface" water molecules present that might contribute low-energy CT excitations. On the other hand, Hartree—Fock exchange is incredibly expensive to evaluate in a plane-wave basis. Gaussian-orbital-based electronic structure theory, in conjunction with MM point charges to model an extended solvent network, thus represents a useful, affordable alternative.

C. Electronic Absorption Spectra. To this point, all calculations have used geometries taken from the same MD snapshot, which allows us to discuss trends with respect to cluster size. Solvent and chromophore geometry, however, play important roles in modulating the excitation energies, modifying the order and relative intensities of the valence excitations at least, and possibly the CT excitations as well. In order to take these effects into account, we next discuss electronic absorption spectra simulated as averages over a

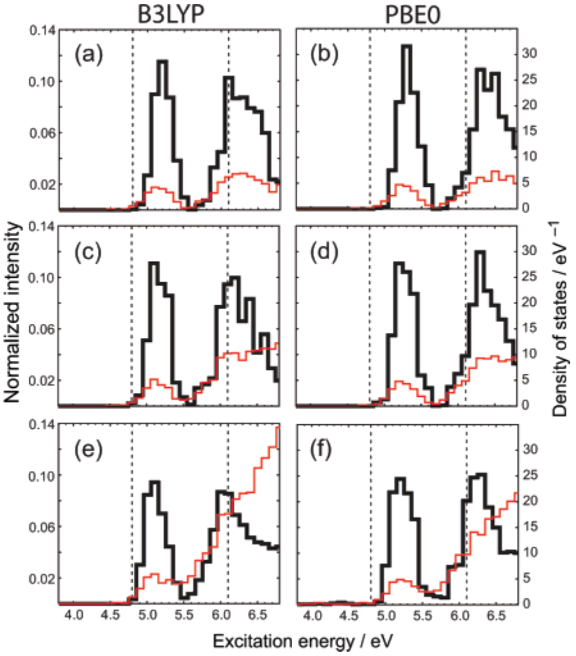


Figure 6. Electronic absorption spectra (thick black lines, scale on the left) and densities of states (thin red lines, scale on the right) from QM/MM calculations at the TD-B3LYP/6-31+G* and TD-PBE0/6-31+G* levels. As described in the text, the calculations in (a) and (b) utilize a uracil-only QM region, (c) and (d) employ a microhydrated QM region, while (e) and (f) use a full solvation shell for the QM region. Dotted vertical lines show the positions of the first two band maxima in the experimental absorption spectrum of aqueous uracil (from ref 49).

total of 100 configurations extracted from an MD simulation at intervals of 1.0 ps. (A 2-fold reduction in the number of configurations has a modest effect on the spectra, but the rough spectral envelopes appear to be converged with respect to configuration sampling.) As with the QM/MM calculations discussed above, water molecules near the uracil (according to criteria described below) are included in the QM region, while additional water molecules out to 20.0 Å are incorporated as TIP3P point charges.

Absorption spectra are obtained by constructing histograms of the TD-DFT excitation energies (using a bin width of 0.1 eV), wherein the excitations are summed according to their oscillator strengths; these spectra are plotted, at both the TD-B3LYP and TD-PBE0 levels, in Figure 6. Electronic densities of states (also plotted in the figure) are obtained in similar fashion, by assigning equal weight to each excited state. Once again PBE0 predicts slightly fewer low-energy CT states, but its overall behavior is very similar to that of B3LYP.

Far more important is the size of the QM region, and we compare three criteria for making the QM/MM separation. In Figure 6(a),(b)—representing B3LYP and PBE0, respectively—the QM region consists only of uracil, and all water molecules are modeled as point charges. In Figure 6(c),(d), the QM region includes all water molecules having at least one atom within 2.5 Å of one of uracil's hydrogen-bonding sites (i.e., the N−H hydrogen atoms and the C═O oxygen atoms). This amounts to an average of 5.5 water molecules in the QM region and so we refer to this case as the "microhydrated" QM region. Finally, in Figure 6(e),(f), the QM region contains all water molecules having at least one atom within 5.0 Å of the uracil center of mass, for an average of 18.7 water molecules in the QM region. This is sufficient to form a full solvation shell around uracil, so we refer to these calculations as the "full solvation shell" QM region.

In order to obtain sensible averages, it is important that each individual TD-DFT calculation determine enough excited states to reach a given energy threshold, at which the histograms will terminate. For the largest QM region, the first 40 excited states consistently reach 6.8 eV, which we thus choose as our energy cutoff. For the uracil-only and microhydrated QM regions, 10 and 20 excited states, respectively, are required to reach 6.8 eV.

In the case of a uracil-only QM region [Figure 6(a),(b)], both the first and second absorption bands appear to be mostly free of CT contamination. It is not obvious a priori that this should be the case, despite the fact that there are no explicit solvent molecules present to support long-range CT states, as there remains the possibility of anomalously low Rydberg excitations. Using BLYP, one does in fact observe Rydberg states below 6 eV for a uracil-only QM region (see the $d = 1.5$ Å results in Table 4). For B3LYP and PBE0, such states are completely absent within the first electronic absorption band, as evident from the density of states, which has a value of approximately 4 states/eV over the span of the first absorption band, which is about 0.5 eV wide. On average, then, this band must consist of two states, namely, the first $n\pi^*$ and $\pi\pi^*$ states. This is consistent with multireference calculations for gas-phase uracil that find one $n\pi^*$ state and one $\pi\pi^*$ state below 5.5 eV.[45] Within the second absorption band, the density of states ranges from 4−7 states/eV over a band that is about 1 eV wide, indicating that on average there are 5 or 6 states within this band. This is also consistent with the aforementioned multireference calculations, which find a total of six $n\pi^*$ and $\pi\pi^*$ states in the 5.5−7.0 eV range.[45]

Examining next the results for the microhydrated QM region, Figure 6(c),(d), we see that the density of states within the first absorption band is largely unchanged and, importantly, decays nearly to zero around 5.5 eV, in between the first and second absorption bands. (The experimental spectrum also decays nearly to zero around 5.5 eV.)[49] Unlike the case of a uracil-only QM region, however, the density of states shows no sign of dropping in the tail of the *second* absorption band and, in the B3LYP case at least, appears to be increasing above 6.5 eV, even as the spectral intensity decays. Embedding the QM region in an MM solvent pushes the CT threshold up to about 6.0 eV, with many more spurious states above 6.5 eV.

Finally there is the QM region consisting of a full solvation shell, Figure 6(e),(f). Here, the threshold for observing a substantial number of CT states creeps down somewhat from the 6.0 eV observed above, and consequently the density of states no longer decays to zero at 5.5 eV. In addition, a small number of system configurations exhibit CT states around 4.5 eV, below the first absorption band.

Regarding the solvatochromatic shifts, we note that gas-phase TD-B3LYP/6-31+G* and TD-PBE0/6-31+G* calcu-

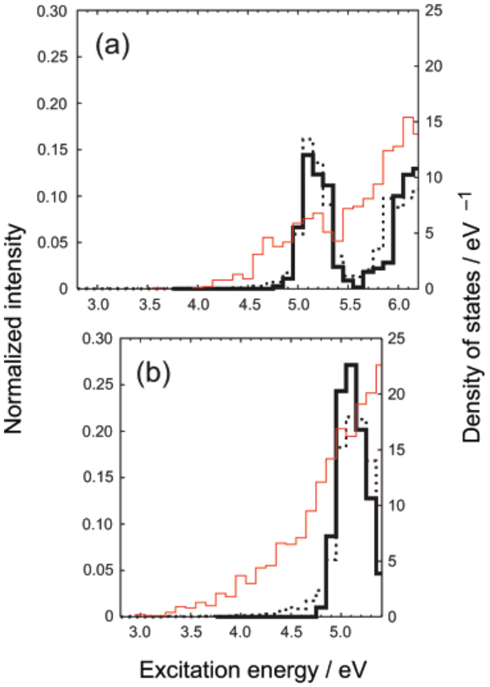


Figure 7. TD-B3LYP/6-31+G* calculations on uracil−water clusters using (a) a microhydrated QM region and (b) a full solvation shell in the QM region. Electronic absorption spectra (scale on the left) are plotted for both the QM region only (thick, solid line) and for the QM/MM calculation (dotted line). The thin, solid line in red is the density of states (scale on the right) for the QM-only cluster calculation.

lations put the $\pi \rightarrow \pi^*$ excitation energy at 5.4 and 5.5 eV, respectively. Based on the absorption band maxima in Figure 6(a),(b), addition of a point-charge-only solvent environment induces, on average, a 0.2 eV redshift in this excitation energy. (For comparison, a polarizable continuum model induces a 0.1 eV redshift.)[48] A further redshift of about 0.1 eV is obtained using a microhydrated QM region, but incorporating a full solvation shell into the QM region does not change the position of the band maximum, though it does slightly increase the relative intensity of the low-energy side of the distribution. Given that TD-DFT calculations of acetone−$(H_2O)_N$ clusters[10] find that the lowest $n \rightarrow \pi^*$ excitation energy changes by <0.05 eV between $N = 10$ and $N = 250$, the remaining 0.2 eV solvatochromatic shift not recovered in our largest QM/MM calculations probably represents the intrinsic accuracy limit of TD-DFT.

Finally, we revisit the role of the MM point charges, this time in the context of configurationally averaged absorption spectra and densities of states. Figure 7 compares the QM/MM absorption spectra to those obtained upon removal of the MM solvent. In the latter case, of course, the configurational averages contain a far larger number of spurious CT states, as is clear from the densities of states plotted in Figure 7, which extend down to 4.0 eV for the microhydrated QM region and down to 3.0 eV for the larger QM region, whereas the QM/MM calculations have virtually no states below 4.8 eV (cf. Figure 6). Nevertheless, absorption profiles obtained with and without the MM charges are quite similar. Partly this reflects the fact that most of the CT excitations are spectroscopically dark, but it furthermore reflects the fact that summing the oscillator strengths (as opposed, say, to finding the configurationally averaged excitation energy of the most intense transition) gathers up the intensity of any CT states that may borrow intensity from the $\pi\pi^*$ state.

D. Truncation of the TD-DFT Excitation Space. Although an MM embedding provides a simple and affordable means to reduce CT contamination (and for large systems is a method of choice in its own right), in some cases one might be interested in a gas-phase cluster rather than a proper liquid. In this section, we examine a separate (though compatible) procedure, whereby CT states are eliminated by ansatz, by removing from the linear-response eigenvalue equation those occupied-to-virtual ($|i\rangle \rightarrow |a\rangle$) excitation amplitudes that correspond to long-range CT. Automated criteria for performing this truncation of the excitation space have been developed by Besley,[9] whose procedure we adopt here. Truncation of the excitation space has also been explored, within the context of plane-wave DFT, by Odelius et al.[8]

According to Besley's procedure,[9] one first identifies a subset of the atoms as belonging to the chromophore, for which we choose all of the uracil atoms. Excitation amplitudes x_{ia} are then removed unless the occupied Kohn−Sham orbital $|i\rangle$ contains a significant contribution from basis functions centered on chromophore atoms, as measured by the contribution that these basis functions make to the Mulliken population of $|i\rangle$. We denote the threshold contribution as κ_{occ}; if the chromophore-centered basis functions do not contribute at least κ_{occ} electrons to $|i\rangle$, then the x_{ia} are omitted, for all a. (Values given for κ_{occ} in this work are total populations, including both spins.)

Besley[9] suggests additional truncation based on a second threshold κ_{virt}, according to which the sum of squares of the MO coefficients, $\sum_\mu |c_{\mu a}|^2$, is used to measure the contribution that the set of chromophore-centered basis functions $\{|\mu\rangle\}$ makes to the virtual orbital $|a\rangle$. As a result of the diffuse functions present in our basis set, however, we find that this sum is quite similar for each of the low-lying virtual MOs, whether or not they are localized around the uracil molecule. One way to circumvent this problem is to employ a mixed basis set, eliminating diffuse functions on the chromophore so that uracil-centered basis functions no longer contribute significantly to virtual MOs localized on the solvent. In practice, we find that useful results can be obtained without any truncation of the virtual space, so we retain 6-31+G* for all atoms and truncate the excitation space based solely on the occupied orbital criterion.

Table 5 lists TD-PBE0 excitation energies for the first and second $\pi\pi^*$ states of two different uracil−water clusters, using several different values of κ_{occ} ranging from $\kappa_{occ} = 0$ (a full excitation space) to $\kappa_{occ} = 1$ (the value used in Besley's benchmark calculations).[9] Results are presented both with and without MM point charges. In the smaller of these two clusters, uracil−$(H_2O)_7$ ($d = 2.5$ Å), we find that the accuracy of the excitation energies degrades rather slowly as a function of κ_{occ}. Already at $\kappa_{occ} = 0.2$, all CT states below the second $\pi\pi^*$ state are eliminated, meanwhile no significant error is incurred in the $\pi \rightarrow \pi^*$ excitation energies. Errors of $\lesssim 0.04$ eV were also reported by Besley[9] using κ_{occ}

Table 5. TD-PBE0/6-31+G* Excitation Energies Obtained Using Truncated Excitation Spaces

			first $\pi\pi^*$		second $\pi\pi^*$	
d/Å	includes TIP3P charges?	κ_{occ}	ω/eV	state no.[a]	ω/eV	state no.[a]
2.5	no	0.0	5.28	5	6.13	13
2.5	no	0.2	5.28	2	6.13	4
2.5	no	0.4	5.32	2	6.14	4
2.5	no	1.0	5.39	2	6.20	4
2.5	yes	0.0	5.22	2	6.18	8
2.5	yes	0.2	5.24	1	6.20	5
2.5	yes	0.4	5.27	1	6.26	4
2.5	yes	1.0	5.34	1	6.29	4
4.5	no	0.0	5.06	18		>60
4.5	no	0.2	5.18	5	6.18	17
4.5	no	0.4	5.21	3	6.18	11
4.5	no	1.0	5.21	3	6.18	11
4.5	yes	0.0	5.10	1	6.04	13
4.5	yes	0.2	5.21	1	6.23	6
4.5	yes	0.4	5.33	1	6.28	6
4.5	yes	1.0	5.52	1	6.37	6

[a] Indicates where the state appears in the TD-DFT excitation manifold.

$= 1.0$ and $\kappa_{virt} = 0.8$ for a formamide$-$(H$_2$O)$_4$ cluster; larger clusters were not considered in that study.

The larger of the two clusters is uracil$-$(H$_2$O)$_{37}$ ($d = 4.5$ Å), and in this case the valence excitation energies are more sensitive to the value of κ_{occ}. Even $\kappa_{occ} = 0.2$ engenders errors of 0.1 and 0.2 eV, respectively, in the first and second $\pi \rightarrow \pi^*$ excitation energies. As before, $\kappa_{occ} = 0.2$ is sufficient to remove the water-to-uracil CT states; therefore, further increase of κ_{occ} is of no benefit. The spurious states that remain involve water-to-water and uracil-to-water CT, and elimination of these states would require truncation based on the virtual orbitals.

To assess the accuracy of truncation over a range of geometries, we recalculate the optical spectrum for the large (full solvation shell) QM/MM calculations, using a truncation threshold of $\kappa_{occ} = 0.2$, and in Figure 8(a) we compare this spectrum to that obtained using a full excitation space. Truncation produces virtually no change in the overall absorption envelope, except that it shifts the entire spectrum (both the first and second absorption bands) to slightly higher energy. The magnitude of this overall shift is something less than the bin width of the histogram, 0.1 eV. (Test calculations on smaller systems indicate that additional configurations are required in order to achieve better than 0.1 eV resolution.) We conclude that truncation affords a consistent level of accuracy across many system configurations.

The density of states for this calculation, Figure 8(b), shows that truncation does not remove any CT states within the first absorption band—these were removed already by the introduction of MM point charges. As the excitation energy increases, however, calculations in the full excitation space predict an increasingly large number of spurious states, relative to results obtained with $\kappa_{occ} = 0.2$. The practical upshot is that the latter calculations consistently require only 15 excited states to reach 6.8 eV (the energy cutoff in Figure 8), whereas 30$-$40 states are required when a full excitation space is employed. This represents an approximately 2-fold reduction in the memory required for the Davidson iterations, which is roughly proportional to the number of excited states requested.

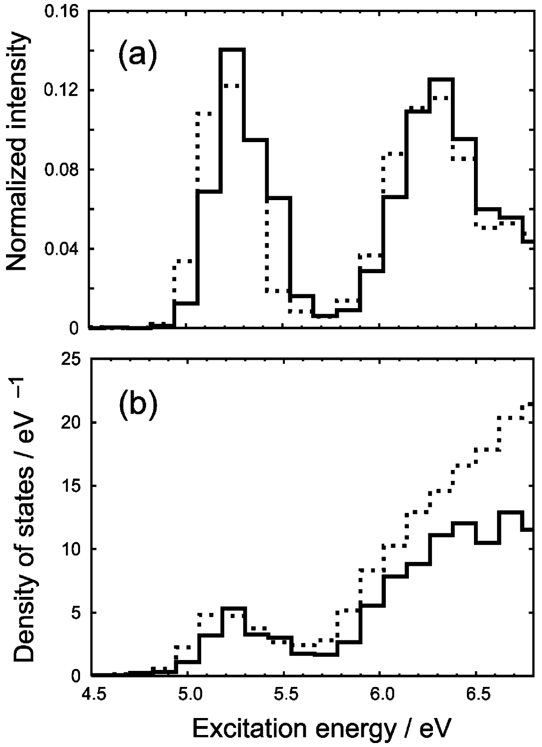


Figure 8. (a) Absorption spectra and (b) densities of states, from TD-PBE0/6-31+G* QM/MM calculations with a full QM solvation shell, using $\kappa_{occ} = 0.2$ (solid lines) and $\kappa_{occ} = 0.0$ (broken lines).

Truncation of the excitation space is similarly accurate for gas-phase clusters, as demonstrated when we remove the MM point charges from the QM/MM calculations discussed above. Figure 9(a) compares spectra obtained with $\kappa_{occ} = 0.2$ to those calculated with a full excitation space, while Figure 9(b) compares the corresponding densities of states. In the absence of truncation, we obtain excited states all the way down to 3 eV, and 40 excited states are required just to reach $\omega = 5.4$ eV. Thus the spectra in Figure 9(a) include only the first absorption band. When $\kappa_{occ} = 0.2$, this energy cutoff is reached consistently with only the first 25 excited states. As before, the spectrum calculated with the truncated excitation space is shifted to higher energy by $\lesssim 0.1$ eV, with little change in the overall absorption envelope.

Finally, we note that the success of orbital truncation as a means to reduce CT contamination is contingent upon use of a hybrid density functional. Even for a fairly small, $d = 3.0$ Å cluster, with MM point charges included and using a truncation threshold of $\kappa_{occ} = 0.2$, a TD-BLYP calculation yields 17 excited states below the first $\pi\pi^*$ state. Keeping $\kappa_{occ} = 0.2$ but omitting the point charges, CT becomes so prevalent that is impossible to discern the identity of the $\pi\pi^*$ state simply on the basis of oscillator strengths, as so much of this intensity has bled into the spurious CT states.

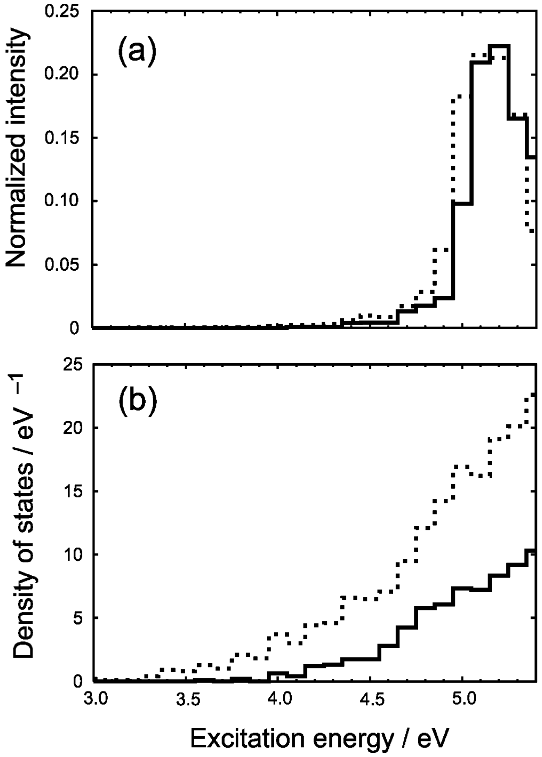


Figure 9. (a) Absorption spectra and (b) densities of states, from TD-B3LYP/6-31+G* calculations of the full QM solvation shell in the absence of MM point charges, using $\kappa_{\mathrm{occ}} = 0.2$ (solid lines) and $\kappa_{\mathrm{occ}} = 0.0$ (broken lines).

IV. Summary and Conclusions

TD-DFT calculations in both small and large molecular clusters are beset by a legion of spurious CT excitations at or below the lowest valence excitation energies, even when hybrid functionals such as B3LYP or PBE0 are employed. These spurious excitations proliferate rapidly as cluster size increases.

If only the optically bright states are of interest, TD-DFT still affords useful information, with the proviso that one should sum oscillator strengths over states with comparable excitation energies, since near-degeneracies between a bright state and one or more spurious CT states may lead to anomalous intensity borrowing, robbing the bright state of intensity. If one is interested in optically forbidden transitions, however, then the spurious CT states present a formidable problem, as they are often difficult to distinguish from real excited states whose oscillator strengths are small. In any case, the spurious CT states introduce a severe memory bottleneck.

Our results demonstrate that CT contamination is substantially reduced (though not eliminated) using either of two simple measures, neither of which requires modification of existing density functionals. One option is to embed the system of interest within a larger electrostatic medium, via a QM/MM calculation. This procedure substantially reduces CT contamination by stabilizing occupied Kohn−Sham orbitals on the edge of QM region and, to a lesser extent, destabilizing virtual MOs at the QM/MM interface. (A dielectric or polarizable continuum model for the surroundings would probably have a similar effect.) The QM/MM embedding removes most low-energy CT states, even when the QM region is rather large (>120 atoms), but works *only* in conjunction with hybrid functionals. Although the same effect can be achieved without MM point charges by using plane-wave DFT with periodic boundary conditions, the plane-wave calculations are prohibitively expensive because hybrid functionals are still required.[6]

As an alternative to, or in conjunction with, a QM/MM embedding, CT states can also be eliminated by removing certain excitation amplitudes from the TD-DFT linear response equations, according to an automated procedure.[9] This procedure must be used with extreme caution, as it eliminates CT states (real or spurious) by ansatz. In cases where no real CT is expected, however, this technique significantly reduces the number of spurious states while introducing errors in the valence excitation energies that are typically smaller than 0.1 eV, at least for the examples considered here. Once again, the success of this technique is contingent upon use of a hybrid functional. Using BLYP, serious CT contamination persists, despite either of the aforementioned measures.

Finally, we note that the aforementioned procedures are intended only to remove CT "contamination", that is, the appearance of *spurious* CT states at low energies. Where *real* CT states are present (whose energies will of course be grossly underestimated by standard TD-DFT), the prescribed truncation of the excitation manifold will eliminate these as well. Electrostatic embedding, on the other hand, will modulate the energetics of real CT states, but it cannot be expected to compensate for the fundamentally incorrect way in which these states are described by contemporary TD-DFT. The simple procedures described here are therefore most applicable to studies of optically bright, valence excitations in large molecular systems.

Acknowledgment. This work was supported by grant no. PAS0291 from the Ohio Supercomputer Center and by start-up funds from The Ohio State University.

References

(1) Chiba, M.; Tsuneda, T.; Hirao, K. *Chem. Phys. Lett.* **2006**, *420*, 391.

(2) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.

(3) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andrés, L. *Mol. Phys.* **1999**, *97*, 859.

(4) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199.

(5) Bernasconi, L.; Sprik, M.; Hutter, J. *J. Chem. Phys.* **2003**, *119*, 12417.

(6) Bernasconi, L.; Sprik, M.; Hutter, J. *Chem. Phys. Lett.* **2004**, *394*, 141.

(7) Bernasconi, L.; Sprik, M. *J. Phys. Chem. B* **2005**, *109*, 12222.

(8) Odelius, M.; Kirchner, B.; Hutter, J. *J. Phys. Chem. A* **2004**, *108*, 2044.

(9) Besley, N. A. *Chem. Phys. Lett.* **2004**, *390*, 124.

(10) Neugebauer, J.; Louwerse, M. J.; Baerends, E. J.; Wesołowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115.

(11) Neugebauer, J.; Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2006**, *124*, 214102.

(12) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697−12699.

(13) Liao, M.-S.; Lu, Y.; Scheiner, S. *J. Comput. Chem.* **2003**, *24*, 623.

(14) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.

(15) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(16) Sundholm, D. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4265.

(17) Magyar, R. J.; Tretiak, S. *J. Chem. Theory Comput.* **2007**, *3*, 976.

(18) Hirata, S.; Zhan, C.-G.; Apra, E.; Windus, T. L.; Dixon, D. A. *J. Phys. Chem. A* **2003**, *107*, 10154.

(19) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425.

(20) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.

(21) Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2004**, *121*, 655.

(22) Peach, M. J. G.; Helgaker, T.; Sałek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558.

(23) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(24) Jacquemin, D.; Perpéte, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(25) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(27) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

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

(29) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.

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

(31) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(32) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889.

(33) Wathelet, V.; Preat, J.; Bouhy, M.; Fontaine, M.; Perpete, E. A.; André, J.-M.; Jacquemin, D. *Int. J. Quantum Chem.* **2006**, *106*, 1853.

(34) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(35) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.

(36) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(37) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291.

(38) Shao, Y.; Fusti-Molnar, L.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; Jr., R. A. D.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F., III; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(39) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics Modell.* **1996**, *14*, 33.

(40) Chien, S.-H.; Gill, P. M. W. *J. Comput. Chem.* **2006**, *27*, 730.

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

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Imprey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(43) *Tinker, v 4.2*. http://dasher.wustl.edu/tinker, downloaded on 10/03/06 (accessed Oct 3, 2006).

(44) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *302*, 375.

(45) Lorentzon, J.; Fülscher, M.; Roos, B. O. *J. Am. Chem. Soc.* **1995**, *117*, 9265.

(46) Head-Gordon, M.; Grana, A. M.; Maurice, D.; White, C. A. *J. Phys. Chem.* **1995**, *99*, 14261.

(47) Furche, F. *J. Chem. Phys.* **2001**, *114*, 5982.

(48) Improta, R.; Barone, V. *J. Am. Chem. Soc.* **2004**, *126*, 14320.

(49) Callis, P. R. *Annu. Rev. Phys. Chem.* **1983**, *34*, 329.

CT700125V

JCTC Journal of Chemical Theory and Computation

Nuclear Magnetoelectric Shieldings for Chiral Discrimination in NMR Spectroscopy. Theoretical Study of (R_a)-1,3-Dimethylallene, (2*R*)-2-Methyloxirane, and (2*R*)-*N*-Methyloxaziridine Molecules

Stefano Pelloni* and Paolo Lazzeretti

Dipartimento di Chimica, Università degli Studi di Modena e Reggio Emilia,
via G. Campi 183, 41100 Modena, Italy

Riccardo Zanasi

Dipartimento di Chimica dell'Università degli Studi di Salerno, via Ponte don Melillo,
84084 Fisciano (SA), Italy

Received May 17, 2007

Abstract: Dynamic magnetoelectric shieldings at the nuclei, having the same magnitude but opposite sign in D and L enantiomers, have been evaluated at the random-phase approximation level of accuracy for three chiral molecules of medium size. For frequencies normally operated in nuclear magnetic resonance spectroscopy, calculated values are probably too small to be detectable in disordered phase. Within the same experimental conditions, the isotropic part of nuclear magnetic shielding polarizability and a related pseudoscalar provide contributions 3 orders of magnitude bigger than the average magnetoelectric shieldings to (i) the magnetic field induced at a resonant nucleus and (ii) the induced electric dipole of electrons rotating at the Larmor frequency; therefore, nuclear magnetic shielding polarizabilities are probably more suitable than nuclear magnetoelectric shieldings for chiral discrimination in nuclear magnetic resonance spectroscopy.

1. Introduction

There has been interest in the possible applications of optical techniques in nuclear magnetic resonance (NMR)[1−7] and in electron spin resonance (ESR)[8,9] spectroscopies. Evans had argued that a circularly polarized laser beam could shift NMR frequencies to gigahertz values,[10−12] but his point was questioned by Barron via group-theoretical considerations.[13] Experimental evidence for the resonance shift of ≈1 Hz in an NMR spectrum, operating at 270 MHz, by optical irradiation at transparent wavelengths, was reported by Warren et al.[14,15] This estimate was revised in a successive paper, taking heating effects into account.[5]

Buckingham and Parlett[3,4] discussed mechanisms producing induced magnetic moment via the inverse Faraday effect[16] and magnetic field at a resonant nucleus (and consequent resonance shift) for a sample irradiated by circularly polarized light. The shifts so determined are small, but they are reversed by a change of the handedness of light.[3] For circularly polarized radiation, Harris and Tinoco[1,2] found negligible effects of light intensities on chemical shifts. The achiral fractional shift change due to a single proton absorption at resonance is estimated as big as 0.1 Hz. Chiral contributions would be 3 orders of magnitude too small to be observed.[17]

The molecular property emphasized by Buckingham and Parlett[4] is the antisymmetric polarizability induced by a nuclear magnetic moment interacting with the optical field. A different intrinsic response property—the nuclear magnetoelectric shielding (axial) tensor—has been considered to rationalize magnetic effects at the nuclei of a molecule in the presence of an electromagnetic field.[18−20]

* Corresponding author e-mail: pelloni.stefano@unimo.it.

10.1021/ct700117y CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/25/2007

All the (nonvanishing) components (and the trace) of the magnetoelectric shielding at the nuclei could in principle be measured for a system in an ordered medium, e.g., in a crystal or in a liquid crystal matrix, although theoretical predictions for some molecules are extremely small.[18,21,22] The components (and the trace) of the shielding tensor have the identical absolute value but change sign for the same atom in two enantiomeric molecules.[23] Such a property could be useful for practical applications, according to recent studies. Buckingham[24] and Buckingham and Fischer[25] speculated on the possibility of chiral discrimination by applying an additional electric field to conventional NMR spectrometers. They showed that in chiral liquids there is an additional signal—an electric polarization—that is generated only by chiral centers and that precesses at the resonance frequency of that nucleus. Observation of the polarization would lift the chiral blindness of NMR.

The present paper sets out to evaluate the magnitude of average nuclear magnetoelectric shieldings in three optically active molecules, (R_a)-1,3-dimethylallene, (2*R*)-2-methyloxirane, and (2*R*)-*N*-methyloxaziridine. A theoretical overview is given in section 2, and calculated values at the random-phase approximation[26] (RPA) level, equivalent to the coupled Hartree−Fock (CHF) approach,[27] are reported in section 3.

2. Response Tensors for Chiral Discrimination

For a molecule with *n* electrons and *N* nuclei, charge, mass, position with respect to an arbitrary origin, canonical, and angular momentum of the *i*th electron are indicated by $-e$, m_e, $\mathbf{r}_i$, $\mathbf{p}_i$, $\mathbf{l}_i = \mathbf{r}_i \times \mathbf{p}_i$, $i = 1,2 \ldots n$. Analogous quantities for nucleus *I* are $Z_I e$, M_I, $\mathbf{R}_I$, etc. Capital letters are used for global electronic operators, e.g., $\hat{\mathbf{R}} = \sum_{i=1}^{n} \mathbf{r}_i$, $\hat{\mathbf{P}} = \sum_{i=1}^{n} \mathbf{p}_i$, etc. The Einstein convention of implicit summation over two repeated Greek subscripts is in force, and $\epsilon_{\alpha\beta\gamma}$ denotes the Levi-Civita unit pseudotensor. The electronic reference state $|a\rangle \equiv |\Psi_a^{(0)}\rangle$ and the excited states $|j\rangle \equiv |\Psi_j^{(0)}\rangle$ of the molecule are eigenfunctions of the unperturbed time-independent Hamiltonian $H^{(0)}$, i.e., $H^{(0)}\,|\Psi_j^{(0)}\rangle = E_j^{(0)}|\Psi_j^{(0)}\rangle$. The natural transition frequencies are indicated by $\omega_{ja} = \hbar^{-1}(E_j^{(0)} - E_a^{(0)})$. SI units are employed.

Using a formalism previously established,[18−20,23] the electric dipole operator is written

$$\hat{\mu} = -e\hat{\mathbf{R}} \tag{1}$$

and the operator for the magnetic field of *n* electrons on nucleus *I*, carrying an intrinsic magnetic dipole $\mathbf{m}_I$, is

$$\hat{\mathbf{B}}_I^n = -\frac{\mu_0}{4\pi}\frac{e}{m_e}\sum_{i=1}^{n}\frac{\mathbf{r}_i - \mathbf{R}_I}{|\mathbf{r}_i - \mathbf{R}_I|^3} \times \mathbf{p}_i \tag{2}$$

The expressions for first-order perturbing Hamiltonians used within the time-dependent perturbation scheme[28] are

$$\hat{H}^{\mathbf{E}} = -\hat{\mu}_\alpha E_\alpha \tag{3}$$

$$\hat{H}^{\mathbf{m}_I} = -\hat{B}_{I\alpha}^n m_{I\alpha} \tag{4}$$

The operator for the electric field acted on nucleus *I* by electron *i* is

$$\hat{\mathbf{E}}_I^i = \frac{1}{4\pi\epsilon_0}\, e\,\frac{\mathbf{r}_i - \mathbf{R}_I}{|\mathbf{r}_i - \mathbf{R}_I|^3} \tag{5}$$

and

$$\hat{\mathbf{E}}_I^n = \sum_{i=1}^{n} \hat{\mathbf{E}}_I^i \tag{6}$$

is the operator for the total field of *n* electrons. The force of the *N* nuclei on the *n* electrons is

$$\hat{\mathbf{F}}_n^N = -e\sum_{I=1}^{N}\sum_{i=1}^{n} Z_I \hat{\mathbf{E}}_I^i \tag{7}$$

Adopting the notation used by Orr and Ward[29] and Bishop[30] (OWB), the polarization propagator[31] for two operators $\hat{A}$ and $\hat{B}$ is defined

$$\langle\langle \hat{A}; \hat{B}\rangle\rangle_\omega = -\sum_P \sum_{j\neq a} \frac{\langle a|\hat{A}|j\rangle\langle j|\hat{B}|a\rangle}{E_j^{(0)} - E_a^{(0)} - \hbar\omega_\sigma} \tag{8}$$

where $\sum_P$ indicates the sum over permutations of the pairs $(\hat{A}/-\omega_\sigma)$ and $(\hat{B}/\omega_1)$, and $\omega_\sigma = \omega_1 \equiv \omega$. The electric dipole magnetoelectric shielding at nucleus I[18,19,21−23] is obtained by the real and imaginary contributions to the propagator $\langle\langle \hat{B}_{I\alpha}^n; \hat{\mu}_\beta\rangle\rangle_\omega$

$$\lambda_{\alpha\beta}^I(-\omega;\omega) = -\mathscr{R}\langle\langle \hat{B}_{I\alpha}^n; \hat{\mu}_\beta\rangle\rangle_\omega = \frac{1}{\hbar}\sum_{j\neq a}\frac{2\omega_{ja}}{\omega_{ja}^2 - \omega^2}\,\mathscr{R}(\langle a|\hat{B}_{I\alpha}^n|j\rangle\langle j|\hat{\mu}_\beta|a\rangle) \tag{9}$$

$$\lambda_{\alpha\beta}^{\prime I}(-\omega;\omega) = \mathscr{I}\langle\langle\, \hat{B}_{I\alpha}^n; \hat{\mu}_\beta\rangle\rangle_\omega = -\frac{1}{\hbar}\sum_{j\neq a}\frac{2\omega}{\omega_{ja}^2 - \omega^2}\,\mathscr{I}(\langle a|\hat{B}_{I\alpha}^n|j\rangle\langle j|\hat{\mu}_\beta|a\rangle) \tag{10}$$

Equations 9 and 10 hold within the dipole length formalism.[19,23] Alternative definitions for the magnetoelectric shielding (10) are found in the dipole velocity gauge

$$\lambda_{\alpha\beta}^{\prime I}(-\omega;\omega) = -\frac{e}{m_e\hbar}\sum_{j\neq a}\frac{2\omega}{\omega_{ja}(\omega_{ja}^2 - \omega^2)}\,\mathscr{R}(\langle a|\hat{B}_{I\alpha}^n|j\rangle\langle j|\hat{P}_\beta|a\rangle) \tag{11}$$

and in the dipole acceleration gauge

$$\lambda_{\alpha\beta}^{\prime I}(-\omega;\omega) = -\frac{e}{m_e\hbar}\sum_{j\neq a}\frac{2\omega}{\omega_{ja}^2(\omega_{ja}^2 - \omega^2)}\,\mathscr{I}(\langle a|\hat{B}_{I\alpha}^n|j\rangle\langle j|\hat{F}_{n\beta}^N|a\rangle) \tag{12}$$

allowing for the off-diagonal hypervirial relationships

$$\langle a|\hat{R}_\alpha|j\rangle = \frac{i}{m_e}\,\omega_{ja}^{-1}\langle a|\hat{P}_\alpha|j\rangle = -\frac{1}{m_e}\,\omega_{ja}^{-2}\langle a|\,\hat{F}_{n\alpha}^N|j\rangle$$

$$= \frac{e}{m_e}\,\omega_{ja}^{-2}\sum_{I=1}^{N} Z_I\langle a|\hat{E}_{I\alpha}^n|j\rangle \tag{13}$$

Equations 13 for the matrix elements of the electric dipole operator are exactly satisfied by optimal variational wavefunctions.[32] In this ideal case, the quantum mechanical definitions 10−12 are theoretically and computationally equivalent.

Within the algebraic approximation, molecular orbitals (MO) are expressed as a linear combination of atomic orbitals (LCAO). In the limit of a complete set of expansion, the same numerical results would be obtained via (10)−(12). Therefore, closeness of the numerical results from different gauges (10)−(12) provides a measure of basis set quality and completeness.

The magnetoelectric shieldings 9 and 10−12 are equivalent to the polarizabilities introduced by Buckingham[24]

$$\lambda^I_{\alpha\beta} \equiv -\xi^I_{\beta\alpha}, \quad \lambda'^I_{\alpha\beta} \equiv -\xi'^I_{\beta\alpha}$$

2.1. Magnetic Field Induced at a Nucleus by an Electric Field. In NMR spectroscopy a strong static magnetic field with flux density $\mathbf{B}^{(0)}$ produces a torque on a nuclear magnetic dipole $\mathbf{m}_I$, inducing Larmor precession with angular frequency ω. If we take $\mathbf{B}^{(0)}$ parallel to the z axis, the magnetization precesses around this axis. A magnetic field $\mathbf{B}^{(1)}$ perpendicular to $\mathbf{B}^{(0)}$, formally required to be circularly polarized and synchronized with the precession of $\mathbf{m}_I$ about $\mathbf{B}^{(0)}$, so that there is a coherence in the phase of the precessing nuclei, rotates the bulk magnetization of the precessing nuclei from the z direction to the xy plane.[33,34]

However, a linearly polarized field is adequate, as it can be thought of as the superimposition of two circularly polarized fields rotating in opposite directions. Only the component having the same sense as the precession synchronizes with the nuclear magnetic dipole; the other component is far from the resonance condition and has no effect.[33,34] In practice, a radiowave of frequency ω is put in by passing a current through a coil at right angles to $\mathbf{B}^{(0)}$ so as to induce an oscillating magnetic field $\mathbf{B}^{(1)}$ along the coil axis, say in the x direction. If ω coincides with the resonance frequency ω_I, for nucleus I in a given molecular domain, the magnetic dipole will absorb energy from the coil, causing the macroscopic magnetization to rotate toward the xy plane and an emf to be induced in a receiver coil in the y direction.[34]

In modern NMR spectrometers, a $\pi/2$ pulse rotates the equilibrium magnetization from the z direction of the magnetic field into the xy plane. Its free induction decay is recorded by the receiver coil. The NMR spectrum is obtained through Fourier transformation of the signal.

The total effective field acting upon $\mathbf{m}_I$ is the vector sum $\mathbf{B}(\omega) = \mathbf{B}^{(0)} + \mathbf{B}^{(1)}(\omega)$. The response property involved is $\sigma^I_{\alpha\beta}$, the magnetic shielding of nucleus I, so that the magnetic field induced at the probe nucleus is[19,20] $\Delta\langle B^{n'}_{I\alpha}\rangle = -\sigma^I_{\alpha\beta}B_\beta$. In an isotropic medium we take the average $\sigma_{\mathrm{av}} = (1/3)\sigma^I_{\alpha\alpha}$.

In principle, a time-dependent electric field $\mathbf{E}(\omega)$ can be used to induce a magnetic field at nucleus I.[18−20,23] Time-dependent perturbation theory (TDPT)[28,29,35,36] can be applied to discuss the phenomenology. The expectation value of the magnetic field induced at the position of nucleus I by the n-electron cloud of a molecule responding to an external electric field, described as a monochromatic wave with pulsation ω, is given by the expression

$$\Delta\langle \hat{B}^n_{I\alpha}(\omega)\rangle = \lambda^I_{\alpha\beta}(-\omega;\omega)E_\beta(\mathbf{0},\omega t) + \lambda'^I_{\alpha\beta}(-\omega;\omega)\dot{E}_\beta(\mathbf{0},\omega t)\omega^{-1} \quad (14)$$

where $E_\alpha(\mathbf{0}, \omega t)$ is the time-varying electric field at the origin of the coordinate system. A dot denotes partial time derivative.

The terms on the rhs of expression 14 have been obtained assuming spatially uniform electric field[37] over the molecular domain, within the Goeppert-Mayer dipole approximation.[38,39] One can also regard the magnetic field evaluated by eq 14 as the contribution provided by a Fourier component of the perturbing electric field. According to an obvious interpretation of eq 14, the second-rank tensors $\lambda^I_{\alpha\beta}(-\omega;\omega)$ and $\lambda'^I_{\alpha\beta}(-\omega;\omega)$ have been called magnetoelectric nuclear shieldings.[18−20,23]

Within the assumption of harmonic fields, the electric field at the origin of the coordinate system is $E_\alpha = E^{(0)}_\alpha \cos(\omega t)$, and its partial time derivative in the second addendum on the rhs of eq 14 can be replaced by an out-of-phase field allowing for the identity

$$\dot{E}_\alpha\omega^{-1} = -E^{(0)}_\alpha \sin(\omega t) = E^{(0)}_\alpha \cos(\omega t + \pi/2) \quad (15)$$

Therefore the magnetic field induced at the nucleus becomes

$$\Delta\langle \hat{B}^n_{I\alpha}(\omega)\rangle = \lambda^I_{\alpha\beta}(-\omega;\omega)E_\beta(\mathbf{0},\omega t) + \lambda'^I_{\alpha\beta}(-\omega;\omega)E_\beta(\mathbf{0},\omega t + \pi/2) \quad (16)$$

In an isotropic phase, the magnetic field, induced by linear response in the direction of the perturbing electric field at nucleus I of a (time-even) diamagnetic molecule, is evaluated via the second term of the rhs of eq 16

$$\Delta\langle \hat{B}^n_{I\beta}(\omega)\rangle = \frac{1}{3}\lambda'^I_{\alpha\alpha}(-\omega;\omega)E_\beta(\mathbf{0},\omega t + \pi/2) \quad (17)$$

Within the model[34] quoted above, an alternating current is passed through a coil (mounted perpendicularly to $\mathbf{B}^{(0)}$), which in turn gives rise to a linearly polarized electromagnetic field. The oscillating magnetic (electric) field is polarized along the x (y) direction. Relationship 17 implies that a magnetic field $\mathbf{B}^{(1)}$ linearly polarized along the y direction (regarded as the vector sum of two circularly polarized fields on the xy plane) can be generated by an out-of-phase, time-varying electric field linearly polarized in the same direction. Then the effect of the magnetic field described by eqs 16 and 17, induced by the electric field, could, if large enough, be detected by a receiver coil as a $\pi/2$ out-of-phase signal.

2.2. Electric Dipole Induced by a Nuclear Magnetic Dipole. Allowing for the perturbing Hamiltonian 4 within the TDPT, an expression is obtained for the electric dipole induced by the precession of a permanent magnetic dipole at nucleus I [20,24,25]

$$\Delta\langle \mu_\alpha(\omega)\rangle = -\lambda^I_{\beta\alpha}(-\omega;\omega)m_{I\beta} - \lambda'^I_{\beta\alpha}(-\omega;\omega)\,\dot{m}_{I\beta}\omega^{-1} \quad (18)$$

Relationship 18 can be rearranged, expressing the contribution of the second addendum on the rhs to the induced electric dipole as an out-of-phase term, as for relation 16.

Table 1. Nonvanishing Components of the Axial Tensor $\lambda'^{I}_{\alpha\beta}$ for Various Nuclear Site Symmetries

	number of nonvanishing components		
nuclear site symmetry	total	unique	nonvanishing components
C_i, C_{2h}, D_{2h}, C_{4h}, D_{4h}, S_6, D_{3d}, C_{3h}, C_{6h}, D_{3h}, D_{6h}, $D_{\infty h}$, T_h, T_d, O_h	0	0	
C_1	9	9	*xx*, *yy*, *zz*, *xy*, *yx*, *xz*, *zx*, *yz*, *zy*
C_2	5	5	*xx*, *yy*, *zz*, *xy*, *yx*
C_s	4	4	*xz*, *yz*, *zx*, *zy*
D_2	3	3	*xx*, *yy*, *zz*
C_{2v}	2	2	*xy*, *yx*
C_4, C_3, C_6	5	3	$xx = yy$, zz, $xy = -yx$
S_4	4	2	$xx = -yy$, $xy = yx$
D_4, D_3, D_6	3	2	$xx = yy$, zz
C_{4v}, C_{3v}, C_{6v}, $C_{\infty v}$	2	1	$xy = -yx$
D_{2d}	2	1	$xx = -yy$
T, O	3	1	$xx = yy = zz$

The induced magnetic field $\Delta\langle\hat{B}^{n}_{I\alpha}(\omega)\rangle$ and electric dipole $\Delta\langle\hat{\mu}_{\alpha}(\omega)\rangle$ are invariant in a translation of coordinate system,[19,40] as the magnetoelectric shieldings $\lambda^{I}_{\alpha\beta}$ and $\lambda'^{I}_{\alpha\beta}$ are origin independent.

$\lambda^{I}_{\alpha\beta}$ is odd under time reversal $\hat{T}$ and inversion $\hat{P}$, and it vanishes for a closed-shell molecule in the absence of magnetic field.[19,20,24,25] However, it has a linear dependence on an applied magnetic field.[20,24,25] This case is best treated by taking into account quadratic response in terms of a third-rank tensor $\sigma^{I}_{\alpha\beta\gamma}$, referred to as polarizability of nuclear magnetic shielding.[41] The first term of eqs 14 and 18 suggests that, in the presence of an external magnetic field, we could replace the $\lambda^{I}_{\beta\alpha}$ magnetoelectric shielding by the shielding polarizability via the scheme[24,25]

$$\lambda^{I}_{\beta\alpha} \rightarrow \sigma^{I}_{\beta\gamma\alpha}B_{\gamma} \tag{19}$$

so that, for instance, its contribution to the induced electric dipole becomes

$$\Delta\langle\mu_{\alpha}\rangle = -\sigma^{I}_{\beta\gamma\alpha}m_{I\beta}B_{\gamma} \tag{20}$$

The isotropic part $\overline{\sigma^{(1)I}}\,\epsilon_{\alpha\beta\gamma}$ and the pseudoscalar $\overline{\sigma^{(1)I}} = (1/6)\epsilon_{\alpha\beta\gamma}\sigma^{I}_{\alpha\beta\gamma}$ have the same magnitude but opposite sign for enantiomeric molecules.[24,25] They are origin independent in the static case. The transformation law of dynamic polarizability of nuclear magnetic shielding in a gauge translation of the vector potential has been studied.[40]

2.3. Magnetoelectric Shielding and Site Symmetry. $\lambda'^{I}_{\alpha\beta}$ is even under $\hat{T}$ and odd under $\hat{P}$, and it is equal and opposite for the *I*th nucleus of D and L enantiomers, see the discussion after eq 921 of ref 23.

The number of unique parameters needed to describe this tensor, reported in Table 1, is obtained via group-theoretical methods,[42] taking into account the site symmetry. As the nuclear magnetoelectric shielding transforms like the optical rotatory power under symmetry operations, this number is the same as in the next-to-last column of Table 1 of ref 35.

Table 2. RPA Magnetoelectric Shieldings[a] of the (R_a)-1,3-Dimethylallene Molecule via Dipole Length (*R*), Velocity (*P*), and Acceleration (*F*) Formalisms, for $\omega = 10^{-7}$ au[49]

atom	form.	$xx \times 10^8$	$yy \times 10^8$	$zz \times 10^8$	av $\times 10^8$
C_1	*R*	8.111	−1.863	−6.427	−0.05952
	P	8.165	−1.873	−6.459	−0.05577
	F	8.239	−1.828	−6.562	−0.05018
C_3	*R*	−11.27	2.012	9.692	0.1442
	P	−11.23	2.077	9.594	0.1452
	F	−11.76	1.986	10.17	0.1325
C_5	*R*	0.2200	−0.1096	−0.3274	−0.07232
	P	−0.01894	−0.02162	−0.2214	−0.08733
	F	0.6061	0.9649	−1.764	−0.06420
H_1	*R*	−2.371	4.294	−0.9102	0.3377
	P	−2.383	4.308	−0.9161	0.3363
	F	−2.613	4.343	−0.8347	0.2986
H_3	*R*	2.512	−5.446	1.729	−0.4015
	P	2.521	−5.453	1.728	−0.4012
	F	2.687	−5.478	1.722	−0.3563
H_5	*R*	1.919	2.828	−1.525	1.074
	P	1.926	2.828	−1.528	1.075
	F	2.017	2.791	−1.487	1.107
H_7	*R*	−0.07287	−3.047	−0.1551	−1.092
	P	−0.06639	−3.050	−0.1607	−1.092
	F	−0.1514	−3.014	−0.1581	−1.108

[a] The magnetoelectric shieldings in SI atomic units are obtained from the values in the table multiplying by $\alpha^2 = 5.325 \ldots \times 10^{-5}$, see section 4. The conversion factor from SI au to SI units is $1/\alpha c = 4.571 \ldots \times 10^{-7}$ m^{-1} s.

The trace $\lambda'^{I}_{\alpha\alpha}$ vanishes for achiral molecules, but, depending on site symmetry, diagonal and off-diagonal tensor components can be different from zero,[18,21,22] see Table 1. Therefore, measurements of the average value $(1/3)\lambda'^{I}_{\alpha\alpha}$ may, in principle, serve for chiral discrimination in an isotropic phase.[24,25] However, the estimated magnitude is small, e.g., $|B_z\epsilon_{z\alpha\beta}\lambda^{I}_{\alpha\beta}/2\lambda'^{I}_{\gamma\gamma}| \approx 10^4$ in a field $B_z \approx 10$ T for $CH_3CH(OH)C_6H_5$.[24]

3. Theoretical Estimates of Average Magnetoelectric Shieldings

In the present study, a computational scheme[18,21,22] based on the random-phase approximation[26,27] was applied. Values of diagonal components of the $\lambda'^{I}_{\alpha\beta}$ tensor were calculated by the SYSMO package[43] for three chiral systems, (R_a)-1,3-dimethylallene, (2*R*)-2-methyloxirane, and (2*R*)-*N*-methyloxaziridine.

Molecular geometries have been optimized at the HF/6-31G level using GAUSSIAN03,[44] as described in previous calculations of shielding polarizabilities[45] related to $\lambda^{I}_{\alpha\beta}$, see section 2.2. Other data needed to reproduce our results are available in that paper.[45] The same basis sets were tested to check convergence of calculated $\lambda'^{I}_{\alpha\beta}$. The best results obtained via the basis sets developed in refs 18, 21, and 22 are shown in Tables 2 −4.

As a rule, reliable estimates are obtained via basis sets containing diffuse Gaussian functions, which provide accurate representation of the electric dipole, eq 1, rather than the magnetic field operator, eq 2. Therefore, Sadlej basis

Table 3. RPA Magnetoelectric Shieldings[a] of the (2*R*)-2-Methyloxyrane Molecule via Three Formalisms, for $\omega = 10^{-7}$ au[49]

atom	form.	$xx \times 10^8$	$yy \times 10^8$	$zz \times 10^8$	av $\times 10^8$
C_1	*R*	−5.533	2.683	0.4494	−0.8001
	P	−5.444	2.678	0.3844	−0.7939
	F	−4.557	1.917	0.2985	−0.7804
C_2	*R*	−8.867	−1.486	7.655	−0.8994
	P	−8.813	−1.535	7.704	−0.8813
	F	6.886	−1.418	6.002	−0.7638
C_3	*R*	−0.9702	0.5765	−0.3359	−0.2432
	P	−0.9461	0.5818	−0.3194	−0.2279
	F	−1.017	0.4915	−0.2989	−0.2749
O_1	*R*	65.88	−3.932	−46.88	5.023
	P	65.31	−3.597	−46.37	5.112
	F	50.79	−2.109	−34.68	4.667
H_1	*R*	0.9140	−1.042	0.2766	0.04961
	P	0.9168	−1.045	0.2807	0.05075
	F	0.8967	−1.006	0.1087	0.0
H_2	*R*	−1.836	3.382	−0.8728	0.2243
	P	−1.840	3.386	−0.8790	0.2221
	F	−1.827	3.211	−0.9490	0.1448
H_3	*R*	1.603	−2.686	0.7583	−0.1080
	P	1.613	−2.689	0.7574	−0.1062
	F	1.503	−2.587	0.9580	−0.04198
H_4	*R*	1.065	0.1711	−0.1803	0.3518
	P	1.067	0.1698	−0.1789	0.3526
	F	1.066	0.1377	−0.09872	0.3683
H_5	*R*	−1.030	0.1972	−0.4732	−0.4353
	P	−1.031	0.1973	−0.4751	−0.4361
	F	−0.9853	0.2025	−0.5056	−0.4295
H_6	*R*	−0.5083	0.9850	−0.1978	0.09295
	P	−0.5096	0.9857	−0.1977	0.09278
	F	−0.5314	0.9932	−0.1671	0.09826

[a] See footnote *a* to Table 2.

Table 4. RPA Magnetoelectric Shieldings[a] of the (2*R*)-*N*-Methyloxaziridine Molecule via Three Formalisms, for $\omega = 10^{-7}$ au[49]

atom	form.	$xx \times 10^8$	$yy \times 10^8$	$zz \times 10^8$	av $\times 10^8$
C_1	*R*	1.666	−4.467	−1.765	−1.522
	P	1.713	−4.496	−1.809	−1.531
	F	3.500	−5.590	−2.195	−1.428
C_2	*R*	3.427	−2.511	0.7644	0.5602
	P	3.422	−2.495	0.7569	0.5610
	F	4.372	−3.448	1.285	0.7362
N_1	*R*	−12.40	−37.93	30.73	−6.534
	P	−12.20	−38.57	30.73	−6.681
	F	51.83	−81.83	29.90	−0.03377
O_1	*R*	3.656	70.92	−33.24	13.78
	P	3.133	73.89	−32.90	14.71
	F	−65.43	174.2	−20.88	29.31
H_1	*R*	−0.8194	2.425	−0.8505	0.2516
	P	−0.8217	2.421	−0.8490	0.2502
	F	−0.5015	2.266	−0.8754	0.2963
H_2	*R*	2.886	−1.970	0.2514	0.3892
	P	2.889	−1.968	0.2596	0.3934
	F	3.112	−1.726	0.1230	0.5031
H_3	*R*	1.794	−0.5595	0.6338	0.6229
	P	1.792	−0.5639	0.6333	0.6203
	F	2.106	−0.9393	0.9611	0.7094
H_4	*R*	−1.120	0.1689	−0.6674	−0.5396
	P	−1.115	0.1720	−0.6724	−0.5386
	F	−0.8996	0.2889	−0.6002	−0.4036
H_5	*R*	0.1827	0.4023	−0.2735	0.1038
	P	0.1816	0.4120	−0.2752	0.1061
	F	0.1921	0.3414	−0.05236	0.1604

[a] See footnote *a* to Table 2.

sets[46] are well suited to calculate magnetoelectric shieldings within dipole length and velocity formalisms. They provide results close to those from the bigger aug-cc-pVTZ basis sets.[47,48]

The ω frequency assumed in the calculation is 10^{-7} au, close to the value of the ^{1}H resonant frequency in a field of 14.1 T, see section 4. By comparison with the natural frequencies ω_{ja}, it is seen that this value of ω can be neglected in the denominator of eq 10. Therefore the magnitude of $\lambda'^{I}_{\alpha\beta}$ for a given optical $\bar{\omega}$, up to $\approx 10^{-3}$ au, can approximately be estimated[49] multiplying the values of Tables 2−4 by $\bar{\omega} \times 10^7$.

According to previous experience, within the acceleration formalism, Gaussian functions are not suitable to describe accurately the electron charge density in the vicinity of a nucleus heavier than hydrogen.[18,21,22] Actually, the *F*-values obtained for the magnetoelectric shielding of oxygen and nitrogen nuclei are very poor in most cases. Strong discrepancies in magnitude and sign with *R*- and *P*-predictions can be observed in Tables 3 and 4.

Only for the (R_a)-1,3-dimethylallene molecule are the *F*-estimates for hydrogen and carbon quite close to *R*- and *P*-, see Table 2. However, the ability of Gaussian functions to represent the force operator can be improved by ad hoc procedures for "polarized" basis sets, providing a better description of the charge distribution near the nuclei.[50]

A preliminary attempt has been made in the present work, developing a "semipolarized" basis set from the Sadlej basis sets[46] for the (2*R*)-*N*-methyloxaziridine molecule. To keep the basis size as small as possible, the set of hydrogen functions was left unchanged, as it is large enough to provide almost the same results within different gauges in Tables 2−4. The basis sets of C, N, and O were polarized via the recipe of ref 50. The predictions obtained in the *F*-formalism, see Table 2 of the Supporting Information, are significantly improved compared to Table 4. The results from *R*- and *P*-approaches also agree quite nicely.

Other procedures based on R12 basis sets[51] can be applied according to previous investigations[52] to improve predictions within the acceleration formalism. The *F*-results obtained for the water molecule, reported in Table 1 of the Supporting Information, are close to convergence, i.e., fairly similar to those from *R*- and *P*-formalisms. They show that the problem can, at least in principle, be solved.

Therefore, the accuracy of the present calculation was assessed via closeness of results within dipole length and dipole velocity gauges from the basis sets employed[18,21,22] (We recall that they should be the same for complete basis sets.). It can reasonably be assumed that calculated *R*- and *P*-values for the molecules examined here are close to the limit for the RPA scheme.

4. Units and Magnitude of Observable Properties

The calculations were carried out setting to 1 the base units[53] of mass m_e = 9.109 381 88 × 10^{-31} Kg, charge e = 1.602 176 462 × 10^{-19} C, action $\hbar$ = 1.054 571 596 × 10^{-34} J s, permittivity $\kappa_0 = 4\pi\epsilon_0$, with ϵ_0 = 8.854 187 817 . . . × 10^{-12} F m^{-1}, and speed of light c = 299 792 458 m s^{-1}.

Derived units, e.g., the bohr a_0 = 0.529 177 2083 × 10^{-10} m and the hartree $E_h = m_e e^4/\kappa_0^2\hbar^2 = e^2/\kappa_0 a_0$ = 4.359 743 81 × 10^{-18} J, are accordingly set to 1. The magnetic constant is $\mu_0 = 1/\epsilon_0 c^2$ and $\mu_0/4\pi = 1 \times 10^{-7}$ N A^{-2}. The CODATA recommended values for the base and derived units are taken from ref 54.

The magnetoelectric shieldings 10 in (SI) atomic units are obtained from the values in the tables by expressing $\mu_0/4\pi = \alpha\hbar/e^2c$ in the same units via the fine structure constant $\alpha = e^2/\kappa_0\hbar c$ = 7.297 352 533 × 10^{-3}. The conversion factor is α^2 = 5.325 135 399 × 10^{-5}. Within the SI system of units, the magnetoelectric shielding has dimension T V^{-1} m ≡ m^{-1} s, i.e., inverse velocity, see eq 14. Therefore the conversion factor from SI au to SI units is $1/\alpha c$ = 4.571 028 927 × 10^{-7} m^{-1} s.

Within the cgs system of units, the magnetoelectric shielding is dimensionless. Its definition contains c in the denominator, instead of $\mu_0/4\pi$, see, for instance, eq 33 of ref 18. Then the value in the cgs system is obtained multiplying by α the results in the tables. If the electric field is expressed in statV cm^{-1}, with 1 V m^{-1} = $10^4/c$ statV cm^{-1}, the induced magnetic field is obtained in gauss (1 G = 10^{-4} T).

The largest value obtained for the average magnetoelectric shielding of ^{17}O in (2*R*)-*N*-methyloxaziridine is ≈14 × 10^{-8} in the units of Table 4, corresponding to 7.5 × 10^{-12} au, that is, 3.4 × 10^{-18} T mV^{-1}, for a pulsation $\omega \approx 1 \times 10^{-7}$ au of the order typically available in proton magnetic resonance experiments[49] (e.g., for ^{1} H, in a magnetic field B = 14.1 T, resonance occurs at 600 MHz, equivalent to 9.12 × 10^{-8} au). However, the magnetic field density ≈114 T, needed for ^{17}O resonance at $\omega = 1 \times 10^{-7}$ au, is unrealistic. Therefore, we take a scaled $\lambda'^{O}_{av} \approx 4.6 \times 10^{-19}$ T V^{-1} m for the resonance frequency ν = 89.23 MHz of ^{17}O in a field of 15.45 T.[49]

This value is used to estimate the magnetic field induced at the oxygen nucleus by an electric field and to make a comparison with a corresponding estimate of magnetic field induced via the isotropic part $\overline{\sigma^{(1)O}}\ \epsilon_{\alpha\beta\gamma}$ of the magnetic hypershielding at the oxygen nucleus in the same molecule.[45] Equations 16 and 19 show that an electric field can be applied to observe an induced magnetic field at ^{17}O. A calculated value[45] for the pseudoscalar[25] $\overline{\sigma^{(1)O}}$ is ≈7.8 × 10^{-17} mV^{-1}, and then the electric field should be as big as ≈1.3 × 10^8 V m^{-1} to induce a magnetic field corresponding to 0.01 ppm,[45] normal to the strong magnetic field of an NMR spectrometer, that is 1.545 × 10^{-7} T, if we assume to operate at 15.45 T.[49] The same electric field would induce a magnetic field of magnitude 6.0 × 10^{-11} T at ^{17}O in (2*R*)-*N*-methyloxaziridine, allowing for the second term in eq 16. The ratio of the first term to the second in this relationship is ≈2.6 × 10^3.

The magnitude of the electric dipole induced by the precession of ^{17}O nuclear magnetic dipole, ≈1.132 × 10^{-26} JT^{-1}, is analogously estimated via eq 18. In a magnetic field of 15.45 T, the contribution from the first term on the rhs of this relationship, allowing for scheme 19, is[45] ≈1.4 × 10^{-41} Cm. The second term contributes ≈5.2 × 10^{-45} Cm.[55] The ratio is ≈2.6 × 10^3. Therefore, the contribution arising from the magnetoelectric shielding $\lambda'^{O}_{av} \approx 4.6 \times 10^{-19}$ T mV^{-1} is negligible, compared with that from $\overline{\sigma^{(1)O}} \approx 7.8 \times 10^{-17}$ mV^{-1} in magnetic fields normally available in NMR,[49] unless higher ω values were used via experimental techniques that cannot be foreseen at the present time.

It can be asked if any effect could be detected in other spectral regions, e.g., for a molecule absorbing infrared radiation. Equation 18 would imply that the vibration of a nucleus carrying a permanent magnetic dipole moment also induces an electric dipole in the electrons, oscillating with the same frequency as that of the nuclear motion. The average electric dipole vanishes in isotropic phase. However, if a strong magnetic field B_z is applied, there is a magnetization along the z axis, which arises from magnetic moments of nuclei precessing at the Larmor frequency. The macroscopic magnetization should oscillate, following the nuclear vibration. The frequency ω of the oscillating induced electric dipole would be ≈10^4−10^6 times bigger than the Larmor frequency, of the order 10^8 Hz in NMR experiments. As the vibrational frequency ω in the denominator of eq 10 is negligible compared to the natural transition frequencies ω_{ja}, the magnetoelectric shieldings in Tables 1−3 would therefore increase by a factor of ≈10^5 due to ω in the numerator. The magnitude of the induced dipole 18 would increase to the same extent.

5. Concluding Remarks

The magnetic field induced at the nuclei of a molecule by a time-varying electric field and the rotating electric dipole induced by the precession of nuclear magnetic dipoles have the same magnitude but different sign for D and L enantiomers. The dynamic magnetoelectric shieldings at the nuclei of three molecules, (R_a)-1,3-dimethylallene, (2*R*)-2-methyloxirane, and (2*R*)-*N*-methyloxaziridine, have been calculated for a frequency $\omega = 10^{-7}$ au, close to the ^{1}H resonant frequency in a field of 14.1 T, at the random-phase approximation level via extended basis sets.

The accuracy of the theoretical predictions was established via closeness of corresponding results within dipole-length and dipole-velocity gauges. The average values, defined as one-third the trace of the tensor, are usually much smaller than the diagonal tensor components. The latter are characterized by a different sign, so that partial cancellation occurs—the situation is analogous to that observed for the optical rotatory power tensor.

The calculations show that, for the molecules considered, the order of magnitude of the frequency dependent average magnetoelectric shielding, in the most favorable case, i.e., for oxygen shielding in (2*R*)-*N*-methyloxaziridine, is approximately 4.6 × 10^{-19} T m V^{-1} at the resonance frequency 89.23 MHz in a magnetic field of 15.45 T. Therefore, in the disordered phase, the contributions from the average magnetoelectric shielding to the magnetic field induced at ^{17}O and to the rotating orbital electric dipole are negligible compared with those arising from the isotropic part of the nuclear magnetic shielding polarizability in magnetic fields operated in NMR spectroscopy.

Acknowledgment. The authors would like to thank Prof. A. D. Buckingham, Dr. P. Fischer, and Prof. R. A. Harris for helpful comments. Financial support to the present research from the Italian MURST (Ministero dell'Università e della Ricerca Scientifica e Tecnologica), via PRIN funds, is gratefully acknowledged.

Supporting Information Available: Test of convergence of nuclear magnetoelectric shieldings in the acceleration formalism for the water molecule and estimates for the (2*R*)-*N*-methyloxaziridine molecule via a semipolarized Sadlej basis set. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Harris, R. A.; Tinoco, I., Jr. *Science, N. Y.* **1993**, *259*, 835−836.

(2) Harris, R. A.; Tinoco, I., Jr. *J. Chem. Phys.* **1994**, *101*, 9289−9294.

(3) Buckingham, A. D.; Parlett, L. C. *Science, N. Y.* **1994**, *264*, 1748−1750.

(4) Buckingham, A. D.; Parlett, L. C. *Mol. Phys.* **1997**, *91*, 805−813.

(5) Warren, W. S.; Goswami, D.; Mayr, S. *Mol. Phys.* **1993**, *93*, 371−375.

(6) Li, L.; He, T.; Chen, D.; Wang, X.; Liu, F.-C. *J. Phys. Chem. A* **1998**, *102*, 10385−10390.

(7) Jaszuński, M.; Rizzo, A. *Mol. Phys.* **1999**, *96*, 855−861.

(8) Buckingham, A. D.; Parlett, L. C. *Chem. Phys. Lett.* **1995**, *243*, 15−21.

(9) Li, L.; He, T.; Wang, X.; Liu, F.-C. *Chem. Phys. Lett.* **1998**, *268*, 549−556.

(10) Evans, M. W. *J. Phys. Chem.* **1991**, *95*, 2256−2260.

(11) Evans, M. W. *Physica B (Amsterdam)* **1992**, *182*, 227−236.

(12) Evans, M. W. *Adv. Chem. Phys.* **1993**, *51*, 85.

(13) Barron, L. D. *Physica B (Amsterdam)* **1992**, *190*, 307−309.

(14) Warren, W. S.; Mayr, S.; Goswami, D.; West, A. P., Jr. *Science, N. Y.* **1992**, *255*, 1683−1685.

(15) Warren, W. S.; Mayr, S.; Goswami, D.; West, A. P., Jr. *Science, N. Y.* **1993**, *259*, 836.

(16) van der Ziel, J. P.; Pershan, P. S.; Malmstrom, L. D. *Phys. Rev. Lett.* **1965**, *15*, 190−193.

(17) According to a private communication quoted in section IV of ref 2, W. S. Warren eventually agreed that chiral shifts are unobservable in the conditions discussed in previous references.[14,15]

(18) Lazzeretti, P.; Zanasi, R. *Phys. Rev. A* **1986**, *33*, 3727−3741.

(19) Lazzeretti, P. *Adv. Chem. Phys.* **1987**, *75*, 507−549.

(20) Lazzeretti, P. *Chem. Phys.* **1989**, *134*, 269−278.

(21) Lazzeretti, P.; Zanasi, R. *J. Chem. Phys.* **1987**, *87*, 472−480.

(22) Lazzeretti, P.; Zanasi, R.; Bursi, R. *J. Chem. Phys.* **1988**, *89*, 987−996.

(23) Lazzeretti, P. Electric and Magnetic Properties of Molecules. In *Handbook of Molecular Physics and Quantum Chemistry*; John Wiley & Sons, Ltd.: Chichester, 2003; Vol. 3, Part 1, Chapter 3.

(24) Buckingham, A. D. *Chem. Phys. Lett.* **2004**, *398*, 1−5.

(25) Buckingham, A. D.; Fischer, P. *Chem. Phys.* **2006**, *324*, 111−116.

(26) Rowe, D. J. *Rev. Mod. Phys.* **1968**, *40*, 153−166.

(27) Jørgensen, P.; Simons, J. *Second Quantization-Based Method in Quantum Chemistry;* Academic Press: New York, 1981; p 149.

(28) Langhoff, P. W.; Epstein, S. T.; Karplus, M. *Rev. Mod. Phys.* **1972**, *44*, 602−644.

(29) Orr, B. J.; Ward, J. F. *Mol. Phys.* **1971**, *20*, 513−526.

(30) Bishop, D. M. *Rev. Mod. Phys.* **1990**, *62*, 343−374.

(31) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235−3264.

(32) Epstein, S. T. *The Variation Method in Quantum Chemistry;* Academic Press: New York, 1974; pp 92−104.

(33) Abragam, A. *Principles of Nuclear Magnetism;* Oxford University Press: Amen House, London, 1961; Chapter II, p 19.

(34) Emsley, J. W.; Feeney, J.; Sutcliffe, L. H. *High Resolution Nuclear Magnetic Resonance Spectroscopy;* Pergamon Press: Oxford, 1967; pp 10−13.

(35) Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107−142.

(36) Bogaard, M. P.; Orr, B. J. Electric Dipole Polarisabilities of Atoms and Molecules. In *International Review of Science. Molecular Structure and Properties. Physical Chemistry Series Two*; Buckingham, A. D., Ed.; Butterworths: London, 1975; Vol. 2.

(37) In the radiation gauge the scalar potential vanishes. Within the dipole approximation, the vector potential is assumed uniform over the molecular dimensions, and then the electric field is also spatially uniform.

(38) Goeppert-Mayer, M. *Ann. Phys. (Leipzig)* **1931**, *9*, 273.

(39) Cohen-Tannoudji, C.; Dupont-Roc, J.; Grynberg, G. *Photon & Atoms;* John Wiley & Sons: New York, 1989; p 269.

(40) Lazzeretti, P.; Soncini, A.; Zanasi, R. *Theor. Chem. Acc.* in press. DOI:10.1007/s00214-006-0184-3.

(41) Buckingham, A. D. *Can. J. Chem.* **1960**, *38*, 300−307.

(42) McWeeny, R. *Symmetry;* Pergamon Press: Oxford, 1963; pp 211−218.

(43) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Technical Report on Project "Sistemi Informatici e Calcolo Parallelo"*; Research Report 1/67; CNR: 1991.

(44) Frisch, M. J.; Trucks, G. W.; et al. *Gaussian 2003, Revision B.05;* Gaussian, Inc.: Pittsburgh, PA, 2003.

(45) Zanasi, R.; Pelloni, S.; Lazzeretti, P. *J. Comput. Chem.* **2007**, *28*, 2159−2163.

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

(47) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(48) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572−4585.

(49) As 1 au of ω is 6.579 683 921 × 10^{15} Hz from ref 54, and the resonance frequency of a nucleus with magnetogyric ratio γ_I is $\nu_I = \gamma_I B/2\pi$, the magnetic field density B for proton resonance is ≈15.45 T for $\omega = 1 \times 10^{-7}$ au. However, if one takes this value for a proton, then, allowing for the corresponding magnetogyric ratios,[56] ω is only 0.2514 times this frequency for ^{13}C, 0.1014 times for ^{15}N, and 0.1356 times for ^{17}O (for $B = 15.45$ T). It is unrealistic to use $\omega = 1 \times 10^{-7}$ au for these heavier nuclei; therefore, the magnetoelectric shieldings of ^{13}C, ^{15}N, and ^{17}O in the tables should be scaled by the same factors.

(50) Wolinski, K.; Roos, B. O.; Sadlej, A. J. *Theor. Chim. Acta* **1985**, *68*, 431−444.

(51) Bakken, V.; Helgaker, T.; Klopper, W.; Ruud, K. *Mol. Phys.* **1999**, *96*, 653−671.

(52) Soncini, A.; Lazzeretti, P.; Bakken, V.; Helgaker, T. *J. Chem. Phys.* **2004**, *120*, 3142−3151.

(53) McWeeny, R. *Nature* **1973**, *243*, 196−198.

(54) Mohr, P. J.; Taylor, B. N. *Rev. Mod. Phys.* **2000**, *72*, 351−495, Table XXVIII.

(55) In the case of ^{17}O with $I = 5/2$ there are also nuclear electric quadrupole contributions that are not present on the $I = 1/2$ nuclei. Therefore, the ^{17}O nucleus, although displaying the largest isotropic magnetoelectric shielding in the compounds studied here, is most likely all but an ideal probe, due to low receptivity and unfavorable nuclear spin, which may cause undesirable lifetime broadening of signals.

(56) www.webelements.com.

CT700117Y

JCTC Journal of Chemical Theory and Computation

Finite-Field Spin-Flip Configuration Interaction Calculation of the Second Hyperpolarizabilities of Singlet Diradical Systems

Ryohei Kishi,*,† Masayoshi Nakano,*,† Suguru Ohta,† Akihito Takebe,†
Masahito Nate,† Hideaki Takahashi,† Takashi Kubo,‡ Kenji Kamada,§ Koji Ohta,§
Benoît Champagne,‖ and Edith Botek‖

Department of Materials Engineering Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan, Department of Chemistry, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan, Photonics Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Ikeda, Osaka 563-8577, Japan, and Laboratoire de Chimie Théorique Appliquée, Facultés Universitaires Notre-Dame de la Paix (FUNDP), rue de Bruxelles, 61, B-5000 Namur, Belgium

Received May 17, 2007

Abstract: Ab initio spin-flip configuration interaction (SF-CI) methods with the finite-field (FF) scheme are applied to the calculation of static second hyperpolarizabilities (γ) of several singlet diradical systems, i.e., the model H_2 molecule under dissociation, *p*-quinodimethane, *o*-quinoid five-membered ring, and 1,4-bis(imidazole-2-ylidene)cyclohexa-2,5-diene (BI2Y) models. The SF-CI method using the UHF reference wave function provides the qualitatively correct diradical character (*y*) dependence of γ in a wide range of a diradical character region for H_2 under dissociation and *p*-quinodimethane as well as *o*-quinoid five-membered ring models. For BI2Y, which is a real diradical system, a non-negligible spin contamination is found in the spin-unrestricted Hartree−Fock (UHF) triplet state, which results in overestimations (SF-CIS) or underestimations (SF-CIS(D)) of γ. Such deficiencies are significantly reduced when using the pure spin state, i.e., the restricted open-shell HF (ROHF) triplet wave function as the reference wave function. These results indicate the applicability of the FF-SF-CI method starting with a pure or a nearly pure high-spin state to provide qualitative or semiquantitative γ for large-size diradical systems. For selected systems, these SF-CI results are also compared to the SF equation of motion coupled cluster singles and doubles (SF-EOM-CCSD) and to SF time-dependent density functional theory (SF-TDDFT) schemes. In particular, large amounts of Hartree−Fock exchange in the functional are required to obtain qualitatively correct dependence of γ on *y* in the case of *p*-quinodimethane.

1. Introduction

There have been several guidelines for designing molecules presenting large first and second hyperpolarizabilities, β and γ, the molecular properties at the origin of the macroscopic second- and third-order nonlinear susceptibilities, $\chi^{(2)}$ and $\chi^{(3)}$.[1−11] For instance, π-conjugated systems with donor−acceptor substitutions[1−3] and charged compounds[4−9] as well as two- or three-dimensional supramolecular architectures[10,11] have been proposed as fundamental units to display high nonlinear optical (NLO) effects. In spite of such persistent pursuit, most candidates examined up to now have been restricted to closed-shell systems. In our previous studies, as a novel class of NLO systems, we have theoretically

* Corresponding author e-mail: rkishi@cheng.es.osaka-u.ac.jp (R.K.), mnaka@cheng.es.osaka-u.ac.jp (M.N.).

† Graduate School of Engineering Science, Osaka University.

‡ Graduate School of Science, Osaka University.

§ National Institute of Advanced Industrial Science and Technology (AIST).

‖ Facultés Universitaires Notre-Dame de la Paix (FUNDP).

10.1021/ct700118q CCC: $37.00
Published on Web 08/23/2007

proposed singlet diradical systems with intermediate diradical character, which exhibit significantly enhanced second hyperpolarizabilities relative to conventional closed-shell molecules.[12−18] Very recently, on the basis of these theoretical and computational studies, two-photon absorption (TPA) measurements have been performed on some of these diphenalenyl compounds, and it has been found that those are among the pure hydrocarbon systems with the largest TPA cross sections.[19]

Theoretical and computational analyses of the NLO properties of open-shell singlet systems in the sum-over-state (SOS) approach require the correct description of their ground- and excited-state electronic structures, which have to be described, in principle, by the multireference (MR) based electron correlation methods, e.g., the MR Møller−Plesset perturbation (MRMP) and MR coupled cluster (MRCC) methods if the high-order excitations involved in the single-reference spin-unrestricted CC (UCC) methods are contracted. Because these MR methods could not be applied to real large-size open-shell molecules due to their demanding computational resources, we need an alternative class of single-reference based methods, e.g., spin-projected methods using the spin-unrestricted solutions including low-order electron correlations such as the spin-projected UMP perturbation (PUMP) method.[20] Indeed, the PUMP methods reproduce well γ values for open-shell π-conjugated linear chain molecules calculated by the UCC method including the single, the double, and, perturbatively, the triple excitations [UCCSD(T)].[21,22] Still, these PUMP approaches are hardly applicable to large systems. On the other hand, the spin-unrestricted hybrid density functional theory (DFT) method employing a hybrid exchange-correlation functional with a large amount of HF exchange, in particular the BHandHLYP XC functional, is efficient to qualitatively describe the variation in static γ of the *p*-quinodimethane model as a function of the diradical character.[13] However, DFT methods based on conventional exchange-correlation functionals lead to catastrophic behavior when computing the linear and nonlinear responses of extended systems and therefore are not suitable to study the size effects on γ.[23] This failure has been related to the shortsightedness of the conventional XC functionals, which are not able to describe the ultranonlocality of the electronic response to electric fields. Nowadays, several solutions have been proposed and mostly correct the wrong behavior of the exchange component.[24] Although they are very promising, extensions are needed to include a balanced and consistent correlation term or to retrieve the computational advantages of DFT over high-order ab initio methods.[25]

In this paper, we address the potential of another single reference scheme, the spin-flip configuration interaction (SF-CI) method developed by Krylov,[26] which has been shown to describe the potential energy surface of the bond dissociation process with high precision in the single reference based theory.[26a] This foresees that the SF approach could be a reliable method to determine the properties including the second hyperpolarizabilities of diradical molecules. Then, compared to the conventional MRCI methods, in view of applications to large diradical systems, this method presents the advantage of significantly reducing the need in computer resources. The reliability of the SF-CI method is examined here in combination with the finite-field (FF) approach. Four types of systems have been selected: the H_2 molecule under dissociation,[12a,17] the *p*-quinodimethane (PQM) model undergoing an aromatic-to-quinoid transformation,[13] *o*-quinoid five-membered ring models, and 1,4-bis(4,5-diphenylimidazol-2-ylidene)cyclohexa-2,5-diene (BI2Y).[14] The basis set dependence of these FF-SF-CI γ is also investigated for PQM and BI2Y. From the comparison of the FF-SF-CI results of γ to several conventional ab initio molecular orbital (MO) electron correlation and the DFT methods, we discuss the reliability and applicability of the FF-SF-CI method to determine the static γ of large-size singlet diradical molecules.

2. Calculation Methods

Within the FF approach, the longitudinal component of the static electronic γ_{iiii} (γ) is calculated using the fourth-order numerical differentiation expression[5b]

$$\gamma_{iiii} = \frac{1}{36(F^i)^4}\{E(F^i) - 12E(F^i) + 39E(F^i) - E(0) + 39E(-F^i) - 12E(-2F^i) + E(-3F^i)\} \quad (1)$$

where $E(F^i)$ represents the total energy of a system in an electric field with an amplitude of F^i (the *i*th component of the field). The numerical stability on the derivatives was checked by using several values of F^i ranging from 0.0010 to 0.0360 au. The convergence on the total energy is fixed to 10^{-10}au. From these field amplitudes and the magnitudes of the total energies, the numerical errors of the FF method are estimated to be about 1% at most, except for the ROHF-SF-CIS(D)/6-31G*+*p* result of BI2Y (∼10%). Similar field amplitudes were already used in previous investigations.[13−18] The advantage of the FF method lies in the simplicity of the scheme, which only requires the evaluation of field-dependent energies, allowing its use with a broad range of methods including electron correlation and therefore the use of many program packages.

Details of the SF technique are presented in ref 26, so that we are restricted here to a brief explanation. In the SF single excitation CI (SF-CIS) method, the initial reference wave function is taken to be the open-shell Hartree−Fock (HF) triplet wave function, $|\cdots \phi_b\alpha\phi_a\alpha\rangle$, where ϕ_b and ϕ_a represent a pair of bonding (b) and antibonding (a) spatial orbitals, respectively. Allowing the spin-flipping operation, $\alpha \rightarrow \beta$, in the construction of single excitation configurations, the lowest spin-unrestricted CIS (UCIS) state corresponding to the ground state contains two configurations, $|\cdots \phi_b\alpha\phi_b\beta\rangle$ and $|\cdots \phi_a\alpha\phi_a\beta\rangle$, the latter being essential for describing the bond dissociation (singlet diradical state). This feature indicates that the SF-CIS method includes the static electron correlation in singlet diradical systems effectively within the single electron excitation scheme. The SF-CIS(D) method, which takes into account the double excitation effects in a perturbative manner, can correct for the lack of dynamical electron correlation. All (field-dependent) SF-CI calculations in this study were performed using the Q-CHEM 3.0 program package.[27]

For comparison, we also apply the approximate spin-projection (AP) scheme based on the ab initio electron correlation methods. The APUHF and APUMP*n* low-spin (LS) energy of a (2*S*+1) spin multiplet is given by[28]

$$^{LS}E_{APUX} = {}^{LS}E_{UX} + \frac{^{LS}\langle S^2\rangle_{UX} - S(S+1)}{^{HS}\langle S^2\rangle_{UX} - {}^{LS}\langle S^2\rangle_{UX}}[{}^{LS}E_{UX} - {}^{HS}E_{UX}] \qquad (2)$$

where HS represents the high-spin solution (a triplet for a diradical system) and X denotes the ab initio MO methods: HF, MP*n* ($n = 2, 4$), etc.

The diradical character y_i related to HOMO $-$ *i* and LUMO + *i* is defined by the weight of the doubly excited configuration in the multiconfigurational (MC)-SCF theory and is formally expressed in the case of the spin-projected UHF (PUHF) theory as[28,29]

$$y_i = 1 - \frac{2T_i}{1 + T_i^2} \qquad (3)$$

where T_i is the orbital overlap between the corresponding orbital pairs (χ_{HOMO-i} and η_{HOMO-i}) and can also be represented by using the occupation numbers (n_i) of UHF natural orbitals (UNOs):

$$T_i = \frac{n_{HOMO-i} - n_{LUMO+i}}{2} \qquad (4)$$

The diradical character y_i represents the instability of the chemical bond since the y_i amount to 0% and 100% for closed-shell and pure diradical states, respectively. The present calculation scheme using the UNOs is the simplest, but it reproduces the diradical characters calculated using highly correlated configuration interaction (CI) methods.[30]

3. Results and Discussion

3.1. H_2 Dissociation Model. For this simplest model, using the 6-31G**+*sp* ($\zeta_{s,p} = 0.0406$ on H atoms) basis set, the variations in γ are examined as a function of the diradical character *y*, while the reference triplet wave function is obtained at the UHF level. As shown in Figure 1(a), *y* remains equal to zero from the equilibrium bond distance ($r = 0.746$ Å at the CISD/6-31G**+*sp* level) to the triplet instability point[31] ($r = 1.2$ Å), where it starts to increase. As expected, it approaches 1.0 in the bond dissociation limit, i.e., the pure diradical state. Figure 1(b) sketches the variation curves of γ with *y* obtained at different levels of approximation. The reliability of the approximate methods is assessed by comparison to the full CI results, i.e., for a two-electron system, the CI singles and doubles (CISD) method. The full CI results show that γ increases with the increase of diradical character *y*, then it attains a maximum in the intermediate *y* region, and finally it decreases in the large *y* region. Due to triplet instability both the spin-restricted HF and low-order electron correlation methods, e.g., RMP*n* ($n = 2-4$), cannot reproduce the qualitative diradical character dependence of γ, whereas both the UHF and UMP*n* ($n = 2-4$) methods provide incorrect behavior of γ for the diradical character unless removing the spin contamination.[13] The APUHF and APUMP*n* ($n = 2,4$) results reproduce semiquantitatively the full CI results except in the small *y* value region ($y < 0.2$),[13] where the AP scheme leads to an incorrect behavior owing to the lack of ionic configuration in the UHF spin-polarized solution. Indeed, as discussed in ref 32, the resonating HF method including ionic configurations in addition to the UHF spin polarized configurations is known to better reproduce the multireference based results in the whole region of *y*. Moreover, the SF-CIS and SF-CIS(D) curves nearly coincide with the APUHF and APUMP2 results, respectively, except in the small *y* region. This good agreement originates in the fact that the PUHF (~APUHF) ground-state wave function eliminating the triplet component from the singlet UHF is expressed by the UNO CASCI(2,2) form, whose configurations correspond to the dominant configurations included in the SF-CIS method. Therefore, the APUMP2 and SF-CIS(D) methods, which include the dynamical correlation effects at the second-order level, provide nearly the same behaviors of γ. The APUMP4 method, which involves the fourth-order correlation effects, further improves the amplitudes of γ and makes them closer to the full CI results.

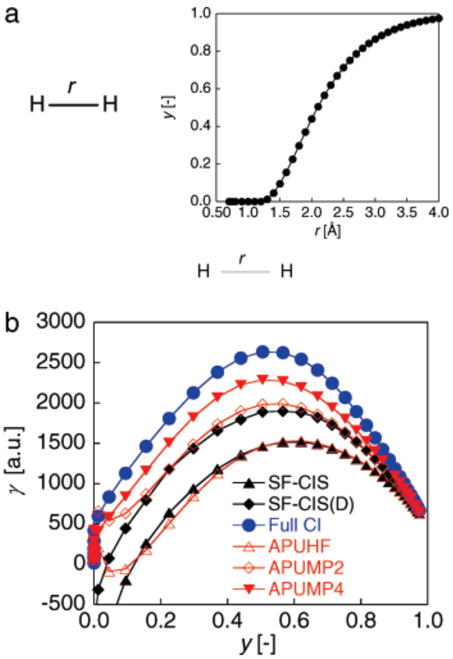


Figure 1. Diradical character (*y*) versus bond distance *r* [Å] for an H_2 dissociation model calculated from the UNOs using the 6-31G**+*sp* basis set (a) and evolution of the longitudinal γ [au] with *y*. The SF-CIS, SF-CIS(D), APUHF, and APUMP*n* ($n = 2, 4$) results are shown as well as the full CI results using the 6-31G**+*sp* basis set.

The *y* value associated with the maximum in γ is slightly smaller when going from the SF-CIS (≈ APUHF) to the full CI: $y_{max} = 0.621$ [SF-CIS(≈ APUHF)], $y_{max} = 0.566$ [SF-CIS(D)(≈ APUMP2)], $y_{max} = 0.505$ (APUMP4 and full CI). On the other hand, the maximum γ value (γ_{max}) increases when going from the SF-CIS (≈ APUHF) to the full CI: $\gamma_{max} = 1514$ (57%) [SF-CIS(≈ APUHF)], $\gamma_{max} = 1889$ (72%) [SF-CIS(D)], $\gamma_{max} = 1989$ (75%) (APUMP2), $\gamma_{max} = 2288$ (87%) (APUMP4), and $\gamma_{max} = 2635$ (100%) (full CI). In the pure diradical regions ($y \approx 1$), all methods give similar γ values since the static correlation is dominant.

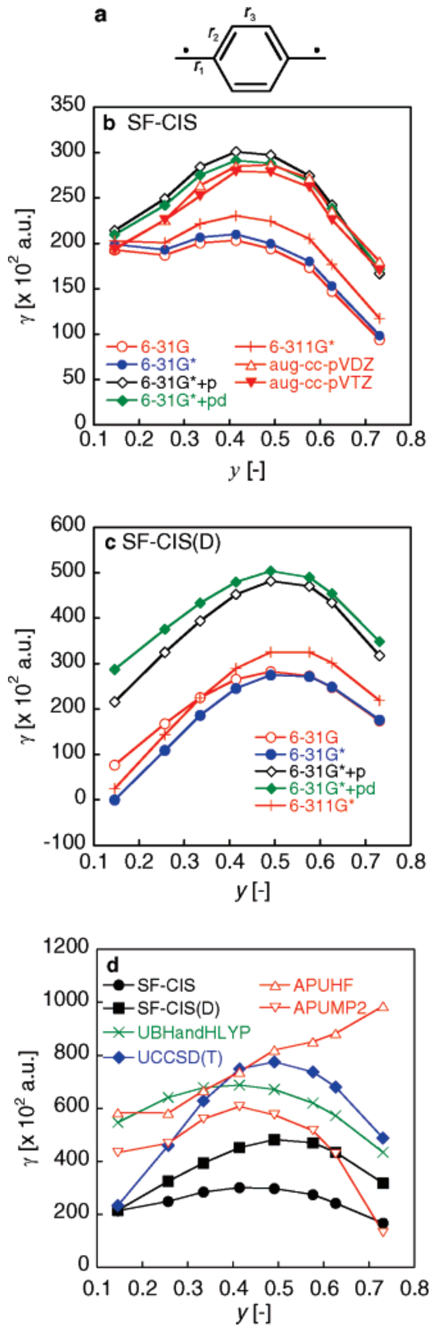


Figure 2. Key geometrical parameters of the *p*-quinodimethane (PQM) model (a) and basis set dependence of the longitudinal γ [$\times$ 10^2 au] at the SF-CIS (b) and SF-CIS(D) (c) levels of approximation as well as the relationship between γ [$\times$ 10^2 au] of PQM and its diradical character as determined using the SF-CIS, SF-CIS(D), UBHandHLYP, UCCSD(T), APUHF, and APUMP2 methods and the 6-31G*+*p* basis sets (d).

In summary, these results suggest that the FF-SF-CIS(D) scheme effectively includes the static and dynamic electron correlation effects on γ like the PUMP*n* schemes.

3.2. *p*-Quinodimethane Model. *p*-Quinodimethane (Figure 2(a)) is one of the simplest π-conjugated singlet diradical systems. In a previous study,[13] the variation in γ was monitored as a function of stretching the r_1 C−C bond length from 1.35 to 1.7 Å while constraining r_2 = r_3 = 1.4 Å. Indeed, varying the geometry in this way leads to variations in the diradical character from 0.146 to 0.731. Since the choice of basis set is important for calculating γ, we first examined the basis set dependence of γ values of PQM at the FF-SF-CI level of approximation. Figure 2(b) shows the SF-CIS γ values calculated using the 6-31G, 6-31G*, 6-31G*+*p*, 6-31G*+*pd* ($\zeta_{p,d}$ = 0.0523 on C atoms), 6-311G* and aug-cc-pVDZ and aug-cc-pVTZ basis sets as a function of *y*. The results using the double-ζ basis sets including diffuse functions, 6-31G*+*p* and 6-31G*+*pd*, are almost the same as those obtained with the aug-cc-pVDZ and aug-cc-pVTZ basis sets, though the latter ones provide slightly smaller γ values than the former ones: when increasing *y*, γ increases, attains a maximum for intermediate diradical character, and then decreases. Similarly, the γ values and their variations with *y* are quasi identical when employing the standard 6-31G, 6-31G*, and 6-311G* basis sets, where the triple-ζ basis set only slightly increases the γ values compared to double-ζ basis sets. On the other hand, the γ values are smaller than those using basis sets containing diffuse functions, except for *y* = 0.146 (equilibrium geometry). The SF-CIS(D) methods (Figure 2(c)) display a more pronounced diradical character dependence of γ than the SF-CIS scheme though the results with and without diffuse functions show similar differences. Judging from these results, the 6-31G*+*p* basis set is adopted for assessing the reliability of the FF-SF-CI schemes in comparison with other methods.

The most reliable and therefore reference results are obtained at the UCCSD(T) level, which significantly removes the spin contamination in the UHF solution. At this level, the behavior of γ with *y* mostly resembles the *y*-dependence of γ in the H_2 model (Figure 1(b)). In ref 13, it was found that all the spin-restricted-based post-HF and DFT methods cannot reproduce the UCCSD(T) results due to triplet instability in the intermediate and large *y* regions, whereas the UHF and UMP*n* (*n* = 2−4) methods provide an incorrect variation in γ, i.e., a monotonous decrease with *y*. On the other hand, the APUMP2 and the UBHandHLYP methods reproduce at least qualitatively the variations in γ except for small values of *y* where the γ values overshoot the reference values (see Figure 2(d)). The SF-CIS method also qualitatively reproduces the correct behavior of γ with *y* though the γ values are significantly underestimated, e.g., $\gamma_{\max}$ = 30 090 au (39%) at *y* = 0.414 (SF-CIS) vs $\gamma_{\max}$ = 77 540 au (100%) at $y_{\max}$ = 0.491 [UCCSD(T)]. The SF-CIS(D) method improves the γ estimate in the intermediate *y* value region, which amounts to 62% (48 150 au at $y_{\max}$ = 0.491). In contrast to the H_2 dissociation model, the SF-CIS(D) and APUMP2 results are different, and the same is true for SF-CIS and APUHF. In the intermediate *y* value region, the APUMP2 γ values are closer to the UCCSD(T) γ values than the SF-CIS(D), the feature of which suggests the more effective inclusion of the dynamical electron correlation by the APUMP2 method. For small values of *y*, the variations of γ are better described by the SF-CIS(D) method than using APUMP2. The APUHF scheme does not reproduce the behavior of γ with the diradical character. Thus, considering PQM, the SF-CI methods are able to describe the diradical

Table 1. Comparison of CPU Times (s) for Single Point Calculations on PQM Using the SF-CIS, SF-CIS(D), and UCCSD(T) Methods

basis set	6-31G	6-31G*	6-31G*+*p*
number of basis functions	88	136	160
CPU times : SF-CIS [s]	11	60	180
CPU times : SF-CIS(D) [s]	111	731	1501
CPU times : UCCSD(T) [s]	5669	37253	85025

character dependence of γ, at least at a comparable level to the APUMP2 and UBHandHLYP methods.

As shown in Table 1, the CPU time required to perform SF-CIS(D) calculations are 1−2 orders of magnitude smaller than for UCCSD(T) calculations, and the computational efforts for SF-CIS calculations are further reduced by about a factor of 10. These CPU times correspond to single point calculations (at equilibrium geometry) performed on Pentium D 3.6 GHz workstation using the Q-CHEM 3.0 program package. Furthermore, the CPU time scaling, estimated by considering larger basis sets, is also smaller for SF-CIS(D) than for UCCSD(T).

The SF concept has also been applied to several other calculation methods, e.g., the SF equation of motion CCSD (SF-EOM-CCSD)[33] and the SF time-dependent DFT (SF-TDDFT),[34] and was then combined here with the FF method to estimate γ values. Figure 3(a) shows the SF-EOM-CCSD γ values as a function of y in comparison with the SF-CIS, SF-CIS(D), UCCSD, and UCCSD(T) results. The SF-EOM-CCSD γ values are in good agreement with those of the UCCSD method, though the maximum in γ is displaced toward large y values, while in the small and intermediate y regions, they tend to undershoot the UCCSD(T) γ values.

Turning to the SF-TDDFT scheme, several exchange-correlation functionals have been considered: B3LYP (20% HF + 8% Slater + 72% Becke88 for exchange, 19% VWN + 81% LYP for correlation), the BHandHLYP functional (50% HF + 50% Becke88 for exchange, 100% LYP for correlation), and a series of modified BHandHLYP functionals including 60, 70, and 80% of HF exchange (60, 70, 80% HF + 40, 30, 20% Becke88 for exchange, 100% LYP for correlation) (see Figure 3(b)). Using the traditional BHandHLYP functional, the diradical character dependence of γ is similar to the UB3LYP results, while the SF-TDDFT-B3LYP data display negative γ values for small diradical character and then a rapid increase with y. Similar shifts of maximum point were reported at the AP-UBHandHLYP level for the H_2 dissociation model, and this was related to overprojection judging from the fact that the spin contaminations at the level of DFT are significantly smaller than those at the level of UHF based approximations. Similarly, one can relate the PQM γ overestimation at the SF-TDDFT-BHandHLYP level to this overprojection as well as to shortsightedness of the DFT exchange functional.[35] Indeed, successive increase of the HF exchange leads to qualitative and quantitative improvement in the description of the variations in γ as a function of y. Nevertheless, when the percentage of HF exchange is large (80%), this approach leads to negative γ values at equilibrium geometry.

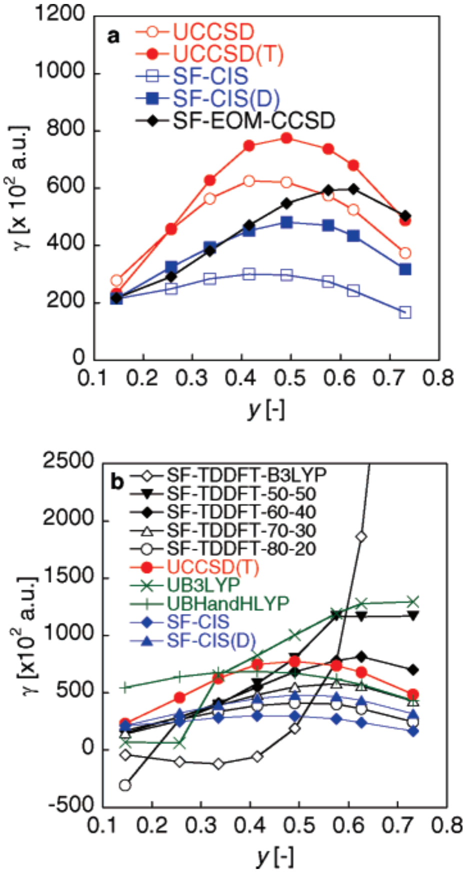


Figure 3. Diradical character dependence of longitudinal γ [× 10^2 au] of PQM for the SF-EOM-CCSD (a) and the SF-TDDFT (b) level of approximations. For SF-TDDFT, we employ the B3LYP (20% HF + 8% Slater + 72% Becke88 for exchange, 19% VWN + 81% LYP for correlation), the BHandHLYP functional (50% HF + 50% Becke88 for exchange, 100% LYP for correlation), and a series of modified BHandHLYP functionals including 60, 70, and 80% of HF exchange (60, 70, 80% HF + 40, 30, 20% Becke88 for exchange, 100% LYP for correlation). These are represented by SF-TDDFT-B3LYP, −50−50, −60−40, −70−30, and −80−20, respectively.

3.3. *o*-Quinoid Five-Membered Ring Models. To further investigate the reliability of the FF-SF-CI calculation, the γ values of two *ortho*-quinoid five-membered ring structures were evaluated. Model compounds A and B are built from a pyrrole or thiophene ring bearing two methylene groups in the β position, respectively (Figure 4(a)). Following UB3LYP/6-31G** geometry optimization, the y values calculated by the UHF/6-31G*+p method amount to 0.487 for model A and 0.694 for model B. Thus, y increases when the ring aromaticity decreases, starting from y = 0.146 for PQM (benzene ring). Despite their larger y values, models A and B present smaller γ values than PQM, at least when considering the direction connecting the radical sites (corresponding to the horizontal axis in Figure 4(a)). This is most probably associated with their smaller extent—and therefore electron delocalization—in this direction.

Figure 4(b) compares the γ values calculated at the SF-CIS, SF-CIS(D), UBHandHLYP, and UCCSD(T) levels of approximation using the 6-31G*+p basis set. The

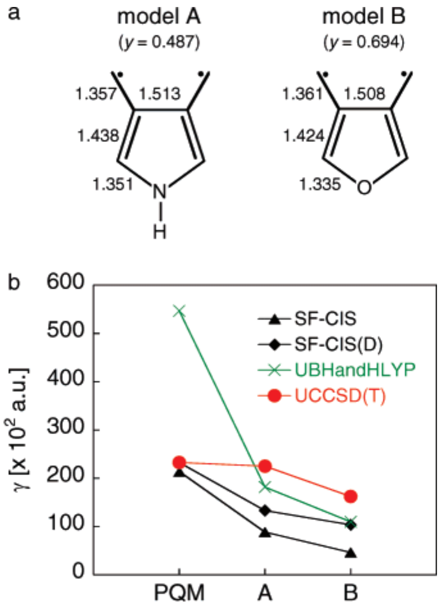


Figure 4. Optimized geometries of *o*-quinoid five-membered ring models involving NH (model A) and O (model B) (a) and calculated γ (along the horizontal axis) [× 10^2 au] of these models in comparison with that of PQM at equilibrium geometry (b).

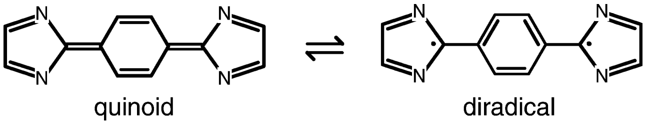


Figure 5. Resonance structures (quinoid and diradical forms) of BI2Y.

Table 2. Longitudinal γ Values [× 10^2 au] of BI2Y[a]

method	6-31G	6-31G*+*p*
UHF	1736	2002
UMP2	9387	9962
UMP3	7747	–
UMP4SDTQ	7484	–
PUHF (*l* = 3)	−862	–
PUMP2 (*l* = 3)	6944	–
PUMP3 (*l* = 3)	4682	–
UCCSD	4474	–
UCCSD(T)	5244	–
UBHandHLYP	4844	6534
UHF-SF-CIS	6278	5709
ROHF-SF-CIS	3570	3370
UHF-SF-CIS(D)	2681	2594
ROHF-SF-CIS(D)	4370	60 × 10^2

[a] γ calculations at the UMP3, UMP4SDTQ, PUHF(l = 3), PUMP2(l = 3), PUMP3(l = 3), UCCSD, and UCCSD(T) levels of approximation could not be performed due to storage needs exceeding our computer capabilities.

UCCSD(T) γ decreases in the following order: PQM (233 × 10^2 au) > model A (225 × 10^2 au) > model B (162 × 10^2 au). Compared to UCCSD(T), the UBHandHLYP method significantly overshoots the difference of γ value for PQM (547 × 10^2 au; 245%), while it gives similar values for the *o*-quinoid models A (181 × 10^2 au; 80%) and B (109 × 10^2 au; 67%). This can be understood by the fact that the latter two models lie in the intermediate y region, where the UCCSD(T) γ is well reproduced by the UBHandHLYP γ, whereas the PQM lies in the small y region, where the UBHandHLYP method overshoots the UCCSD(T) γ (see Figure 2(d)). The SF-CIS method underestimates the UCCSD(T) values, model A (89 × 10^2 au; 40%) and model B (46 × 10^2 au; 28%), while adding dynamical electron correlation effects at the SF-CIS(D) level improves the agreement with the reference data: model A (133 × 10^2 au; 59%) and B (103 × 10^2 au; 64%). In summary, the SF-CIS and the SF-CIS(D) methods can describe at least qualitatively the variations of γ due to changing the nature of the aromatic ring of diradical species.

3.4. 1,4-Bis(imidazole-2-ylidene)cyclohexa-2,5-diene. In order to assess the applicability of the SF-CI method to systems of practical interest, we examined the γ value of BI2Y (Figure 5), a thermally stable π-conjugated diradical molecule.[36] We use the geometry of ref 14, which was optimized under the constraint of D_{2h} symmetry by the B3LYP/6-31G** method. This molecule displays an intermediate diradical character [y = 0.4227 as calculated by the UHF/6-31G*+*p* (ζ = 0.0523 for C atoms and 0.0582 for N atoms)] in the singlet ground state. Table 2 further addresses the basis set dependence of γ at the UHF, UMP2, UBHandHLYP, SF-CIS, and SF-CIS(D) levels of approximation. Because the $<S^2>$ values of the UHF triplet states of BI2Y are relatively large, 2.36 (6-31G) and 2.28 (6-31G*+*p*), SF-CI calculations based on the UHF triplet reference state are expected to give a spin contaminated ground state. Indeed, the $<S^2>$ values of the SF-CIS ground state are 0.80 (6-31G) and 0.69 (6-31G*+*p*), which lead to the significant over- or underestimation of γ values. Therefore, we also consider another type of reference state having no spin contamination, i.e., the restricted open-shell HF (ROHF) triplet state. We perform the SF-CI calculations starting with the restricted open-shell HF (ROHF) triplet wave function as the reference state (ROHF-SF-CI) in addition to those using the UHF triplet one (UHF-SF-CI) in order to examine the spin contamination effects on SF calculations.

It is found that the 6-31G results are smaller than the 6-31G*+*p* results: 87% (UHF), 94% (UMP2), and 74% (UBHandHLYP). For the SF-CI approach, the 6-31G results are slightly larger than the 6-31G*+*p* results: 110% (UHF-SF-CIS), 103% (UHF-SF-CIS(D)), and 106% (ROHF-SF-CIS), except for the 73% of the ROHF-SF-CIS(D) results. As a result, the 6-31G basis set was further used for semiquantitatively examining the dependence of γ on the applied calculation methods. Similarly to the PQM model, the UMP*n* (n = 2−4) methods tend to overshoot the UCCSD(T) γ value, while the UBHandHLYP result (γ = 4844 au; 92%) is very close (see Table 2). The PUMP2 (l = 3) (γ = 6944 au; 132%) and PUMP3 (l = 3) (γ = 4682 au; 89%), which use the l-fold spin-projection scheme by Löwdin,[20] well reproduce the UCCSD(T) γ value. The UHF-SF-CIS method (γ = 6278 au; 120%) seems to be a good approximation to the UCCSD(T) γ values, whereas the UHF-SF-CIS(D) (γ = 2681 au) undershoots these by a factor of 2. This deficiency of the SF-CIS(D) results is due to the spin contamination described above. Indeed, it is significantly improved by using the ROHF triplet reference as the initial wave function. Using the 6-31G basis set, the ROHF-SF-CIS and -CIS(D) approaches reproduce about 68% and 84% of the UCCSD(T) γ value. These ROHF-SF-CIS/UCCSD(T) and ROHF-SF-CIS(D)/UCCSD(T) ratios are even better than for the H_2 model system (53 and 70% for $y = 0.44$) and PQM (40 and 60% for $y = 0.41$) models. The ROHF-SF-CIS(D) results are also very similar to the UBHandHLYP results for both basis sets.

4. Concluding Remarks

The SF-CI approach has been employed to evaluate the static second hyperpolarizability of diradical systems covering a wide range of diradical character, and these results have been compared to those of reference ab initio methods as well as of hybrid DFT schemes. For the H_2 molecule under dissociation, the SF-CIS and SF-CIS(D) methods provide similar γ values to the APUHF and APUMP2 methods, respectively, while they reproduce qualitatively or semiquantitatively the variation in γ with the diradical character obtained by reference high-level electron correlation methods. These conclusions further extend to γ of the model of the *p*-quinodimethane molecule of which the diradical character is externally modified by stretching as well as to the trend in γ among related quinoid compounds, *p*-quinodimethane and *o*-quinoid five-membered ring models. In the case of *p*-quinodimethane, these SF calculations have been extended to the EOM-CC and TDDFT levels of approximation. Though the SF-EOM-CCSD approach reproduces qualitatively the *y*-dependence of γ obtained at the UCCSD level of approximation, the SF-TDDFT approach only provides qualitatively correct results when the exchange-correlation functional contains a large (>70%) percentage of Hartree−Fock exchange. In the case of *p*-quinodimethane, a basis set investigation has shown the need for including a set of *p* diffuse functions, particularly at the SF-CIS(D) level.

Treating the BI2Y molecule, a larger system of practical interest, has evidenced other phenomena. Indeed, the *usual* UHF triplet reference presents a non-negligible spin contamination, which transfers to the SF-CI ground state as shown by $\langle S^2 \rangle$ values larger than 0.6 at the SF-CIS level of approximation. This leads to an underestimation of the SF-CIS(D) γ value, whereas the SF-CIS result is in better agreement. On the other hand, by employing the spin contamination-free ROHF method to define the reference wave function, the SF-CIS(D) γ value gets in good agreement with the reference UCCSD(T) result. Since the ROHF-based SF-CIS and SF-CIS(D) approaches also perform well for the H_2 and PQM model, they turn out to be promising calculation schemes, especially when dealing with large compounds presenting diradical character, such as diphenalenyl compounds. This reliability has also to be contrasted with the difficulties of conventional DFT methods, i.e., without long-range corrections, to address size-dependence effects on γ. Future investigations will concern the determination of both geometries and hyperpolarizabilities of diradical species since the γ values of conjugated systems are sensitive to the geometry and particularly to the bond length alternation along the conjugated backbone. Indeed, very recently, the importance of spin-projection was highlighted for geometry optimization at the UDFT level of approximation,[37] whereas the SF-TDDFT approach was also shown to reproduce highly correlated MO methods in the case of singlet diradical systems. Further investigations will also tackle the efficiency of the SF-TDDFT approach as a function of the nature of the exchange-correlation functional.

Acknowledgment. The authors thank an anonymous referee for pertinent suggestions. This work is supported by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) fellows (No. 1809170), for Scientific Research (No. 18350007) from JSPS, and Grant-in-Aid for Scientific Research on Priority Areas (No. 18066010) from the Ministry of Education, Science, Sports and Culture of Japan. E.B. thanks the IUAP program No. P6-27 for her postdoctoral grant. B.C. thanks the Belgian National Fund for Scientific Research for his Research Director position.

References

(1) (a) Prasad, P. N.; Williams, D. J. *Introduction to Nonlinear Optical Effects in Molecules and Polymers*; Wiley: New York, 1990. (b) Bosshard, Ch.; Sutter, K.; Prêtre, Ph.; Hulliger, J.; Flörsheimer, M.; Kaatz, P.; Günter, P. Organic Nonlinear Optical Materials. In *Advances in Nonlinear Optics*; Garito, A. F., Kajzar, F., Eds.; Gordon & Breach Science Pub.: Basel, Switzerland, 1995. (c) Champagne, B.; Kirtman, B. In *Handbook of Advanced Electronic and Photonic Materials and Devices*; Nalwa, H. S., Ed.; Academic Press: New York, 2001; Vol. 9, Chapter 2, p 63. (d) *Non-Linear Optical Properties of Matter − From Molecules to Condensed Phases*; Papadopoulos, M. G., Sadlej, A. J., Leszczynski, J., Eds.; Springer: Dordrecht, The Netherland, 2006. (e) Datta, A.; Pati, S. K. *Chem. Soc. Rev.* **2006**, *35*, 1305.

(2) (a) Kanis, D. R.; Ratner, M. A.; Marks, T. J. *Chem. Rev.* **1994**, *94*, 195. (b) Morley, J. O. *J. Phys. Chem.* **1995**, *99*, 10166. (c) Verbiest, T.; Houbrechts, M.; Kauranen, M.; Clays, K.; Persoons, A. *J. Mater. Chem.* **1997**, *7*, 2175. (d) Coe, B. *Chem. Eur. J.* **1999**, *5*, 2464.

(3) (a) Brédas, J. L.; Adant, C.; Tackx, P.; Persoons, A.; Pierce, B. M. *Chem. Rev.* **1994**, *94*, 243. (b) Kirtman, B.; Champagne, B. *Int. Rev. Phys. Chem.* **1997**, *16*, 389. (c) Brunel, J.; Mongin, O.; Jutand, A.; Ledoux, I.; Zyss, J.; Blanchard-Desce, M. *Chem. Mater.* **2003**, *15*, 4139. (d) Garcia, M. H.; Royer, S.; Robalo, M. P.; Dias, A. R.; Tranchier, J. P.; Chavignon, R.; Prim, D.; Auffrant, A.; Rose-Munch, F.; Rose, E.; Vaissermann, J.; Persoons, A.; Asselberghs, I. *Eur. J. Inorg. Chem.* **2003**, *21*, 3895. (e) Sliwa, M.; Létard, S.; Malfant, I.; Nierlich, M.; Lacroix, P. G.; Asahi, T.; Masuhara, H.; Yu, P.; Nakatani, K. *Chem. Mater.* **2005**, *17*, 4727. (f) Ferrighi, L.; Frediani, L.; Cappelli, C.; Salek, P.; Ågren, H.; Helgaker, T.; Ruud, K. *Chem. Phys. Lett.* **2006**, *425*, 267. (g) Sanguinet, L.; Pozzo, J. L.; Guillaume, M.; Champagne, B.; Castet, F.; Ducasse, L.; Maury, O.; Soulié, J.; Mançois, F.; Adamietz, F.; Rodriguez, V. *J. Phys. Chem. B* **2006**, *110*, 10672. (h) Gradinaru, J.; Forni, A.; Druta, V.; Tessore, F.; Zecchin, S.; Quici, S.; Garbalau, N. *Inorg. Chem.* **2007**, *46*, 884.

(4) (a) de Melo, C. P.; Silbey, R. *J. Chem. Phys.* **1988**, *88*, 2567. (b) Villesuzanne, A.; Hoarau, J.; Ducasse, L.; Olmedo, L.; Hourquebie, P. *J. Chem. Phys.* **1992**, *96*, 495.

(5) (a) Nakano, M.; Yamaguchi, K. *Chem. Phys. Lett.* **1993**, *206*, 285. (b) Nakano, M.; Shigemoto, I.; Yamada, S.; Yamaguchi, K. *J. Chem. Phys.* **1995**, *103*, 4175. (c) Fujita, H.; Nakano, M.; Takahata, M.; Yamaguchi, K. *Chem. Phys. Lett.* **2002**, *358*, 435.

(6) (a) Kamada, K.; Ohta, K.; Iwase, Y.; Kondo, K. *Chem. Phys. Lett.* **2003**, *372*, 386. (b) Kishi, R.; Nakano, M.; Yamada, S.; Kamada, K.; Ohta, K.; Nitta, T.; Yamaguchi, K. *Chem. Phys. Lett.* **2004**, *393*, 437.

(7) (a) Champagne, B.; Spassova, M.; Jadin, J. B.; Kirtman, B. *J. Chem. Phys.* **2002**, *116*, 3935. (b) Oliveira, L. N.; Amaral, O. A. V.; Castro, M. A.; Fonseca, T. L. *Chem. Phys.* **2003**, *289*, 221.

(8) Botek, E.; Spassova, M.; Champagne, B.; Asselberghs, I.; Persoons, A.; Clays, K. *Chem. Phys. Lett.* **2005**, *412*, 274, (E) *ibid.* **2006**, *417*, 282.

(9) Chen, W.; Li, Z. R.; Wu, D.; Li, Li, Y.; Sun, C. C.; Gu, F. L. *J. Am. Chem. Soc.* **2005**, *127*, 10977.

(10) (a) Moylan, C. R.; Ermer, S.; Lovejoy, S. M.; McComb, I.-H.; Leung, D. S.; Wortmann, R.; Krdmer, P.; Twieg, R. J. *J. Am. Chem. Soc*. **1996**, *118*, 12950. (b) Chang, S. J.; Kim, K. S.; Lin, T. C.; He, G. S.; Swiatkiewicz, J.; Prasad, P. N. *J. Phys. Chem. B* **1999**, *103*, 10741. (c) Le Bozec, H.; Renouard, T. *Eur. J. Inorg. Chem.* **2000**, 229. (c) Nakano, M.; Fujita, H.; Takahata, M.; Yamaguchi, K. *J. Am. Chem. Soc.* **2002**, *124*, 9648. (d) Yang, M.; Champagne, B. *J. Phys. Chem. A* **2003**, *107*, 3942. (e) Botek, E.; Castet, F.; Champagne, B. *Chem. Eur. J.* **2006**, *12*, 8687.

(11) Yokoyama, S.; Nakahama, T.; Otomo, A.; Mashiko, S. *J. Am. Chem. Soc.* **2000**, *122*, 3174.

(12) (a) Nakano, M.; Nagao, H.; Yamaguchi, K. *Phys. Rev. A* **1997**, *55*, 1503. (b) Nakano, M.; Kishi, R.; Ohta, S.; Takahashi, H.; Kubo, T.; Kamada, K.; Ohta, K.; Botek, E.; Champagne, B. *Phys. Rev. Lett.* **2007**, *99*, 033001.

(13) Nakano, M.; Kishi, R.; Nitta, T.; Kubo, T.; Nakasuji, K.; Kamada, K.; Ohta, K.; Champagne, B.; Botek, E.; Yamaguchi, K. *J. Phys. Chem. A* **2005**, *109*, 885.

(14) Nakano, M.; Kishi, R.; Nakagawa, N.; Ohta, S.; Takahashi, H.; Furukawa, S.; Kamada, K.; Ohta, K.; Champagne, B.; Botek, E.; Yamada, S.; Yamaguchi, K. *J. Phys. Chem. A* **2006**, *110*, 4238.

(15) Nakano, M.; Kubo, T.; Kamada, K.; Ohta, K.; Kishi, R.; Ohta, S.; Nakagawa, N.; Takahashi, H.; Furukawa, S.; Morita, Y.; Nakasuji, K.; Yamaguchi, K. *Chem. Phys. Lett.* **2006**, *418*, 142.

(16) Ohta, S.; Nakano, M.; Kubo, T.; Kamada, K.; Ohta, K.; Kishi, R.; Nakagawa, N.; Champagne, B.; Botek, E.; Umezaki, S.; Takebe, A.; Takahashi, H.; Furukawa, S.; Morita, Y.; Nakasuji, K.; Yamaguchi, K. *Chem. Phys. Lett.* **2006**, *420*, 432.

(17) Nakano, M.; Kishi, R.; Ohta, S.; Takebe, A.; Takahashi, H.; Furukawa, S.; Kubo, T.; Morita, Y.; Nakasuji, K.; Yamaguchi, K.; Kamada, K.; Ohta, K.; Champagne, B.; Botek, E. *J. Chem. Phys.* **2006**, *125*, 074113.

(18) Nakano, M.; Takebe, A.; Kishi, R.; Ohta, S.; Nate, M.; Kubo, T.; Kamada, K.; Ohta, K.; Champagne, B.; Botek, E.; Takahashi, H.; Furukawa, S.; Morita, Y.; Nakasuji, K. *Chem. Phys. Lett.* **2006**, *432*, 473.

(19) Kamada, K.; Ohta, K.; Kubo, T.; Shimizu, A.; Morita, Y.; Nakasuji, K.; Kishi, R.; Ohta, S.; Furukawa, S.; Takahashi, H.; Nakano, M. *Angew. Chem., Int. Ed.* **2007**, *46*, 3544.

(20) Löwdin, P. O. *Phys. Rev.* **1955**, *97*, 1509.

(21) Champagne, B.; Botek, E.; Quinet, O.; Nakano, M.; Kishi, R.; Nitta, T.; Yamaguchi, K. *Chem. Phys. Lett.* **2005**, *407*, 372.

(22) Champagne, B.; Botek, E.; Nakano, M.; Nitta, T.; Yamaguchi, K. *J. Chem. Phys.* **2005**, *122*, 114315.

(23) (a) Champagne, B.; Perpète, E. A.; van Gisbergen, S. J. A.; Baerends, E.; Snijders, J. G.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Chem. Phys.* **1998**, *109*, 10489; erratum **1999**, *110*, 11664. (b) van Gisbergen, S. J. A.; Schipper, P. R. T.; Gritsenko, O. V.; Baerends, E. J.; Snijders, J. G.; Champagne, B.; Kirtman, B. *Phys. Rev. Lett.* **1999**, *83*, 694.

(24) (a) Faassen, M. V.; de Boeij, P. L.; Leeuwen, R. v.; Berger, J. A.; Snijders, J. G. *J. Chem. Phys.* **2003**, *118*, 1044. (b) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425. (c) Kamiya, M.; Sekino, H.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *122*, 234111. (d) Bulat, F. A.; Toro-Labbé, A.; Champagne, B.; Kirtman, B.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 014319. (e) Peach, M. J. G.; Helgaker, T.; Salek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558. (f) Champagne, B.; Bulat, F. A.; Yang, W.; Bonness, S.; Kirtman, B. *J. Chem. Phys.* **2006**, *125*, 194114.

(25) (a) Bartlett, R. J.; Grabowski, I.; Hirata, S.; Ivanov, S. *J. Chem. Phys.* **2005**, *122*, 034104. (b) Staroverov, V. N.; Scuseria, G. E.; Davidson, E. R. *J. Chem. Phys.* **2006**, *125*, 081104.

(26) (a) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *338*, 375. (b) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *350*, 522. (c) Krylov, A. I. *Acc. Chem. Res*. **2006**, *39*, 83.

(27) Shao, Y.; Fusti-Molnar, L.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; Distasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F., III ; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(28) Yamaguchi, K. In *Self-Consistent Field Theory and Applications*; Carbo, R., Klobukowski, M., Eds.; Elsevier: Amsterdam, The Netherlands, 1990; p 727.

(29) Yamanaka, S.; Okumura, M.; Nakano, M.; Yamaguchi, K. *J. Mol. Struct. (THEOCHEM*) **1994**, *310*, 205.

(30) Herebian, D.; Wieghardt, K. E.; Neese, F. *J. Am. Chem. Soc.* **2003**, *125*, 10997.

(31) Okumura, M.; Yamanaka, S.; Mori, W.; Yamaguchi, K. *J. Mol. Struct. (THEOCHEM*) **1994**, *310*, 177.

(32) (a) Fukutome, H. *Prog. Theor. Phys.* **1988**, *80*, 417. (b) Takeda, R.; Yamanaka, S.; Yamaguchi, K. *Int. J. Quantum Chem.* **2006**, *106*, 3303.

(33) Levchenko, S. V.; Krylov, A. I. *J. Chem. Phys.* **2004**, *120*, 175.

(34) (a) Shao, Y.; Head-Gordon, M.; Krylov, A. I. *J. Chem. Phys.* **2003**, *118*, 4807.

(35) (a) Bulat, F.; Toro-Labbé, A.; Champagne, B.; Kirtman, B.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 014319. (b) Champagne, B.; Bulat, F. A.; Yang, W.; Bonness, S.; Kirtman, B. *J. Chem. Phys.* **2006**, *125*, 194114.

(36) (a) Kikuchi, A.; Iwahori, F.; Abe, J. *J. Am. Chem. Soc.* **2004**, *126*, 6526. (b) Kikuchi, A.; Iwahori, F.; Abe, J. *J. Phys. Chem. B* **2005**, *109*, 19448.

(37) Kitagawa, Y.; Saito, T.; Masahide, I.; Shoji, M.; Koizumi, K.; Yamanaka, S.; Kawakami, T.; Okumura, M.; Yamaguchi, K. *Chem. Phys. Lett.* **2007**, *442*, 445.

CT700118Q

JCTC Journal of Chemical Theory and Computation

Nitrogen Fixation by a Molybdenum Catalyst Mimicking the Function of the Nitrogenase Enzyme: A Critical Evaluation of DFT and Solvent Effects

Alessandra Magistrato,*,† Arturo Robertazzi,† and Paolo Carloni

CNR-INFM-Democritos Modelling Center for Atomistic Simulations, International School for Advanced Studies (SISSA/ISAS) and Italian Institute of Technology (IIT) via Beirut 2-4, Trieste, Italy

Received April 23, 2007

Abstract: Compounds mimicking the enzyme nitrogenase represent promising alternative routes to the current Haber-Bosch industrial synthesis of ammonia from molecular hydrogen and nitrogen. In this work, we investigated the full catalytic cycle of one of such compounds, Mo-(HIPTN3N) (with HIPT = hexaisopropylterphenyl), by means of DFT calculations. Our results suggest these large ligands to exert mainly a steric influence on the structural properties of the catalyst. In addition, we provided a structural and electronic characterization of the putative reaction intermediates along with a picture of the electronic mechanism of molecular nitrogen N−N bond breaking. A large discrepancy was observed between calculated and experimental reaction free energies, suggesting that in the present case the predictability of DFT reaction energies is limited. Investigation of explicit solvation of specific catalytic intermediates as well as of the protonation and reducing agents reveal the crucial role played by the solvent molecules (benzene and heptane) particularly for protonation steps. Furthermore, the analysis of several DFT functionals indicates that these have to be carefully chosen in order to reproduce the energetic profile of reduction steps. This study shows how DFT calculations may be a powerful tool to describe structural and electronic properties of the intermediates of the catalytic cycle, yet, due to the complexity of the system, reaction energies cannot be easily reproduced without a careful choice of the solvation model and the exchange-correlation functional.

Introduction

Ammonia is the sixth largest chemical produced in the world.[1−4] Since the industrial synthesis from molecular nitrogen and hydrogen requires a drastic condition (500 °C at 150−200 atm),[1−4] a considerable effort is being devoted in discovering alternative routes under milder conditions of temperature and pressure.[1−4] A promising strategy is based on biomimetics (i.e., compounds that perform the same chemical reaction of natural enzymes at the same mild conditions and with no energy loss) of Fe, Mo-containing nitrogenase enzyme expressed by soil bacteria.[5−7] A major step in this direction[8−10] is the synthesis of complex **I**, in which a Mo ion is bound to a chelating triamidoamine ligand ($(RNCH_2CH_2)_3N^{3-}$) with R = hexaisopropylterphenyl or HIPT) with four nitrogen donor atoms[8−10] (Figures 1A and 2). Remarkably, the slow addition of a proton source (such as 2,6-lutidiniumBAr$_4$ where Ar is 3,5-$(CF_3)_2C_6H_3$) and a reducing agent (such as decamethylcromocene) to a solution of **I** in heptane in the presence of molecular nitrogen allows the production of ammonia with high efficiency (63−66%).[8] The proposed reaction mechanism, based on catalytic intermediates experimentally characterized,[8−11] is a Chatt-like mechanism,[8−10] in which molecular nitrogen binds the metal ion in a linear end-on-fashion (Figure 2).[11−13] However, an exhaustive understanding of this molecular processes is still lacking.[2,14]

* Corresponding author e-mail: alema@sissa.it.

† These authors equally contributed to this work.

10.1021/ct700094y CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/04/2007

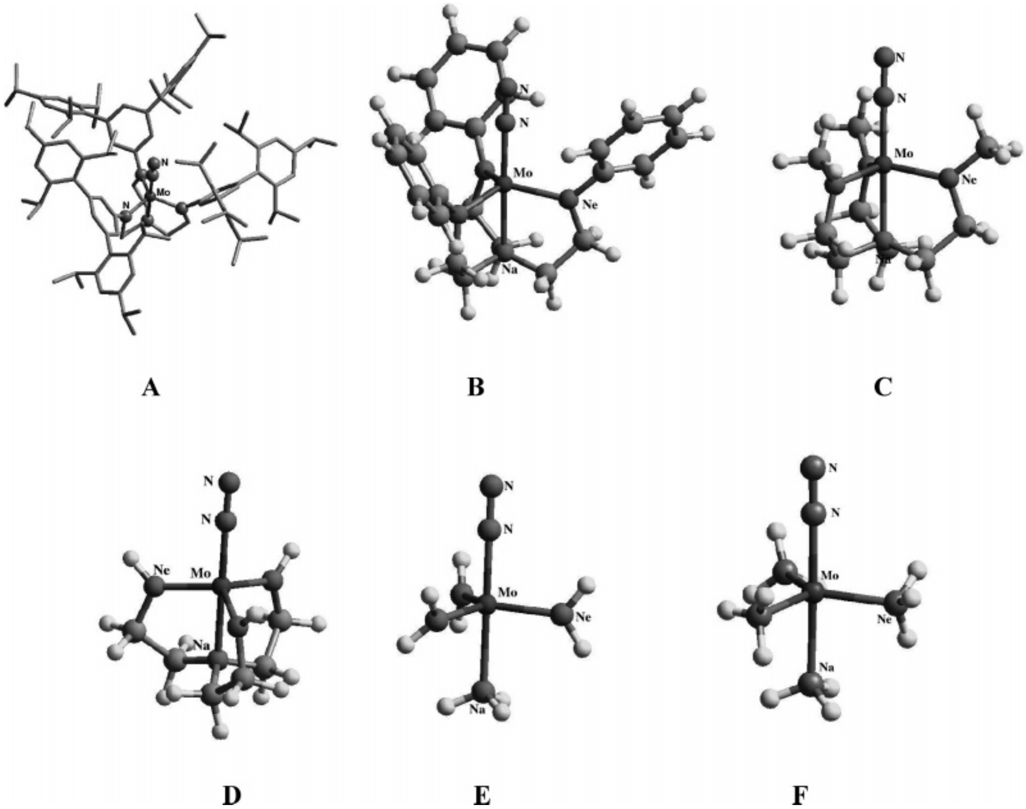


Figure 1. Computational models of **I**. **A** is the entire catalyst (hydrogen atoms are not shown for clarity). **B**−**F** are computational models of different complexity of **I**.

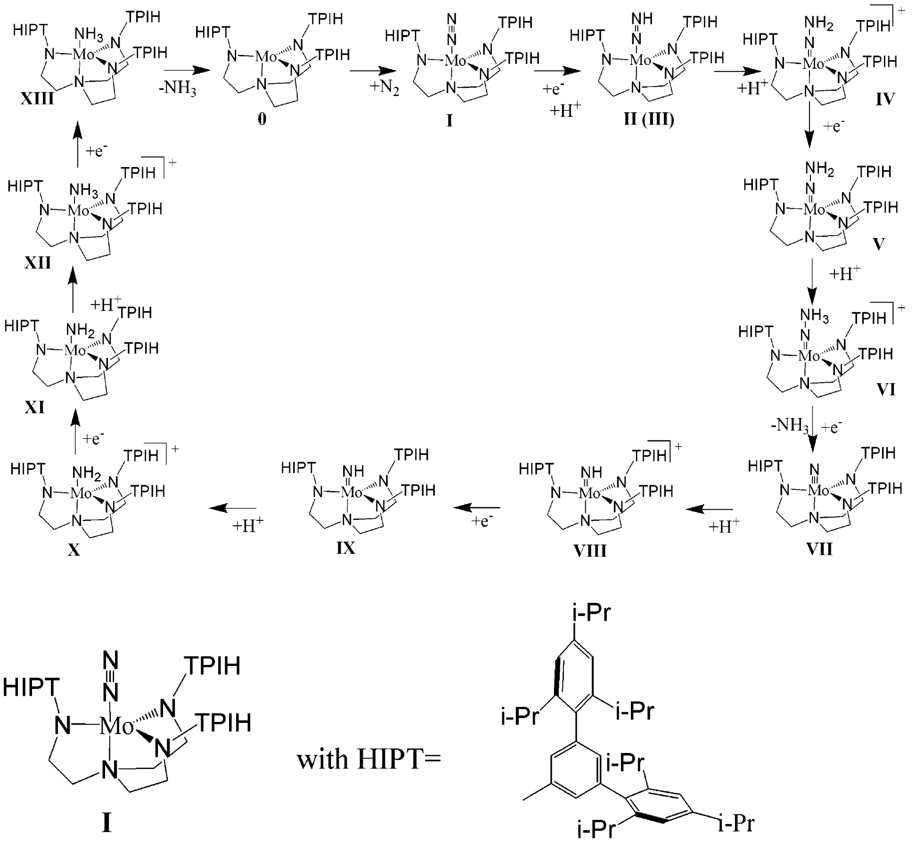


Figure 2. Schematic view of the mechanism for nitrogen fixation at a single Mo center.

In an effort to further improve the efficiency of the catalyst, the transition-metal ion has been changed,[15,16] and the R group has been modified (R = 3,5-(2,4,6-t-$Bu_3C_6H_2)_2C_6H_3$, 3,5-(2,4,6-$Me_3C_6H_2)_2C_6H_3$, 4-Br-3,5-(2,4,6-i-$Pr_3C_6H_2)_2C_6H_3$

(p-Br(HIPT)).[17] Unfortunately, none of the modified complexes has turned out to show a stronger catalytic power than that of the original compound.

Recent density functional theory (DFT) calculations with the BLYP[18,19] and B3LYP[19,20] exchange-correlation functionals, and using the PCM implicit solvation model[21] (which is widely applied to study processes involving inorganic catalysts),[22] were used to investigate the reaction mechanism.[23−26] Theoretical predictions of the catalytic intermediates and reaction energies show that geometries of molybdenum complexes are well reproduced,[25−29] while reaction energies estimated both in vacuo and with the PCM implicit solvent model turned out to be remarkably differ from the experimental free energies.[10,23,25,26] The discrepancy may be caused by several factors, including (i) the choice of the exchange-correlation functional, (ii) the absence of explicit solvation in the calculations (as solvation has been treated previously with continuum solvation models), and (iii) the lack of rigorous treatment of entropic effects.

In order to address some of these issues, we performed extensive DFT calculations on the reaction mechanism proposed in Figure 2. These include (i) static DFT calculations and (ii) dynamic (Car−Parrinello, CP)[30] DFT calculations, with PW basis sets and BP[18,31] exchange-correlation functional. Our choice for this computational method was dictated by its reliability in describing reactions catalyzed by inorganic compounds.[32,33] Subsequently, selected intermediates obtained with the latter approach were compared with DFT calculations using localized basis sets (6-31+G-(d) and 6-311++G(d,p)) and B3LYP,[19,20] B3P,[20,31] PW91,[34] BHandHLYP,[19,20] and BHandH[20] exchange-correlation functionals. Calculations were carried out both in vacuo and in the presence of either an implicit solvation model (such as PCM)[18] or with a qualitative model of explicit solvation. This was obtained by including *n* solvent molecules (i.e., methane, heptane, benzene, and fluorobenzene (PhF),[35]with $1 < n < 4$).

Our calculations confirm that the nitrogen fixation promoted by **I** may proceed through a Chatt-like mechanism and that a back-donation from a filled d_{xz} Mo orbital to the empty π^*-N_2 orbital may be at the basis of the electronic mechanism responsible for the N_2 cleavage. However, in line with previous DFT studies, our calculations reveal large discrepancies between experimental free energies and calculated reaction energies.[10,23,25] Such evident failures may be prevented by explicitly accounting for solvent effects (even for apolar solvents) for the protonation steps, whereas a careful choice of the DFT exchange-correlation functional is required for reductions steps.[36] Thus, this work reveals DFT potentialities and limitations, suggesting that the predictive abilities of DFT techniques may be limited for this intricate catalytic system and that solvent effects (typically neglected) may play a significant role.

Computational Details

Several models of the catalyst of different complexity were taken into account, i.e., from the entire catalyst (**A**) to the molybdenum complexes containing triphenylamidoamine, trimethylamidoamine, triamidoamine, amide, and ammonia ligands (**B**−**F**, respectively in Figure 1).

DFT calculations were initially performed with the Amsterdam Density Functional program ADF2000.01.[37,38]A spin unrestricted formalism was applied to open shell systems, and structures of each catalytic intermediate were optimized for all possible spin states. Calculations were performed with the gradient-corrected model developed by Becke (B),[18] combined with the Perdew's correlation term (P).[31] The electronic configurations of molecular systems were described by a triple-STO basis set on the Mo center for the ns, np, nd, (n+1)s, and (n+1)p valence shells; a double-STO basis set was used for C (2s 2p), N (2s, 2p), and H(1s). Inner shells of the atoms were treated within the frozen core approximation. ZORA[39] relativistic corrections were added to the total energy of Mo.

Static and dynamic density functional calculations were performed with Car−Parrinello molecular dynamics simulations.[30] These simulations combine a classical molecular dynamics scheme with an electronic structure calculation in the framework of density functional theory (DFT), a pseudopotential formalism, periodic boundary conditions, and a basis set of plane waves (PW). Calculations reported in this work were performed with the program CPMD.[40] In this study, we employed analytical pseudopotentials for Mo and Cr of the Goedecker type[41] and nonlocal, norm conserving pseudopotentials of the Martins−Trouiller type[42] for all the other elements except for H. For these an analytical local pseudopotential was used. Moreover, 14 valence electrons and a reference configuration of $4s^24p^64d^55s^1$ and $3s^23p^6$-$3d^54s^1$ were considered for Mo and Cr, respectively. Pseudopotentials for N and C were transformed to a fully nonlocal form by adopting the scheme developed by Kleinman and Bylander.[43] A kinetic energy cutoff of 70 Ry was applied to all the calculations. In order to properly describe charged systems, periodic images were decoupled using the scheme of Hockney.[44] A face centered cubic (FCC) super cell of the edge of 25 Å was employed for **A**. In addition, a simple cubic (SC) cell of edge $a = 16$ Å was used for all other calculations. Classical equations of motion were integrated with a velocity Verlet algorithm with a time step of 0.145 fs and a fictitious mass for the electronic degrees of freedom of $\mu = 800$ au. Geometry optimization runs were also carried out within the CPMD code.[40] These calculations were performed with a preconditioned conjugate gradient procedure.[40] All the calculations were performed within a spin unrestricted formalism.

Calculations on selected reaction intermediates (**III**, **IV**, **V**, **VII**, **VIII**, and **IX**, Figure 2), the proton source, and the reductant were also carried out using the Gaussian03 suite of programs[45] in vacuo and in the presence of implicit and explicit solvent.

6-31+G(d) and 6-311++G(d,p) basis sets were employed with the B3LYP,[19,20] B3P,[18,31] PW91,[34] BHandHLYP,[19,20] and BHandH[20] exchange-correlation functionals. In order to evaluate the effect of implicit solvent molecules, the polarized continuum model, PCM, was employed.[21,22] In addition, calculations were carried out in the presence of *n* solvent molecules (i.e., methane, heptane, benzene, and fluorobenzene (PhF),[35] with $1 < n < 4$). We are aware that such a static first solvation shell is only a rough representation of solvation, which is a very complex dynamical phenomenon. Due to the size of the solute/solvent adducts, geometries were fully optimized with a reduced threshold, which sets the optimization convergence criteria to a maximum step size of 0.01 au and an rms force of 0.0017 au.

Energetics for Each Reaction Step. Reaction energies were calculated for each step of the catalytic cycle as reported below. We would like to remark that no entropic effects have been considered.

(i) Reaction energy of protonation ($\Delta E_{R,prot}$): $\Delta E_{R,prot} = E_{product} + E_{Lut} - E_{reactant} - E_{LutH+}$, where $E_{product}$ and $E_{reactant}$ are the energies of the protonated and neutral catalyst, respectively; E_{Lut} and E_{LutH+} are the energies of the protonation agent (lutidinium, Lut), in the unprotonated and protonated state, respectively.[8−10]

(ii) Reaction energy of reduction ($\Delta E_{R,red}$): $\Delta E_{R,red} = E_{product} + E_{dmCr}{}^{+} - E_{reactant} - E_{dmCr}$, where $E_{product}$ and $E_{reactant}$ are the energies of the catalyst in the reduced and oxidized form, respectively; $E_{dmCr}{}^{+}$ and E_{dmCr} are the energies of the reducing agent (decamethylcromocene, dmCr) in the oxidized and reduced state, respectively.

Solvent effects and exchange-correlation influence were considered for the calculated reaction energies of the fourth, fifth, eighth, and ninth steps of the catalytic cycle for which experimental free energies are available.[10] Steps leading to the formation of an intermediate (i.e., **I**) were labeled as Sn, with *n* equal to the number of the intermediate (i.e., **SI**).

(iii) Reaction energies were estimated with an implicit solvent ($\Delta E_R{}^{S}{}_{PCM}$). Thus, reaction energies in an implicit solution were calculated as follows: $\Delta E_R{}^{S}{}_{PCM} = E^{S}{}_{PCM}$(product) + $E^{S}{}_{PCM}$(LutH$^+$ or dmCr$^+$) − $E^{S}{}_{PCM}$(reagent) − $E^{S}{}_{PCM}$(Lut or dmCr). Solvation energies were also estimated with an explicit solvent ($\Delta E^{S}{}_{exp}$) as $\Delta E^{S}{}_{exp} = E_{solute/solvent} - E_{solute} - nE_{solvent}$, where $E_{solute/solvent}$ is the energy of the solute in the presence of *n* ($1 < n < 4$) solvent molecules, E_{solute} is the energy of the solute in vacuo, and $E_{solvent}$ is the energy of each single solvent molecule. Since solvation is shown to have a larger impact on relative energies of Lut/LutH$^+$ and dmCr/dmCr$^+$ and to a minor extent on relative energies on reaction intermediates (vide infra), explicit solvation effects on reaction intermediates were neglected. Thus, reaction energy upon explicit solvation was given by $\Delta E_R{}^{S}{}_{exp} = E$(product) + $E^{S}{}_{exp}$(LutH$^+$ or dmCr$^+$) − E(reagent) − $E^{S}{}_{exp}$(LutH or dmCr). Errors with respect to experimental free energies (ΔG_{mes})[10] are given as $\Delta\Delta E_R$, $\Delta\Delta E_R{}^{S}{}_{PCM}$, and $\Delta\Delta E_R{}^{S}{}_{exp}$ where $\Delta\Delta E_R = |\Delta E_R - \Delta G_{mes}|$, $\Delta\Delta E_R{}^{S}{}_{PCM} = |\Delta E_R{}^{S}{}_{PCM} - \Delta G_{mes}|$, and $\Delta\Delta E_R{}^{S}{}_{exp} = |\Delta E_R{}^{S}{}_{exp} - \Delta G_{mes}|$.

Calculated Properties. Bond orders and Boys orbitals were calculated for all the complexes studied in this work.[46,47] Based on Boys orbitals, the bond ionicity BI_{AB} was calculated as in ref 48, namely $BI_{AB} = (d_A)/(d_{AB})$, where d_A is the distance between atom A and the Boys orbital along the AB bond; d_{AB} is the length of the bond between A and B. The analysis of BI_s is a useful tool to individuate lone pairs and provides an estimation of the ionicity of chemical bonds.[48]

To analyze intermolecular solute/solvent interactions, the Atoms In Molecules (AIM) theory was employed.[49,50] Topological analysis of computed electron densities (ρ) was performed using the AIM2000 package.[51] AIM is based upon those critical points (CPs) where the gradient of the density, $\nabla\rho$, vanishes. Such points are classified by the curvature of the electron density, for example bond critical points (BCPs) have one positive curvature (in the internuclear direction) and two negative (perpendicular to the bond). Properties evaluated at such BCPs characterize the bonding interactions present[52] and have been widely used to study intermolecular interactions. Recently, the AIM based method has been applied for quantifying π-stacking interactions.[53]

Results

1. Structural and Electronic Properties of I. Structural parameters of **I** (**A**, in Figure 1), calculated with PW basis set and BP[18,31] gradient corrections, compare well with those of the X-ray structure.[9,10] In particular, all bond lengths range within $\Delta d \sim 0.03$ Å, except the N_1-N_2 and $Mo-N_a$ bond distances, for which differences are of $\Delta d \sim 0.09$ Å and ∼0.06 Å, respectively. However, the estimated N_1-N_2 bond length is in line with previous theoretical calculations.[23,25,28,54] Furthermore, test calculations performed for an isolated N_2 molecule give a N_1-N_2 distance in excellent agreement with the experimental value ($N_1-N_2 = 1.11$ Å and 1.0975 Å,[11] respectively) suggesting that our computational setup is reliable for the dinitrogen moiety. Thus, the error observed in the N_1-N_2 distance of **A** may be partly due to our theoretical description and partly caused by strong N_2 thermal motions which may render the experimental N_1-N_2 distance mistakenly shorter.[8−10]

In contrast, discrepancies observed for $Mo-N_a$ may be ascribed to the chelating triamidoamine ligand backbone rigidity, i.e., the small error of calculated CH_2-CH_2 and CH_2-N_a bonds ($\Delta d = +0.01$ Å and $\Delta d = +0.02$ Å compared to the X-ray structure, respectively) may force the $Mo-N_a$ bond to elongate in order to avoid angular strain.

2. Choice of the Model System. Our calculations on **A** reproduce fairly well the experimentally available structural properties of **I**. However, calculations of the reaction energies of the entire catalytic cycle (Figure 2) as well as the estimation of solvation effects employing model **A** are not feasible due to the large size of the system. We attempted to identify a more computationally suitable model which retains the most relevant chemical features of **I** by comparing the main structural and electronic properties of **A** with those of simpler models, from **B** to **F**. In these models the hexaisopropylterphenyl aminoamine ligand was replaced by triphenylaminoamine (**B**), trimethylaminoamine (**C**), triamidoamine (**D**), by tree amine and one ammonia ligands (**E**), and finally by four ammonia ligands (**F**) (Figure 1). The accuracy of using model **B** for our calculations was tested performing calculations with PW, STO, and Gaussian basis sets with BP[18,31] (see Table S1, in the Supporting Information). These confirm a negligible dependence of the structural properties on the basis set used. Notably, the agreement between **A** and **B** is excellent ($\Delta d = +0.01$ Å with respect to **A**, Table 1). Furthermore, the HOMO orbital of **B** is characterized by a back-bonding from the occupied d_{xz} of

Table 1. Selected Experimental and Calculated Bond Lengths of **A** and **B** (Figure 1)

bond (Å)	X-ray [9,10]	**A**	**B**
$Mo-N_1$	1.96	1.98	1.99
N_1-N_2	1.06	1.15	1.15
$Mo-N_a$	2.19	2.24	2.25
$Mo-N_e$	1.98	2.01	2.00
$N_e-C(Ph)$	1.45	1.42	1.41
N_e-CH_2	1.47	1.47	1.48
CH_2-CH_2	1.51	1.52	1.53
CH_2-N_a	1.47	1.49	1.50

Mo to the empty π^* orbital of the nitrogen molecule (Figure S2a).[54]

Comparison of the structural parameters of **A** and **B** suggests the coordination sphere of **I** to be hardly affected by HIPTs. This may be consistent with experimental findings of HIPTs avoiding the formation of catalyst dimers.[2,8−10] Thus, our calculations confirm the experimental hypothesis that the structural properties of the catalyst are mainly dictated by the rigid geometry of the chelating triamidoamine ligand,[2,8−10] the HIPT substituents playing only a steric role.[2,8]

From a comparison of structural and electronic properties of models **B**−**F** (Supporting Information, Table S2), we conclude that **B** represents a good tradeoff between accuracy and computational cost to explore the entire energetic profile of dinitrogen reduction at a single Mo center. Therefore, model **B** was employed for all calculations reported in this work.

3. The Catalytic Mechanism. A detailed electronic and structural analysis of possible catalytic intermediates was carried out using **B** as model of **I**. Typically, reaction intermediates are more stable in the low spin state than in the higher spin state with the only exception of **XII**, for which the triplet state is the most likely.

Albeit not observed experimentally,[2,8−10] the isolated Mo complex (**0** in Figure 2) turns out to be a stable intermediate. Comparison with **I** shows that the absence of N_2 causes a decrease of the $Mo-N_a$ axial bond by $\Delta d = -0.11$ Å accompanied by a bond order (BO) increase, $\Delta BO = +0.1$ (Tables 2 and 3, respectively). Notably, the analysis of the Kohn−Sham orbitals supports experimental hypotheses that the $Mo-N_a$ weakening upon N_2 coordination may be caused by the *trans* influence of the N_2 ligand itself (Figure S2b).[55]

Formation of **I** starting from N_2 and **0** is exergonic ($\Delta E \sim -29$ kcal/mol) (Figure 3).[56] The bond order[46] of coordinated N_2 (BO = 2.3) is significantly smaller with respect to that of free nitrogen (BO = 3.0). This may be caused by the back-bonding from the occupied d_{xz} of the metal ion to the empty π^* orbital of the nitrogen molecule, as described by the HOMO orbital (Figure S2a). This strengthens the $Mo-N_1$ bond, in turn weakening the N_1-N_2 bond. Furthermore, the lone pair localized on N_1, identified by its Boys orbital,[47,48] is polarized upon the formation of the $Mo-N_1$ bond (Figure S5a), and it can no longer be exclusively attributed to N_1 ($BI_{MoN} = 0.78$) (Table 4.). Notably, a triple covalent bond is present between the two nitrogens as the bond ionicity indexes (which provide a picture of the polarity of chemical bonds) are close to 0.5 (BI_{N1-N2} of 0.40, 0.48, 0.49) (Figure S5a).[48] However, one bond is slightly polarized, in agreement with a N_1-N_2 bond order lower than three.

According to experimental suggestions,[8−10] the next step of the cycle involves the simultaneous protonation and reduction of the catalyst with formation of a diazenido complex. However, formation of a positively charged diazenido complex has not been completely ruled out.[2,8−10] Yet, this intermediate has been characterized by previous theoretical calculations,[25,28] and the existence of similar intermediates has been shown for different catalysts.[12] Indeed, we do find a stable positively charged diazenido (**II**) species (Figure 2), but the protonation process **SII** is endoergonic (Figure 3). The N_1-N_2 bond length increases ($\Delta d = +0.08$ Å), and in turn its bond order decreases ($\Delta BO = -0.5$) compared to **I**, N_1-N_2 becoming only double ($BI_{N1N2} = 0.57, 0.58$). Hybridization changes of both N_2 nitrogen atoms cause a bending of $Mo-N_1-N_2$ by a few degrees (Table 2). The $Mo-N_1$ distance decreases by $\Delta d = -0.14$ Å, and it is associated to an increase of bond order ($\Delta BO = +0.5$) with the formation of a metal−nitrogen double bond ($BI_{MoN} = 0.65, 0.65$).

Subsequent reduction, with formation of a neutral diazenido complex (**III**), is exergonic (Figure 3). Comparison with **II** shows that reduction weakens N_1-N_2 ($\Delta d = +0.02$ Å, $\Delta BO = -0.2$) but strengthens $Mo-N_1$ ($\Delta d = -0.04$, $\Delta BO = +0.2$). The reaction proceeds with a further protonation and subsequent formation of a positively charged hydrazido intermediate (**IV**), with $Mo-N_1$ being stronger than that of **III** ($\Delta d = -0.04$ Å, $\Delta BO = +0.2$). Indeed the BO analysis reveals that $Mo-N_1$ is triple with a slight ionic degree in **IV** ($BI_{Mo-N1} = 0.66, 0.65, 0.68$). A further weakening of the N_1-N_2 bond ($\Delta d = +0.07$ Å, $\Delta BO = -0.3$) occurs, showing an ionic and a covalent bond ($BI_{N1N2} = 0.84, 0.52$). A lengthening of the $Mo-N_a$ bond parallels these structural transformations ($\Delta d = +0.04$ Å, $\Delta BO = -0.1$), while the $Mo-N_1-N_2$ angle bends to 171°. Unlike **SII**, **SIV** is exergonic (Figure 3). Remarkably, comparison of calculated ΔE[36] and experimental reaction free energies of **SIV** shows a large discrepancy ($\Delta\Delta E_R \sim 17$ kcal/mol). Previous theoretical calculations (carried out with different models of the catalyst and solvated with an implicit solvent model)[21] showed a slightly smaller and yet significant error ($\Delta\Delta E_R{}^S{}_{PCM} \sim 10$ kcal/mol).[25] Possible sources of error for this large discrepancy are addressed below (vide infra). **SV** leads to reduction of **IV** to a neutral hydrazido complex (**V**), the process being exergonic (Figure 3). In contrast to the observed trend, reduction weakens the $Mo-N_1$ bond ($\Delta d = +0.06$, $\Delta BO = -0.3$), while it hardly affects N_1-N_2. BI_{N1N2} values indicate that a covalent (0.42) bond and an ionic (0.84) bond are present. As the $Mo-N_1-N_2$ angle further bends to 159°, one of the electron pairs forming the triple $Mo-N_1$ bond gains a stronger ionic character ($BI_{Mo-N1} = 0.88$), while the other ones show a double covalent character (0.65, 0.60). As for **SIV**, a large discrepancy is observed between calculated and experimental reaction energies of **SV** ($\Delta\Delta E_R \sim 9$ kcal/mol). Yet, the error is sensibly smaller than that of the previous protonation step. As seen for **SIV**, inclusion of an implicit solvent model improves the agreement, $\Delta\Delta E_R{}^S{}_{PCM} \sim 3$ kcal/mol.[25]

Table 2. Selected Bond Lengths (Å) and Angles (deg) of Each Catalytic Intermediate

	Mo **0**	MoN_2 **I**	$MoNNH^+$ **II**	MoNNH **III**	$MoNNH_2^+$ **IV**	$MoNNH_2$ **V**	$MoNNH_3^+$ **VI**
$Mo-N_1$		1.99	1.85	1.81	1.77	1.83	1.84
N_1-N_2		1.14	1.22	1.24	1.31	1.31	1.51
$Mo-N_{e1}$	2.02	2.02	2.00	2.00	2.00	2.04	1.98
$Mo-N_{e2}$	2.01	2.03	2.01	2.02	2.01	2.05	2.02
$Mo-N_{e3}$	2.02	2.02	1.99	2.01	2.02	2.03	2.02
$Mo-N_a$	2.14	2.25	2.30	2.28	2.32	2.31	2.37
$Mo-N_1-N_2$		180	177	176	171	159	131

	MoN **VII**	$MoNH^+$ **VIII**	MoNH **IX**	$MoNH_2^+$ **X**	$MoNH_2$ **XI**	$MoNH_3^+$ **XII**	$MoNH_3$ **XIII**
$Mo-N_1$	1.68	1.74	1.75	1.93	1.94	2.22	2.30
$Mo-N_{e1}$	2.04	1.99	2.05	1.99	2.02	1.99	2.02
$Mo-N_{e2}$	2.03	2.00	2.04	2.00	2.01	2.01	2.03
$Mo-N_{e3}$	2.04	2.00	2.09	1.99	2.03	1.99	2.04
$Mo-N_a$	2.45	2.37	2.37	2.34	2.30	2.25	2.20

Table 3. Selected Bond Orders[46] for Characteristic Bond Lengths of Each Catalytic Intermediate

bond order	Mo **0**	MoN_2 **I**	$MoNNH^+$ **II**	MoNNH **III**	$MoNNH_2^+$ **IV**	$MoNNH_2$ **V**	$MoNNH_3^+$ **VI**
$Mo-N_1$		0.5	1.0	1.2	1.3	1.0	1.1
N_1-N_2		2.3	1.7	1.5	1.2	1.2	0.8
$Mo-N_{e1}$	0.5	0.5	0.6	0.5	0.6	0.5	0.5
$Mo-N_{e2}$	0.5	0.5	0.6	0.5	0.6	0.5	0.6
$Mo-N_{e3}$	0.5	0.5	0.6	0.5	0.6	0.5	0.5
$Mo-N_a$	0.3	0.2	0.2	0.1	0.2	0.1	0.1

bond order	MoN **VII**	$MoNH^+$ **VIII**	MoNH **IX**	$MoNH_2^+$ **X**	$MoNH_2$ **XI**	$MoNH_3^+$ **XII**	$MoNH_3$ **XIII**
$Mo-N_1$	2.1	1.5	1.5	0.7	0.7	0.1	0.1
$Mo-N_{e1}$	0.5	0.6	0.6	0.6	0.5	0.5	0.5
$Mo-N_{e2}$	0.5	0.6	0.6	0.6	0.5	0.6	0.5
$Mo-N_{e3}$	0.5	0.6	0.6	0.6	0.5	0.5	0.5
$Mo-N_a$	0.1	0.1	0.1	0.1	0.1	0.2	0.2

Upon protonation of **V**, formation of a positively charged hydrazidium complex (**VI**) is exergonic (Figure 3). The additional proton leads to a further weakening of $Mo-N_1$ and N_1-N_2 bonds, $\Delta d = +0.01$ Å, $\Delta BO = -0.1$ and $\Delta d = +0.2$ Å, $\Delta BO = -0.4$, respectively. As a result (i) no electron pairs are present along the N_1-N_2 bond, while one pair is localized on N_2 ($BI_{N1N2} \sim 1$); (ii) the $Mo-N_1$ bond exhibits a double bond character ($BI_{MoN1} = 0.68, 0.69$); and (iii) the $Mo-N_1-N_2$ angle bends up to 131°.

SVII involves reduction of **VI** with the formation of a nitrido complex and the release of NH_3. The stability of a neutral hydrazidium complex was tested by adding one electron to **VI**. In line with the experiments, release of ammonia spontaneously occurs during DFT-based molecular dynamics simulations on the reduced form of **VI** at 0 K, suggesting that no energy barrier is expected for the formation of **VII**. Moreover, reduction is very exergonic (Figure 3). Interestingly, **VII** is characterized by a strengthening of $Mo-N_1$ ($\Delta d = -0.16$ Å, $\Delta BO = +1.0$), which gains the character of a triple bond ($BI_{Mo-N1} = 0.57, 0.57, 0.50$) (Figure S5b), along with a further weakening of $Mo-N_a$ ($\Delta d = +0.08$ Å, $\Delta BO = -0.5$).

The further stepwise protonation and reduction leads to formation of a positively charged and a neutral imido intermediate (**VIII** and **IX**). In both complexes, $Mo-N_1$ is weaker ($\Delta d = +0.07$ Å, $\Delta BO = -0.6$) with a strong ionic character ($BI_{Mo-N1} = 0.68, 0.68, 0.66$), compared to that of **VII**. The $Mo-N_e$ bond is stronger in **VIII** (average value $\Delta d = -0.04$ Å, $\Delta BO = 0.0$) than in **IX**. Moreover, protonation is exergonic, while reduction is endoergonic (Figure 3). Unfortunately, despite the satisfactory agreement between calculated and experimental structures of the intermediates,[9,10] estimated reaction energies differ remarkably from measured ones ($\Delta\Delta E_R \sim 16$ and $\Delta\Delta E_R \sim 10$ kcal/mol, for **SVIII** and **SIX**, respectively). As for **SIII** and **SIV**, inclusion of an implicit solvent model slightly improves the agreement ($\Delta\Delta E_R{}^S{}_{PCM} \sim 12$ and $\Delta\Delta E_R{}^S{}_{PCM} \sim 6$ kcal/mol, for **SVIII** and **SIX**, respectively).[25]

The cycle proceeds with the formation of a positively charged and a neutral amido intermediates (**X** and **XI**, respectively). These exhibit a significant weakening of the $Mo-N_1$ bonds ($\Delta d = +0.19$ Å, $\Delta BO = -0.8$). However, a difference of $\Delta d \sim 0.06$ Å ($\Delta BO \sim -0.8$) is observed in the $Mo-N_e$ bonds, of **X** and **XI**, with **X** showing the

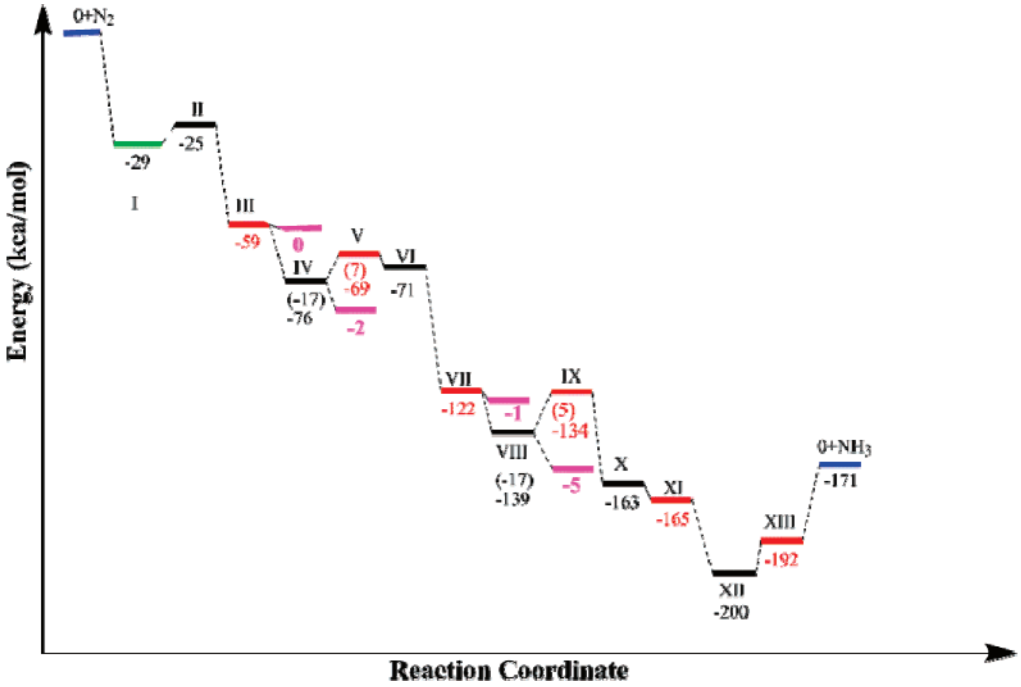


Figure 3. Reaction energy profiles. The green line represents the binding energy of N_2 to the catalyst. The black lines represent protonation steps. Red lines are reduction steps. The blue line represents the release of NH_3. Magenta lines refer to experimental ΔG.[10] Reaction energies are given with respect to 0 intermediate. For those steps for which experimental reaction energies are available we report calculated reaction energies and experimental free energies relative to the preceding intermediate in parentheses.

Table 4. Bond Ionicity for Mo−N_1 and N_1−N_2 Bonds for Each Catalytic Intermediate[a]

BI_{AB}	MoN_2 **I**	$MoNNH^+$ **II**	MoNNH **III**	$MoNNH_2^+$ **IV**	$MoNNH_2$ **V**	$MoNNH_3^+$ **VI**
Mo−N_1	0.78	0.65	0.65	0.66	0.88	0.96
Mo−N_1		0.65	0.63	0.65	0.62	0.69
Mo−N_1				0.60	0.61	0.68
N_1−N_2	0.40	0.58	0.56	0.84	0.84	1
N_1−N_2	0.48	0.57	0.55	0.51	0.42	
N_1−N_2	0.49					

BI_{AB}	MoN **VII**	$MoNH^+$ **VIII**	MoNH **IX**	$MoNH_2^+$ **X**	$MoNH_2$ **XI**	$MoNH_3^+$ **XII**	$MoNH_3$ **XIII**
Mo−N_1	0.57	0.68	0.73	0.78	0.78	0.77	0.79
Mo−N_1	0.57	0.68	0.73	0.71	0.71		
Mo−N_1	0.56	0.66	0.65				

[a] BI_{AB} is defined as $BI_{AB} = (d_A)/(d_{AB})$ where d is the displacement of the Boys orbitals along the bond and A = Mo and N_1 and B = N_1 and N_2, for Mo−N_1 and N_1−N_2 bonds, respectively.[47,48]

strongest ones. Furthermore, **SX** is exergonic, while **SXI** is endoergonic (Figure 3). Formation of the amido complex provokes a rearrangement of the charge density to accommodate the additional proton. This implies a double polarized Mo−N_1 bonds (BI $_{Mo-N1}$ ∼ 0.78, 0.71) in both the charged and the neutral compounds.

A final protonation leads to formation of a positively charged ammine intermediate (**XII**), in which the Mo−N_1 bond is weaker ($\Delta d = +0.28$ Å, $\Delta BO = -0.5$) with a single polar Mo−N_1 bond ($BI_{Mo-N1} = 0.77$).

The final reduction step determines the formation of a neutral ammine intermediate (**XIII**), which is characterized by an even weaker Mo−N_a bond ($\Delta d = +0.08$) with respect to **XII** and by a single Mo−N_1 ionic ($BI_{MoN1} = 0.79$) bond character. **SXII** is exergonic, and formation of **XIII** proceeds with no gain in reaction energy.

At the end of the cycle, a second NH_3 is released. The process is endoergonic by 21 kcal/mol (Figure 3).[56] This suggests that the activation energy of the process must be larger than this relatively high value, indicating that dissociation of the second ammonia molecule may be a slow step of the catalytic cycle. Indeed, the exchange between NH_3 and N_2 is exergonic by $\Delta E_{exc} = -8$ kcal/mol. However, it is not clear whether the reaction proceeds through an associative or a dissociative mechanism.[14]

Discussion

General Structural and Electronic Features of the Catalytic Cycle. Our calculations confirm experimental hypoth-

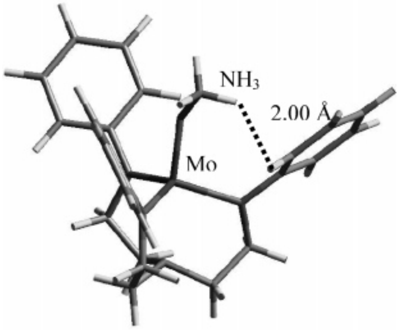


Figure 4. $MoNNH_3^+$ intermediate (**VI**). Steric contacts between a hydrogen of the NNH_3 ligand and the phenyls substituents are shown.

eses that the presence of a bulky substituent such as HIPT imposes the end-on-fashion coordination of N_2 and in turn allows for an efficient back-donation.[2,8]

In addition, our results suggest that the molecular nitrogen activation performed by the Mo catalyst is mainly caused by a back-bonding from the filled Mo d_{xz} orbital to the empty π^*-$N_1-N_2H_n$ (with $n = 0-2$) orbital, which is ubiquitous in the reaction intermediates **I**−**V** (Figure S2a). The transfer of electron density from the metal center to the π^*-$N_1-N_2H_n$ orbital strengthens Mo−N_1 and weakens the dinitrogen bond, leading to its cleavage.

The Mo−N_1 bond length does not decrease monotonically in the first half of the catalytic cycle. This may be ascribed to two factors: (i) the bending of the Mo−N_1−N_2 angle (initially collinear) causes the newly formed NH_3 group to experience steric repulsion from the phenyl group (minimum distance as short as 2.0 Å, Figure 4)[57] and the steric repulsion hampers the shortening of the Mo−N_1 bond and (ii) the back-bonding from the occupied d_{xz} metal orbital to the π^*-N_2-orbital becomes progressively less effective as the triple and double dinitrogen bonds are broken. This may also hamper the strengthening of the Mo−N_1 bond.

Experimental vs Calculated Reaction Energies of SIV, SV, SVIII, and SIX

Solvation Effects on Protonation Steps. As shown in Figure 3 the exergonic character of the protonation steps **SIV** and **SVII** for which experimental measurements are available is largely overestimated by DFT calculations ($\Delta\Delta E_R = 17$, 16 kcal/mol, respectively).[10] In order to explore possible sources of such discrepancy, DFT calculations with BP,[18,31] B3LYP,[19,20] and BHandH[20] exchange-correlation functionals have been performed on both the catalyst (model **B)** and the proton source.

In line with previous findings,[24−28] our in vacuo calculations confirm that geometries of molybdenum complexes and reaction energies of protonation steps do not depend significantly on the exchange-correlation functional used, Tables 5, S2, and S3. However, **SIV** and **SVIII** reaction energies do differ from experimental free energies, Table 5.[25] Taking into account solvent effects with an implicit solvent model,[21] as in previous studies,[25,26] only slightly reduces the discrepancy to $\Delta\Delta E_R{}^S{}_{PCM} \sim 10$ and ~ 12 kcal/mol, for **SIV** and **SVIII**, respectively.[25] This error is well beyond that typical of DFT calculations and confirms that other factors may be at the basis of such a discrepancy.[58]

Since reaction energies are rather independent of the exchange-correlation functional used,[24−27] and PCM[21] only slightly improves the agreement, the observed error may be due to a lack of explicit description of solvent effects. Albeit experimental studies demonstrated that the solubility of the reactive species is limited in heptane solution,[2,59] the aromatic rings of both molybdenum species and proton source suggest the presence of solute/solvent hydrophobic interactions. Thus, we explored the effect of an increasing number of explicit solvent molecules (from one methane molecule up to four heptane and benzene molecules) on the energetics of **SIV** and **SVIII**. Since solute/solvent interactions should mainly be C−H...π and π...π, the use of the most popular DFT functionals is not appropriate.[60] In contrast, correlated ab initio methods (such as MP2[61] and CCSD),[62] which account for dispersion interactions, are computationally too demanding for such systems. A valuable alternative is constituted by the DFT-BHandH[20] as it has recently been shown to describe dispersion interactions with the similar accuracy of CCSD[62] calculations but at an affordable computational cost.[63] Indeed, the BHandH functional[20] has been successfully applied to a variety of systems governed by dispersion forces, showing excellent results.[63,64] Interestingly, preliminary data (not reported) indicate that interaction energies of a prototypical example of a C−H...π bonded system (CH_4...benzene) obtained with BHandH/6-31+G(d) are in good agreement with those estimated with high level calculations.[65] Moreover, test calculations reported in the Supporting Information confirm the reliability of the BHandH[20]/6-31+G(d) calculations to reproduce the geometries of the molybdenum complexes (Tables S2−S4). Thus, we fully optimize solutes (reaction intermediates as well as $LutH^+$ and Lut) in the presence of explicit solvent molecules (Figure 5 and Table 6).

A first test to estimate the impact of C−H...π interactions was carried out with methane. The interaction of catalytic intermediates with a methane molecule turns out to similarly stabilize molybdenum complexes involved in **SIV** and **SVIII** as the difference in solvation energies, $\Delta\Delta E_{exp\ (MoN/MoNH+)}$ and $\Delta\Delta E^S{}_{exp\ (MoNNH/MoNNH2+)}$, is less than 0.5 kcal/mol, Table 6. In contrast, Lut and $LutH^+$ solvation energies differ by $\sim$3 kcal/mol as the former gains $\sim$2 kcal/mol upon solvation but the latter $\sim$5 kcal/mol.

Since the complete catalytic cycle takes place in heptane,[2,8−10] we estimated $\Delta E_R{}^S{}_{exp}$ with an increasing number of heptane molecules. In the presence of one heptane, molybdenum complexes are similarly affected by the inclusion of solvent ($\Delta\Delta E^S{}_{exp\ (MoN/MoNH+)} \approx \Delta\Delta E^S{}_{exp(MoNNH/MoNNH2+)} \sim 1$ kcal/mol), while $\Delta\Delta E^S{}_{exp\ (Lut/LutH+)}$ is ~ 3 kcal/mol. The addition of a second heptane molecule determines a further increase of both $\Delta\Delta E^S{}_{exp(MoNNH/MoNNH2+)}$ and $\Delta\Delta E^S{}_{exp(Lut/LutH+)}$, whereas $\Delta\Delta E^S{}_{exp(MoN/MoNH+)}$ is hardly changed. The addition of other solvent molecule(s) would render molybdenum species so large that DFT calculations become prohibitive even with considerable computational resources. Thus, solvent corrections for molybdenum species were estimated with two heptane molecules at the most. This is enough to

Table 5. ΔE_R (kcal/mol) of Processes **SIV** and **SVIII**, DFT vs Experimental Results

					in solution						
	in vacuo				C_6H_6			C_7H_{16}			
	BP[a]	BP[b]	B3LYP[c]	BH&H[d]	B3LYP (PCM)[b]	BH&H[b,e]	BH&H[d,e]	B3LYP (PCM)[c]	BH&H[b,e]	BH&H[d,e]	expt (C_6H_6)
SIV	−17.0	−18.9	−18.9	−16.3	−8.7	4.7	2.9	−10.2	−5.3	−4.7	∼0
SVIII	−17.2	−17.8	−15.3	−14.6	−7.9	6.4	4.6	−13.2	−3.6	−3.0	∼ −1

[a] With plane waves. [b] With 6-31+G(d). [c] B3LYP/TZVP ref 25. [d] With BH&H/6-311++G(d,p). [e] Considering solvated LutH$^+$ and Lut with four solvent molecules.

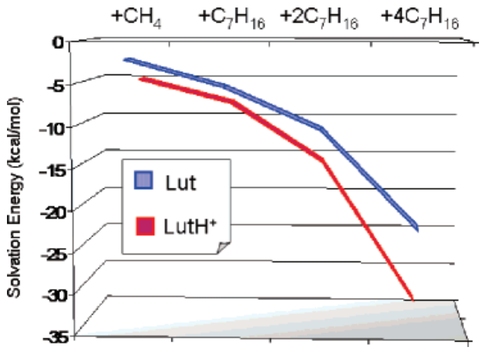


Figure 5. Solvation energies ΔE^{S}_{exp} vs number of solvent molecules (methane and heptane) for LutH$^+$ (red) and Lut (blue).

provide a rough idea that solvent effects ($\Delta E^{S}_{exp(MoNNH/MoNNH2+)} < \Delta E^{S}_{exp(MoN/MoNH+)} < 4$ kcal/mol)[66] are similar for reactants and products and much smaller[67] than that of the proton source. Instead, calculations including four heptane molecules, that virtually represent a first solvation shell, are still affordable for LutH$^+$ and LutH.[68]

As displayed in Table 6 and Figure 5, the solvent effect for the protonation source is quite large, as $\Delta\Delta E^{S}_{exp(Lut/LutH+)} \sim 11$ kcal/mol. This figure is slightly enhanced if the larger basis set is used (Table 6). Therefore, even an apolar solvent like heptane may play a relevant role for the energetic profile of **SIV** and **SVIII**.[59]

Since measurements of reaction free energies have been carried out in the presence of benzene,[10] this solvent was also taken into account. Considering implicitly benzene solvent reduces $\Delta\Delta E_R{}^{S}{}_{PCM}$ to ∼9 and ∼7 kcal/mol for **SIV** and **SVIII**, respectively. Yet, solvent effects are much larger when 2 and 4 benzene molecules are included explicitly in the calculation as $\Delta\Delta E^{S}_{exp(Lut/LutH+)}$ is ∼ 9 and ∼20 kcal/mol, respectively. Unfortunately, no more than four benzene molecules can be considered as the size of the system becomes rapidly prohibitive even for DFT calculations. Nevertheless, we are confident that four solvent molecules may represent a virtual first solvation shell.[69]

In fact, although the method is highly approximated, several features confirm that a first solvation sphere was achieved for both the benzene and heptane solvation models: (a) solvation energies showed no significant changes when the fourth molecule was added (less than ∼3 kcal/mol); (b) one of the four solvent molecules does not interact directly with the solutes (Figure 6); and (c) four solvent molecules occupy most of the space around both forms of the protonation agent, reducing the probability of a fifth solvent molecule directly interacting with the solutes (Figures 6 and S6). We are, hence, confident that four solvent molecules may represent a good tradeoff between computational costs and the fair description of solvation effects.

Once solvation effects are included in the estimation of reaction energies, discrepancy between experimental free energies and calculated reaction energies is drastically reduced to less than ∼5 kcal/mol for both **SIV** and **SVII** when the large basis set is used (Table 6).[68] Therefore, inclusion of explicit solvent molecules significantly improves the agreement between experimental and calculated reaction energies.

In order to understand the reason why LutH$^+$ undergoes a larger stabilization than Lut in the presence of explicit solvent, the Atoms In Molecules (AIM)[51] analysis was employed (Table 6). In particular, Figure 6 displays a schematic view of C−H...π attractive interactions between four heptane molecules and both LutH$^+$ (Figure 6A′) and Lut (Figure 6B′). Lut(heptane)$_4$ has seven Bond Critical Points (BCPs) corresponding to C−H...π interactions with a total electron density (ρ_{TOT}) equal to 0.0497 au, where LutH$^+$(heptane)$_4$ shows nine BCPs with $\rho_{TOT} = 0.0550$ au. Notably, one of the four heptane molecules in both Lut(heptane)$_4$ and LutH$^+$(heptane)$_4$ is not directly interacting with the solute via attractive interactions. Consistently, π...π interactions are present between benzene molecules and a protonation agent. In particular, for Lut(benzene)$_4$ we observe six BCPs, corresponding to π...π interactions with $\rho_{TOT} =$ 0.137 au. On the other hand, LutH$^+$(benzene)$_4$ shows seven BCPs with $\rho_{TOT} = 0.160$ au. Indeed, electron densities reported in Table 6 confirm that LutH$^+$ interacts more strongly with the solvent molecules than Lut. Furthermore, the strength of such interactions depends quite significantly on the solvent employed, providing a rationale for the energies reported in Tables 5 and 6.

In summary, inclusion of explicit apolar solvent molecules sensibly influences the energetic profile of **SIV** and **SVIII**.

Solvation Effect on Reduction Steps. In order to verify whether disagreement between experimental and calculated reaction energies of **SV** and **SIX** may be due to the lack of solvation effects as shown for protonation steps, similar calculations were performed also for the reduction steps. First, reaction energies of **SV** and **SIX** were estimated considering an implicit solvent model. Since experimental energies reported in the literature are measured in PhF,[10] while the reaction actually takes place in heptane, both solvents were considered. With the inclusion of implicit PhF and heptane solvents, the description of reduction steps does not improve. For instance, $\Delta\Delta E_R{}^{S}{}_{PCM} \sim 3$ and ~ 6 kcal/mol

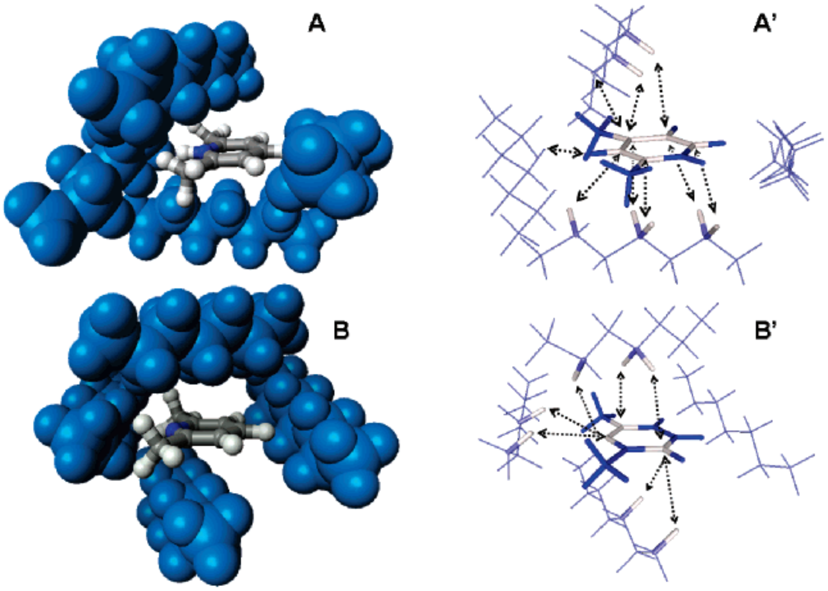


Figure 6. Views of solvation models with four heptane molecules of (A) $LutH^+$ and (B) Lut and schematic view of C−H...π interactions (black arrows) in A′ and B′, respectively.

Table 6. (a) and (b) Solvation Energies (ΔE^S_{exp}) (kcal/mol) and Total Electron Densities (au) of π...π and C−H...π Interactions with Explicit Solvent Molecules

(a)

	+ 1xCH_4		+ 1xC_7H_{16}		+ 2xC_7H_{16}		+ 4xC_7H_{16}		+ 4xC_7H_{16}
	ΔE^{Sa}	ρ_{TOT}	ΔE^{Sa}	ρ TOT	ΔE^{Sa}	ρ TOT	ΔE^{Sa}	ρ TOT	ΔE^{Sb}
Lut	−2.2	0.0081	−5.4	0.0183	−10.5	0.0419	−22.5	0.0497	−25.2
$LutH^+$	−5.1	0.0212	−8.1	0.0300	−15.6	0.0440	−33.5	0.0550	−36.8
MoN	−3.6	0.0160	−2.0	0.0170	−5.8	0.0390			
$MoNH^+$	−3.9	0.0171	−0.6	0.0075	−2.0	0.0320			
MoNNH	−1.7	0.0110	−0.4	0.0200	−0.9	0.0198			
$MoNNH_2^+$	−2.0	0.0140	−1.0	0.0120	−1.9	0.0286			

(b)

	+ 2xC_6H_5F		+ 2xC_7H_{16}		+ 2xC_6H_6		+ 4xC_6H_6		+ 4xC_6H_6
	ΔE^{Sa}	ρ_{TOT}	ΔE^{Sa}	ρ TOT	ΔE^{Sa}	ρ TOT	ΔE^{Sa}	ρ TOT	ΔE^{Sb}
dmCr	−1.45	0.0212	−2.7	0.0290					
$dmCr^+$	−6.1	0.0510	−4.0	0.0250					
Lut					−7.8	0.0290	−13.2	0.137	−19.8
$LutH^+$					−17.2	0.0890	−34.2	0.160	−39.0

[a] Calculated solvation energies with BHandH/6-31+G(d). [b] Calculated solvation energies with BHandH/6-311++G(d,p).

for **SV** and **SIX** in heptane and $\Delta\Delta E_R{}^S{}_{PCM} \geq 20$ kcal/mol for both reactions in PhF.[35]

Unlike protonation reactions, a correct evaluation of explicit solvation effects for reduction steps present several issues. First, the size of decamethylcromocene is rather large, and calculations of dmCr(heptane)$_n$ complexes become prohibitive already with $n > 2$. Most importantly, the ionization energy of $dmCr^+$/dmCr calculated with BHandH,[20] which is the only exchange-correlation functional among those used in this work that accounts for dispersion interactions,[53] is underestimated (Table 7). Nevertheless, DFT-BHandH[20] calculations were performed to estimate the influence of solvation on the overall reaction energy of reduction steps. As shown in the previous section, solvation effects on reaction intermediates were negligible. Only calculations on $dmCr^+$ and dmCr were carried out by considering two heptane and two PhF molecules explicitly. Interestingly, the effect of heptane solvation is small, for instance $\Delta E^S_{exp}(heptane)_2$ are ∼ −3 and ∼ −4 kcal/mol for dmCr and $dmCr^+$, respectively (compared with $\Delta E^S_{exp}(heptane)_2$ ∼ −10 and ∼ −15 kcal/mol of Lut and $LutH^+$). Accordingly, ΔE^S_{exp} $(PhF)_2$ is ∼ −1.5 and ∼ −6 kcal/mol, for dmCr and $dmCr^+$, respectively, confirming that solvation effects are indeed much smaller than those of the protonation agent (ΔE^S_{exp} $(benzene)_2$ ∼ −8 and ∼ −17 kcal/mol for Lut and $LutH^+$, respectively). Unlike Lut and $LutH^+$, interactions with heptane are similarly strong for both $dmCr^+$ and dmCr (5 BCPs, ρ_{TOT} = 0.0290 and 5 BCPs, ρ_{TOT} =

Table 7. Dependence of the Ionization Energy and the ΔE_R for **SV** and **SIX** on the Exchange-Correlation Functionals

		MoNNH$_2^{+/0}$		MoNH$^{+/0}$	
	dmCr$^{+/0}$ IE	IE	ΔE_R(**SV**)	IE	ΔE_R(**SIX**)
expt.[70]	113.7		~ −2		~ −5
B3LYP	100.7	106.5	−5.9	109.0	−8.3
B3LYP[a]	101.6	105.2	−3.6	109.6	−8.0
BLYP	102.5	97.2	5.3	99.6	+2.9
BHandH	89.7	109.9	−20.2	112.0	−22.3
BHandHLYP	89.6	115.6	−26.0	116.7	−27.2
BP	108.5	101.6	6.9	104.0	+4.5
BP[§]	108.5	101.7	6.8	105.5	3.0
B3P	114.4	118.9	−4.5	121.3	−6.9
B3P[§]	114.7	117.4	−2.7	122.2	−7.5
PW91	98.4	107.5	−9.2	110.0	−11.6

[a] Full optimization with localized basis sets, while all other data refer to single point calculations on BP[18,31] optimized geometry with PW basis sets.

0.0250, for dmCr and dmCr$^+$, respectively, Figure S7), resulting in a $\Delta\Delta E^S_{exp}$ (heptane)$_2$ ≤ 1.5 kcal/mol. In contrast, two PhF molecules cause a larger difference, yet not significant, between reduced and oxidized dmCr, as $\Delta\Delta E^s_{exp}$ (PhF)$_2$ ≤ 4.6 kcal/mol (3 BCPs, $\rho_{TOT} = 0.0212$ and 7 BCPs, $\rho_{TOT} = 0.0510$ for dmCr and dmCr$^+$, respectively). Therefore, no strong solvent effect is expected on the energetics of reduction steps, and in vacuo calculations should provide reasonable reaction energies. Thus, the discrepancy between experimental and calculated reaction energies cannot be solely attributed to the lack of explicit solvent.

Performance of DFT Exchange-Correlation Functionals for Reduction Steps. Since ionization energy (IE) of dmCr/dmCr$^+$ strongly depends on the exchange-correlation functional, its effect on ΔE_R of **SV** and **SIX** was monitored. Ionization energy for dmCr/dmCr$^+$ has been measured as 113.7 kcal/mol.[70] As summarized in Table 7, several DFT functionals provide remarkably discrepant results (ranging between 89 and 114 kcal/mol). IEs were evaluated either minimizing the structure or performing SP calculations on the BP[18,31] optimized structures.[71]

Among all functionals, B3P[20,31] and BP[18,31] give the best agreement with experimental IE,[70] while those estimated with BHandH[20] and BHandLYP[20,19] are strongly underestimated.

Similarly, IEs of the catalytic intermediates redox couples largely depend on the DFT functionals, with IE(MoNH$^+$/MoNH) and IE(MoNNH$_2^+$/MoNNH$_2$) ranging between ~97/~117 kcal/mol and ~99/~122, respectively. As a consequence, calculated ΔE_R(**SV**) and ΔE_R**(SIX)** range between −26.0/+6.9 kcal/mol and −27.2 /+4.5 kcal/mol. Thus, reduction steps $\Delta\Delta E_R$ is mainly affected by the errors in the estimation of the ionization energies.

Our results indicate that DFT functionals B3LYP[19,20] and B3P[20,31] are the only ones able of reproducing **SIV** and **SVIII** reaction energies within an error of 2−3 kcal/mol. However, the correct value produced by B3LYP[20,19] is likely due to a fortuitous cancellation of error, as IE(dmCr/dmCr$^+$) calculated with B3LYP was quite different from the experimental data (ΔIE ~ 12 kcal/mol). Thus, B3P[20,31] is the only exchange-correlation functional reproducing both IE and ΔE_R correctly.

Conclusions

We performed a structural and electronic characterization of Mo(HIPTN$_3$N) (with HIPT = hexaisopropylterphenyl) a compound that performs the same catalytic function of the enzyme nitrogenase.[2,8] Our calculations suggest that the chelating ligand rigidity is mainly responsible for the structural properties of the catalyst. In agreement with experimental findings,[2,8−10] we showed that the large HIPT substituents hardly affect the structural and electronic properties of the catalyst, playing only a steric role that hampers formation of catalyst dimers.[8−10] At the same time, the bulky HIPT substituents create a cage in which the molecular nitrogen binds in an end-on-fashion.

Using a relatively small computational model (**B**, Figure 1), which reproduces the structural and electronic properties of the full complex (**A**, Figure 1), we described the electronic and structural features of possible catalytic intermediates of the Chatt-like mechanism in Figure 2. A visual inspection of the Kohn−Sham orbitals reveals that the HOMO orbital represents a π*-back-donation from the filled d_{xz} metal orbital to the empty π* orbital of N$_2$[28,54] This orbital is common to all reaction intermediates **I**−**V** suggesting the back-donation as the electronic mechanism responsible for the activation of N$_2$.

An increase of the N$_1$−N$_2$ bond lengths (with a concomitant decrease of the bond order) is observed for the first half reaction, while Mo−N$_1$ bond lengths do not follow any trend because of the combination of steric and electronic effects that come along with the change in hybridization of both nitrogen atoms. In contrast, the Mo−N$_1$ bond constantly decreases in the second half of the catalytic cycle.

Notably, calculated reaction energies are remarkably different from experimental ones even considering implicit solvent models.[25] However, ΔE_R, corrected by taking into account explicit solvent molecules around the protonation agent, reduces the discrepancy for protonation steps to less than roughly 5 kcal/mol. Thus, a lack of explicit solvent may be one of the reasons of the observed discrepancy for protonation steps.[59]

In contrast, an explicit solvation does not improve the agreement for reduction steps, where the discrepancy is mainly associated with the dependence of the ionization energies on the DFT exchange-correlation functionals.

In conclusion, our results show that DFT calculations are a powerful tool to unveil structural and electronic properties of the intermediates of the catalytic cycle. However, due to the complexity of the catalytic system, reaction energies cannot be easily reproduced, limiting the predictability of such calculations.

Acknowledgment. We thank the CINECA computing center, Bologna, Italy and the CNR-INFM grants for the computational resources. Financial support has been provided by the MURST grant ‘Fondo per il sostegno dei giovani e per favorire la mobilita’ degli studenti’.

Supporting Information Available: Test calculations on the dependence of the geometries on the basis sets and the exchange-correlation functionals and figures illustrating the electronic structure and the solvation model employed. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Modak, J. M. *Resonance* **2002**, 69.

(2) Schrock, R. R. *Acc. Chem. Res.* **2005**, *38*, 955−962.

(3) Schrock, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17087−17087.

(4) Howard, J. B.; Rees, D. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, 1*03*, 17088−17093.

(5) Rees, D. C.; Howard, J. B. *Curr. Opin. Chem. Biol.* **2000**, *4*, 559−566, and references therein.

(6) Burgess, B. K. *Chem. Rev.* **1990**, *90*, 1377−1406.

(7) Burgess, K. B.; Lowe, D. J. *Chem. Rev.* **1996**, *96*, 2983−3011.

(8) Yandulov, D. V.; Schrock, R. R *Science* **2003**, *301*, 76−78.

(9) Yandulov, D. V.; Schrock, R. R.; Rheingold, A. L.; Ceccarelli, C.; Davis, W. M. *Inorg. Chem.* **2003**, *42*, 796−813.

(10) Yandulov, D. V.; Schrock, R. R. *Inorg. Chem.* **2005**, *44*, 1103−1117.

(11) Mackay, B. A.; Fryzuk, M. D. *Chem. Rev.* **2004**, *104*, 385−401.

(12) Pickett, C. J. *J. Biol. Inorg. Chem.* **1996**, *1*, 601−606.

(13) Chatt, J.; Dilworth, J. R.; Richards, R. L. *Chem. Rev.* **1978**, *78*, 589−625.

(14) Weare, W. W.; Dai, X.; Brynes, M. J.; Chin, J. M.; Schrock, R. R. Mueller, P. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17099−17106.

(15) Smythe, N. C.; Schrock, R. R, Muller, P.; Weare, W. W. *Inorg. Chem.* **2006**, 45, 9197−9205.

(16) Smythe, N. C.; Schrock, R. R.; Muller, P.; Weare, W. W. *Inorg. Chem.* **2006**, *45*, 7111−7118.

(17) (a) Ritleng, V.; Yandulov, D. V.; Weare, W. W.; Schrock, R. R.; Hock, A. S.; Davis, W. M. *J. Am. Chem. Soc.* **2004**, *126*, 6150−6163. (b) Weare, W. W.; Schrock, R. R.; Hock, A. S.; Muller, P. *Inorg. Chem.* **2006**, *45*, 9185−9196.

(18) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(19) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785−789.

(20) (a) Becke, A. D. *J. Chem. Phys.* **1993**, 98, 5648−5652. (b) Becke; A. D. *J. Chem. Phys.* **1993** *98*, 1372−1377.

(21) Cammi, R.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2000**, *103*, 9100- 9108.

(22) (a) Mo, S. J.; Vreven, T.; Mennucci, B.; Morokuma, K.; Tomasi, J. *Theor. Chem. Acc.* **2004**, *111*, 154−161. (b) Balcells, D.; Carbo, J. J.; Maseras, F.; Eisenstein, O. *Organometallics* **2004**, *23*, 6008−6014. (c) Belkova, N. V.; Besora, M.; Epstein, L. M.; Lledos, A.; Maseras, F.; Shubina, E. S. *J. Am. Chem. Soc.* **2003**, *125*, 7715−7725. (d) De Abreu, H. A.; Guimaraes, L.; Duarte, H. A. *J. Phys. Chem. A* **2006**, *110*, 7713−7718. (e) Kovacs, G.; Papai, I. *Organometallics* **2006**, *25*, 820−825. (f) Begue, D.; Carbonniere, P.; Barone, V.; Pouchan, C. *Chem. Phys. Lett.* **2005**, *416*, 206−211.

(23) Zexing, C.; Zhou, Z.; Wan, H.; Zhang, Q. *Int. J. Quantum Chem.* **2005**, 103, 344−353.

(24) Neese, F. *Angew. Chem., Int. Ed.* **2006**, *45*, 196−199.

(25) Studt, F.; Tuczek, F. *Angew. Chem., Int. Ed.* **2005**, *44*, 5639−5642.

(26) Studt, F.; Tuczek, F. *J. Comput. Chem.* **2006**, *27*, 1278−1291.

(27) Reiher, M.; Le Guennic, B.; Kirchner, B. *Inorg. Chem.* **2005**, *44*, 9640−9642.

(28) Le Guennic, B.; Kirchner, B.; Reiher, M. *Chem. Eur. J.* **2005**, *11*, 7448−7460.

(29) Hölscher M.; Leitner, W. *Eur. J. Inorg. Chem.* **2006**, 4407−4417.

(30) Carr, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(31) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822−8824.

(32) (a) Magistrato, A.; Woo, T. K.; Togni, A.; Rothlisberger, U. *Organometallics* **2004**, *23*, 3218−3227. (b) Magistrato, A.; Togni, A.; Rothlisberger, U. *Organometallics* **2006**, *25*, 1151−1157.

(33) Maurer, P.; Magistrato, A.; Rothlisberger, U. *J. Phys. Chem. A* **2004**, *108*, 11494−11499.

(34) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1986**, *33*, 8800−8802.

(35) PhCl has been considered for PCM calculations since PhF is not available in the Gaussian03 package.

(36) We have to remark that our calculated Δ*E* do not account for entropic effects.

(37) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41−51.

(38) Baerends, E. J.; Ros, P. *Chem. Phys.* **1973**, *2*, 52−59.

(39) (a) van Lenthe, E.; Ehlers, A. E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 8943−8953. (b)van Lenthe, E.; Baerends, E. J. Snijders J. G. *J. Chem. Phys.* **1993**, *99*, 4597−4610.

(40) Hutter, J.; Ballone, P.; Bernasconi, M.; Focher, P.; Fois, E.; Goedecker, S.; Parrinello, M.; Tuckerman, M. Max-Planck-Institut für Festkörperforschung and IBM Zurich Research Laboratory, 1995−1996.

(41) (a) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703−1710. (b) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641−3662.

(42) Trouillier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993−2006.

(43) Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425−1428.

(44) Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136−211. (b) Barnett, R. N.; Landmann, U. *Phys. Rev. B* **1993**, *48*, 2081−2097.

(45) Frisch, M. J. T.; 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) Magistrato, A.; Maurer, P.; Fassler, T.; Rothlisherger, U. *J. Phys. Chem. A* **2004**, *108*, 2008−2013.

(47) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847−12856.

(48) Alber, F.; Folkers, G.; Carloni, P. *J. Phys. Chem. B* **1999**, *103*, 6121−6126.

(49) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893−928.

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

(51) Bieglerkonig, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *3*, 317−328.

(52) Bader, R. F. W.; Essen, H. *J. Chem. Phys.* **1984**, *80*, 1943−1960.

(53) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 491−504.

(54) (a) Cui, Q.; Musaev, D. G.; Svensson, M.; Sieber, S.; Morokuma, K. *J. Am. Chem. Soc.* **1997**, *117*, 12366−12367. (b) Khoroshun, D. V.; Musaev, D. G.; Morokuma, K. *Mol. Phys*. **2002**, *100*, 523−532.

(55) Cotton, F. A.; Wilkinson, G. *Advanced Inorganic Chemistry*; John Wiley and Sons Inc.: New York, 1980.

(56) It has not been possible to localize an energy barrier for this reaction step.

(57) It is likely that, in the presence of the large HIPT substituents, these steric effects are further enhanced.

(58) Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6648−6653.

(59) Also the effect of the counterions, totally neglected in our calculation, may have an important role for the reaction energies.

(60) (a) Ye, X. Y.; Li, Z. H.; Wang, W. N.; Fan, K. N.; Xu, W.; Hua, Z. Y. *Chem. Phys. Lett.* **2004**, *397*, 56−61. (b) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334−338. (c) VandeVondele, J.; Lynden-Bell, R.; Meijer, E. J.; Sprik, M. *J. Phys. Chem. B* **2006**, *110*, 3614−3623.

(61) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503.

(62) (a) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35. (b) Hobza, P.; Sponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802−11808. (c) Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608−15613.

(63) (a) Robertazzi, A.; Platts, J. A. *J. Phys. Chem. A* **2006**, *110*, (11), 3992−4000. (b) Robertazzi, A.; Platts, J. A. *Chem. Eur. J.* **2006**, *12*, 5747−5756.

(64) Burda, J.; Sokalski, A.; Leszczynski, J. *J. Mol. Model.* **2007**, *13*, 335−345.

(65) Shibasaki, K.; Fujii, A.; Mikami, N.; Tsuzuki, S. *J. Phys. Chem. A* **2006**, *110*, 4397−4404.

(66) The effect of explicit solvation on reaction intermediates is not considered since it is similar for neutral and positively charged reaction intermediates, thus it is expected not to significantly improve the calculation of reaction energies.

(67) Molybdenum complexes employed in experiments are larger than the model used, and the metal center is less accessible to solvent molecules. Thus, the effects of solvent may be even smaller than those obtained in this work.

(68) Solvation energies of Lut and $LutH^+$ estimated with three and four heptane molecules are very similar (data not shown), suggesting that adding more solvent molecules would not largely affect solvation energies. Thus, four heptane molecules may constitute the first solvation shell.

(69) Firstly, four benzene molecules occupy the entire space around the solute; therefore, a further solvent molecule would interact with one of the benzenes rather than with the solute. Secondly, solvation energies estimated with three and four benzenes are very similar (data not shown).

(70) Ryan, M. F.; Richardson, D. E.; Lichtenberger, D. L.; Gruhn, N. E. *Organometallics* **1994**, *13*, 1190−1199.

(71) A rough idea of exchange-correlation influence is estimated also by performing SP calculations on the BP[18,31] optimized geometries. Indeed, performing SP calculations on the BP[18,31] optimized geometry with a different functional we observe a maximal error of $\Delta E \sim 1.5$ kcal/mol, with respect to the total energy we would obtain relaxing geometry with the functional chosen.

CT700094Y

JCTC Journal of Chemical Theory and Computation

TIP5P-Consistent Treatment of Electrostatics for Biomolecular Simulations

Sarah M. Tschampel, Michael R. Kennerty, and Robert J. Woods*

Complex Carbohydrate Research Center, 315 Riverbend Road, Athens, Georgia 30602

Received February 23, 2007

Abstract: The inclusion of zero-mass point charges around electronegative atoms, such as oxygen, within molecular mechanical force fields is known to improve hydrogen-bonding directionality. In parallel, inclusion of lone-pairs (LPs) in the TIP5P water model increased its ability to reproduce both gas-phase and condensed-phase properties over its non-LP predecessor, TIP3P. Currently, most biomolecular parameter sets compute partial atomic charges via fitting of the classical molecular electrostatic potential (MEP) to the quantum mechanical MEP. Application of this methodology to optimize lone-pair description is therefore consistent with the current approach to modeling electrostatics and is straightforward to implement. Here, we present an atom-type specific lone-pair model, which leads to the most optimal LP placement for each atom type, and, notably, results in reproduction of the lone-pair description present in TIP5P. Carbohydrates are rich in hydroxyl groups, and development of a lone-pair inclusive carbohydrate force field for use with a lone-pair containing water model, such as TIP5P, ensures the compatibility between these two models. Implementation of this lone-pair model improves the geometry and energetics for a series of hydrogen-bonded clusters and the properties of several small molecule crystals over the non-LP containing force field.

Introduction

Partial charges are nonphysical entities that are nevertheless convenient to employ in the computation of the nonbonded Coulomb interaction in molecular mechanics force fields, eq 1.[1] The common use of atom-centered partial charges (monopoles) exclusively is based on the approximation that the higher order terms (dipole, quadrupole, etc.) can be ignored, due to the rapid rate at which the higher order contributions diminish with respect to internuclear distance. A variety of protocols have been developed to obtain atom-centered partial atomic charges for biomolecules. Several groups have pioneered the general method that employs the computed quantum mechanical molecular electrostatic potential (MEP) at a grid or shell of points around a given molecule to derive partial charges, eq 2.[1−10] Partial charges are fit to the atomic centers in a given molecule, so as to optimize the agreement between the classical MEP arising from these partial charges and the quantum mechanical MEP, eq 3.[1,7] Least-squares fitting yields a minimized error (χ), which is utilized here to gauge the quality of the fit between the classical and quantum mechanical MEP.

$$V_{\mathrm{electrostatic}} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1}$$

$$\mathrm{MEP}_{\mathrm{QM}}(r) = \sum_A \frac{Z_A}{r - R_A} - \sum_{\mu,\upsilon} P_{\mu},\upsilon \int \frac{\varphi_\mu \varphi_\upsilon}{|r - r'|}\,\mathrm{d}r' \tag{2}$$

$$\mathrm{MEP}_{\mathrm{classical}}(i) = \sum_{j=1}^{n} \frac{q_j}{r_{ij}} \tag{3}$$

Despite subtle differences in the derivation of partial atomic charges, most of the current MEP-based techniques have been shown to perform well in practice and are widely employed. In contrast, the inclusion of point charges around oxygen, nitrogen, and sulfur, to mimic lone-pair electrons, continues to be controversial. The option of including lone-

* Corresponding author phone: (706)542-4454; fax: (706)542-4412; e-mail: rwoods@ccrc.uga.edu.

10.1021/ct700046j CCC: $37.00
Published on Web 07/12/2007

pairs in molecular dynamics (MD) simulations with the AMBER program has been available since 1984.[9,11−13] Since 2000, the AMBER package has employed an analytical treatment of lone-pairs with zero-mass and rigid relative positions.[14] Previously, lone-pairs were treated as pseudoatoms with both a mass, a partial charge, and associated valence force constants.[14] Along with the conversion to an analytical treatment of lone-pairs ("Extended Points" (EPs) in AMBER nomenclature) the distance of the lone-pairs from the corresponding oxygen nucleus was only marginally altered from 0.20 Å (PARM81) to 0.35 Å (PARM02EP).[12,15−17] The method used to determine the lone-pair oxygen bond length in PARM02EP was based on the location of critical points in the charge density.[18−20] Critical points were identified for a test set of 21 compounds comprised of sp, sp,^{2} and sp^3 nitrogen as well as sp^2 and sp^3 oxygen atom types. Due to the insensitivity of critical point position to chemical environment, it was inferred that an average value of 0.35 Å for the LP−O bond length was transferable to all relevant oxygen and nitrogen atom types. More recently, a similar method was utilized for preliminary lone-pair placement to develop a polarizable lone-pair inclusive model.[21] However, the critical point location in the electron density is dependent on the level of theory and basis set employed in the quantum mechanical calculation.[9] Furthermore, the charge density has not been shown to lead to good reproduction of the MEP. Therefore, since fitting partial charges to the quantum mechanical MEP has been shown to be appropriate for condensed-phase MD simulations, a LP placement derived from this approach was examined for application within the AMBER force field for the AMBER and GLYCAM parameter sets.

Empirically adjusting partial charges to reproduce experimentally known condensed-phase properties, such as density and enthalpy of vaporization, from liquid-state MD simulations is an alternative method often used to compute partial charges for pure liquids, such as water, giving rise to the TIP5P water model.[22−25] In the case of TIP5P, the LP placement is specific for the hydroxyl-type oxygen present in the water molecule. This technique was also applied to the derivation of the LP−O distance for a model compound, methanol, in which the oxygen was approximated to be representative for all oxygen atom types. A distance of 0.47 Å was determined and applied to all oxygen atom types in a polarizable version of the OPLS force field.[26] This empirical treatment cannot be applied to amino acids, nucleic acids, or carbohydrates since they do not exist as pure liquids, and it is unclear to what extent values determined for model liquids are transferable to solutes.

Due to the number of hydroxyl groups as well as the presence of a ring oxygen atom carbohydrates inherently possess a large number of potential LP sites; in a typical hexopyranoside, such as methyl α-D-glucopyranoside (α-D-Glc*p*-OMe), there are 12 possible LP placements. Therefore, effects observed for the addition of LPs to one electronegative atom may be amplified in carbohydrates. Due to the number of hydroxyl groups, carbohydrates have the ability to form multiple inter- and intraresidue as well as solute−solvent hydrogen bonds. As a result, inclusion of LPs may have a beneficial effect on hydrogen-bond directionality in carbohydrate-containing systems, such as glycoproteins and carbohydrate−protein complexes. In these systems, the orientation of the glycan relative to the protein surface is directly influenced by intermolecular hydrogen bonds and may further include interactions with bridging water molecules.[27,28]

Derivation of an approach that can predict the LP description in TIP5P, and yet is applicable to biomolecular solutes, would result in a TIP5P-consistent approach to developing a LP-containing biomolecular force field. Currently, the GLYCAM parameter set is consistent with the AMBER methodology for partial charge derivation, which involves determination of partial charges that produce a classical MEP with the best fit to the quantum mechanical HF/6-31G(d) MEP.[16,29−32] Extension of this technique to determine the LP−O distance as well as the LP partial charges will ensure consistency with the current methodology, as long as the fit is to the quantum mechanical MEP computed with the same HF/6-31G(d) wave function. Nevertheless, the effect of the level of theory and basis set utilized to obtain the quantum mechanical MEP on the determination of the LP−O separation will be examined. In contrast to the results from critical point analysis of the charge density, analysis of the fits to the MEPs for a variety of test compounds, which include sp^2 and sp^3 oxygen atoms, show a correlation between the LP descriptors (LP−O distance and partial charge) and chemical environment. Therefore, different oxygen atom types can have different LP descriptors. Lone-pair placement around nitrogen atoms can also be derived with this approach, but addition of lone-pairs attached to oxygen atoms has previously been shown to have a more profound effect on the ability to reproduce characteristics of hydrogen bonding interactions and will only be examined here.[26,33]

Once implemented in the AMBER/GLYCAM parameter set, an analysis of the ability of this model to reproduce experimental and theoretical properties for gas-phase hydrogen-bonded clusters as well as monosaccharides in their crystalline form will be presented.[34,35]

Methods

Quantum Mechanical Computations. All quantum mechanical computations were performed with the Gaussian 98 suite of programs, version A.11.3.[36] The optimized geometry of water was computed at the HF/6-31G(d) and B3LYP/6-31G(d) levels.[37−40] The quantum mechanical MEPs were computed with the grid-based CHarges from the ELectrostatic Potentials (ChelpG) algorithm as implemented in Gaussian 98 at the HF and B3LYP level with the 6-31G(d) basis set.[7] Partial charges were computed utilizing the Restrained Electrostatic Potential (RESP[32]) scheme implementing a weighting factor of 0.010.[35]

A series of model compounds representing the oxygen atom types found in alcohols, ethers, ketones, amides, and carboxylate compounds was optimized, and the quantum mechanical MEPs were determined at the HF/6-31G(d), B3LYP/aug-cc-pVTZ, and MP2/aug-cc-pVTZ levels.[41,42] The use of this basis set with the B3LYP functional is to maintain

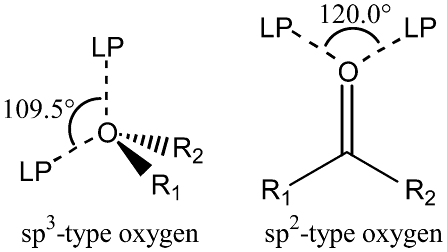


Figure 1. The lone-pair geometry for sp^3 and sp^2 oxygen atoms.

consistency with previously developed models implemented in AMBER.[43] For all sp^3-type oxygen atoms the LP geometry was constrained to prefer a tetrahedral placement, while for all sp^2-type oxygen atoms the LPs were constrained to be in the plane of the carbonyl group and its two substituents, forming a 120° angle, except for the carboxylate compounds, Figure 1.[15,26]

Molecular Mechanics. The GLYCAM06 parameters were utilized in conjunction with the PARM94 parameter set, with the addition of LPs that are defined and fixed in relation to their attached oxygen atom type and have no van der Waals radii. Energy minimizations were performed with the SANDER module of AMBER7 with a dielectric constant of unity for 10 000 cycles (9000 steepest descent followed by 1000 conjugate gradient).[14] All atoms were included in the calculation of nonbonded interactions, and each energy minimization was initiated from the ab initio geometry to which it was compared. Two types of systems were investigated, in which the first type of system consisted of a series of small hydrogen-bonded clusters that did not employ any geometrical restraints and the second investigated the approach of a water molecule toward either methanol or *N*-methyl acetamide (NMA) in which additional restraints were necessary.[44,45] In each of the approach trajectories, the intermolecular O···O distance was restrained at the QM value along the potential energy surface. In addition, the methyl group was restrained from rotating away from the ab initio value in the NMA−water complex with a harmonic restraint. The only additional restraints were employed as needed to ensure the water molecule remained in the QM orientation relative to the methanol or NMA molecule.

Crystal Simulations. The unit cells for the orthorhombic crystals of α-D-Glc*p* (GLUCSA10), α-D-Glc*p*-OMe (MGLUCP11), α-D-Man*p*-OMe (MEMANP), and *β*-D-Gal*p*-OMe (MBDGAL02) were each transformed using the $P2_12_12_1$ symmetry operators to generate crystal lattices containing 64 monosaccharides.[46−48] Similarly, the $P2_1$ symmetry operators were utilized to generate a 2 × 4 × 2 system for α-D-Glc*p* monohydrate (GLUCMH11) and a 4 × 4 × 4 system for α-D-Glc*p*NAc (ACGLUA11).[49,50] The experimental monosaccharide conformation was employed as the initial configuration. Subsequently, no position or symmetry restraints were applied, and the box dimensions were allowed to vary over the course of the MD simulations. Each system governed by the $P2_12_12_1$ symmetry operator was obtained via neutron diffraction allowing for direct determination of the proton positions, while the protons in the $P2_1$ structures were obtained by experimental crystal density difference experiments.

Table 1. Partial Atomic Charges (au) for α-D-Glc*p*

GLYCAM	2000a	06	06-LP
C1	0.151	0.509	0.292
C2	0.131	0.246	0.170
C3	0.211	0.286	0.109
C4	0.160	0.255	0.146
C5	0.085	0.283	0.227
C6	0.244	0.277	0.138
O1/LP	−0.612	−0.639	0.000/−0.220[a]
O2/LP	−0.632	−0.713	0.000/−0.218
O3/LP	−0.668	−0.699	0.000/−0.215
O4/LP	−0.665	−0.710	0.000/−0.214
O5/LP	−0.404	−0.574	0.000/−0.183
O6/LP	−0.671	−0.682	0.000/−0.210
H1	0.153	0.000	0.000
H2	0.103	0.000	0.000
H3	0.061	0.000	0.000
H4	0.081	0.000	0.000
H5	0.086	0.000	0.000
H6/H6′	0.031/0.031	0.000/0.000	0.000/0.000
HO1	0.432	0.445	0.327
HO2	0.415	0.437	0.281
HO3	0.430	0.427	0.280
HO4	0.429	0.436	0.278
HO6	0.417	0.418	0.272

[a] The partial charge on a single LP is listed; note there are 2 LPs per oxygen atom.

For each monosaccharide, ensemble-averaged charges were derived utilizing the same 100 conformers employed previously in the generation of ensemble averaged charges in the GLYCAM2000a parameter set.[31,51] Unlike GLYCAM-2000a, in GLYCAM06 aliphatic hydrogen atoms are excluded from the charge fitting.[30] Exclusion of the aliphatic hydrogen atoms during the charge fitting led to a robust partial charge set, with nearly identical values for the same atom types within the monosaccharide, Table 1.[52] A similar charge fitting scheme was utilized to determine corresponding partial charges for the ala_2 zwitterion.[53] The unit cell of the alanine dipeptide zwitterion contains 8 molecules, as determined by experimental crystal density difference diffraction, and was transformed to a 128 molecule cell utilizing the *I*4 symmetry operators (2 × 2 × 4). The experimental conformation was utilized in the initial configuration, which contained all hydrogen atoms except one methyl hydrogen atom. The missing methyl hydrogen atom was added based on the standard tetrahedral configuration of methyl groups.

Unit scaling factors for 1,4 nonbonded interactions were employed during the MD crystal simulations of carbohydrates, which is consistent with the GLYCAM2000a and GLYCAM06 parameter sets.[30,31] The Particle Mesh Ewald algorithm was implemented for treatment of long-range interactions as each system was heated from 5 to 300 K (experimental temperature range was 283−303 K) over 50 ps and subsequently maintained at 300 K for 1 ns utilizing the Berendsen temperature coupling scheme.[54,55] The effect of both isotropic and anisotropic pressure scaling was implemented to establish the effects these pressure models had on the lattice structure. The deviation from experiment was amplified with the anisotropic model, but the same

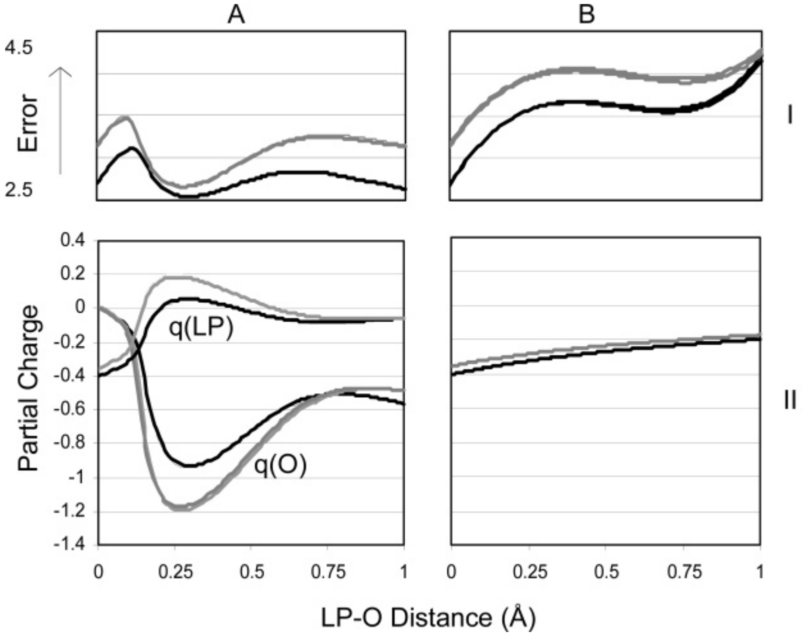


Figure 2. Correlation of the error function (I) and the partial charges (II) in scheme **A** (q(O) and q(LP)) and in scheme **B** (q(LP), q(O) = 0) from fitting to the B3LYP/6-31G (gray) and HF/6-31G(d) (black) MEPs for a water molecule.

relative trends in cell distortion were observed with both models. Due to program limitations, anisotropic scaling was not implemented in the crystals governed by $P2_1$ symmetry. Data from only the last 500 ps were utilized to determine the average cell dimensions and hydrogen-bond distances.[34,35,48]

Results and Discussion

Determination of the Optimal LP−O Distance. Initially, the effect of fitting to the electrostatic potential to determine the placement of LPs around oxygen was assessed by determining the quality of the fit between the classical and quantum mechanical MEPs in a water molecule as a function of lone-pair position. The two LP−O distances were scanned from 0.01 to 1.00 Å, in 0.01 Å increments, with both LP−O distances and the partial atomic charge on the lone-pairs constrained to be equal. Two different approaches were taken during the RESP charge fitting stage. In the first scheme (**A**) a charge on oxygen was permitted, while in the second scheme (**B**) the charge on oxygen was constrained to zero, as in the TIP5P water model. Scheme **A** resulted in a minimum at a short LP−O distance of approximately 0.3 Å but possessed unintuitive partial charges; the LP charge became increasingly positive as the error function was minimized. Application of scheme **B** led to a minimum at a longer LP−O distance of approximately 0.75 Å, with negatively charged lone-pairs of approximately −0.2 e, Figure 2. Notably, the LP description obtained from scheme B is very similar to that in the TIP5P water model, which has an LP−O distance of 0.70 Å, and q(O) = 0 e, and q(LP) = −0.24 e. Thus, the well-defined MEP fitting approach reproduces the electrostatic description of the TIP5P water model, despite the fact that the TIP5P water model was originally obtained via empirical fitting to bulk liquid properties.

Schemes **A** and **B** were applied to a series of test molecules, including alcohols and ethers, to determine the optimal LP−O descriptors for sp^3 type oxygen atoms, Table 2. The average LP−O distance at the best fit was 0.69 Å from the HF/6-31G(d) MEP, 0.72 Å from the B3LYP/aug-cc-pVTZ MEP, and 0.71 Å from the MP2/aug-cc-pVTZ MEP with scheme **A**. Similarly, application of scheme **B** led to average optimal LP−O distances of 0.71, 0.75, and 0.71 Å, respectively. Notably in scheme **A** only a small partial charge was observed on the oxygen atom in the alcohols and ethers. Further, in some cases this small charge was positive. Each of the LP−O distances derived thus far was based on the approximation that the LP−O distances are equivalent, which is consistent with previous studies and molecular symmetry.[15,26,33] When both LP−O distances were varied independently for each of the alcohol and ether model compounds, the optimal locations converged to the symmetric results, Figure 3. Therefore, the results for the sp^3 type oxygen atoms in water, alcohols, and ethers show that the most consistent and intuitive fit between the classical and quantum mechanical MEP was obtained when the partial charge on oxygen was constrained to zero (scheme **B**) and the LPs should have equivalent negative charges, which are determined directly from the RESP-fitting.

In order to extend this model to sp^2-type oxygen atoms, a test set consisting of ketones and amides was compiled. For the ketones, the fit resulted in a very flat error function surface, and, at longer LP−O distances, the fit deteriorated and the partial charges at the lone-pair sites became positive. An optimal fit was maintained at shorter LP−O distances, of less than 0.5 Å, with a shallow minimum at approximately 0.3 Å. In addition, at shorter LP−O distances the charge on oxygen was nearly zero, even though it was allowed to vary, and the lone-pairs maintained a negative charge in scheme **A**. Therefore, the optimal lone-pair placement for ketones

Table 2. LP−O Distance, q(O), and q(LP) at the Best Fit between the Classical and ab Initio MEP for Each Molecule in the sp^3 Oxygen Atom Test Set at the HF/6-31G(d) (I), B3LYP/aug-cc-pVTZ (II), and MP2/aug-cc-pVTZ (III) Levels

	q(O) ≠ 0 (scheme **A**)			q(O) = 0 (scheme **B**)	
	LP−O	q(O)	q(LP)	LP−O	q(LP)
methanol					
I	*0.70*[a]	−0.029	−0.178	0.80	−0.177
II	*0.70*	0.017	−0.169	0.88	−0.149
III	0.89	−0.046	−0.146	0.84	−0.163
ethanol					
I	0.65	−0.047	−0.195	0.57	−0.221
II	0.74	−0.043	−0.168	0.68	−0.188
III	0.71	−0.048	−0.179	0.64	−0.202
2-butanol					
I	0.53	0.008	−0.240	0.53	−0.237
II	0.58	−0.005	−0.205	0.56	−0.209
III	0.53	0.001	−0.228	0.52	−0.229
2-propanol					
I	*0.70*	−0.097	−0.169	0.70	−0.199
II	*0.70*	−0.074	−0.157	0.77	−0.173
III	*0.70*[a]	−0.094	−0.163	0.72	−0.190
3-pentanol					
I	0.59	0.031	−0.232	0.61	−0.218
II	0.65	0.013	−0.197	0.66	−0.191
III	0.60	0.014	−0.216	0.60	−0.213
t-butanol					
I	*0.70*	−0.080	−0.173	0.76	−0.192
II	*0.70*	−0.076	−0.160	0.80	−0.174
III	*0.70*	−0.093	−0.170	0.77	−0.192
dimethyl ether					
I	0.92	−0.011	−0.105	0.91	−0.115
II	0.90	0.110	−0.129	0.70	−0.106
III	0.81	0.183	−0.168	0.70	−0.116
ethyl methyl ether					
I	0.70	0.052	−0.150	0.80	−0.126
II	0.79	0.158	−0.177	0.88	−0.109
III	0.74	0.188	−0.188	0.91	−0.114
methyl isopropyl ether					
I	0.65	0.047	−0.163	0.67	−0.145
II	0.72	0.097	−0.167	0.79	−0.131
diisopropyl ether					
I	0.73	0.041	−0.160	0.74	−0.147
II	0.75	0.049	−0.158	0.76	−0.143

[a] Italics indicate systems in which the resolution was not fine enough to detect a precise minimum, so the partial charges at an LP−O distance of 0.70 Å are shown.

was found to be 0.3 Å from the oxygen atom and the partial charge on the oxygen set to zero.

In contrast to all the model compounds investigated thus far, inclusion of symmetric lone-pairs around the carbonyl oxygen of the amide group slightly deteriorated the fit to the MEP containing molecules. Allowing each lone-pair to adopt a unique distance from the carbonyl oxygen atom improved the fit for the amides, with the best fit at a combination of a longer (0.7 Å) LP1−O separation and a shorter (0.3 Å) LP2−O separation, Figure 4. The shorter LP2−O distance is located on the side of the carbonyl containing the amide nitrogen.

The charge model for anionic carboxylate groups, such as those in aspartate and glutamate, did not incorporate lone-pairs in the PARM02EP parameter set.[15] In order to examine the validity of this approximation several different lone-pair placements were investigated for acetate, propanoic acid anion, 2-methylbutanoic acid anion, and 2-methylpropanoic anion. Initially, the carboxylate group was treated as containing two ketone oygen atoms, and the standard sp^2 geometrical placement was applied, the LP−O−LP angle of 120° as in Figure 5 (I). The two lone-pairs located between the oxygen atoms in the carboxylate group (LP$_{\text{inner}}$) and the remaining two lone-pairs (LP$_{\text{outer}}$) were constrained to have equivalent partial charges and separation from their respective oxygen atoms to maintain molecular symmetry. Second, the two LP$_{\text{inner}}$ partial charge sites were replaced by a single partial charge, Figure 6(II).

Third, removal of both LP$_{\text{outer}}$ sites from the second model, leaving only a single LP bisecting the O−C−O angle was investigated, Figure 5(III), which would theoretically help to alleviate the imbalance observed during MD simulations. Commonly, a hydrogen bond formed with a carboxylate entity is bifurcated in nature, but during an MD simulation at 300 K the hydrogen bond donor typically associates with a single oxygen atom at a given point in time. Overall, the majority of the negative charge is located on the two oxygen atoms, while the charge on LP$_{\text{inner}}$ is only slightly negative, ranging only to −0.03 e for the HF and B3LYP levels, and becomes positive for all LP$_{\text{inner}}$−C distances less than 2 Å at the MP2 level for this model, Figures 5 and 6(III). Since the inclusion of LPs in anionic carboxylate groups did not lead to an improvement in the fit between MEPs, nor to a minimum in the error function, lone-pairs were not included in carboxylate groups.

These atom-type specific lone-pair placements for oxygen will be applied to GLYCAM06, and ensemble averaged partial charges will be derived for each pyranoside to yield GLYCAM06-LP.

Analysis of Hydrogen-Bonded Clusters. In order to assess the accuracy of the GLYCAM06-LP model, the geometry of several hydrogen-bonded clusters were examined in vacuo. Each cluster contained at least one water molecule, which, if no LPs were present, was modeled as TIP3P or, when the new LP model was employed, as TIP5P. The B3LYP/6-31+G(d) optimized geometry was used as the starting point for each cluster geometry force field energy minimization.[45] The root-mean-square deviations (RMSDs) for both the heavy atoms and for all atoms in each neutral cluster were determined between the energy minimized molecular mechanics structure and the quantum mechanically optimized geometry. Overall, for the 21 clusters examined, the RMSDs for the non-LP containing model and the new LP model were 0.58 Å (0.30 Å) and 0.49 Å (0.21 Å), respectively (heavy-atom RMSDs shown in parentheses).

Examination of the relative energies for the approach of water to a hydrogen acceptor or donor-containing molecule provides a useful test of the electrostatic model and illustrates the applicability of the chosen water model. For the methanol−water cluster, the approach in which methanol is the hydrogen bond donor (MdW) and the approach in which

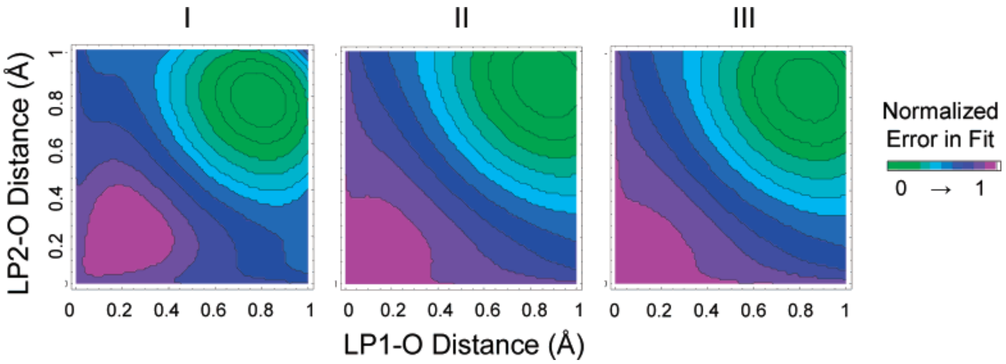


Figure 3. Variation of χ with respect to the LP−O distance at the HF/6-31G(d) (I), B3LYP/aug-cc-pVTZ (II), and MP2/aug-cc-pVTZ (III) levels with the partial charge on oxygen set to zero for methanol.

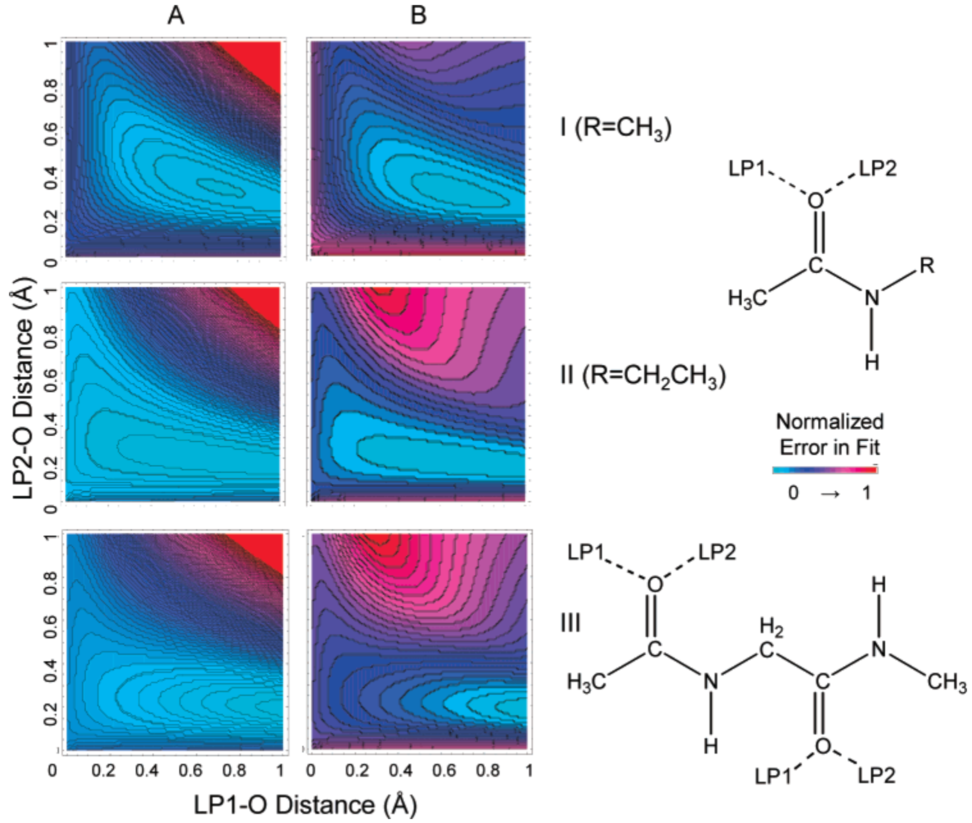


Figure 4. Variation of χ with respect to the LP−O distance (Å) in *N*-methyl acetamide (I), and *N*-ethyl acetamide (II), and the glycine dipeptide (III). The partial charge on oxygen was freely determined in scheme **A**, while the oxygen partial charge was set to zero during the RESP fit in scheme **B**.

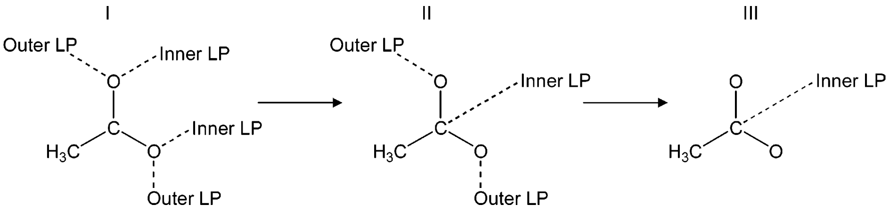


Figure 5. LP placement for anionic carboxylate groups.

water is the hydrogen bond donor (WdM) were compared at the HF/6-31G(d) level, which is the same level at which the MEPs were determined, Figure 7. At this level of theory there is no difference for the two configurations.[44] At the global minima, determined at the MP2/aug-cc-pVQZ level, the interaction energy for the WdM dimer (−5.72 kcal•mol^{-1}) was more favorable than for MdW (−4.95 kcal•mol^{-1}).[44] Although both the non-LP and LP models reversed the relative ranking of the two configurations, the approach trajectory of the LP model was energetically much closer than the non-LP model. In both cases the energy difference was less than 1 kcal•mol^{-1}, specifically, 0.6 and 0.3 kcal•mol^{-1} for the non-LP and LP models, respectively. Further work, such as the inclusion of a van der Waals term

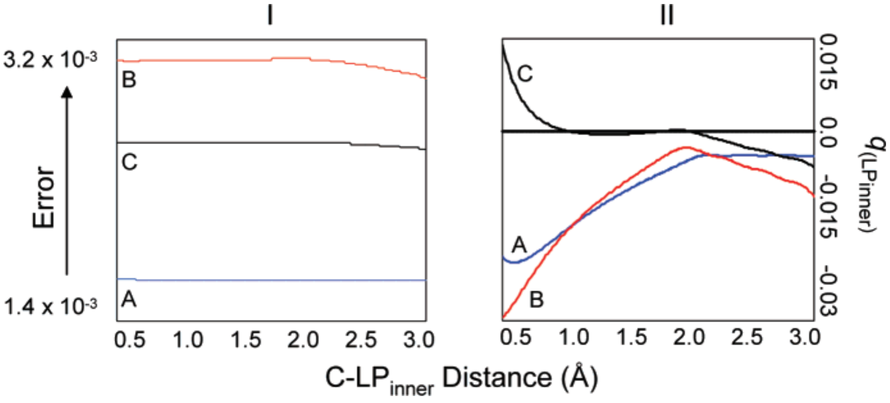


Figure 6. Variation of χ (I) and charge on LP_{inner} (II) with respect to the LP_{inner}−C distance when the QM level was HF/6-31G(d) (A, blue), B3LYP/aug-cc-PVTZ (B, red), or MP2/aug-cc-PVTZ (C, black) for acetate for charge arrangement III in Figure 5.

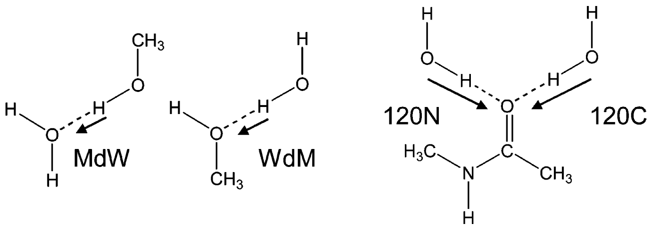


Figure 7. Approach of H_2O to the symmetrical sp^3-type lone-pair hydroxyl group in methanol (left) and the approach of H_2O from the each of the asymmetric sp^2-type lone-pair axes of NMA (right).

on the hydroxyl hydrogen may be necessary to reproduce the high-level quantum results. For example, inclusion of a van der Waals parameter on the hydrogen atoms increased the ability of the TIP3P water model to reproduce bulk water properties when the particle mesh ewald algorithm was employed.[56]

The comparison of the relative energies associated with the approach of a water molecule along each of the LP−O axes in *N*-methyl acetamide ensures that the asymmetric lone-pair model degrades neither the geometry nor the energetics of the hydrogen-bonded cluster. Approach along the LP−O axes from the methyl side is more favorable than the approach from the *N*-methyl side (120N), by 1.3 kcal·mol^{-1} at the B3LYP/6-31+G(d, p) level. Starting from the quantum mechanically determined structures, energy minimization utilizing the non-LP containing model with TIP3P reverses the ranking obtained with density functional theory, while inclusion of LPs and the TIP5P water model reproduces the quantum mechanically determined ranking, Figure 7.

MD Crystal Simulations. MD simulations of monosaccharide crystals provide a sensitive method for testing nonbonded parameters in the condensed phase and have been performed previously for carbohydrates.[34,35] If the nonbonded parameters are too attractive, then it leads to a more tightly packed crystal, while if the interactions are underestimated, then the crystal cell will expand. The addition of lone-pairs is purely a variation in the electrostatic model, and monitoring the behavior of cell lengths over the course of MD simulations, for both the non-LP and LP models, provides a sound test for the effects of lone-pair inclusion.

Initially, for each of three methyl glycosides, α-D-Glc*p*-OMe, *β*-D-Gal*p*-OMe, and α-D-Man*p*-OMe, a 64 and a 256 molecular unit system was generated. Two different size systems were employed with the GLYCAM06 parameter set to determine if the smaller system size introduced any artifacts in the distortion of the cell dimensions. The average difference in cell lengths for the three methyl glycosides in the large (256 molecules) cell relative to the 64 molecule cell was 0.01%. Nonbonded cutoffs of 8, 9, and 10 Å were investigated within the 256 molecule cells, yielding less than a 0.02 and 0.03% difference for the 9 and 10 Å cutoff, respectively, relative to the 8 Å cutoff for the three methyl glycosides. A time step of 0.5 fs was implemented as a standard to determine if longer time steps would be appropriate. Due to the negligible difference in the results obtained with a time step of 1 fs relative to the 0.5 fs initial time step, the larger time step was employed along with an 8 Å cutoff and the smaller lattice size.

The majority of the improvement observed from utilization of the GLYCAM2000a to the GLYCAM06 and 06-LP force fields can be accounted for mainly by the change in nonbonded parameters for the hydroxyl oxygen atom. The radius of the hydroxyl oxygen atom in GLYCAM2000a (1.961 Å) is larger than the OPLS value of 1.7210 Å. The OPLS van der Waals radii are implemented in the AMBER parameter sets as well as GLYCAM06 and here in GLYCAM06-LP.[22,31] The larger van der Waals radius as well as the smaller well depth (0.14 versus 0.21 × 10^{-3} kcal·mol^{-1}) both contributed to the expansion and elongation of hydrogen bonds observed in GLYCAM2000a versus 06 and 06-LP, Tables 3 and 4.

The different partial charge arrangement in each of these force fields would be expected to have a relatively small effect on the overall change in cell dimensions since each charge set was derived from the same set of conformers for the ensemble with fitting to the HF/6-31G(d) MEP and RESP weighting of 0.01. In order to observe the sensitivity purely due to electrostatics, the OPLS hydroxyl oxygen atom van der Waals parameters were implemented in GLYCAM2000a to yield GLYCAM2000b, Table 3. Again, GLYCAM06-LP yields the smallest average unit-cell deviation and reduces the deviation by over 20% from that reported for the GROMOS and HGFB force fields, Table 3.[34] Despite being fit to the same electrostatic potential, the addition of lone-pairs improves the hydrogen bonding interactions over the

Table 3. Effect of Different Force Field Parameters in GLYCAM and Pressure Scaling on the Crystallographic Cell Dimensions for α-D-Glc*p* and the Methyl Glycosides of α-D-Glc*p*, *β*-D-Gal*p*, and α-D-Man*p*

			cell dimensions (Å)			deviation (%)			
		pressure scaling[a]	*A*	*B*	*C*	Δ*A*	Δ*B*	Δ*C*	mean |Δ|
α-D-Gl*cp*									
	2000a	a	21.25	29.86	23.73	4.3	0.5	19.2	8.0
		i	22.14	32.28	21.63				8.7
	2000b	a	20.74	29.70	20.57	1.8	0.0	3.4	1.7
		i	20.68	30.16	20.21				1.5
	06	a	21.04	30.14	19.65	3.3	1.5	−1.3	2.0
		i	20.65	30.10	20.17				1.4
	06-LP	a	20.80	30.03	19.76	2.1	1.1	−0.7	1.3
		i	20.57	29.99	20.10				1.0
	GROMOS[34]	a				−0.8	−3.0	−0.9	1.6
	CHARMMHGFB[34]	a				−3.9	0.8	−0.6	1.8
expt[48]			20.37	29.70	19.90				
α-D-Gl*cp*-OMe									
	2000a	a	24.73[b]	32.48	20.00	9.3	9.9	−5.3	8.2
		i	23.76	31.05	22.19				5.0
	06	a	22.91	21.19	20.55	1.3	1.6	−2.7	1.8
		i	22.69	29.65	30.02				0.3
	06-LP	a	22.70	30.40	20.25	0.3	2.8	−4.1	2.4
		i	22.64	29.58	21.14				0.1
expt[47]			22.62	29.56	21.12				
β-D-Ga*lp*-OMe									
	2000a	a	33.95	16.52	28.96	9.1	−3.2	10.3	7.5
		i	32.94	18.07	27.80				5.8
	06	a	32.05	16.87	26.09	3.0	−1.1	−0.6	1.6
		i	31.30	17.17	26.42				0.6
	06-LP	a	31.89	16.69	26.26	2.5	−2.2	0.0	1.6
		i	31.21	17.12	26.34				0.3
expt[46]			31.12	17.07	26.26				
α-D-Man*p*-OMe									
	2000a	a	19.81	38.91	21.78	5.1	4.4	8.3	5.9
		i	21.34	39.54	21.34				6.1
	06	a	19.28	37.22	20.17	2.2	−0.1	0.3	0.9
		i	18.99	37.52	20.25				0.7
	06-LP	a	19.16	37.03	20.07	1.6	−0.6	−0.2	0.8
		i	18.89	37.32	20.14				0.2
expt[47]			18.86	37.26	20.11				

[a] Isotropic, i; anisotropic, a. [b] Standard deviations were all within 0.15% of the mean.

more elongated hydrogen bonds found in the models sans lone-pairs. Within each model, there is a weak correlation between the magnitude of the standard deviation and the difference between the calculated and experimental hydrogen bonds, with the smallest deviation of 0.11 Å corresponding to simulated hydrogen bonds that are on average within 0.1 Å of the experimental value. This is a reassuring occurrence in all the models, that they do not adopt the incorrect minimum but are fluctuating over several low energy states. Overall, GLYCAM06-LP has the lowest standard deviation and yields the best reproduction of the hydrogen-bonding environment in the crystal.

Previous studies have examined the ability of carbohydrate force fields to reproduce solvent−solute properties, revealing that GLYCAM2000a underestimates the hydrogen bonding interaction between the hydroxyl groups in the pyranoside and the TIP3P water model.[57] Simulation of the monohydrate glycoside crystal structure required the inclusion of a water molecule, which directly assessed the compatibility of the carbohydrate force field with the chosen solvent model. Here, the TIP series of models was implemented, with the smallest deviation from experiment observed when GLYCAM06-LP was utilized with the TIP5P water model, Table 5. Notably, GLYCAM2000b and GLYCAM06 yield the smallest distortion of the cell dimensions when the TIP5P model is implemented, with TIP3P being the worst, although still in good agreement with the experimental values. Therefore, it is clear not only that under these conditions TIP5P performs better than TIP3P and TIP4P but also that the inclusion of lone-pairs in the carbohydrate force field results in a substantial improvement in the model as well.

In order to assess the performance of the asymmetric LP model, crystal-phase MD simulations were performed on a small peptide, the Ala−Ala (ala_2) zwitterion shown in Figure 8. The *I*4 space group, inherent to the crystal structure of the ala_2 zwitterion, allows MD simulation with both isotropic

Table 4. Effect of Different Force Field Parameters and Pressure Scaling on Selected Intermolecular Hydrogen Bonds (Å) in α-D-Glc*p*, *β*-D-Gal*p*, and α-D-Man*p*

			interatomic distances		
α-D-Glc*p*	force field	pressure scaling[a]	O2···H3O	O3···H6O	O6···H2O
	GLYCAM 2000a	a	4.05 ± 0.31	2.44 ± 0.36	2.36 ± 0.23
		i	4.08 ± 0.39	2.56 ± 0.65	2.33 ± 0.25
	GLYCAM 06	a	2.13 ± 0.29	1.91 ± 0.18	1.79 ± 0.12
		i	2.05 ± 0.25	1.88 ± 0.16	1.79 ± 0.11
	GLYCAM 06-LP	a	1.97 ± 0.22	1.87 ± 0.14	1.79 ± 0.11
		i	1.91 ± 0.18	1.86 ± 0.14	1.77 ± 0.11
expt[47]			1.770	1.772	1.738
β-D-Gal*p*	force field	pressure scaling[a]	O2····H6O	O6···H4O	O4···H2O
	GLYCAM 2000a	a	4.03 ± 0.87	3.06 ± 0.68	2.39 ± 0.24
		i	4.38 ± 1.33	3.90 ± 1.34	2.98 ± 0.76
	GLYCAM 06	a	3.07 ± 0.47	1.84 ± 0.13	2.03 ± 0.20
		i	2.81 ± 0.72	1.83 ± 0.13	1.97 ± 0.19
	GLYCAM 06-LP	a	2.40 ± 0.44	1.84 ± 0.12	1.98 ± 0.17
		i	2.03 ± 0.27	1.85 ± 0.12	1.89 ± 0.14
expt[46]			1.860	1.739	1.773
α-D-Man*p*	force field	pressure scaling[a]	O3···H6O	O4···H3O	O5···H4O
	GLYCAM 2000a	a	2.32 ± 0.21	2.19 ± 0.17	2.41 ± 0.25
		i	2.35 ± 0.22	2.16 ± 0.15	2.39 ± 0.24
	GLYCAM 06	a	1.92 ± 0.17	1.82 ± 0.11	2.10 ± 0.24
		i	1.90 ± 0.16	1.82 ± 0.12	2.19 ± 0.26
	GLYCAM 06-LP	a	2.00 ± 0.21	1.85 ± 0.12	2.07 ± 0.20
		i	1.96 ± 0.19	1.86 ± 0.13	2.15 ± 0.21
expt[47]			1.917	1.810	2.052

and anisotropic pressure scaling. The deviation of the crystal cell dimensions was similar with all models if isotropic scaling was employed, Table 6. Notably, the asymmetric LP model presented here (LP) yielded the lowest deviation, while the 02EP model was the next best model. Inclusion of polarization into the PARM99 parameter set, PARM02, actually resulted in a worse reproduction of the crystal unit cell dimensions. All models implemented here reproduce the interresidue hydrogen bond distances and angles to within 5% of the experimental values, with PARM02EP yielding the closest agreement, with ≤0.01 Å average deviation in the hydrogen bond distances. Anisotropic pressure scaling on the ala_2 zwitterion had only a subtle, worsening effect on the reproduction of the experimental cell dimensions and hydrogen bond distances, which is in contrast to the large effect anisotropic scaling had on the crystal simulations of monosaccharides.

Inclusion of lone-pairs into a pre-existing molecular mechanics force field may require subsequent refitting of the torsion terms. The partial atomic charges in GLYCAM06-LP were fit to the same MEP as in GLYCAM06, minimizing the impact of the LP-model on existing rotation potentials. This is illustrated clearly for the rotational profiles of the C−O−C−O torsion angle, which is common to all oligosaccharides, and exemplified by axial and equatorial tetrahydro-2-methoxy-2H-pyran, corresponding to α- and *β*-linkages, respectively, Figure 9. Here, addition of LPs does not have a substantial impact, and the shape of torsional energy curves retains the original optimized shape obtained without LPs.

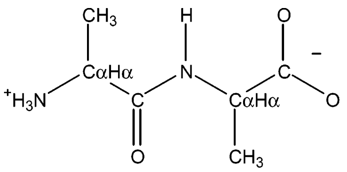


Figure 8. Structure of the Ala−Ala (ala_2) zwitterion.

A highly sensitive measure of the balance between internal rotational energies and external solvent influences is provided by the rotamer population distribution for the exocyclic C5−C6 bond in hexopyranoses. In order to examine the robustness of this LP model, a 10 ns condensed phase MD simulations was performed in conjunction with the TIP5P.[59] Rotamers of the primary alcohol group are populated to varying extents in different monosaccharides, as determined by NMR spectroscopy. The three different rotamers that are populated are defined by the gauche and trans orientation of both the O5−C5−O6−C6 and C4−C5−O6−C6 angles, respectively. All three rotamers are populated for α-D-Gal*p*OMe, with experimentally determined populations *gg*:*gt*:*tg* of 14:47:39,[60] 16:75:9,[61] and 21:61:18,[62] and are reproduced with GLYCAM06 (100 ns) at 8:75:18[30] and here with GLYCAM06-LP at 13:81:6. When 10 ns is too short to ensure statistical convergence for this rotation, longer

Table 5. Effect of Force Field and Water Model on the Crystallographic Unit Cell Parameters and Intermolecular Hydrogen Bond Geometries of α-D-Glc*p*·H_2O

		water model		
force field GLYCAM2000a	expt[49]	TIP3P	TIP4P-EW[58]	TIP5P-EW[59]
A^a	17.61	18.90 ± 0.03	18.76 ± 0.02	18.73 ± 0.02
B^a	20.34	21.83 ± 0.04	21.68 ± 0.02	21.64 ± 0.02
C^a	19.42	20.84 ± 0.04	20.69 ± 0.02	20.66 ± 0.02
ΔA, B, and C (%)^b		7.33	6.57	6.39
mean $\|\Delta\|$ HB_{dist} (Å)		0.49 ± 0.18	0.36 ± 0.14	4.17 ± 0.73
mean $\|\Delta\%\|$ HB_{dist} (%)		23.68	16.97	203.46
mean $\|\Delta\|$ HB_{angle} (°)		72 ± 34	15 ± 17	29 ± 25
mean $\|\Delta\%\|$ HB_{angle}		44.25	9.54	16.50

		water model		
force field GLYCAM2000b	expt[49]	TIP3P	TIP4P-EW[58]	TIP5P-EW[59]
A		17.77 ± 0.01	17.74 ± 0.01	17.66 ± 0.01
B		20.54 ± 0.01	20.49 ± 0.01	20.40 ± 0.01
C		19.60 ± 0.01	19.56 ± 0.01	19.47 ± 0.01
ΔA, B, and C		0.96	0.75	0.29
mean $\|\Delta\|$ HB_{dist}		0.17 ± 0.09	0.19 ± 0.06	0.20 ± 0.11
mean $\|\Delta\|$ HB_{dist}		7.78	8.38	9.17
mean $\|\Delta\|$ HB_{angle}		8 ± 12	6 ± 10	7 ± 12
mean $\|\Delta\|$ HB_{angle}		4.70	3.78	4.68

		water model		
force field GLYCAM06	expt[49]	TIP3P	TIP4P-EW[58]	TIP5P-EW[59]
A		17.75 ± 0.01	17.72 ± 0.01	17.62 ± 0.01
B		20.50 ± 0.01	20.48 ± 0.01	20.36 ± 0.01
C		19.57 ± 0.01	19.55 ± 0.01	19.44 ± 0.01
ΔA, B, and C		0.81	0.67	0.11
mean $\|\Delta\|$ HB_{dist}		0.18 ± 0.06	0.17 ± 0.07	0.16 ± 0.08
mean $\|\Delta\|$ HB_{dist}		8.41	8.00	7.55
mean $\|\Delta\|$ HB_{angle}		7 ± 11	6 ± 10	6 ± 10
mean $\|\Delta\|$ HB_{angle}		4.42	3.81	3.56

		water model		
force field GLYCAM06-LP	expt[49]	TIP3P	TIP4P-EW[58]	TIP5P-EW[59]
A		17.62 ± 0.01	17.67 ± 0.01	17.61 ± 0.01
B		20.36 ± 0.01	20.42 ± 0.01	20.35 ± 0.01
C		19.44 ± 0.01	19.49 ± 0.01	19.42 ± 0.01
ΔA, B, and C		0.11	0.39	0.02
mean $\|\Delta\|$ HB_{dist}		0.19 ± 0.07	0.17 ± 0.09	0.16 ± 0.09
mean $\|\Delta\|$ HB_{dist}		8.68	8.06	7.53
mean $\|\Delta\|$ HB_{angle}		7 ± 12	7 ± 11	7 ± 11
mean $\|\Delta\|$ HB_{angle}		4.60	4.13	4.56

[a] In Å. [b] Isotropic pressure scaling.

simulations would be required in order to determine whether this torsion term should be refit. The presence of lone-pairs on both solvent and solute is likely to be particularly influential in modeling dynamic processes, such as conformational lifetimes, bound water occupancies, diffusion rates, and autocorrelation times. It is hoped that the present model will be useful in providing further insight into these phenomena.

Conclusion

Utilizing the quantum mechanical MEP to determine the distance of the lone-pairs from the respective oxygen atoms leads to a description of the molecular electrostatics that is consistent with the currently available TIP5P water model for the hydroxyl and ether type oxygen atoms. The aforementioned sp^3-type oxygen atoms each have a LP−O distance of 0.7 Å and a charge of zero on the oxygen atom.

Table 6. Effect of Force Field Parameter Set on the Crystallographic Unit Cell Dimensions (Å) for the ala$_2$ Zwitterion[53]

	isotropic scaling[a]						
AMBER parameter set	*A*	*B*	*C*	Δ*A*	Δ*B*	Δ*C*	mean |Δ|
expt	35.97	35.97	20.62				
94	35.47	35.47	20.33	−0.50	−0.50	−0.28	0.43
96	35.50	35.50	20.34	−0.47	−0.47	−0.27	0.41
99	35.52	35.52	20.36	−0.45	−0.45	−0.26	0.39
02	35.44	35.44	20.31	−0.53	−0.53	−0.30	0.45
02EP	35.59	35.59	20.40	−0.38	−0.38	−0.22	0.33
GLYCAM06-LP	36.22	36.22	20.77	0.25	0.25	0.15	0.22

	anisotropic scaling[a]						
AMBER parameter set	*A*	*B*	*C*	Δ*A*	Δ*B*	Δ*C*	mean |Δ|
94	35.63	35.63	20.08	−0.34	−0.34	−0.54	0.41
96	35.64	35.64	20.12	−0.33	−0.33	−0.49	0.38
99	35.64	35.64	20.17	−0.33	−0.33	−0.45	0.38
02	36.17	36.16	19.24	0.20	0.19	−1.38	0.59
02EP	37.79	35.10	19.54	1.82	−0.87	−1.07	1.25
GLYCAM06-LP	35.89	35.89	21.23	−0.08	−0.08	0.61	0.26

[a] All simulations with isotropic pressure scaling resulted in standard deviations of 0.01 Å in cell dimensions, while the maximum increase observed with anisotropic scaling was 0.05 Å.

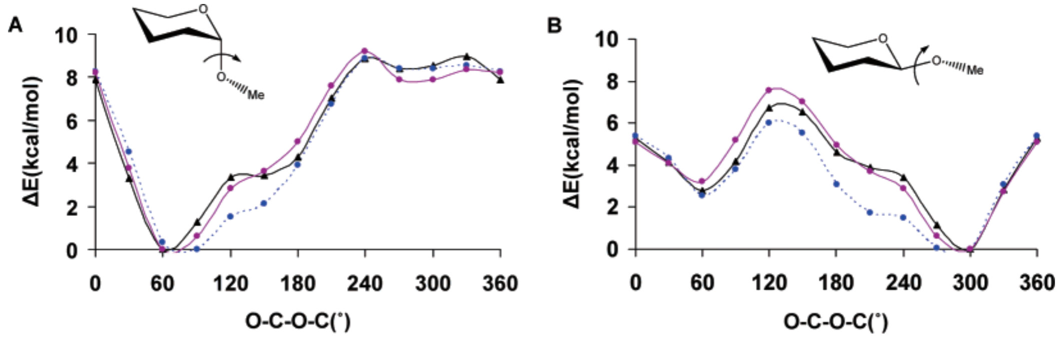


Figure 9. Rotation around the O−C−O−C angle in axial (A) and equatorial (B) tetrahydro-2-methoxy-2*H*-pyran determined at the B3LYP/6-31++G(2d,2p) level (black triangles), GLYCAM06 (purple circles), and single point calculations with GLYCAM06-LP (dashed line, blue circles).

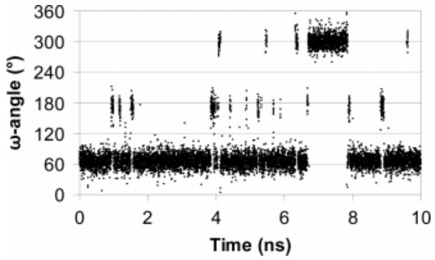


Figure 10. Rotation around the O5−C5−C6−O6 ω-angle in α-D-Gal*p*OMe.

A shorter LP−O distance of 0.3 Å for ketones and an asymmetric LP arrangement, with LP1−O and LP2−O distances of 0.7 and 0.3 Å, for amides yielded the most optimal fit between the quantum mechanical and classical MEP. For each fitting, constraining the oxygen as well as the aliphatic hydrogen atoms to zero charge led to a robust partial charge set with very similar charges for the various atom types in similar environments, i.e., each lone-pair in a secondary alcohol group of α-D-Glc*p* has a partial charge of −0.2 *e*. Inclusion of the new lone-pair model with TIP5P consistently increased the accuracy for MM energy minimized geometries of complexes over those of the non-LP containing model with TIP3P. In addition, the crystal MD simulations clearly illustrate the improved reproduction of electrostatic interactions when LPs are included. With respect to future applications, including LPs may be advantageous in the examination of ligand−receptor complexes, in which water molecules mediate the hydrogen bonding interactions between the ligand and receptor.[63] In addition, it is expected that GLYCAM06-LP will display improved bulk properties such as diffusion rates, rotational correlation times, and radial distribution functions. Extension to a polarizable lone-pair model is currently underway.

Acknowledgment. The authors thank the National Institutes of Health (RR05357 and GM55230) for funding.

References

(1) Leach, A. R. *Molecular Modelling principles and applications*; Addison Wesley Longman Limited: Essex, 1996.

(2) Smit, P. H.; Derissen, J. L.; van Duijneveldt, F. B. *Mol. Phys.* **1979**, *37*, 521.

(3) Scrocco, E.; Tomasi, J. Electronic molecular structure, reactivity and intermolecular forces: An Euristic interpretation by means of electrostatic molecular potentials. In *Advances in Quantum Chemistry;* Löwdin, P., Ed.; 1978; Vol. 11, pp 116−193.

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

(5) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833−1840.

(6) Besler, B. H.; Mertz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431−439.

(7) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361−373.

(8) Chirlian, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *8*, 894−905.

(9) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129−145.

(10) Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H., Jr. *J. Comput. Chem.* **1990**, *11*, 297−310.

(11) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D. M.; Seibel, G. L.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(12) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

(13) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230−252.

(14) 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.

(15) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632−1646.

(16) 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. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(17) Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H. X.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 994−*994.*

(18) Malcolm, N. O. J.; Popelier, P. L. A. *Faraday Discuss.* **2003**, *124*, 353−363.

(19) Bader, R. F. W.; Matta, C. F. *J. Phys. Chem. A* **2004**, *108*, 8385−8394.

(20) Popelier, P. L. A. *Chem. Phys. Lett.* **1994**, *228*, 160−164.

(21) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587−1597.

(22) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638−6646.

(23) Mackerell, A. D.; Wiorkiewiczkuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.

(24) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910−8922.

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

(26) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(27) Imperiali, B. *Acc. Chem. Res.* **1997**, *30*, 452−459.

(28) Coltart, D. M.; Royyuru, A. K.; Williams, L. J.; Glunz, P. W.; Sames, D.; kuduk, S. D.; Schwarz, J. B.; Chen, X.-T.; Danishefsky, S. J.; Live, D. H. *J. Am. Chem. Soc.* **2002**, *124*, 9833 -9844.

(29) Woods, R. J.; Dwek, R. A.; Edge, C. J. *J. Phys. Chem.* **1995**, *99*, 3832−3846.

(30) Kirschner, K. N.; Yongye, A.; Tschampel, S. M.; Daniels, C.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2007**, Submitted for publication.

(31) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541−10545.

(32) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620−9631.

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

(34) Kouwijzer, M. L. C. E.; van Eijck, B. P.; Kroes, S. J.; Kroon, J. *J. Comput. Chem.* **1993**, *14*, 1281−1289.

(35) Woods, R. J.; Chapelle, R. J. *J. Mol. Struct.* **2000**, *527*, 149−156.

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, (Revision A.9)*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(37) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200−206.

(38) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785−789.

(39) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(40) Roothan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69−89.

(41) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358−1371.

(42) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618−622.

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

(44) Kirschner, K. N.; Woods, R. J. *J. Phys. Chem. A* **2001**, *105*, 4150−4155.

(45) Tschampel, S. M.; Woods, R. J. *J. Phys. Chem. A* **2003**, *107*, 9175−9181.

(46) Takagi, S.; Jeffrey, G. A. *Acta Crystallogr., Sect. B: Struct. Sci.* **1979**, *35*, 902−906.

(47) Jeffrey, G. A.; McMullan, R. K.; Takagi, S. *Acta Crystallogr., Sect. B: Struct. Sci.* **1977**, *33*, 728−737.

(48) Brown, G. M.; Levy, H. A. *Acta Crystallogr., Sect. B: Struct. Sci.* **1979**, *35*, 656−659.

(49) Hough, E.; Neidle, S.; Rogers, D.; Troughton, P. G. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **1973**, *B 29*, 365−367.

(50) Mo, F.; Jensen, L. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **1975**, *31*, 2867−2873.

(51) Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22*, 1125−1137.

(52) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367−383.

(53) Fletterick, R.; Tsai, C. C.; Hughes, R. E. *J. Phys. Chem.* **1971**, *75*, 918-&.

(54) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(55) Darden, T. A.; York, D.; Pederson, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(56) Price, D. J.; Brooks, C. L. *J. Chem. Phys.* **2004**, *121*, 10096−10103.

(57) Corzana, F.; Motawia, M. S.; Du Penhoat, C. H.; Perez, S.; Tschampel, S. M.; Woods, R. J.; Engelsen, S. B. *J. Comput. Chem.* **2004**, *25*, 573−586.

(58) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665−9678.

(59) Rick, S. W. *J. Chem. Phys.* **2004**, *120*, 6085−6093.

(60) Bock, K.; Duus, J. O. *J. Carbohydr. Chem.* **1994**, *13*, 513−543.

(61) Nishida, Y.; Ohrui, H.; Meguro, H. *Tetrahedron Lett.* **1984**, *25*, 1575−1578.

(62) Nishida, Y.; Hori, H.; Ohrui, H.; Meguro, H. *J. Carbohydr. Chem.* **1988**, *7*, 239−250.

(63) Loris, R.; Stas, P. P. G.; Wyns, L. *J. Biol. Chem.* **1994**, *269*, 26722−26733.

CT700046J

JCTC Journal of Chemical Theory and Computation

Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity

Michael Feig*

Department of Biochemistry and Molecular Biology, and Department of Chemistry, Michigan State University, East Lansing, Michigan 48824-1319

Received March 20, 2007

Abstract: Kinetic properties of alanine dipeptide, the B1 domain of streptococcal protein G, and ubiquitin are compared between explicit solvent and implicit solvent simulations with the generalized Born molecular volume (GBMV) method. The results indicate that kinetics from explicit solvent simulations and experiments can be matched closely when the implicit solvent simulations are combined with Langevin dynamics and a friction coefficient near 10 ps^{-1}. Smaller and larger friction coefficients accelerate and slow down conformational sampling. It is found that local conformational exploration without the crossing of significant barriers can be accelerated by a factor of 4−5; however, the acceleration of barrier crossings is limited to about a factor of 2. The use of a Nosé−Hoover thermostat instead of Langevin dynamics greatly enhances local conformational sampling but slows down the crossing of barriers by at least an order of magnitude because of the lack of solute−solvent stochastic collisions.

Introduction

Computer simulations are often involved in the exploration of how biomolecular structure and dynamics give rise to biological function. Conventional simulation methods employ explicit representations of the solvent environment which can provide a high level of realism but usually at substantial computational costs. As an alternative, implicit solvent models have become increasingly popular in the simulation of biological macromolecules in order to be able to reach larger system sizes and longer time scales.[1,2] Implicit solvent models rely on the assumption that ensemble averages of instantaneous interactions between a solute and explicit solvent molecules may be approximated through a mean-field formalism.[3,4] Explicit solvent molecules can then be omitted from the system, thereby reducing the computational cost of such simulations because of smaller system sizes and the absence of solvent relaxation.

Physically motivated implicit solvent models often decompose the solvation free energy into electrostatic and nonpolar contributions.[5] The electrostatic component is commonly calculated on the basis of a continuum electrostatic model, where the solvent is represented as a homogeneous, high-dielectric medium that surrounds a low-dielectric solute cavity with explicit partial charges at the atomic centers of the solute.[6,7] The Poisson−Boltzmann (PB) equation rigorously describes such a model and can be solved numerically for the electrostatic potential throughout space.[2,8,9] The electrostatic component of the free energy of solvation is readily calculated from the electrostatic potential and the solute partial charges. Direct application of PB theory in biomolecular simulations is possible but hindered by the lack of sufficiently efficient and accurate numerical PB solvers and by difficulties in obtaining continuous first derivatives without altering the solute−solvent dielectric boundary.[2,10−13] Alternatively, generalized Born formalisms provide a convenient and efficient analytical approximation of electrostatic solvation free energies based on the same continuum electrostatic model described by Poisson theory.[14] When recent methodological advances are followed, the latest generation of generalized Born models can reproduce electrostatic solvation energies from Poisson theory accurately at a fraction of the cost that it would take to solve the Poisson equation.[15−21] The nonpolar contribution to the solvation free energy can be approximated with a term proportional to the solvent-accessible surface area (SASA).[5]

* Corresponding author. Phone: (517) 432-7439. Fax: (517) 353-9334. E-mail: feig@msu.edu.

10.1021/ct7000705 CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/14/2007

Implicit solvent models address the thermodynamic aspects of solvation but neglect hydrodynamic effects that become relevant in the simulation of kinetic processes. Particularly important are stochastic collisions with solvent molecules and frictional forces which directly impact kinetic rates and the magnitude of conformational fluctuations. Both of these contributions can be included with Langevin dynamics in combination with a given implicit solvent model.[22−24] In Langevin dynamics, a modified equation of motion is applied for a single particle i:

$$m_i\mathbf{a}_i(t) = \mathbf{F}_i(\mathbf{r}) - f_i m_i \mathbf{v}_i + \mathbf{F}_{\mathrm{random}}(t) \tag{1}$$

where m_i, $\mathbf{a}_i$, $\mathbf{v}_i$, and $\mathbf{F}_i$ are mass, acceleration, velocity, and the force due to the interactions with the rest of the system. f is the friction coefficient, and $\mathbf{F}_{\mathrm{random}}$ is a stochastic force simulating random collisions with solvent molecules.[22] Langevin dynamics effectively provide a thermostat with a temperature that is controlled through the magnitude of the random forces. It may be compared with other commonly employed thermostats,[24] in particular, the Nosé−Hoover thermostat[25,26] with the following equations of motion:

$$\dot{\mathbf{r}}_i = \mathbf{p}_i/m_i$$

$$\dot{\mathbf{p}}_i = -\nabla V(\mathbf{r}) - \zeta \mathbf{p}_i$$

$$\dot{\zeta} = \frac{1}{q}\left(\sum_i \frac{\mathbf{p}_i^2}{m_i} - gkT\right)$$

where $\mathbf{r}_\mathrm{i}$, $\mathbf{p}_\mathrm{i}$, and m_i are the position, momentum, and mass of particle i. V is the interaction potential according to a given force field; g is the number of degrees of freedom; k is the Boltzmann constant; and T is the temperature of the thermal bath. The extended variable ζ acts like a friction constant that is coupled to the temperature bath according to the coupling constant q. The Nosé−Hoover thermostat resembles Langevin dynamics without stochastic forces, but ζ may assume both positive and negative values and fluctuates during a simulation.

A number of simulation studies with implicit solvent have been reported during recent years.[27−37] The general conclusion from these studies is that is possible to obtain stable trajectories with implicit solvent simulations that exhibit conformational sampling comparable to explicit solvent simulations.[16,21,28,37−40] Some discrepancies, in particular with respect to the stability of salt bridges, have also been reported, but they appear to be resolved at least in part through careful adjustment of the dielectric interface and/or force field reparameterization.[40−43]

A question that has not been fully addressed to date is to what extent the kinetic properties can be reproduced correctly in implicit solvent simulations and to what extent simulations can be accelerated when solvent viscosity is reduced or omitted. Previous studies have found that conformational transitions in the context of protein folding can be accelerated substantially and predictably in Langevin dynamics with reduced friction coefficients,[23] but other evidence also suggests that the choice of the implicit solvent itself may also have a significant impact on the simulated dynamics.[28]

In this study, this question is addressed further with the application of the thermodynamically highly accurate generalized Born molecular volume (GBMV) implicit solvent model[17] to the following two cases: First, kinetic rates of transitions between dominant conformational basins in alanine dipeptide are compared between implicit and explicit solvent simulations. Second, native state conformational sampling of the B1 domain of streptococcal protein G and ubiquitin is compared with experimental data from X-ray crystallography and NMR spectroscopy. The results for both the dipeptide kinetics and protein structure dynamics indicate that kinetic properties can in fact be accurately reproduced with implicit solvent when the effects of solvent friction are included appropriately. Furthermore, conformational sampling can be accelerated by a moderate amount when reduced friction coefficients are applied.

In the following, the simulation methodology is described in more detail before the simulation results are presented and compared to experimental data.

Methods

Explicit and implicit solvent molecular dynamics simulations of blocked alanine dipeptide, the B1 domain of streptococcal protein G, and ubiquitin were performed as summarized in Table 1. The alanine dipeptide simulations were started with peptide torsion angles of $\phi = -65°$ and $\psi = -40°$ corresponding to the α basin. The simulations of protein G and ubiquitin used the experimentally determined structures in aqueous solvent according to NMR spectroscopy (PDB codes: 3GB1[44] and 1D3Z,[45] respectively) as the initial conformations. The all-atom CHARMM force field[46] was used in all of the simulations in conjunction with the CMAP cross-correlation potential[47,48] to improve the sampling of ϕ/ψ backbone torsion angles.

In the explicit solvent simulations, the solute was solvated in a cubic box filled with TIP3P water molecules. Four sodium counterions were added to neutralize the negative charge of protein G. Counterions were not needed for the neutral ubiquitin, and no extra salt was added. The resulting box sizes are 27.23 Å^3 for alanine dipeptide, 61.00 Å^3 for protein G, and 69.30 Å^3 for ubiquitin. Periodic boundary conditions were applied, and particle-mesh Ewald summation with a real-space cutoff of 9 Å was used to calculate electrostatic interactions. The explicit solvent system was simulated in an NVT ensemble at 300 K with a Nosé−Hoover thermostat.[25] A time step of 2 fs was employed in conjunction with SHAKE[49] in order to constrain bonds between heavy atoms and hydrogen atoms. A standard equilibration protocol was applied with initial steepest descent and adopted-basis Newton−Raphson minimization followed by slow heating over 42 ps to the final temperature and slight adjustment of the simulation box size in order to obtain correct bulk water densities corresponding to a temperature of 300 K and a pressure of 1 atm.

Implicit solvent simulations were run with the GBMV variant[20,50] of the generalized Born formalism. The dielectric constant inside the solute cavity was set to $\epsilon = 1$ and to $\epsilon =$ 80 for the surrounding medium. GBMV parameters defined in the original references[20,51] were set to $\beta = -12$, $S_0 =$

Table 1. Summary of Simulations Analyzed in This Study

system	solvent	CMAP	length
alanine dipeptide	explicit	original	200 ns
alanine dipeptide	implicit $f = 50, 10, 5, 1, 0.5$ ps^{-1}	original	500 ns
alanine dipeptide	implicit $f = 50, 10, 5, 1, 0.5$ ps^{-1}	modified	500 ns
alanine dipeptide	implicit $f = 5$ ps^{-1} (C), 10 ps^{-1} (N), 20 ps^{-1} (O)	modified	1000 ns
alanine dipeptide	implicit Nosé−Hoover, $q = 1, 10, 100$ kcal/(mol ps^2)	modified	500 ns
protein G	explicit	original	50 ns
protein G	implicit $f = 5$ ps^{-1} (C), 10 ps^{-1} (N), 20 ps^{-1} (O)	original	50 ns
protein G	implicit $f = 50, 5$ ps^{-1}	modified	50 ns
protein G	implicit Nosé-Hoover, $q = 10$	modified	50 ns
protein G	implicit $f = 5$ ps^{-1} (C), 10 ps^{-1} (N), 20 ps^{-1} (O)	modified	50 ns
ubiquitin	explicit	original	22 ns
ubiquitin	implicit $f = 5$ ps^{-1} (C), 10 ps^{-1} (N), 20 ps^{-1} (O)	original	22.5 ns
ubiquitin	implicit $f = 50, 5$ ps^{-1}	modified	22.5 ns
ubiquitin	implicit Nosé-Hoover, $q = 10$ kcal/(mol ps^2)	modified	22.5 ns
ubiquitin	implicit $f = 5$ ps^{-1} (C), 10 ps^{-1} (N), 20 ps^{-1} (O)	modified	22.5 ns

0.65, $C_0 = -0.1$, and $C_1 = 0.9$ in order to obtain stable trajectories.[38] The GBMV method also provides an estimate of the SASA that was used to calculate the hydrophobic solvation free energy according to γ·SASA with $\gamma = 5.42$ cal/(mol Å^2). Electrostatic interactions were switched to zero from 16 to 18 Å in the simulations of ubiquitin and protein G. No cutoff was applied in the alanine dipeptide simulations. The implicit solvent simulations were run with the same force field and CMAP torsion potential as with the explicit solvent simulations. In addition, a slightly modified CMAP potential was also used in order to improve agreement with the potential of mean force of alanine dipeptide from the explicit solvent simulation (see below). An integration time step of 1.5 fs was used in the protein G and ubiquitin simulations in conjunction with SHAKE[49] to constrain bonds involving hydrogen atoms. Alanine dipeptide was simulated without SHAKE and with a time step of 1 fs. In order to maintain a constant temperature during the simulations, either the Nosé−Hoover algorithm[25] (with coupling constants of 1, 10, or 100 kcal/mol·ps^2 in a single thermostat applied to the entire system) or a Langevin heat bath was applied.[22] Langevin dynamics also include the effects of viscosity. In this study, friction coefficients f of 0.5, 1, 5, 10, and 50 ps^{-1} were applied to non-hydrogen atoms. No memory function was employed in the Langevin algorithm that represents the overdamped limit.

All of the simulations were performed with the CHARMM program.[52] Version c30b2 was used for the explicit solvent simulation of alanine dipeptide, c32a2 for the implicit solvent simulations of alanine dipeptide, and c33a1 for the implicit and explicit solvent simulations of protein G and ubiquitin. The MMTSB Tool Set[53] was used to facilitate and analyze the simulations in conjunction with CHARMM.

Results

Kinetic Transitions in Alanine Dipeptide. The alanine dipeptide system (Figure 1A) is used as a model for peptide backbone dynamics. A potential of mean force map from the 200 ns explicit solvent simulation (Figure 2A) shows that there are essentially four distinct states: α at (−60,−45) to (−100,0), β/PPII at (−60,150) to (−160,160), α_L at (60,50), and C7$_{ax}$ at (50,−150). With the CHARMM force field used here, the α and β basins are practically equienergetic while the other two minima are slightly higher by a few kilocalories per mole. Implicit solvent simulations with the GBMV model and the same force field (Figure 2B) result in a very similar free energy map, but some differences are noticeable. The main minimum in the β basin is shifted from the poly proline II (PPII) conformations at (−60,150) to fully extended conformations (−160,160), and the second minimum in the α basin at (−100,0) is enhanced. Furthermore, the barriers from the α_L basin to the β and C7$_{ax}$ basins are elevated compared to the explicit solvent simulations. In order to understand the source of these deviations better, it is instructive to examine the energetics of alanine dipeptide conformations with implicit solvent conformations in more detail. Table 2 compares the relative energies for selected conformations between the explicit solvent PMF and adiabatic free energies from the continuum dielectric implicit solvent model. It can be seen that the deviations are relatively small when electrostatic solvation energies are obtained directly from solutions to the Poisson equation. However,

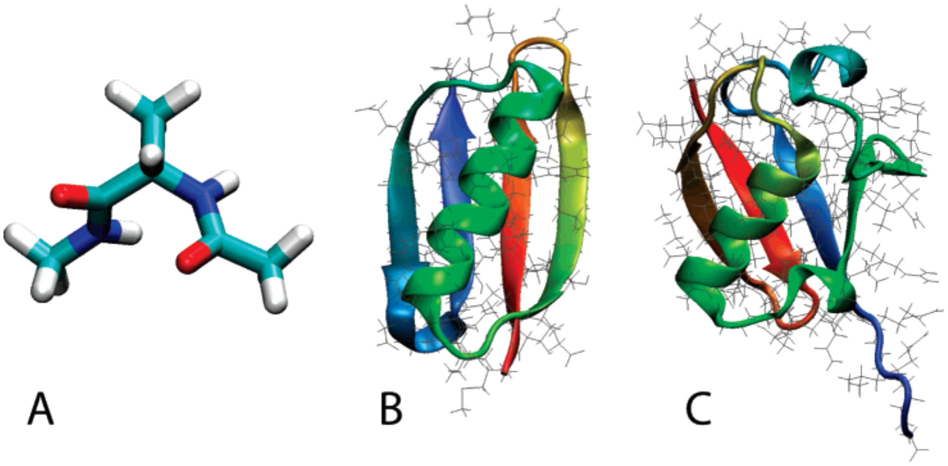


Figure 1. Structures of blocked alanine dipeptide (A), B1 domain of streptococcal protein G (PDB code: 3GB1[44]) (B), and ubiquitin (PDB code: 1D3Z[45]) (C). Graphics were generated with VMD.[69]

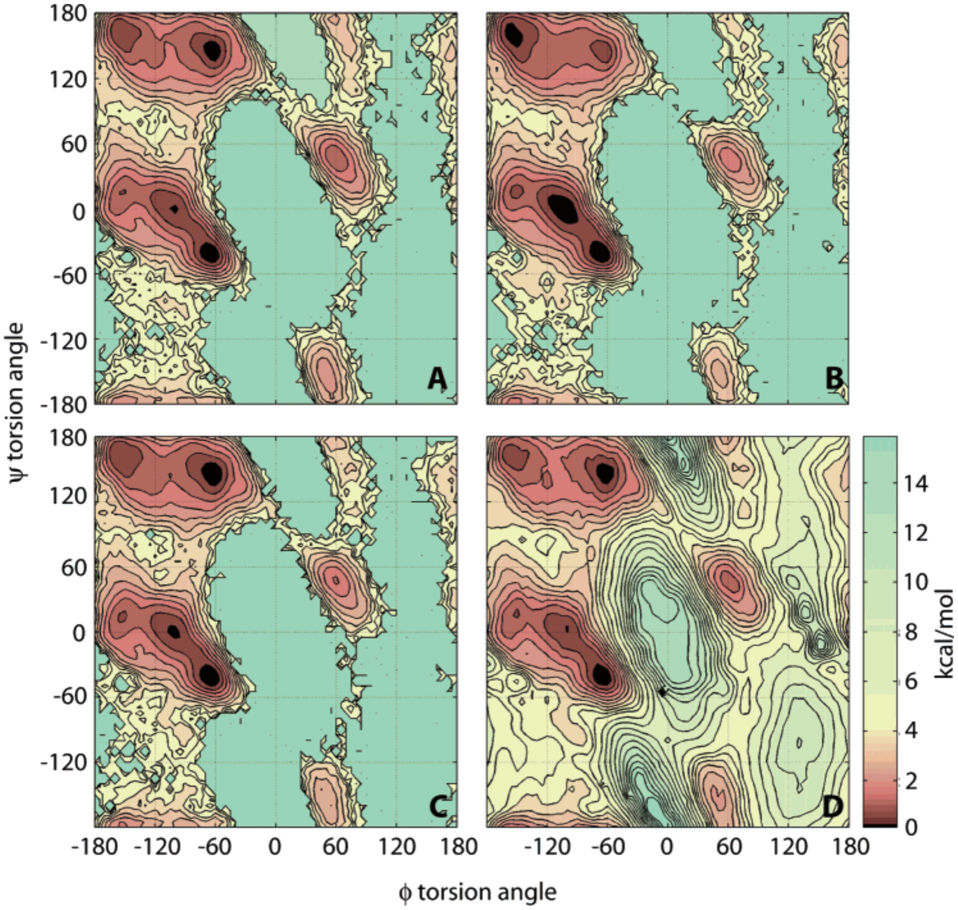


Figure 2. Potential of mean force from molecular dynamics simulations of blocked alanine dipeptide with explicit solvent and original CMAP torsion potential (A), implicit solvent and original CMAP potential (B), and implicit solvent and modified CMAP potential (C). A friction coefficient of 10 ps^{-1} was used in the implicit solvent simulation shown. The results from the simulations are compared with an adiabatic map (D) where implicit solvent energies with the modified CMAP potential are evaluated directly after minimization at different values of ϕ/ψ.

the deviations are amplified when the GBMV formalism is used to approximate the Poisson solutions. Therefore, the (small) deviations between the free energy maps shown in Figure 2A and B represent mostly errors due to the generalized Born approximation rather than the implicit solvent model itself.

In order to focus on a comparison of kinetic properties between implicit and explicit solvent, it is desirable that the underlying thermodynamics is as similar as possible. Most of the differences described above should be resolved in principle with a more accurate generalized Born method. However, further improvements of the already very accurate

Table 2. Free Energies of Selected Conformations of Alanine Dipeptide with Implicit and Explicit Solvent Relative to β Conformation[a]

	f (C−N−Cα−C)	y (N−Cα−C−N)	GBMV [kcal/mol]	PB [kcal/mol]	explicit PMF [kcal/mol]
$\boldsymbol{\beta}$	−155	160	0	0	0
PPII	−65	145	0.47	0.12	−0.45
$\beta \rightarrow \alpha_L$	0	105	4.82	4.25	4.02
α_L	60	45	1.34	0.75	0.40

[a] Implicit solvent energies are obtained under the adiabatic approximation by minimizing alanine dipeptide with a harmonic restraint on the ϕ and ψ torsion angles, each with a force constant of 1000 kcal/mol/Å^2, for different values of ϕ and ψ on a grid with a spacing of 15°. Explicit solvent energies are obtained as a potential of mean force from the sampling probabilities in the explicit solvent molecular dynamics simulation. The original CMAP potential was used in all cases. Poisson−Boltzmann (PB) results were obtained with the PBEQ finite difference solver in CHARMM using a grid spacing of 0.15 Å.

Table 3. Kinetic Rates in ns^{-1} for Conformational Transitions in Blocked Alanine Dipeptide from Implicit and Explicit Solvent Simulations[a]

solvent	CMAP	friction	α/β	β/α_L	$\alpha_L/C7_{ax}$
expl.	orig.		2.5 (*225*)	0.3 (*26*)	3.8 (*35*)
			2.6 (*224*)	2.7 (*25*)	14.8 (*33*)
impl.	orig.	50 ps^{-1}	1.1 (*316*)	0.02 (*4*)	1.4 (*9*)
			1.6 (*317*)	0.8 (*5*)	8.8 (*8*)
	orig.	10 ps^{-1}	3.8 (*1000*)	0.06 (*12*)	5.5 (*47*)
			5.0 (*1000*)	1.4 (*12*)	25.7 (*47*)
	orig.	5 ps^{-1}	4.8 (*1273*)	0.09 (*19*)	6.7 (*75*)
			6.3 (*1273*)	1.4 (*16*)	38.3 (*71*)
	orig.	1 ps^{-1}	5.4 (*1391*)	0.1 (*25*)	6.5 (*92*)
			6.9 (*1388*)	1.6 (*22*)	38.6 (*89*)
	orig.	0.5 ps^{-1}	5.4 (*1422*)	0.1 (*21*)	8.7 (*79*)
			7.0 (*1421*)	2.9 (*26*)	55.6 (*8.5*)
	mod.	50 ps^{-1}	1.4 (*312*)	0.09 (*21*)	1.7 (*20*)
			1.4 (*311*)	1.6 (*19*)	7.5 (*19*)
	mod.	10 ps^{-1}	4.6 (*893*)	0.2 (*45*)	6.4 (*62*)
			4.3 (*897*)	4.9 (*48*)	22.9 (*64*)
	mod.	5 ps^{-1}	5.7 (*1299*)	0.2 (*53*)	6.5 (*77*)
			5.9 (*1302*)	4.8 (*57*)	29.5 (*80*)
	mod.	1 ps^{-1}	6.5 (*1427*)	0.3 (*58*)	7.0 (*99*)
			6.3 (*1425*)	4.6 (*65*)	27.2 (*107*)
	mod.	0.5 ps^{-1}	6.0 (*1340*)	0.3 (*75*)	7.1 (*121*)
			6.0 (*1346*)	4.4 (*75*)	33.3 (*117*)
	mod.	mixed[b]	4.1 (*1940*)	0.2 (*110*)	4.2 (*146*)
			4.0 (*1944*)	3.4 (*118*)	18.5 (*154*)

[a] Transitions are counted when one of the four basins is reached from one of the other basins. No transitions are recorded if a basin is left and re-entered without visiting another basin. Forward and backward rates are give in the first and second row, respectively. The number of observed transitions is given in parentheses. [b] Friction coefficient of 10 ps^{-1} for nitrogen, 20 ps^{-1} for oxygen, and 5 ps^{-1} for carbon.

GBMV method compared to other generalized Born formalisms are not easily achievable. On the other hand, direct solutions of the Poisson equation at a very high level of accuracy are not practical in simulation applications. A pragmatic solution is the use of a slightly modified CMAP term for the ϕ/ψ torsion angles. The CMAP potential function is obtained from a spline-interpolated grid-based two-dimensional function of ϕ and ψ.[47,48,54] The deviations between the explicit and implicit solvent simulations are easily incorporated by adding the difference between the respective free energy maps to the original CMAP grid data. The modified CMAP is available from the author by request. The use of different ϕ/ψ torsion maps for implicit and explicit solvent has been suggested previously to correct for deficiencies of the implicit solvent model.[40] As a result, it is possible to reproduce the explicit solvent free energy map nearly exactly (see Figure 2C).

Implicit solvent allows the calculation of the complete free energy map under the adiabatic approximation (shown in Figure 2D). As would be expected, the resulting map agrees closely with the potentials of mean force obtained from the simulations, but the adiabatic map also offers insight into higher-energy regions that are not sampled during sub-microsecond simulations. In particular, a clearer view of possible transition pathways between the four main minima is given. Transitions between α and β basins may progress either through the C7eq conformation near (−60,60) or through the second transition state at (−100,−120). Transitions between β and α_L basins proceed through the transition state at (0,90), while transitions between the α_L and C7$_{ax}$ can also follow two routes along $\psi = 80$ either through positive or negative ϕ angles. Additional transitions may occur directly between β and C7$_{ax}$ across the transition state at (120,120) and from α to C7$_{ax}$ across the transition state at (0,−100) or to α_L across (130,−30).

The sub-microsecond simulations described here sample only the α/β, β/α_L, and $\alpha_L/C7_{ax}$ transitions sufficiently to obtain meaningful statistical averages. Table 3 shows the kinetic rates for these transitions in alanine dipeptide (following either pathway in the case of α/β and $\alpha_L/C7_{ax}$ transitions) obtained from explicit and implicit solvent simulations. It can be seen that the implicit solvent simulations are in good qualitative agreement with the explicit solvent results when the modified CMAP is used to match the sampling of ϕ/ψ angles to the explicit solvent and when a friction coefficient between 10 and 50 ps^{-1} is used. With a friction coefficient of 10 ps^{-1}, most transitions are faster by about a factor of 2, except for the transition from β to α_L, which is too slow. On the other hand, the rates are significantly reduced to about half the value found with explicit solvent when a friction coefficient of 50 ps^{-1} is used. If the friction coefficient is reduced further from 10 ps^{-1}, some but not all of the rates are accelerated. Most noteworthy, the α_L-to-β rate appears to be largely unaffected and actually slightly decreases as the friction is reduced. Previous studies have suggested that a single friction coefficient may not be optimal.[34] After trying a number of combinations (data not shown), it was found that a friction coefficient of 10

Table 4. Mean First Passage Time in ps of Conformational Transitions in Blocked Alanine Dipeptide from Implicit Solvent Simulations in Comparison with Results from Explicit Solvent Simulation[a]

solvent	CMAP	friction	α/β	β/α_L	$\alpha_L/C7_{ax}$	$\alpha/C7_{ax}$
expl.	orig.		**426** (*27*)	**6257** (*1130*)	**4441** (*1181*)	**8968** (*1748*)
			459 (*35*)	**430** (*78*)	**602** (*376*)	**744** (*97*)
impl.	orig.	50 ps^{-1}	929 (*56*)	94 617 (*15420*)	53 305 (*29034*)	157 676 (*41942*)
			660 (*40*)	1278 (*599*)	77 (*30*)	2088 (*737*)
	orig.	10 ps^{-1}	282 (*9*)	25 562 (*4537*)	**5710** (*1901*)	25465 (*4536*)
			223 (*8*)	**686** (*188*)	**2216** (*1378*)	**693** (*168*)
	orig.	5 ps^{-1}	220 (*6*)	21 029 (*4522*)	**4115** (*1601*)	24 193 (*5579*)
			178 (*6*)	**608** (*107*)	2219 (*1076*)	693 (*147*)
	orig.	1 ps^{-1}	195 (*6*)	14 487 (*2557*)	**2897** (*851*)	17 714 (*3359*)
			166 (*6*)	**545** (*88*)	2134 (*923*)	**658** (*94*)
	orig.	0.5 ps^{-1}	195 (*5*)	16 350 (*4741*)	**4682** (*1651*)	18 778 (*5473*)
			159 (*5*)	**372** (*59*)	**1176** (*1027*)	518 (*70*)
	mod.	50 ps^{-1}	762 (*42*)	20 901 (*4017*)	14 357 (*4976*)	27 691 (*6156*)
			832 (*46*)	721 (*200*)	4621 (*2288*)	1083 (*235*)
	mod.	10 ps^{-1}	234 (*8*)	**8089** (*1033*)	**5328** (*1161*)	**12473** (*1871*)
			261 (*9*)	247 (*30*)	**950** (*485*)	534 (*67*)
	mod.	5 ps^{-1}	192 (*5*)	**7549** (*750*)	**5160** (*970*)	12526 (*1748*)
			190 (*6*)	254 (*32*)	**670** (*362*)	495 (*53*)
	mod.	1 ps^{-1}	166 (*4*)	**6637** (*665*)	**3708** (*624*)	**9116** (*1013*)
			183 (*6*)	267 (*32*)	**514** (*192*)	451 (*50*)
	mod.	0.5 ps^{-1}	178 (*5*)	**4998** (*605*)	**3121** (*618*)	**7776** (*1230*)
			192 (*6*)	238 (*25*)	**476** (*137*)	437 (*41*)
	mod.	mixed[b]	265 (*6*)	7943 (*691*)	**6203** (*797*)	12667 (*1241*)
			287 (*7*)	**359** (*29*)	**345** (*149*)	**616** (*46*)

[a] Passage times in forward and backward direction are given for each transition. Estimated Statistical errors calculated as $\sigma/\sqrt{N}$ from the standard deviation σ and the number of transitions *N* are given in parentheses. Values that agree with the explicit solvent simulation within the error intervals are shown in boldface. [b] Friction coefficient of 10 ps^{-1} for nitrogen, 20 ps^{-1} for oxygen, and 5 ps^{-1} for carbon.

ps^{-1} applied to the amide nitrogen, 20 ps^{-1} applied to the carbonyl oxygen, and 5 ps^{-1} applied to carbon atoms gave the best results. In that case, the rates are in good quantitative agreement with the explicit solvent simulations (see Table 3), although the transition rates between the α and β basins are still significantly overestimated.

Transition rates obtained with implicit solvent and the original CMAP torsion term do not agree as well with the explicit solvent simulations. As would be expected from the differences in the free energy maps, the rates between α and β basins are more asymmetric with faster rates from β to α reflecting a more favorable free energy of the α basin versus the poly proline II conformation. As a result of the increased barrier height at (0,100), transitions between β and α_L are significantly slowed down. However, despite the differences in detail, the implicit solvent simulations with the original CMAP term also provide an overall reasonable qualitative description of the kinetics of alanine dipeptide if a friction coefficient of 10 ps^{-1} is chosen.

A comparison of mean first passage times is given in Table 4. The mean first passage time between two basins measures the time it takes to reach the second basin after the first basin has been entered. According to this definition, a transition from β to α may involve a simple barrier crossing from β to α or a more complicated path such as β to α_L, α_L to β, and finally β to α. Therefore, transitions between the α and $C7_{ax}$ basins are also considered along with transitions between α/β, β/α_L, and $\alpha_L/C7_{ax}$. An analysis of the mean first passage times provides similar conclusions as for the kinetic rates of single barrier crossings described above. Overall good agreement with the explicit solvent simulation is found when the modified CMAP potential is used in conjunction with the mixed friction coefficients as described above. When a single friction coefficient is used for all non-hydrogen atoms, the data from explicit solvent are reproduced best with friction coefficients between 1 and 10 ps^{-1}. However, larger discrepancies remain for transitions between α and β basins that are too fast with implicit solvent unless a friction coefficient of 50 ps^{-1} is employed.

The distribution of mean first passage times for transitions between α and β shown in Figure 3 provides a more detailed view of how kinetics from implicit and explicit solvent compare. It can be seen that the relative frequency of long passage times becomes more similar to the explicit solvent simulations beyond 500 ps to 1 ns while the differences are largest for the distribution of short-time transition events. Furthermore, it is apparent that the explicit solvent distribution of mean first passage times at short times is represented best with the mixed friction coefficients around 10 ps^{-1}. However, a friction coefficient of 50 ps^{-1} provides significantly better agreement with explicit solvent for the distribution of passage times beyond 500 ps (β-to-α transition) and beyond 800 ps (α-to-β transition). This finding suggests that Langevin dynamics with any choice of friction coefficient do not fully capture the hydrodynamic effects of water throughout the entire range of relevant time scales if compared to explicit solvent.

As discussed briefly above, transitions are accelerated over explicit solvent with small friction coefficients and slowed down with large friction coefficients. Figure 4 shows the

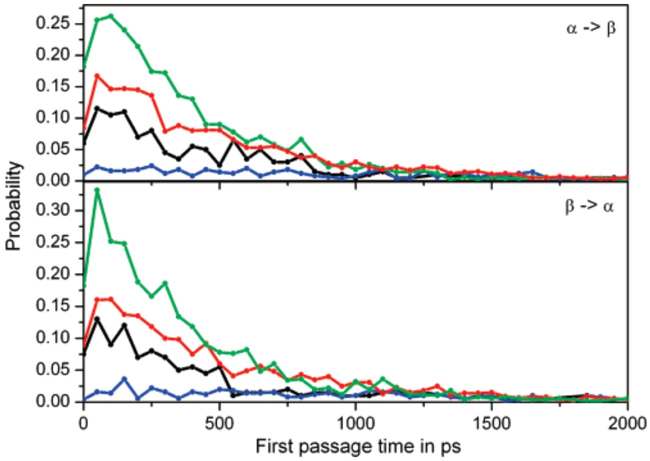


Figure 3. Distribution of first passage times for α → β (top) and β → α (bottom) transitions in blocked alanine dipeptide. Results from explicit solvent simulation are shown in black, from implicit solvent simulations with modified CMAP potential in blue (f = 50 ps^{-1}), red (mixed friction, see text), and green (f = 5 ps^{-1}), respectively.

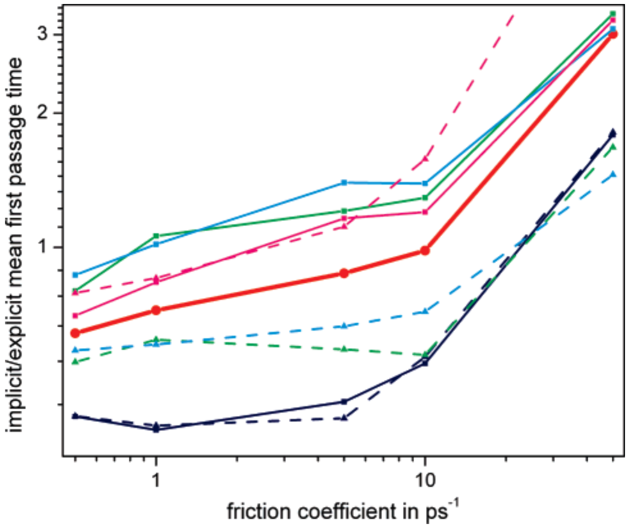


Figure 4. Mean first passage times from implicit solvent simulations with modified CMAP potential relative to explicit solvent values for conformational transitions in blocked alanine dipeptide as a function of friction coefficient. Both axes are shown in logarithmic scale. Individual transitions are indicated by color as follows: α/β (blue), β/α$_L$ (green), α$_L$/C7$_{ax}$ (magenta), and α/C7$_{ax}$ (cyan). Solid (dashed) lines indicate forward (backward) transitions. The red line indicates the average relative mean first passage times from all transitions.

mean first passage times relative to explicit solvent as a function of the friction coefficient for the observed transitions in the alanine dipeptide system. The graph readily identifies two distinct regimes above and below a friction coefficient of 10 ps^{-1}. At 10 ps^{-1}, the mean first passage times are roughly equal between implicit and explicit solvent. Transitions below 10 ps^{-1} are accelerated with decreasing friction according to $r = f \times 0.034 + 0.67$, where r is the speed relative to explicit solvent and f is the implicit solvent friction coefficient in picoseconds^{-1}. The extrapolation of this function to zero friction gives a speedup by less than a factor of 2. A slightly better fit is found with $\log(r) = \log(f) \times 0.137 - 0.34$; however, this function does not extrapolate to zero friction. Solute friction is dominant in this regime, as pointed out previously, and consequently the transitions are only accelerated moderately. On the other hand, friction coefficients larger than 10 ps^{-1} slow down the kinetics much more rapidly, indicative of a solvent friction-dominated regime. Because only two points (10 and 50 ps^{-1}) were simulated, a functional form for this regime cannot be given with confidence.

Table 5. Mean First Passage Time in ps of Conformational Transitions in Blocked Alanine Dipeptide from Implicit Solvent Simulations with Modified CMAP Torsion Potential (see Text) and Nosé−Hoover Thermostat as a Function of the Coupling Constant q[a]

	q = 1	q = 10	q = 100	explicit
α → β	11 367 (*2357*)	85 948 (26 652)	5683 (982)	426 (*27*)
β → α	15 040 (*18*)	44 753 (24 621)	7048 (888)	459 (*35*)

[a] Statistical errors are given in parentheses. Results from the explicit solvent simulations are shown for comparison.

So far, only the implicit solvent simulations with Langevin dynamics have been discussed. Table 5 shows the mean first passage times between α and β basins when a Nosé−Hoover thermostat is used instead of Langevin dynamics. It can be seen that with any of coupling constants, the first passage times are 1−2 orders of magnitude longer than with the explicit solvent. Transitions to the right side of the Ramachandran plot were observed, but they occurred so rarely that significant statistics could not be obtained from 500 ns simulations. Although simulations with a Nosé−Hoover thermostat represent the limit of zero friction, the absence of stochastic collisions with water molecules greatly diminishes the effectiveness of overcoming transitional barriers, which explains the very slow kinetics in the alanine dipeptide system when Langevin dynamics are not used.

Native State Dynamics of Protein G and Ubiquitin. Protein G and ubiquitin were simulated with explicit and implicit solvent over 50 and 22.5 ns, respectively, in order to examine how the conclusions from the alanine dipeptide system apply to native state simulations of proteins.

Average Structural and Dynamic Properties. First, average structural and dynamic properties are compared between implicit and explicit solvent simulations and with experimental data to establish that the implicit solvent model results in a thermodynamically sufficiently accurate description of native state dynamics. Figures 5 and 6 show root-mean-square deviations (rmsd) as a function of time for both systems. Overall, the data demonstrate that the simulated structures remain very close to the experimental structure with both explicit and implicit solvent. The most notable exception is an excursion as far away as about 4 Å for about 5 ns in one of the implicit simulations of protein G (see Figure 5D). Other significant deviations from the experimental structure are observed around 22 ns with explicit solvent and at different times past 20 ns in the implicit solvent simulation with a Nosé−Hoover thermostat. On the basis of the time evolutions, the simulations were, somewhat arbitrarily, considered fully equilibrated after 30ns (protein G) and 10ns (ubiquitin), respectively. The protein G simulations from 30 to 50 ns and the ubiquitin simulations from

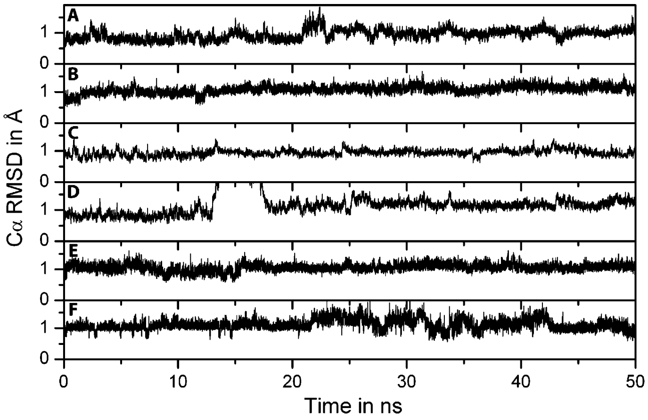


Figure 5. Cα root-mean-square deviations from NMR conformation (3GB1) during molecular dynamics simulations of protein G with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F).

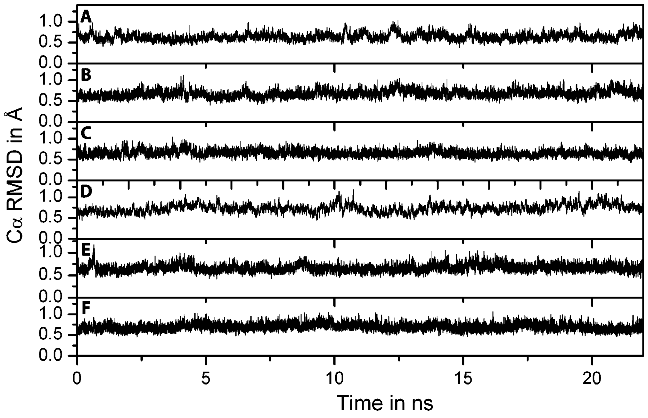


Figure 6. Cα root-mean-square deviations from NMR conformation (1D3Z) during molecular dynamics simulations of ubiquitin (residues 1−72) with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F).

10 to 22.5 ns were subjected to further analysis. Average root-mean-square deviations for Cα and Cβ atoms from both NMR and crystallographic structures are given in Table 6. Relatively few significant differences can be discerned suggesting that implicit and explicit solvent simulations essentially sample very similar, nativelike conformations.

A more rigorous comparison with experimental structures should be based on average structures from the simulations. The corresponding root-mean-square deviations shown in Table 7 are in fact smaller than the average root-mean-square deviations from Table 6, as close as 0.4 Å for ubiquitin Cα atoms and as far as 1 Å for protein G Cβ atoms. This excellent agreement is generally matched with implicit solvent.

Average dynamic structural fluctuations in protein G and ubiquitin can be compared to experimental B factors from crystallography[55] and S^2 order parameters from NMR spectroscopy.[56] B factors are related to atomic mean-square displacements σ_i under the assumption that fluctuations are isotropic according to $B_i = 8\pi^2\sigma_i/3$. The comparison of B factors from simulations and experiments is shown in Figures 7 and 8 for protein G and ubiquitin, respectively. B factors for potein G are available from two different crystal structures (1PGA[57] and 1PGB[57]). The two sets of B factors are substantially different, indicating the degree of uncertainty for this type of data. The data extracted from the explicit solvent simulation agree best with the B factors from 1PGA,[57] although the loop between the first two strands of the β sheet near residue 10 is significantly more flexible in the simulation. A closer look at the crystal structure reveals that these residues are involved in crystal packing interactions to a greater extent in 1PGA than in 1PGB. This corresponds with larger crystallographic B factors from 1PGB in this region. Significant differences from explicit solvent are found in most of the implicit solvent simulations. However, the implicit solvent simulations also do not fully agree with each other in particular with respect to the flexibility around residues 10, 20, and 40. B factors for ubiquitin are also available from two different sources (1UBQ[58] and 1UBI[59]). The differences between these two data sets are much smaller than for protein G. Furthermore, B factors from the explicit

Table 6. Time-Averaged Coordinate Root-Mean-Square Deviations in Å from Explicit and Implicit Solvent Simulations of Protein G and Ubiquitin in Comparison with NMR (3gb1, 1d3z) and X-ray Structures (1pga, 1ubq)[a]

		protein G (30-50ns)		ubiquitin (10−22.5 ns)	
		Cα	Cβ	Cα	Cβ
explicit solvent	NMR	1.01 (*0.11*)	1.20 (*0.13*)	0.65 (*0.09*)	0.77 (*0.10*)
	X-ray	0.83 (*0.11*)	1.00 (*0.14*)	0.68 (*0.09*)	0.79 (*0.09*)
GBMV, orig. CMAP mixed friction	NMR	1.13 (*0.10*)	1.37 (*0.11*)	0.70 (*0.09*)	0.82 (*0.10*)
	X-ray	0.97 (*0.13*)	1.18 (*0.14*)	0.69 (*0.09*)	0.79 (*0.09*)
GBMV, mod. CMAP mixed friction	NMR	1.25 (*0.14*)	1.63 (*0.24*)	0.64 (*0.07*)	0.76 (*0.08*)
	X-ray	1.10 (*0.16*)	1.45 (*0.26*)	0.66 (*0.07*)	0.77 (*0.07*)
GBMV, mod. CMAP $f = 50$ ps^{-1}	NMR	1.20 (*0.11*)	1.39 (*0.12*)	0.76 (*0.10*)	0.87 (*0.11*)
	X-ray	1.12 (*0.15*)	1.29 (*0.16*)	0.73 (*0.10*)	0.84 (*0.10*)
GBMV, mod. CMAP $f = 5$ ps^{-1}	NMR	1.09 (*0.10*)	1.31 (*0.11*)	0.67 (*0.08*)	0.77 (*0.08*)
	X-ray	0.90 (+*0.12*)	1.08 (*0.24*)	0.69 (*0.09*)	0.80 (*0.08*)
GBMV, mod. CMAP Nosé−Hoover	NMR	1.08 (*0.19*)	1.32 (*0.23*)	0.69 (*0.08*)	0.80 (*0.09*)
	X-ray	1.00 (*0.20*)	1.23 (*0.24*)	0.68 (*0.08*)	0.79 (*0.09*)

[a] Standard deviations are given in parentheses. Only residues 1−72 were considered for ubiquitin because of the highly flexible C terminus.

Table 7. Coordinate Root-Mean-Square Deviations in Å of Average Structures from Explicit and Implicit Solvent Simulations of Protein G and Ubiquitin in Comparison with NMR (3gb1, 1d3z) and X-ray Structures (1pga, 1ubq)[a]

		protein G (30-50ns)		ubiquitin (10−22.5 ns)	
		Cα	Cβ	Cα	Cβ
explicit solvent	NMR	0.83	1.00	0.41	0.49
	X-ray	0.60	0.74	0.44	0.52
GBMV, orig. CMAP mixed friction	NMR	0.97	1.16	0.47	0.53
	X-ray	0.77	0.94	0.45	0.49
GBMV, mod. CMAP mixed friction	NMR	1.10	1.42	0.43	0.51
	X-ray	0.92	1.22	0.46	0.52
GBMV, mod. CMAP $f = 50$ ps^{-1}	NMR	1.01	1.18	0.57	0.63
	X-ray	0.93	1.06	0.54	0.60
GBMV, mod. CMAP $f = 5$ ps^{-1}	NMR	0.94	1.12	0.45	0.51
	X-ray	0.71	0.84	0.48	0.54
GBMV, mod. CMAP Nosé−Hoover	NMR	0.83	0.99	0.46	0.52
	X-ray	0.72	0.87	0.45	0.49

[a] Only residues 1−72 were considered for ubiquitin because of the highly flexible C terminus.

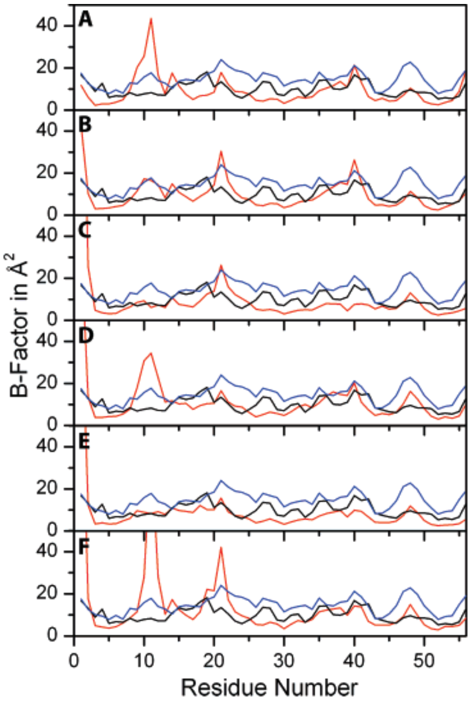


Figure 7. Cα B factors calculated from root-mean-square fluctuations during molecular dynamics simulations of protein G (30-50ns) with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F). Simulation results (red) are compared with crystallographic data from 1PGA (black) and 1PGB (blue).

solvent simulation of ubiquitin agree well with the experimental data. The flexibility between residues 50 and 65 is only slightly underestimated, and flexibility around residue 47 is overestimated. Contrary to the results for protein G, the implicit simulations of ubiquitin show no significant differences from the explicit solvent results.

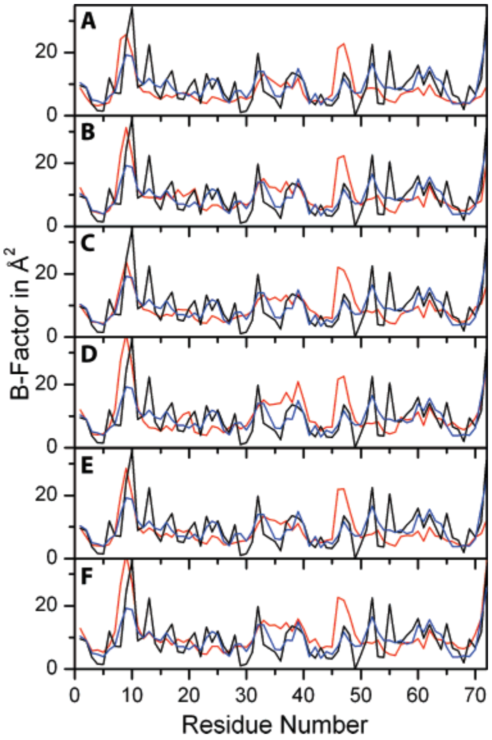


Figure 8. Cα B factors calculated from root-mean-square fluctuations during molecular dynamics simulations of ubiquitin (10−22.5ns) with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F). Simulation results (red) are compared with crystallographic data from 1UBI (black) and 1UBQ (blue).

Generalized S^2 order parameters were calculated for the dynamics of backbone N−H vectors $\mathbf{u}_i$ according to

$$S^2 = \frac{1}{2}[3\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\langle \mathbf{u}_{i,\alpha}\mathbf{u}_{i,\beta}\rangle^2 - 1]$$

where α and β are the Cartesian coordinates of $\mathbf{u}_i$. The results from the simulations are compared with experimental data in Figures 9 and 10 for protein G and ubiquitin, respectively. The data from the explicit solvent simulation of protein G are in fairly good agreement with the experimental S^2 values according to the Lipari−Szabo analysis,[60,61] only slightly overestimating flexibility near residues 10, 17, 22, and 41 (see Figure 9A). The generally better agreement of the explicit solvent simulation with the NMR data than with the crystallographic data near residue 10 further supports the view that the reduced flexibility in this region observed in the crystals is an artifact of crystallization while this region is actually more flexible in solution. Again, larger deviations between implicit and explicit solvent are observed for protein G while the differences between simulated and experimental S^2 values for ubiquitin are negligible (see Figure 10).

Sampling Efficiency. Short-time sampling efficiency with different implicit solvent methods was examined in the context of protein simulations to examine kinetic properties due to differences between thermostats and Langevin friction coefficients. Figures 11 and 12 show how the rmsd difference

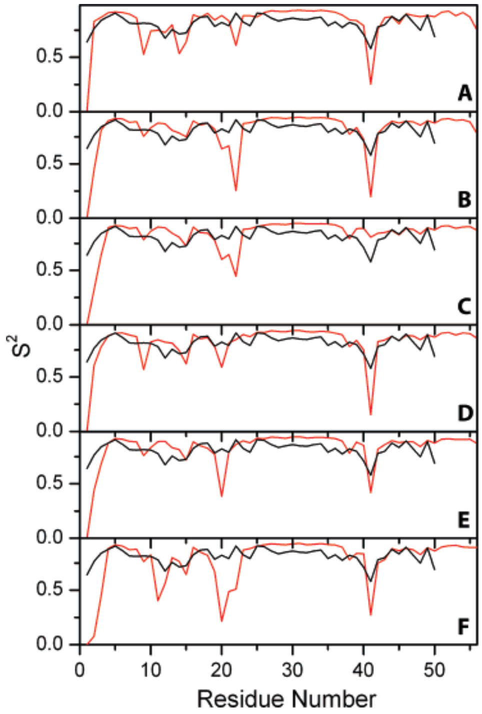


Figure 9. Backbone N−H order parameters S^2 during molecular dynamics simulations of protein G (30−50 ns) with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F). Simulation results (red) are compared with data from NMR spectroscopy (black).[70,71]

between simulated conformations at times t and $t + \delta t$ varies as a function of δt in the simulations of protein G and ubiquitin, respectively. A rapid increase in rmsd as a function of the time interval means that less time is spent in a given basin and that conformational sampling of different conformations is more efficient. A slow rise of rmsd, on the other hand, means that it takes a long time before a given conformational basin is left and that conformational sampling is inefficient. Similar conclusions are found for both, protein G and ubiquitin. Conformational sampling with implicit solvent is less efficient than with explicit solvent only if Langevin dynamics with a friction coefficient of 50 ps^{-1} are employed. When a friction coefficient of 5 ps^{-1} is used, sampling becomes more efficient by about a factor of 4−5 when comparing how much time it takes with implicit solvent to reach the average rmsd at $\delta t = 50$ ps with explicit solvent. The sampling appears to be much more efficient when a Nosé−Hoover thermostat is used instead of Langevin dynamics. In that case, conformations diverge rapidly after a very short time. A curious oscillatory behavior is seen, where conformations initially diverge further but return slightly after about 5 ps before diverging again.

Discussion

This study was carried out in order to examine in detail how the application of implicit solvent affects kinetic properties in simulations of peptides and proteins. Because very long simulations are needed to provide meaningful comparisons, only a limited number of systems and only one implicit solvent model, the GBMV method, with different viscosity parameters could be tested. Nevertheless, interesting observations could be made that are expected to be relevant at least qualitatively in a more general context for other systems and implicit solvent methods.

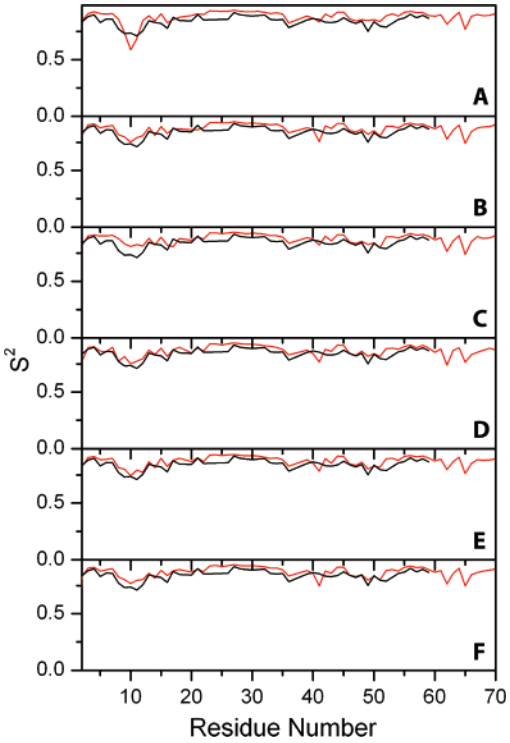


Figure 10. Backbone N−H order parameters S^2 during molecular dynamics simulations of ubiquitin (10−22.5 ns) with explicit solvent (A) or implicit solvent: mixed friction, original CMAP (B); mixed friction, modified CMAP (C); $f = 50$ ps^{-1}, modified CMAP (D); $f = 5$ ps^{-1}, modified CMAP (E); Nosé−Hoover thermostat, modified CMAP (F). Simulation results (red) are compared with data from NMR spectroscopy (black).[72]

The first part of this study focuses on kinetic transitions in the well-studied alanine dipeptide system. It is found that implicit solvent combined with Langevin dynamics can reproduce the kinetic behavior seen in explicit solvent simulations quite well when a suitable friction coefficient of around 10 ps^{-1} is chosen. Physical insight suggests that different friction coefficients should be used for different atom types rather than a single friction coefficient for all non-hydrogen atoms.[34] Results from this study provide some evidence that different friction coefficients may provide results closer to those for the explicit solvent, but it is unclear whether the rather modest improvements seen here would warrant substantial parametrization efforts for all protein side chains.

Not surprisingly, the agreement between implicit and explicit solvent is quite sensitive to the underlying free energy surface. Small deviations in the conformational preferences between the GBMV model and explicit solvent are manifest in significant differences in some of the kinetic rates. Sampling differences in the alanine dipeptide system are

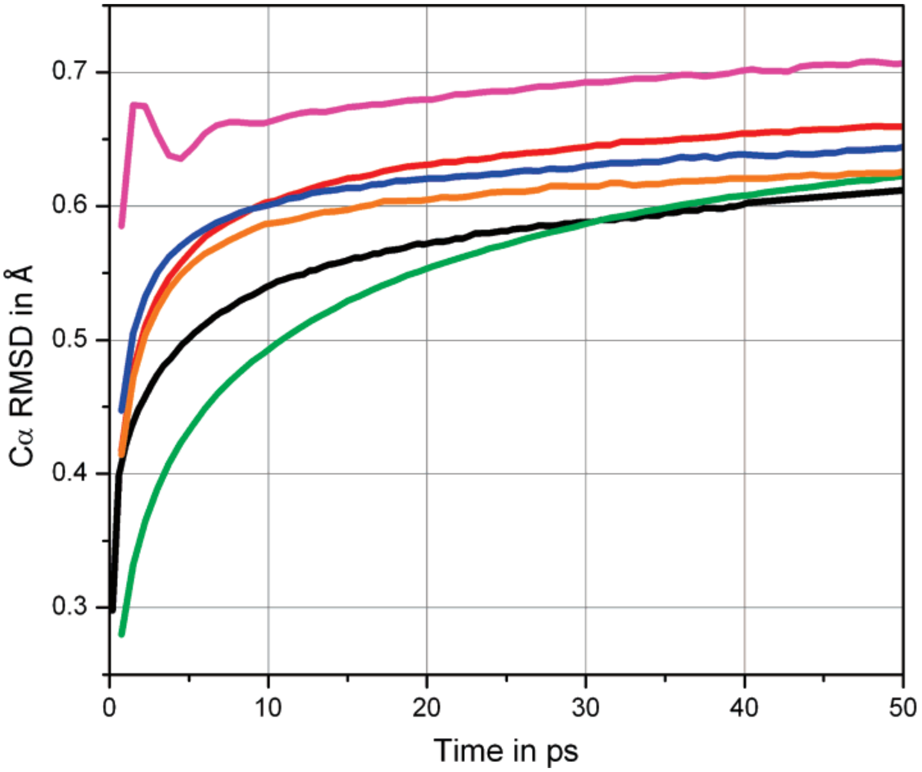


Figure 11. Average Cα rmsd between structures at times t and $t + \Delta t$ during simulations of protein G (30−50 ns) with explicit solvent (black) or implicit solvent: mixed friction, original CMAP (red); mixed friction, modified CMAP (orange); $f = 50$ ps^{-1}, modified CMAP (green); $f = 5$ ps^{-1}, modified CMAP (blue); Nosé−Hoover thermostat, modified CMAP (magenta).

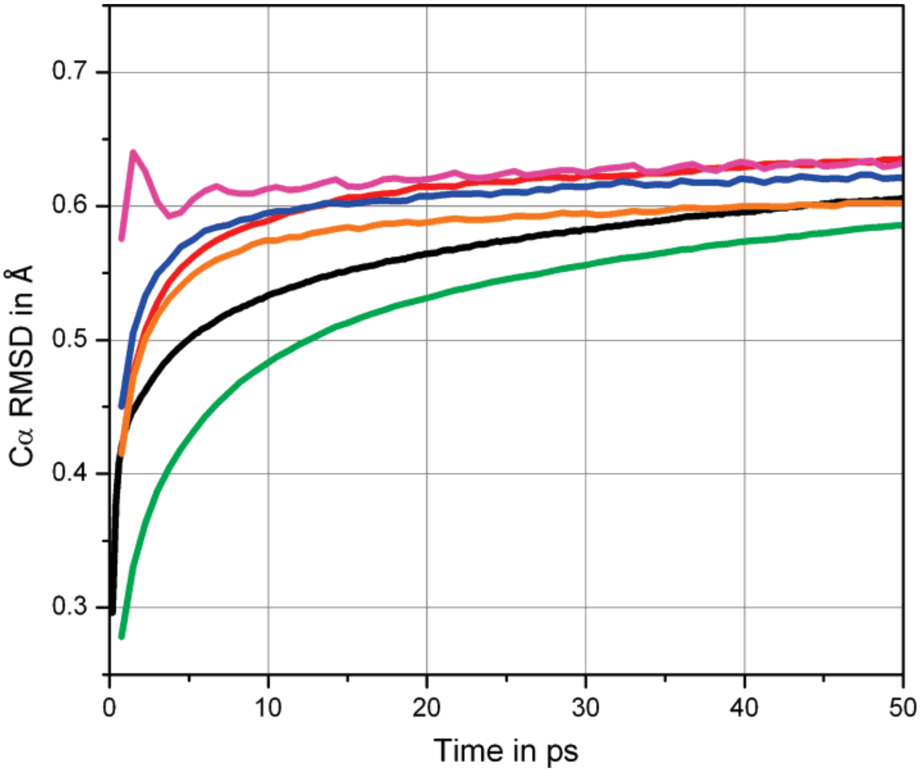


Figure 12. Average Cα rmsd between structures at times t and $t + \Delta t$ during simulations of ubiquitin (10−22.5 ns, residues 1−72) with explicit solvent (black) or implicit solvent: mixed friction, original CMAP (red); mixed friction, modified CMAP (orange); $f = 50$ ps^{-1}, modified CMAP (green); $f = 5$ ps^{-1}, modified CMAP (blue); Nosé−Hoover thermostat, modified CMAP (magenta).

easily remedied by using a modified CMAP torsion potential that accounts for the difference between the implicit and explicit solvent energetics. Previous work has also suggested the use of a modified force field for implicit solvent.[40] However, on the basis of the data presented here, it is impossible to tell whether the modified CMAP potential that improves sampling of the alanine dipeptide surface would also be able to compensate implicit solvent inadequacies in a wider range of systems, and it is noted that application of the modified CMAP term had little effect on the sampling of protein G or ubiquitin compared to the original CMAP.

A number of previous studies have examined how viscosity affects the barrier crossing kinetic rates in isomerization transitions[62] and during protein folding.[23,63−65] An interesting conclusion from some of these studies is the identification of two regimes where either solute−solute or solute−solvent friction is dominant. Similar conclusions can also be made from the data presented here: Small friction coefficients accelerate alanine dipeptide kinetics by a relatively modest amount in the solute−solute friction-dominated regime.[23,63] On the other hand, the kinetic rates are affected much more dramatically with larger friction coefficients in the regime where solvent−solute friction dominates. The ability to speed up kinetics through the use of implicit solvent depends on the characteristics of the solute−solute friction-dominated regime. On the basis of the results presented here, barrier crossings can be accelerated only by a maximum factor of about 2 with the GBMV method. This finding may be compared with results from Pande and Zagrovic[23] that suggest that protein folding kinetics could be accelerated up to about a factor of 20 with low friction coefficients when a different, thermodynamically less-accurate[17,39] generalized Born model is used.

A comparison of how individual kinetic rates are affected at reduced solvent viscosity indicates that different barrier crossings may not be accelerated uniformly. This finding could have implications for transition pathways extracted from such simulations. A recent study of protein folding that looked for such an effect did not find evidence that different pathways may be sampled as a function of solvent viscosity.[65] However, it is possible that folding funnels are sufficiently robust to accommodate subtle changes in individual barrier crossing rates. Further exploration of this aspect is needed to determine more conclusively to what extent reduced friction implicit solvent simulations might affect sampled kinetic pathways.

Temperature control of the solute with a Nosé−Hoover thermostat without any account of solvent viscosity greatly diminishes the ability to cross barriers in the alanine dipeptide system. This can be understood from a complete absence of stochastic collisions with the solvent that typically provide the kinetic energy for crossing a given barrier. From a statistical mechanics point of view, the coupling of only the solute with a Nosé−Hoover thermostat in the absence of explicit solvent also creates a different ensemble than a Nosé−Hoover thermostat applied to a solute plus a surrounding box of explicit water. In the latter case, much larger temperature fluctuations of the solute itself are allowed as kinetic energy is free to transfer between the solute and solvent through atomic collisions. Temperature control of the solute alone suppresses most of these temperature fluctuations in the implicit solvent systems. The result is a dramatic slowdown of the barrier crossing rates.

The kinetic rates and mean first passage times obtained in this study are also interesting in themselves, as there are only very few studies that explore alanine dipeptide kinetics on 100 ns time scales. The results can be compared to other theoretical studies where kinetic rates were obtained with Brownian dynamics methods on an implicit solvent surface[66] or through the extraction of state-to-state transition functions.[67,68] Although differences due to the choice of other force fields are expected, the agreement is reasonable. In the study of Chekmarev et al.,[66] mean first passage times between α and β basins are estimated to be 29 and 293 ps for forward and backward transitions, respectively, compared to 426 and 459 ps with explicit solvent and about half that with implicit solvent and a friction coefficient of 10 ps^{-1} in this study. Transitions between α and C7$_{ax}$, on the other hand, are found to take about 11 ns and 224 ps in the forward and backward directions compared with 9 ns and 744 ps with explicit solvent and 12 ns and 534 ps with implicit solvent in this study, respectively. The simulations presented here also support the preference for indirect transitions between C7$_{ax}$ and α that follow a path through α$_L$ and β rather than a direct transition between C7$_{ax}$ and α due to the significant barrier height near $\phi = 0$ and $\psi = -100$.[66]

The detailed analysis of kinetic rates in alanine dipeptide is contrasted with native state dynamics of two well-studied proteins: protein G and ubiquitin. Essentially, it is found that implicit solvent simulations closely reproduce results from explicit solvent and experiments. To the extent that differences exist, in particular in the protein G simulations, they likely indicate incomplete convergence of conformational sampling of a more flexible system compared to ubiquitin. The choice of different thermostats and Langevin friction coefficients could affect convergence rates of average structural and dynamic properties. However, no systematic differences can be found from the simulations presented here, suggesting that the long-time convergence of native state dynamics in simulations over tens of nanoseconds is not significantly affected by the thermostats tested here.

In contrast, differences in the exploration of conformational space during tens of picoseconds are clearly present. Conformational sampling in this time regime can be more efficient by a factor of 4−5 when a reduced friction coefficient is used in Langevin dynamics. With a Nosé−Hoover thermostat, the decorrelation of conformational sampling is even more rapid. This can be understood from the complete lack of friction and a continuous adjustment of the atomic velocities in the Nosé−Hoover method to maintain coupling with the thermal bath. As a result, it is expected that sampling can be accelerated significantly with the Nosé−Hoover thermostat as long as no significant kinetic barriers are present. This is the presumed case of conformational dynamics within the native basin of most macromolecules. However, if conformational changes over kinetic barriers that are large compared to *kT* are involved, such as in loop rearrangements or protein folding/unfolding, it is expected that significant differences as a function of the thermostat are manifest in a similar way as in the alanine dipeptide system. The practical question of which thermostat would be most suitable for a given system then depends on the particular application and expected dynamic properties. While the present study provides the first insight into the effect of different thermostats on the sampling of biological systems with implicit solvent, further studies are clearly needed to understand the effect of such a methodology on the sampling of larger conformational changes in biological macromolecules where implicit solvent is expected to be most beneficial.

Previous studies have indicated that different implicit solvent methods may significantly contribute to the solute−solute friction.[28] While only one implicit solvent method, the GBMV method, was considered here, it is likely that the general conclusions would apply in a similar fashion for other implicit solvent methods as well. However, there is some evidence that the optimal choice of solvent friction coefficients may be different in those cases.[28] Further studies will be needed to address this point more systematically.

Summary

This paper compares the kinetic properties of biomolecules in simulations with implicit and explicit solvent. The first important conclusion is that it is possible to closely match kinetic properties from explicit solvent by combining an implicit solvent with Langevin dynamics. Friction coefficients near 10 ps^{-1} appear to be optimal in conjunction with the GBMV implicit solvent model studied here. The second conclusion is that conformational sampling can be accelerated with reduced friction coefficients although the degree of acceleration depends on the circumstances. Local conformational exploration without the crossing of significant barriers can be accelerated substantially on the basis of native state simulations of proteins. However, when the crossing of significant kinetic barriers is involved, solvent viscosity appears to play a less important role. Consequently, reduced friction coefficients affect the kinetics only to a limited extent in agreement with previous studies of protein folding kinetics.[23] The use of Nosé−Hoover thermostats instead of Langevin dynamics in conjunction with an implicit solvent has quite dramatic consequences. Local conformational sampling is greatly enhanced over Langevin dynamics because of the lack of friction. However, the crossing of barriers is slowed down by at least an order of magnitude because of tight solute temperature control and the lack of stochastic collisions to provide the necessary kinetic energy to overcome barriers. For practical applications that may involve both the exploration of local conformational minima and the crossing of significant barriers, the results from this study indicate that Langevin dynamics with small friction coefficients of less than 5 ps^{-1} may offer substantially improved sampling over explicit solvent simulations while maintaining a thermodynamically accurate description of the simulated system.

Acknowledgment. Financial support from NSF CAREER grant 0447799 and the Alfred P. Sloan Foundation and access to computational resources through the Michigan Center for Biological Information at Michigan State University is acknowledged.

References

(1) Feig, M.; Brooks, C. L., III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.

(2) Baker, N. A. Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.* **2005**, *15* (2), 137−143.

(3) Roux, B.; Simonson, T. Implicit Solvent, Models. *Biophys. Chem.* **1999**, *78*, 1−20.

(4) Cramer, C. J.; Truhlar, D. G. Implicit solvation models: Equilibria, structure spectra and dynamics. *Chem. Rev.* **1999**, *99*, 2161−2200.

(5) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(6) Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144−1149.

(7) Sharp, K. A.; Honig, B. Electrostatic Interactions in Macromolecules - Theory and Applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301−332.

(8) Warwicker, J.; Watson, H. C. Calculation of the Electric Potentialin the Active Site Cleft due to α-Helix Dipoles. *J. Mol. Biol.* **1982**, *157*, 671−679.

(9) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.* **2002**, *15*, 377−392.

(10) Im, W.; Beglov, D.; Roux, B. Continuum Solvation Model: Computation of Electrostatic Forces from Numerical Solutions to the Poisson-Boltzmann Equation. *Comput. Phys. Commun.* **1998**, *111*, 59−75.

(11) Friedrichs, M.; Zhou, R. H.; Edinger, S. R.; Friesner, R. A. Poisson-Boltzmann analytical gradients for molecular modeling calculations. *J. Phys. Chem. B* **1999**, *103*, 3057−3061.

(12) Lu, B. Z.; Chen, W. Z.; Wang, C. X.; Xu, X.-j. Protein Molecular Dynamics With Electrostatic Force Entirely Determined by a Single Poisson-Boltzmann Calculation. *Proteins* **2002**, *48*, 497−504.

(13) Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson-Boltzmann Calculations for Static and Dynamic Systems. *J. Comput. Chem.* **2002**, *23*, 1244−1253.

(14) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(15) Yu, Z. Y.; Jacobson, M. P.; Friesner, R. A. What role do surfaces play in GB models? A new-generation of surface-generalized Born model based on a novel Gaussian surface for biomolecules. *J. Comput. Chem.* **2006**, *27* (1), 72−89.

(16) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55* (2), 383−394.

(17) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A. Brooks, C. L., III. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J. Comput. Chem.* **2004**, *25*, 265−284.

(18) Grycuk, T. Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **2003**, *119*, 4817−4826.

(19) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf. *J. Comput. Chem.* **2003**, *25*, 238−250.

(20) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116* (24), 10606−10614.

(21) Gallicchio, E.; Levy, R. M. AGBNP: An Analytic Implicit Solvent Model Suitable for Molecular Dynamics Simulations and High-Resolution Modeling. *J. Comput. Chem.* **2004**, *25*, 479−499.

(22) Brooks, C. L.; Berkowitz, M.; Adelman, S. A. Generalized Langevin Theory for Many-Body Problems in Chemical-Dynamics - Gas-Surface Collisions Vibrational-Energy Relaxation in Solids and Recombination Reactions in Liquids. *J. Chem. Phys.* **1980**, *73* (9), 4353−4364.

(23) Zagrovic, B.; Pande, V. Solvent Viscosity Dependence of the Folding Rate of a Small Protein: Distributed Computing Study. *J. Comput. Chem.* **2003**, *24*, 1432−1436.

(24) Hoover, W. G.; Aoki, K.; Hoover, C. G.; De Groot, S. V. Time-reversible deterministic thermostats. *Physica D* **2004**, *187* (1−4), 253−267.

(25) Nose, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52*, 255−268.

(26) Hoover, W. G. Constant-Pressure Equations of Motion. *Phys. Rev. A: At., Mol., Opt. Phys.* **1986**, *34* (3), 2499−2500.

(27) Zagrovic, B.; Sorin, E. J.; Pande, V. β-Hairpin Folding Simulations in Atomistic Detail Using an Implicit Solvent Model. *J. Mol. Biol.* **2001**, *313*, 151−169.

(28) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. Comparative study of generalized Born models: Protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6760−6764.

(29) David, L.; Luo, R.; Gilson, M. K. Comparison of Generalized Born and Poisson Models: Energetics and Dynamics of HIV Protease. *J. Comput. Chem.* **2000**, *21*, 295−309.

(30) Tsui, V.; Case, D. A. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.* **2000**, *122*, 2489−2498.

(31) Calimet, N.; Schaefer, M.; Simonson, T. Protein Molecular Dynamics With the Generalized Born/ACE Solvent Model. *Proteins* **2001**, *45*, 144−158.

(32) Ferrara, P.; Apostolakis, J.; Caflish, A. Evaluation of a Fast Implicit Solvent Model for Molecular Dynamics Simulations. *Proteins* **2002**, *46*, 24−33.

(33) Suenaga, A. Replica-exchange, molecular dynamics simulations for a small-sized protein folding with implicit solvent. *THEOCHEM* **2003**, *634*, 235−241.

(34) Shen, M.-y.; Freed, K. F. Long Time Dynamics of Met-Enkephalin: Comparison of Explicit and Implicit Solvent Models. *Biophys. J.* **2002**, *82*, 1791−1808.

(35) Krol, M. Comparison of Various Implicit Solvent Models in Molecular Dynamics Simulations of Immunoglobulin G Light Chain Dimer. *J. Comput. Chem.* **2003**, *24*, 531−546.

(36) Bursulaya, B. D.; Brooks, C. L., III. Comparative Study of the Folding Free Energy Landscape of a Three-Stranded β-Sheet Protein with Explicit and Implicit Solvent Models. *J. Phys. Chem. B* **2000**, *104* (51), 12378−12383.

(37) Chocholousova, J.; Feig, M. Implicit solvent simulations of DNA and DNA-protein complexes: Agreement with explicit solvent vs experiment. *J. Phys. Chem. B* **2006**, *110* (34), 17240−17251.

(38) Chocholousova, J.; Feig, M. Balancing an Accurate Representation of the Molecular Surface in Generalized Born Formalisms with Integrator Stability in Molecular Dynamics Simulations. *J. Comput. Chem.* **2006**, *27*, 719−729.

(39) Zhu, J.; Alexov, E.; Honig, B. Comparativestudy of generalized Born models: Born radii and peptide folding. *J. Phys. Chem. B* **2005**, *109* (7), 3008−3022.

(40) Chen, J. H.; Im, W. P.; Brooks, C. L. Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. *J. Am. Chem. Soc.* **2006**, *128* (11), 3728−3736.

(41) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. Investigation of salt bridge stability in a generalized born solvent model. *J. Chem. Theory Comput.* **2006**, *2* (1), 115−127.

(42) Zhu, J.; Shi, Y.; Liu, H. Parametrization of a Generalized Born/Solvent-Accessible Surface Area Model and Applications to the Simulation of Protein Dynamics. *J. Phys. Chem. B* **2002**, *106*, 4844−4853.

(43) Nymeyer, H.; Garcia, A. E. Simulation of the folding equilibrium of α-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934−13939.

(44) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

(45) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.

(46) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, J. D.; Evanseck, M. J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(47) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(48) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698−699.

(49) 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* (3), 327−341.

(50) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. New Analytical Approximation to the Standard Molecular Volume Definition And Its Application to Generalized BornCalculations. *J. Comput. Chem.* **2003**, *24*, 1348−1356.

(51) Im, W.; Lee, M. S.; Brooks, C. L., III. Generalized Born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691−1702.

(52) 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.

(53) Feig, M.; Karanicolas, J.; Brooks, C. L., III. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.

(54) Feig, M.; MacKerell, A. D.; Brooks, C. L. Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations. *J. Phys. Chem. B* **2003**, *107* (12), 2831−2836.

(55) Drenth, J. *Principles of Protein X-ray Crystallography*; Springer: New York, 1994; pp 89−91.

(56) Henry, E. R.; Szabo, A. Influence of Vibrational Motion on Solid-State Line-Shapes and Nmr Relaxation. *J. Chem. Phys.* **1985**, *82* (11), 4753−4761.

(57) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. 2 Crystal-Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein-G and Comparison with Nmr. *Biochemistry* **1994**, *33* (15), 4721−4729.

(58) Vijaykumar, S.; Bugg, C. E.; Cook, W. J. Structure of Ubiquitin Refined at 1.8 a Resolution. *J. Mol. Biol.* **1987**, *194* (3), 531−544.

(59) Alexeev, D.; Bury, S. M.; Turner, M. A.; Ogunjobi, O. M.; Muir, T. W.; Ramage, R.; Sawyer, L. Synthetic Structural and Biological Studies of the Ubiquitin System - Chemically Synthesized and Native Ubiquitin Fold into Identical 3-Dimensional Structures. *Biochem. J.* **1994**, *299*, 159−163.

(60) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity. *J. Am. Chem. Soc.* **1982**, *104* (17), 4546−4559.

(61) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. *J. Am. Chem. Soc.* **1982**, *104* (17), 4559−4570.

(62) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N′-Methylamide. *Biopolymers* **1992**, *32* (5), 523−535.

(63) Hagen, S. J.; Qiu, L. L.; Pabit, S. A. Diffusional limits to the speed of protein folding: fact or friction? *J. Phys.: Condens. Matter.* **2005**, *17* (18), S1503−S1514.

(64) Gee, P. J.; van Gunsteren, W. F. Numerical simulation of the effect of solvent viscosity on the motions of a beta-peptide heptamer. *Chem.*—*Eur. J.* **2005**, *12* (1), 72−75.

(65) Jagielska, A.; Scheraga, H. A. Influence of Temperature Friction and Random Forces on Folding of the B-Domain of Staphylococcal Protein A: All-Atom Molecular Dynamics in Implicit Solvent. *J. Comput. Chem.* **2007**, *28*, 1068−1082.

(66) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous, and discrete-state kinetic models. *J. Phys. Chem. B* **2004**, *108* (50), 19487−19495.

(67) Swope, W. C.; Pitera, J. W.; Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B* **2004**, *108* (21), 6571−6581.

(68) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B* **2004**, *108* (21), 6582−6594.

(69) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33.

(70) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. *J. Am. Chem. Soc.* **1999**, *121* (10), 2337−2338.

(71) Idiyatullin, D.; Daragan, V. A.; Mayo, K. H. (NH)-N-15 backbone dynamics of protein GB1: Comparison of order parameters and correlation times derived using various “model-free” approaches. *J. Phys. Chem. B* **2003**, *107* (11), 2602−2609.

(72) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **1998**, *120* (27), 6836−6837.

CT7000705

JCTC Journal of Chemical Theory and Computation

Meta-Optimization of Evolutionary Strategies for Empirical Potential Development: Application to Aqueous Silicate Systems

Brian C. Barnes and Lev D. Gelb*

Department of Chemistry and Center for Materials Innovation, Washington University in St. Louis, St. Louis, Missouri 63130

Received April 12, 2007

Abstract: The use of evolutionary strategy optimizations in fitting empirical potentials against first-principles data is considered. Empirical potentials can involve a large number of interdependent quantities, the number varying with the complexity of the potential, and the optimization of these presents a challenging numerical problem. Evolutionary strategies are a general class of optimization methods that mimic natural selection by stochastically evolving a population of trial solutions according to rules that select for high values of some fitness function. In this work we apply a variety of evolutionary optimization methods to a representative “parametrization problem” in order to determine which such methods are well-suited to such applications. Prior work on the design of evolutionary strategies has generally focused on finding the extrema of relatively simple mathematical functions, and the findings of such studies may not be transferable to chemical applications of very high dimensionality. The test problem consists of parametrization of the Feuston-Garofalini all-atom potential developed for simulation of silicic acid oligomerization in aqueous solution (Feuston, B. P.; Garofalini, S. H. *J. Phys. Chem.* **1990**, *94*, 5351). “Meta-optimization” of the evolutionary method is first considered by fitting this potential against itself, using a wide variety of population sizes, recombination algorithms, mutation-size control methods, and selection methods. Simulated annealing is also considered as an alternative approach. Optimal choices of population size, recombination operator, mutation size control approach, and selection method are discussed, as well as the quantity of data required for the parametrization. It is clear from comparisons of multiple independent optimizations that, even when fitting this potential against itself, there are a considerable number of local extrema in the fitness function. Evolutionary methods are found to be competitive with simulated annealing and are more easily parallelized. Finally, the potential is reparametrized against reference data taken from a Car−Parrinello Molecular Dynamics trajectory of several relevant silicate species in aqueous solution, again using several variant algorithms.

1. Introduction

Empirical potentials (force fields) are widely used in molecular modeling and simulation and usually consist of analytic functions which have been parametrized to reproduce selected reference data. The functional forms are chosen to model specific intermolecular and intramolecular interactions thought to be important for a given application. For instance, in the potentials commonly used for studying the phase behavior of fluids one generally includes terms describing atomic-core repulsions, dispersion forces, bond angles, and torsions; if dipolar or charged species are present, then these may be described using point dipoles or distributions of point charges. By inclusion of higher multipoles and/or polariz-

* Corresponding author e-mail: gelb@wustl.edu.

10.1021/ct700087d CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/16/2007

abilities, such potentials can become quite complex. The design of effective potentials has been discussed extensively in the simulation literature, and the functional forms used vary considerably from problem to problem.[1−5]

Parameters may be fit to a wide range of data, including both experimental results and quantities calculated using first-principles or semiempirical electronic-structure methods. Experimental data often used for this purpose include, among others, crystal structures, thermophysical properties such as melting points and critical parameters, partial radial distribution functions, angular distributions, and diffusion constants. Parametrization against thermophysical quantities requires the use of simulations to determine the corresponding properties of trial parameter sets, which can be computationally expensive.

With first-principles methods one may calculate the energies and associated gradients for selected molecular configurations as well as charge distributions, multipole moments, and structural quantities. Such data may be obtained either for isolated molecules or in the condensed phase. The parametrization of empirical potentials against first-principles reference data is now a popular and widely used approach,[6−16] building on both the broad availability of software for high-quality electronic structure calculations and general interest in multiscale simulation methods.

In all cases, systematic parametrization of the chosen functional form presents a challenging numerical problem. This may be cast as the optimization of an objective function that measures the ability of the empirical potential to reproduce selected reference data and therefore as a minimization in some high-dimensional space where the dimensionality is equal to the number of parameters to be assigned. In general, the properties of the objective function will depend on the physical system under consideration, reference data, potential form, and metric used to compare model results with reference data. For a given parametrization problem there may well exist a multiplicity of possible solutions, as pointed out in the early literature in development of the central force model for liquid water.[17−19]

Many strategies for parametrization of empirical potentials are available, varying in both computational complexity and “philosophy”.[1−5] One significant classification of these strategies is whether all parameters are considered simultaneously or if a sequential, one-parameter-at-a-time (or one-term-at-a-time) approach is used; the latter cases may also be iterated.

Iterated parameter-by-parameter optimizations correspond roughly to direction-set optimization methods[20] and therefore produce local minima of the objective function. Term-by-term optimizations (which may consider a few parameters at a time) are popular because they reflect the additivity of different interactions explicitly built into many potentials. For instance, one may parametrize a torsional motion independently of the associated angular terms by using an electronic structure program to scan over the torsional degree of freedom and then fit that data with some appropriately chosen function. The disadvantage of this approach is that the resulting torsion is then fit at particular values of the associated angles, and any dependence of the torsion on the associated angles will not be described well. To capture such interactions one must have both reference data that explores appropriate deformations of the molecule and additional terms in the potential that depend on both torsions and angles. In such a case, one may choose either to individually fit the torsion-only and angle-only terms and then fit the “cross” term or to fit all three parts simultaneously. The term-by-term approach allows for a better description of the isolated motions with inaccuracies concentrated in the cross-term, whereas the simultaneous fitting will spread inaccuracies more evenly among the three terms. Such issues become particularly important when extending previously developed potentials to include new atomic or molecular species. If the existing potential is not reparametrized to some degree, then the inaccuracies associated with the (necessarily imperfect) description of interaction with the new species will be concentrated in the added terms. Conversely, when all parameters are fit simultaneously this will not be the case, but parts of the potential may not be as accurate as the functional form allows, since global optimization may use them to compensate for some deficiency elsewhere in the functional form.

The ReaxFF family of reactive potentials,[6−11] for instance, is parametrized against small molecule calculations for bond distances and angles and experimental data for heats of formation. A local optimization technique of successive one-parameter optimization (line search) was used.[21] In an alternative approach, Voth et al. have used first-principles simulations of condensed phases to create potentials for water and hydrogen fluoride.[12,13] Their “force matching” technique uses a short-ranged cubic spline and a long-ranged Coulomb form to model site−site interactions. The linearly independent splines enable the use of singular value decomposition to exactly find parameters for a given configuration, and a final set of parameters is then determined by averaging over the results of many configurations.

For potentials describing a small number of degrees of freedom (and therefore either very small systems or species of low structural complexity) electronic structure calculations can be used to “scan” over the complete potential energy surface. These results can then be numerically interpolated, fit to analytical functions, or some combination of both, in order to obtain highly accurate potentials. Recent examples of such parametrizations include the water potential of Bukowski et al.[14] and the nine-dimensional potential for collisions of hydrogen gas and water monomers developed by Faure et al.;[15] there is a considerable literature on the development of such surfaces for use in reaction dynamics calculations.[22−25]

In the 1960s and 1970s, three groups developed independently numerical optimization methods which mimicked the process of evolution.[26,27] Rechenberg and Schwefel created a family of “evolutionary strategies” to solve real-valued problems.[28−32] Fogel researched artificial intelligence problems through an “evolutionary programming” technique.[33] Finally, Holland developed “genetic algorithms” as a general optimization method.[34] De Jong discusses all these methods under a unified framework of “evolutionary computation” and generalizes them as “evolutionary algorithms”.[35]

A brief outline of an evolutionary strategy is as follows. First, a population of trial solutions, called parents, is created. Second, a *recombination* process creates a group of children by averaging or otherwise combining parts of the parents. Third, the children undergo *mutation*, consisting of small random changes. Fourth, those children are *evaluated*. Fifth, a *selection* process is used to select a new group of parents from the current population. The cycle is then repeated, starting with the recombination step.

One important difference between evolutionary strategies and genetic algorithms is in the representation of trial solutions: evolutionary strategies are phenotypic, and genetic algorithms are genotypic.[36] That is, in an evolutionary strategy the individuals are manipulated "as-is", whereas genetic algorithms operate on bitwise representations. This difference in representation requires different operators for recombination and mutation steps. In genetic algorithms, recombination operators exchange strings of bits between two parents in order to generate children, and the basic mutation operator is a random bit flip. For a continuous-valued problem represented phenotypically, the recombination step would involve choosing or averaging values from the parents to create a child, and the simplest mutation would be the random displacement of selected child parameters.

Genetic algorithms (GAs) have been used in potential development in a number of studies, mostly to extend semiempirical methods or to refine popular force fields. Cundari, Deng, and Fu used a GA to parametrize technetium interactions in the semiempirical PM3 method. Their results were fit against crystal structure geometries, and they found that their GA provided significantly better parameters than those obtained by interpolating parameters of the metals to the left and right of technetium in the periodic table.[37] Rossi and Truhlar used a GA to reparametrize the AM1 semiempirical method against quantum mechanical data in order to perform semiquantitative direct dynamics on the $Cl + CH_4$ potential energy surface.[38] Parameters for organic systems containing sodium and transition metals in the AM1 and PM3 methods have also been refit using GAs.[39,40] These targeted reparametrizations can allow semiempirical methods to give substantially improved structures for biochemically relevant systems. Ge and Head used dual genetic algorithms in a study of Si_xH_y clusters, with one GA tasked to iteratively reparametrize the AM1 method, and the other GA to search cluster geometries for a global minimum.[41] GAs have also been used in computer-aided molecular design.[42] As reviewed by Lameijer et al.,[43] in the area of drug design evolutionary algorithms have been applied to the design of molecule libraries, conformational analysis, molecule superposition and pharmacophore detection, quantitative structure−activity relationships (QSAR), ligand docking, de novo design, and "druglikeness" evaluation. In particular, Thomsen investigated the effects of variation operators and local-search hybrid methods on EA/GA performance for ligand docking.[44]

Strassner et al. performed one of the few studies of the influence of GA parameters in the context of developing empirical potentials. They examined the interaction of crossover rates, mutation rates, and selection methods on the overall GA performance for refitting of the MM3 force field for a rhenium complex.[16,45] In this study, different GA parameter sets were compared via the root-mean-squared deviation (rmsd) between experimental (or high-level theoretical) crystal structures and those obtained using the GA-parametrized force field; GAs which produced MM3 parameters with smaller rmsds were judged to be more effective. Results were averaged over only three different independent optimizations for each set of GA parameters, and definite trends in GA performance with different parameters were observed. The most efficient algorithm tested was a simple GA with a tournament selector, 90% crossover rate, and 20% mutation rate. Wolohan et al. reparametrized the MM3 force field for copper complexes[46] using the GA parameters recommended by Strassner et al.[16,45] Other efforts at reparametrizing force fields using GAs include partial reparametrization of the AMBER force field,[47] refitting of the BKS and TTAM potential forms,[48] and refitting of the Stillinger-Weber potential for silicon.[49]

With the exception of the work of Strassner et al.,[16,45] the actual performance of the GAs used in potential parametrization work has rarely been considered in any depth. Many previous studies of the efficiency of evolutionary strategies have considered only the optimization of relatively simple and low-dimensional mathematical functions.[26,27] The behavior of an ES for much more complex problems may be distinctly different.

In this paper we evaluate the performance of a reasonable selection of evolutionary strategy algorithms applied to the problem of optimizing an empirical potential for molecular simulation applications. The process of finding the best algorithm for an optimization is termed a "meta-optimization". The empirical potential we consider is the all-atom, reactive potential for aqueous solutions of silicate oligomers developed by Feuston and Garofalini (FG).[50,51] Reparametrization of the FG potential is a useful test application because the short-ranged nature of the potential makes it inexpensive to evaluate, and optimization of the large number of parameters used poses a difficult numerical problem. The purposes of this work are to provide effective guidelines for future applications of evolutionary strategies in similar parametrization studies and to provide benchmarks for the behavior that can be expected of these algorithms.

2. Methodology

2.1. Evolutionary Strategy Optimizations. A complete evolutionary strategy implementation requires specification of initialization, recombination, mutation, evaluation, selection, and termination algorithms. In this work we evaluate the performance and behavior of a variety of recombination, mutation, and selection methods in the parametrization of an empirical potential against various reference data.

Individuals (parents and children) will be represented as vectors of real numbers $\mathbf{x} = \{x_i, \sigma_i\}$, $i = 1, ..., N$, where N is the number of parameters. The $\{x_i\}$ are the quantities to be optimized (in this case, parameters of an empirical potential), and the $\{\sigma_i\}$ are associated quantities that control the size of mutations applied to each parameter. The $\{\sigma_i\}$ may themselves be subject to evolution. The parts of the evolutionary strategy are presented below.

1. *Initialization.* In this step an initial population of m parents is created. Each parameter x_i of each parent is selected from a continuous uniform distribution within a constrained range, $x_i^{\min}$ to $x_i^{\max}$, which are part of the initial input. The initial values of the $\{\sigma_i\}$ are defined through scaling of an input parameter σ_0: $\sigma_i = \sigma_0 \cdot (x_i^{\max} - x_i^{\min})$. This reflects the fact that the absolute values of the x_i can vary by many orders of magnitude, depending on the units and functional forms used.

2. *Recombination.* Recombination is the process of combining parents to produce children. Following Schwefel, recombination operators are classified as *local* or *global* and also as *discrete* or *intermediate*.[52] Local operators generate a child entirely from two randomly selected parents. Global operators randomly select a new pair of parents *for each parameter* of every child. Discrete operators assign each (x_i, σ_i) pair for the child by setting them equal to the value of the corresponding (x_i, σ_i) pair in one of the randomly chosen parents. Intermediate operators instead assign the average value of the corresponding parent parameters to the child. Selections are made "without replacement", so that it is not possible to create a child from two "copies" of a single parent.

3. *Mutation.* Each parameter x_i in each child $\mathbf{x}$ is displaced with probability p by a random number chosen from a normal distribution of zero mean and standard deviation σ_i, $G(0, \sigma_i)$. This change is represented as

$$x_i = x_i + G(0, \sigma_i) \tag{1}$$

The σ_i control the size of mutations. As discussed in greater detail below, different mutation algorithms may independently evolve the $\{\sigma_i\}$ over the course of the optimization. Alternatively, the $\{\sigma_i\}$ may be controlled through a common reference σ, with $\sigma_i = \sigma \cdot (x_i^{\max} - x_i^{\min})$; various algorithms for evolving σ may then be applied.

4. *Evaluation.* The fitness of each new child is evaluated, as described in the next section.

5. *Selection.* In the selection step, the parents of the next generation are selected from the current population. Selection methods may be categorized according to (a) whether or not they allow overlapping generations and (b) their degree of *elitism*.

Evolutionary strategies are commonly labeled either (m, n)-ES or ($m + n$)-ES, where m is the number of parents and n is the number of children per generation.[53] An (m, n)-ES is nonoverlapping: the m parents of the next generation are chosen only from among the n children of the current generation. An ($m + n$)-ES is overlapping: the m parents of the next generation are chosen from the entire current population of $n + m$ individuals. This allows for the survival of individuals for more than one generation and potentially indefinitely.[52]

Elitism describes the importance placed on fitness when selecting parents. *Truncation* methods are the most elitist, and simply choose the best m individuals from the available population (n or $m + n$). A less elitist method is *binary tournament* selection, in which m random pairs are chosen from the available population, and the "winner" of each pair becomes one of the parents for the next generation.[54] With tournament methods, it is possible that the individual with the highest fitness is not selected. The tournament method may be extended to have competitions between an arbitrary number of children when creating a child, e.g., a three-way tournament instead of a binary (two-way) tournament. The truncation selection method is deterministic, while the tournament method is stochastic. We use the term *semiooverlapping* to refer to selection methods which, when choosing new parents (from either the children or from the full population), always include either the best current parent or the best current individual.

After selection, one generation is complete. The m selected individuals now become the parents, and the algorithm returns to the recombination step.

6. *Termination.* Common termination options include exiting after a certain fitness has been achieved, exiting when the fitness of the fittest individual becomes constant to within a specified tolerance, or exiting after a fixed number of generations. In the studies below, which compare various algorithms, termination criteria are chosen to ensure that the computational costs of the different methods are comparable. For algorithms with the same m and n, this corresponds to termination after a fixed number of generations, but for comparisons of algorithms with different m and n, optimizations are terminated after a fixed number of child evaluations, or "births".

2.2. Fitness Function. Our goal in potential parametrization is to have the empirical potential accurately reproduce some reference data, which we will call the *training set*. Here the training set will consist of the total energies of a series of N_{config} atomic configurations. The *fitness function* is defined as

$$\chi^2(\mathbf{x}) = \frac{1}{N_{\text{config}}} \sum_i^{N_{\text{config}}} [\{E_{\text{emp}}(\mathbf{R}_i, \mathbf{x}) - E_{\text{emp}}(\mathbf{R}_{\text{ref}}, \mathbf{x})\} - \{E_{\text{TS}}(\mathbf{R}_i) - E_{\text{TS}}(\mathbf{R}_{\text{ref}})\}]^2 \tag{2}$$

where $E_{\text{emp}}(\mathbf{R}_i, \mathbf{x})$ is the energy of configuration $\mathbf{R}_i$ determined using the empirical potential with parameters $\mathbf{x}$. $E_{\text{TS}}(\mathbf{R}_i)$ is the energy of configuration $\mathbf{R}_i$ determined using some high quality method, for instance Density Functional Theory (DFT). $\chi^2(\mathbf{x})$ is a measure of the mean-squared difference between the potential energy surfaces sampled by the training set and defined by the chosen empirical functional form and parametrization $\mathbf{x}$.

$\mathbf{R}_{\text{ref}}$ denotes a reference configuration, which is included in the definition of χ^2 because the empirical potential and reference method may differ in ways which make absolute comparisons of their energies impossible. For instance, the energies obtained from typical all-atom empirical potentials cannot be directly compared with the "raw" output of electronic structure calculations. This is because in electronic structure methods even isolated atoms have nonzero total energy due to their internal structure, which is generally not the case for empirical potentials. One possible solution to this problem is to use the energy at the dissociation limit (all atomic separations increased to infinity) to define the energy "zero" in each case, which corresponds to a particular choice of $\mathbf{R}_{\text{ref}}$. However, for many empirical potentials, including nondissociable molecular potentials and potentials that include nonintegral charges, this is an awkward choice. In this work, we chose the lowest-energy configuration in the training set as the reference state $\mathbf{R}_{\text{ref}}$. This choice is applicable regardless of the form of empirical potential used and requires no additional "reference" calculations. Furthermore, it has the appeal of directly including the differences in energy between "relevant" configurations of the reference system, which appear in the Boltzmann factors determining the thermodynamic properties of the system.

2.3. Application. Our test problem for meta-optimization of evolutionary strategies is a reparametrization of the Feuston and Garofalini (FG) potential for aqueous solutions of silicate oligomers.[50,51] The FG potential includes a modified Born-Mayer-Huggins[55,56] functional form and Rahman-Stillinger-Lemberg[18] (RSL) terms for two-body interactions, and three-body terms as introduced by Stillinger and Weber:[57]

$$V_2(\mathbf{r}_i, \mathbf{r}_j) = A_{ij} \exp\left(\frac{-r_{ij}}{\rho_{ij}}\right) + \frac{Z_i Z_j}{r_{ij}} \operatorname{erfc}\left(\frac{r_{ij}}{\beta_{ij}}\right) + \sum_{m=1}^{D_{ij}} \frac{a_{ij,m}}{1 + \exp(b_{ij,m}(r_{ij} - c_{ij,m}))} \quad (3)$$

$$V_3(\mathbf{r}_{ij}, \mathbf{r}_{ik}, \theta_{jik}) = \lambda_{jik} \exp\left[\frac{\gamma_{ij}}{r_{ij} - r_{ij}^{\text{o}}} + \frac{\gamma_{ik}}{r_{ik} - r_{ik}^{\text{o}}}\right] \times (\cos \theta_{jik} - \cos \theta_{jik}^{\text{o}})^2 \quad (4)$$

The two-body part has a damped Coulomb potential, an exponential repulsion, and a soft (and short-ranged) attraction. Note that a different number D_{ij} of RSL terms are used for each type of two-body interaction involving hydrogen (Si−H, O−H, and H−H). The three-body term penalizes deviation from a specified angle θ_{jik}^{o}, controlled by parameters for cutoff distance, magnitude, and rate of decay. This is an all-atom, dissociable potential and can be used to study chemical reactions in solution, including the hydrolysis and condensation of siloxane bonds and the early stages of sol−gel processing.[50,51,58]

The FG potential was fit to thermophysical quantities including the radial distribution functions and angular distribution functions of melt-quenched silica. The short-ranged repulsive term was parametrized using a formula based upon ionic radii and charges. The other parameters were chosen based on hydrogen-bond energies, cluster geometries, and liquid-state properties extracted from molecular dynamics simulations, although how trial parameter sets were chosen for these simulations was not described.

The FG potential has two-body parameters for all combinations of the elements Si, O, and H, and parameters describing four different three-body combinations (Si−O−Si, H−O−H, O−Si−O, and Si−O−H), for a total of 55 adjustable parameters. In this work, 45 were optimized, and 10 were kept at fixed values because of physical arguments. The fixed parameters include the charges on each atom type, five three-body cutoff distances r_{ij}^{o}, and the four preferred angles θ_{jik}^{o}. The atomic charges were kept at their formal values (+1 for hydrogen, +4 for silicon, −2 for oxygen) so that dissociation produced ions with the correct integer charges. The three-body cutoff distances and angles ensure that all silicon and oxygen atoms prefer tetrahedral geometries, except for those oxygens in a waterlike environment, which prefer the experimental angle of 104.5° found in liquid water.

2.4. Training Sets. Two types of training set were used in this paper, both consisting of configurations sampled from molecular dynamics simulations of an aqueous solution of three silicate species. Each configuration in both sets contained one of each of silicic acid, disilicic acid, and cyclotrisilicic acid molecules and 64 water molecules, in a cubic box of 1.4014 nm edge length for a total density of 1.0 g/cm^3.

The first type of reference data, used below in the meta-optimization of the evolutionary strategy, consisted of configurations sampled according to the FG potential and the associated FG energies. These data were generated using a molecular dynamics trajectory thermostatted (via the Gaussian isokinetic method[2]) at 300 K, with configurations sampled at intervals of 2 ps. As in previous studies using this potential, interactions were truncated at 7 Å.

The second training set was generated using Car−Parrinello Molecular Dynamics (CPMD) simulations,[59] also in the canonical ensemble. In these calculations the Perdew-Burke-Ernzerhof (PBE) functional[60] was used with a plane-wave basis with 30 Rydberg cutoff for the wave function and 150 Rydberg cutoff for the density. Vanderbilt ultrasoft pseudopotentials were used for all atoms.[61] The silicon pseudopotential featured a nonlinear core correction. This level of theory was checked by comparing optimized bond distances, bond angles, and hydrogen bond strengths with similar data obtained with the same PBE functional and the 6-31G* and cc-pVTZ basis sets in Gaussian03.[62] The plane-wave results were closer to the 6-31G* basis results, giving bond lengths within 0.005 Å and similar hydrogen bond strengths.

Four visibly and temporally distinct configurations were selected from the first training set. These were used as the starting points for the CPMD simulations. For each configuration, the following procedure was followed. First, each configuration was optimized to a root-mean-square force of 0.005 au. Next, the configuration was relaxed through a series of 11 200-step CPMD simulations using a 3.0 au time step and a 400.0 au fictitious mass for the electrons. A velocity rescaling thermostat was used, with a target temperature of 300 K and rescaling whenever the temperature of the ions was more than 37.5 K away from the target value. After the first six 200-step simulations, the convergence criterion for the gradient of the wave function was tightened from 10^{-5} to 10^{-6} au. Between each 200-step simulation the electrons were quenched back to the Born−Oppenheimer surface. After the relaxation procedure was finished, the production CPMD run was started. The production run used a Nosé-Hoover thermostat for each degree of freedom.[63] The temperature was 300 K with a thermostat frequency of 2500.0 cm^{-1} for the ions and 10 000.0 cm^{-1} for the electrons. These simulations ran for 10 000 steps, giving a total of 242 fs of data in each of the four CPMD simulations or nearly 1 ps total data. From these trajectories, 370 evenly spaced configurations were selected. Single point energies were then calculated for each configuration; these differ slightly from the CPMD energies because during the dynamics run the electrons are not quenched to the Born−Oppenheimer surface at each time step. These configurations and single-point energies make up the second training set. The program "CPMD version 3.9.2", was used for these calculations.[64]

2.5. Implementation. We have developed a computer code to optimize empirical potentials against training sets of the type described above. Our program implements several optimization techniques, including evolutionary strategies, a simple direct search minimizer, an unconstrained Powell line search algorithm, simplex simulated annealing,[65] and Metropolis simulated annealing.[66] Several potentials are implemented, including the Lennard-Jones model, central force water model, FG model, and a charge-transfer model.[67] Additional potentials may be easily added.

The program is parallelized in two ways. In evolutionary strategy optimizations, evaluation of the fitness of the *n* children in each generation is divided over many processors by assigning some number of children to each processor. In other optimization techniques, which do not involve the simultaneous evaluation of many trial solutions, the evaluation of a single χ^2 may be parallelized by the distribution of training-set configurations among multiple processors and the simultaneous evaluation of many of the $E_{\text{emp}}(\mathbf{R}_i)$ terms. Evolutionary strategy speedups were found to be nearly ideal using up to 16 processors, while the training-set decomposition approach is slightly less efficient due to the increased quantity of communication required. The parallel scalability is also different for the two approaches. For algorithms that only evaluate one trial solution at a time, the theoretical maximum number of processors that can be used is equal to the number of configurations in the training set. Evolutionary strategies, on the other hand, evaluate many individuals in parallel, with each processor handling an equal number of individuals. Therefore, if a very large number of processors is available (as is increasingly the case with modern multicore processors), cases where $n > N_{\text{cpu}} > N_{\text{config}}$ allow evolutionary strategies to scale higher than other methods. Finally, evolutionary strategies can be further parallelized by distributing the evaluation of each χ^2 among several processors (as in the single-evaluation methods), which could then be used even for $N_{\text{cpu}} > n$, and for all methods, even the evaluation of the energy of a single configuration could be spread across several processors using either domain-decomposition or replicated-data strategies.

3. Meta-Optimization of Evolutionary Strategies

The evolutionary strategy may itself be optimized for a particular class of problems by selection of appropriate population sizes, recombination methods, mutation size control schemes, and selection methods. In this study this will be accomplished by optimizing the FG functional form against reference data (training sets) generated using the FG potential itself. Since the functional form is unchanged, it is in principle possible for an optimization algorithm to reduce χ^2 to zero (within some numerical tolerance), which would occur at the exact FG parameters; $\chi^2(\mathbf{x}_{\text{FG}}) = 0$. Different ES algorithms will approach this limit more or less quickly and with different "profiles" of χ^2 vs generation.

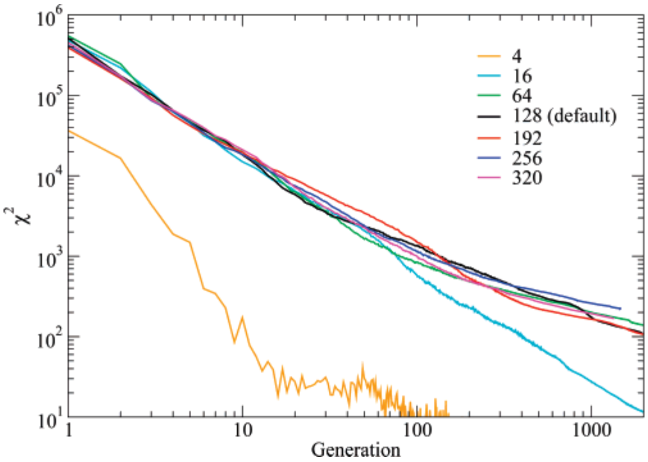


Figure 1. Variation of optimization profile with number of configurations in the training set. Training set sizes used ranged from 4 to 320 configurations. The quantity plotted is the fitness of the fittest (lowest χ^2) member of the current parent population at each generation, averaged over ten independent runs.

Testing different evolutionary strategies is accomplished here by first selecting a "default" combination of population size, recombination method, selection method, etc. and then considering and comparing several alternatives for each of these components. Note that this approach does not consider all possible combinations of methods but does allow for controlled comparisons of different variants of the same operator (for instance, mutation size control schemes).

The default options were selected based on a large number of preliminary trials and recommendations from the literature discussed above. They consist of populations of $m = 8$ and $n = 96$, local discrete recombination, mutation size control using evolving independent σ_i and an initial $\sigma_0 = 0.03$, and nonoverlapping truncation selection.

Unless otherwise stated, all individual optimizations were truncated after 192000 function evaluations, which took roughly 27 wall-clock hours running on two Opteron 250 (2.4 GHz) CPUs. The simulation code was parallelized using MPI. Near-linear scaling was observed in additional tests on up to 16 CPUs; all calculations were performed on a cluster of dual-processor nodes each with 2−4 GB of RAM and networked using Infiniband interconnects.

3.1. Preliminary Studies. In eq 2, each configuration provides only one energy datum. Therefore, the number of configurations in the training set must exceed the number of parameters to be optimized. Training set size may affect the reliability, speed, and smoothness of optimizations. These effects are shown in Figure 1, which compares training sets of varying sizes. Each optimization profile in Figure 1 consists of the lowest parent χ^2 at each generation, averaged over ten independent optimizations (see below). Two of the profiles, using 4 and 16 configurations, are for optimizations against too few independent data to be meaningful. These optimizations have considerably different profiles than the others, rapidly finding parameter sets with very low χ^2, which is perhaps not surprising given that in these cases this can be satisfied in a large fraction of parameter space.

All the other traces are quite similar, both in the shape of the profile and the lowest χ^2 reached after the allotted simulation time. In Figure 1 all optimizations were run to between 1500 and 2000 generations. Based on the similarity of these data, a training set size of 128 configurations was chosen for use in all the calculations that follow. This is significantly greater than the number of free variables (45) and requires less CPU time than the larger sets of 192, 256, or 320 configurations while clearly retaining the same general properties.

Data are plotted in log−log form in this and subsequent figures. It is therefore important to note that the absolute decrease in χ^2 is much larger in the early generations than in later ones. The units of χ^2 are [(kJ/mol) per configuration]^2. The initial values of $\chi^2 > 10^5$ (kJ/mol)^2 correspond to the randomly generated parent populations described above, which are clearly of poor quality. The final values of χ^2 (for meaningfully large training sets) do not converge to zero in the allotted number of generations but instead tend to reach values near 100 (kJ/mol)^2. The meaning of this value can be assessed by performing simple perturbations of various parameters from their original FG values and measuring the resulting change in χ^2. This measure can then be averaged over perturbations of all the parameters. Single-parameter perturbations of ±0.1% increase χ^2 to 0.392 (kJ/mol)^2, on average. Deviations of ±1% increase χ^2 to 39.17 (kJ/mol)^2, on average, and deviations of ±10% increase it to 3812.5 (kJ/mol)^2, on average. Thus, final values near 100 (kJ/mol)^2 correspond roughly to parameters that have converged to within 1% of their optimal values. However, the sensitivity of χ^2 to such deviations varies considerably from parameter to parameter. Sensitive parameters include the ρ_{ij} parameters for the Buckingham exponential repulsions between oxygen and hydrogen atoms and oxygen and silicon atoms, and the position c_{ij} of the second RSL oxygen-hydrogen term (which is important for modeling hydrogen bonding).

In any single optimization run, χ^2 fluctuated strongly because the recombination and mutation steps are stochastic. In order to make meaningful comparisons of different ES algorithms, we therefore present χ^2 profiles averaged over multiple independent runs. “Independent” in this case means differently reseeding the random number generator for each run after generation of the initial population. The different runs therefore have the same “starting point”. We determined that ten independent runs were sufficient to reliably profile different evolutionary strategy variants. This was done by performing 20 runs and then comparing the averaged profiles of two different sets of ten runs with the average profile of all 20 runs. As shown in Figure 2, the average of either set of ten runs is quite similar to the average of all 20 runs. Note that this is not the case for averages over only three independent runs, as used by Strassner et al.[16,45] Each of the 20 individual runs is also plotted in order to illustrate the magnitude of variation between runs. It is clear that the shape of the optimization profile can vary considerably from run to run and also that the final fitness values can vary by approximately one order of magnitude between runs started from the same initial population. As in Figure 1, all subsequent figures will show the χ^2 for the best parent in each generation averaged over ten runs, unless noted otherwise. Further analysis of the variations between individual runs will be presented in section 5, below.

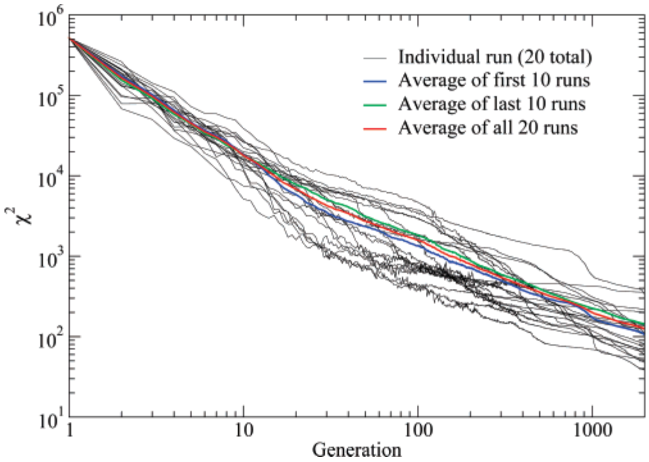


Figure 2. Variation of optimization profile with random number sequence. Twenty independent runs (starting from the same initial population) are shown, along with averages over the full set of 20, the first 10, and the last 10. Run conditions are the “default” algorithm, corresponding to the 128-configuration data shown in Figure 1.

Genetic diversity is a measure of the difference between members of a population. If members of the population differ only slightly, then a population has low genetic diversity. We measure this through a radius of gyration R_g, defined as

$$x_i^{\text{ave}} = \frac{1}{m}\sum_{j=1}^{m} x_{i,j} \tag{5}$$

$$R_g{}^2 = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\frac{x_{i,j}}{x_i^{\text{ave}}} - 1\right)^2 \tag{6}$$

where $x_{i,j}$ is the value of parameter i in parent j. Genetic diversity is an important quantity in ES optimizations. If there is too little genetic diversity, then the entire population will become trapped in a single minimum. While this is generally the end result of an evolutionary optimization, it is important that it not happen too early in the calculation, before a large part of parameter space has been explored. R_g data for the default ES strategy are shown in Figure 3. This is a strongly fluctuating quantity but shows clear structure. The initial R_g is large. After approximately ten generations (corresponding to a reduction of χ^2 from approximately 5×10^5 to around 10^4, see Figure 2) R_g drops to a plateau near 0.3, where it remains for approximately 250 generations. Over this period χ^2 decreases by another two orders of magnitude. After this, R_g begins to diminish quickly, becoming very small by the late generations.

3.2. Population. For an (m, n)-ES, a parent:child (m:n) ratio of 1:4 has been recommended,[26] although many studies use larger ratios.[68] Having a very high ratio of children to

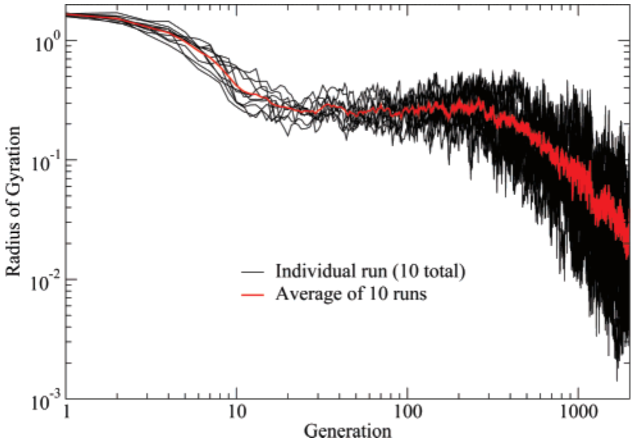


Figure 3. Radius of gyration for 10 individual runs and their average. This calculation corresponds to the 128-configuration data shown in Figure 1.

parents is considered inefficient, since the vast majority of computational time is spent evaluating individuals which do not survive to the next generation. However, in preliminary work we found that a *m*:*n* ratio of 1:12 seemed more effective. The effects of changing the numbers of children and parents, and the ratio *m*:*n*, are therefore of interest in further optimizing the ES approach.

In Figure 4, (*m*, *n*)-ES choices of (8,96), (8,16), (1,8), (48,-96), (8,384), and (8,48) are compared, labeled P-1−P-6, respectively. As explained above, each variant was terminated after a total of 192 000 fitness function evaluations, corresponding here to different numbers of generations. The best initial fitness value among the parents for any (*m*, *n*)-ES with the same number of parents is the same. The profile of P-3 (1,8) has a slightly worse initial best fitness than any $m = 8$ ES, while P-4 (48,96) has an initial best fitness over five times smaller than any m=8 ES. This is not surprising: a population with $m = 48$ instead of $m = 8$ has a much larger probability of generating an initial parent with low χ^2.

Comparing the P-1 (8,96) and P-4 (48,96) data shows the benefit of having a smaller parent:child ratio. In P-4, χ^2 actually increases over the first few generations. This can occur when the fittest parents are either not chosen in the recombination step or chosen so infrequently that a child more fit than those parents is not produced. As the selection method in the default strategy does not allow parents to survive to the next generation, the fitness of the best individual may increase from generation to generation.

P-3 is less effective than the other strategies throughout, but especially at early times. With only one parent, there cannot be recombination. Therefore, fitness can only be improved by random mutation of the single initial parent. Distinct jumps can be seen near generations 200, 600, and 1100, when especially productive mutations occurred. These data are again averaged over ten independent runs, and each of these jumps actually corresponds to a very large drop in χ^2 in an individual run.

Comparing strategies with $m = 8$ shows that an increase in the number of children leads to larger decreases in χ^2 per generation during the early stages of the optimization. P-5 (8,384) has the largest initial decreases in χ^2 per generation, followed by P-1 (8,96), P-6 (8,48), and P-2 (8,16), in that order. However, the use of large numbers of children is generally avoided because it is both computationally more expensive (per generation) and it tends to more quickly reduce genetic diversity. This can be understood as follows. In the (8,384) optimization, there are only 36 unique pairs of parents, each of which will produce, on average, 10.67 children per generation. If the children of a single pair of parents are particularly fit and truncation selection is used (as is the default here), then the *entire* next generation of parents may consist of the offspring of that pair of parents and will have very low genetic diversity. As the ratio of *m* to *n* is increased, more of the current group of parents will likely contribute to the next generation, and genetic diversity will be preserved. Of the populations tested in Figure 4, P-1 (8,96) achieves the lowest χ^2 after the allotted time and appears to make the most effective compromise between genetic diversity and χ^2 reduction per generation. This finding has implications for the use of evolutionary methods on massively parallel computers. Increasing the number of children, *n*, may appear to be an efficient way to utilize many

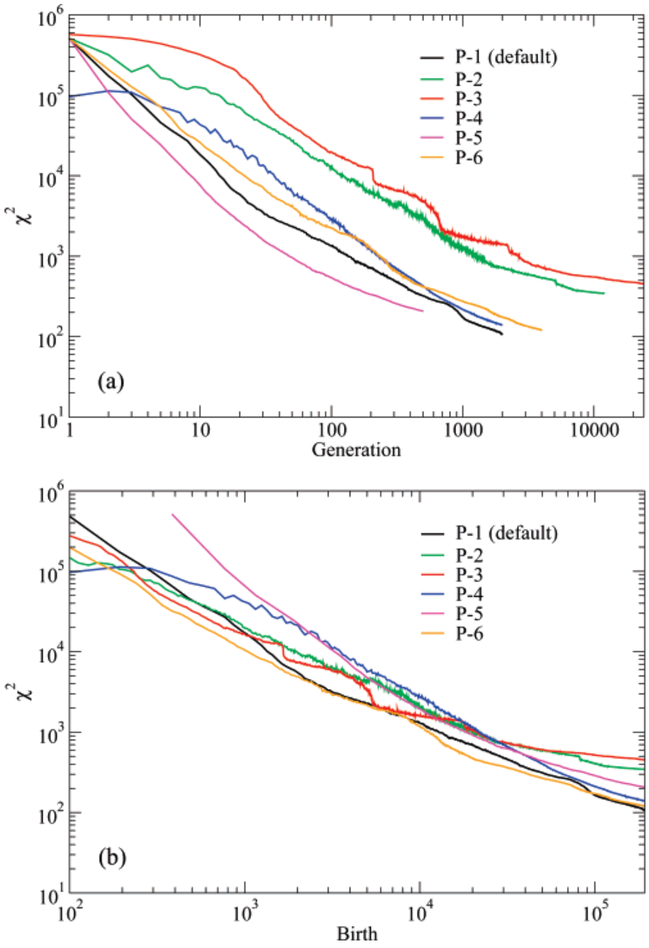


Figure 4. Variation of optimization profile with numbers of parents and children. Tested are (P-1) 8 parents and 96 children, (P-2) 8 parents and 16 children, (P-3) 1 parent and 8 children, (P-4) 48 parents and 96 children, (P-5) 8 parents and 384 children, and (P-6) 8 parents and 48 children. Top: optimization profiles vs number of generations. Bottom: optimization profiles vs number of births.

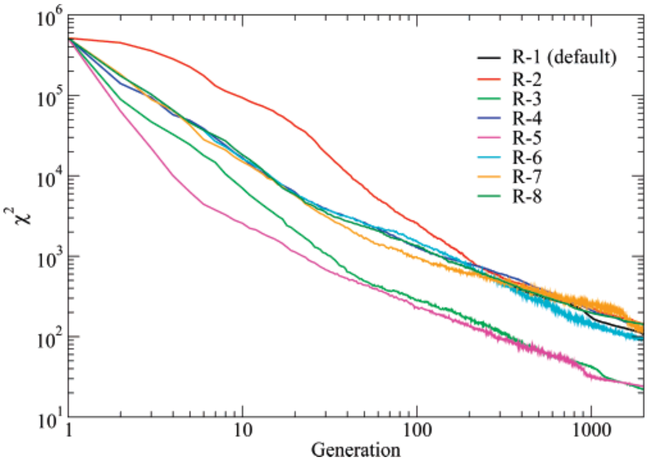


Figure 5. Variation of optimization profile with choice of recombination operator. Operators tested include (R-1) local discrete, (R-2) none, (R-3) local intermediate, (R-4) global discrete, (R-5) global intermediate, (R-6) local discrete for parameters and intermediate for σ, (R-7) global discrete for parameters and intermediate for σ, and (R-8) local discrete for the first 250 generations, none for the subsequent 1750.

processors in an optimization, but then *m* must likewise be increased to prevent loss of diversity. Furthermore, increasing both *m* and *n* does not necessarily improve the rate of convergence of the algorithm in a cost-effective way; this is easily seen in Figure 4b, wherein the performance of method P-6 measured against the number of births is clearly superior at nearly all times to the other algorithms, with P-1 pulling very slightly ahead after 10^5 births.

3.3. Recombination. By default we have used local, discrete recombination. This is the most commonly used recombination operator and is procedurally similar to the method used in genetic algorithms. Various recombination operators are compared in Figure 5. The two intermediate operators (local, R-3, and global, R-5) are seen to provide the most efficient recombination.

After approximately 250 generations, using no recombination at all (R-2) gave results equivalent to local discrete recombination (R-1). This was an unexpected result and suggests that recombination is most effective in the early generations of an optimization. After the first 250 generations, all the optimization profiles have similar slopes, suggesting that after this time the optimization is controlled by mutation instead of recombination. If recombination was still important in the later generations, we would expect the profiles in Figure 5 to differ significantly at late times. Intermediate operators (R-3 and R-5) produce better results overall due to their clear superiority during the early generations; these recombination operators eventually located parameter sets with χ^2 (again, averaged over ten independent runs) only 1/5 that of the typical result of the other operators.

These findings are consistent with the genetic diversity data of Figure 3, where a substantial drop-off in genetic diversity is observed after approximately 250 generations. Once a population is sufficiently inbred, it is unlikely that recombination can lead to substantial improvements in fitness, since the parents are already all very similar. This is investigated by performing an optimization using the default ES parameters (as in R-1) but then disabling all recombination after 250 generations. These results (R-8) overlap with those obtained with the default (R-1) until roughly 1000 generations, after which the default improves very slightly over the modified version, as shown in Figure 5. This behavior is consistent with the hypothesis that recombination is not a substantial contributor to further improvement in fitness after the drop-off in genetic diversity.

It has been suggested that using a discrete operator for the parameters x_i and an intermediate operator for the σ_i is more effective than using either fully discrete or fully intermediate operators.[52] Our results show that this is not the case in this application and that the use of an intermediate operator for the parameters x_i is the key factor. Fully intermediate operators R-3 and R-5 are clearly much more efficient than operators R-6 and R-7, which apply discrete recombination to the x_i and intermediate recombination to the σ_i. The similarity between R-3 and R-5 after the first 50 generations suggests that there is no substantial difference between local and global recombination operators in this application.

3.4. Mutation Size Control. Mutation operators must be included in ES optimizations because recombination operators alone cannot fully search the available parameter space. For instance, when using intermediate operators, the averaging of parameters would mean that children with x_i values outside of the largest and smallest x_i in the current group of parents would never be generated. Likewise, when using discrete recombination operators, the only children that could be created would be combinations of parameters already in the population.

While all mutations involve Gaussian perturbations, the size of these perturbations may be controlled in various ways. It is considered advantageous to have large mutations at the beginning of the optimization, which helps to search across the range of allowed values. However, at later times smaller mutations may be desirable as they can allow near-optimal parents to produce children that are “refinements” of themselves; this is analogous to the very small steps taken by conventional optimization techniques as they approach an extrema. Therefore, the absolute size of mutations should be gradually reduced.[26] The method used for this may also attempt to promote genetic diversity.

The default method used here, labeled M-1, is to use an independent σ_i for each parameter x_i. Following Beyer and Schwefel,[26] the σ_i are generated through a recombination process (as above) and then mutated via

$$\sigma_i^{\mathrm{child}} := \sigma_i^{\mathrm{child}} \cdot \mathcal{S}_g \cdot \mathcal{S}_i \tag{7}$$

where the two mutation operators $\mathcal{S}_g$ and $\mathcal{S}_i$ are

$$\mathcal{S}_g = \exp(\tau_g \cdot G(0, 1)) \quad \tau_g = \frac{1}{\sqrt{2N}} \tag{8}$$

$$\mathcal{S}_i = \exp(\tau_l \cdot G(0, 1)) \quad \tau_l = \frac{1}{\sqrt{2\sqrt{N}}} \tag{9}$$

$\mathcal{S}_g$ is calculated independently for each child and used for all the σ_i; this acts as a global scaling of mutation size, while

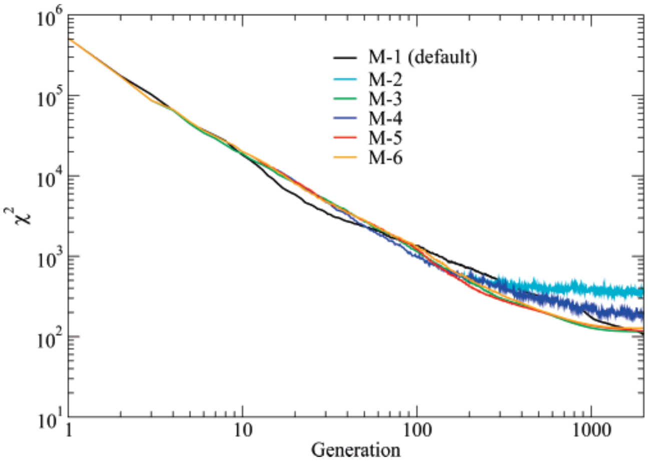


Figure 6. Variation of optimization profile with mutation size control algorithm. Algorithms tested include (M-1) independent $\{\sigma_i\}$, (M-2) constant σ, (M-3) annealing σ by a constant factor, (M-4) adjustment of σ relative to early χ^2, (M-5) a history-dependent, diversity-preserving algorithm, and (M-6) an alternative history-dependent, diversity-preserving algorithm.

the $\mathcal{S}_i$ are calculated independently for each i for each child, allowing for variations in mutation size between parameters.

The simplest mutation size control operator is to fix σ for the entire length of the optimization. Method M-2 demonstrates such a constant global σ.

Method M-3 is referred to as "simple annealing". Here, a global σ is reduced by a constant factor every generation: $\sigma := \sigma \cdot c_\sigma$ where $0 < c_\sigma < 1$. For the profile in Figure 6, $c_\sigma = 0.995$. Note that M-2 may be considered a special case of M-3.

Method M-4 introduces history dependence. It sets σ by scaling σ_0 by the square root of the current average value of the parents' fitness divided by the average value of the parents' fitness after an initial equilibration period. This equilibration period is determined as the end of the initial rapid decrease in χ^2. Specifically, for generation $g > 100$, once $\langle\chi^2\rangle(g) \geq 0.9\langle\chi^2\rangle(g - 100)$, we set $\langle\chi^2\rangle_{\text{ref}} = \langle\chi^2\rangle(g)$ and proceed according to

$$\sigma = \sigma_0 \times \left(\frac{\langle\chi^2\rangle_{\text{parents}}}{\langle\chi^2\rangle_{\text{ref}}}\right)^{(1/2)} \tag{10}$$

where σ_0 is the initial value for σ.

Method M-5 is also history-dependent and attempts to promote genetic diversity while still allowing small mutations near the end of a run. To do this, M-5 compares χ^2_{min} (the lowest χ^2 of the current population) with the χ^2 averaged over the last 100 generations. It uses the following quantities:

$$\chi^2_{\text{scale}} = \left(\frac{10}{g} + 1\right)\cdot\chi^2_{\text{min}}(g) \tag{11}$$

$$\langle\langle\chi^2\rangle\rangle_{100}(g) = \frac{1}{100}\sum_{i}^{i-100}\langle\chi^2\rangle(g) \tag{12}$$

For every tenth generation, if $\langle\langle\chi^2\rangle\rangle_{100}(g) > \chi^2_{\text{scale}}$, then σ is reduced by a multiplicative factor c_σ; else σ is increased by the inverse of the factor c_σ. In this work $c_\sigma = 0.95$. Furthermore, if $\chi^2_{\text{min}}(g) = \chi^2_{\text{min}}(g - 100)$, then we assume that the minimum has been approximately located and reduce σ by c^2_σ. Note that this equality can only be satisfied using overlapping or semioverlapping selection methods.

Last, mutation size control method M-6 uses a history-dependent adjustment of σ which is similar in motivation to M-5 but with a different criterion for changing σ. M-6 tracks the average of the last 10 changes in χ^2_{min} by defining a quantity $\langle\Delta\chi^2_{\text{min}}\rangle_{10}(g)$, which is the average over the 10 most recent nonzero changes in χ^2_{min}. This measures the "step size" of progress toward an optimum solution. Then, if $\langle\langle\chi^2\rangle\rangle_{10}(g) > 4 \cdot \langle\Delta\chi^2_{\text{min}}\rangle_{10}(g)$, then σ is reduced by a multiplicative factor c_σ; else σ is increased by the inverse of c_σ. As in M-5, $c_\sigma = 0.95$, and if $\chi^2_{\text{min}}(g - 100) = \chi^2_{\text{min}}(g)$, then σ is reduced by a factor c^2_σ.

The performance of these different mutation operators is shown in Figure 6. There is no significant impact of mutation size control until roughly 250 generations. It was argued above that recombination methods only had a significant effect in the first 250 generations. It appears that after 250 generations the populations are sufficiently homogeneous that mutation becomes the dominant method of search.

Keeping a constant mutation size prevents parameters from being optimized to values any more precise than the size of Gaussian mutations being applied. This is shown by the fluctuating yet flat fitness of the constant-σ method M-2 from generation 300 onward. The flat fitness profile occurs because the default selection method is nonoverlapping, and the best parent is not carried forward to the next generation. Method M-4 gives results similar to keeping σ constant in the later generations, which is surprising. The scaling factor in M-4 should allow for drops in χ^2 to produce relatively greater drops in σ when the optimization is in its later generations. However, this is not observed, and σ never became small enough to reach the χ^2 values achieved in other methods.

History-dependent, diversity-promoting methods M-5 and M-6 produce results similar to simple annealing, algorithm M-3. Methods M-5 and M-6 did have the desired impact on the genetic diversity of the parent population, but the effect only became noticeable after roughly 1300 generations. At that point, the population had already converged on a single minima, and the diversity was quite low. The likely explanation for the observed behavior is that the diversity-enhancing mutations that were accepted tended to be in the parameters upon which χ^2 did not depend sensitively, such that they would increase the radius of gyration but not lower the fitness. These mutations, therefore, would not contribute strongly to the location of new, lower-χ^2 minima. For such methods to have a significant effect on the optimization, they would have to be tuned to become active closer to the point when mutation takes over from recombination as the dominant form of search, near 250 generations. The default algorithm M-1 performed well but has a somewhat "wavier" profile than the other variants, possibly caused by sporadic large reductions in χ^2 in one of the independent runs. This algorithm ends up very slightly outperforming the other mutation size control algorithms tested.

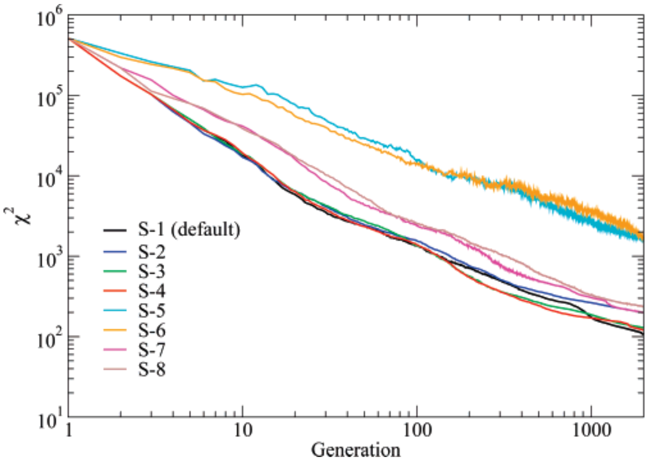


Figure 7. Variation of optimization profile with choice of selection operator. Operators tested include (S-1) nonoverlapping truncation, (S-2) overlapping truncation, (S-3) nonoverlapping truncation plus best parent, (S-4) nonoverlapping truncation plus best-ever individual, (S-5) nonoverlapping 2-way tournament, (S-6) overlapping 2-way tournament, (S-7) nonoverlapping 8-way tournament, and (S-8) overlapping 8-way tournament.

3.5. Selection. Selection methods are compared in Figure 7. The default selection method used, S-1, was the (m, n) nonoverlapping truncation method; S-1 is deterministic, choosing the best m out of n children to be the parents for the next generation. This is compared against overlapping (S-2) and semioverlapping (S-3 and S-4) truncations, and all combinations of overlapping and nonoverlapping two-way and eight-way tournament methods (S-5−S-8). S-1, S-3 and S-4 clearly outperformed all other options in the selection tests. S-1 and S-2 performed similarly until roughly 350 generations into the optimization. S-1 provided a final result with a χ^2 almost 50% better than S-2. Tournament methods are less elitist than truncation methods and also less effective. The two-way tournament methods S-5 and S-6, also called binary tournaments, do not approach the χ^2 value of other methods. Increasing the number of participants in a tournament increases the method's elitism, which makes this method more flexible than truncation methods. However, even eight-way tournament selection methods S-7 and S-8 still lag behind truncation methods.

3.6. Simulated Annealing. For comparison with the evolutionary strategies, we also considered an efficient simulated annealing (SA) algorithm.[66] Simulated annealing is similar to (1+1)-ES, though with different selection and mutation size control operators.

In our SA implementation, a new trial solution (child) is generated by applying Gaussian mutations to parameters of the parent. As this is only done for one child per cycle, we refer to births instead of generations. With probability 0.2 we mutate each parameter x_i by addition of a Gaussian random number $G(0, \sigma_i)$, where σ_i is a global σ scaled by the allowed range of parameter i, as in most of the ES mutation size control variants. Another change made beyond a typical simulated annealing algorithm is that acceptance and rejection of trial solutions are tracked over the past 512 births. If fewer than 20% of children are accepted, then σ is decreased by a factor of $c_\sigma = 0.995$. If more than 20% are accepted, then σ is increased by a factor of $1/c_\sigma$. This is a simple version of the "1/5 rule" sometimes used in (1,1) evolutionary strategies and Monte Carlo simulations.[52] The algorithm has a "temperature" T (with initial value 175.0585 (kJ/mol)^2) which is annealed by a factor $c_T = 0.99994$ after each birth. The child replaces the parent if $U(0,1) \leq \exp(-(\chi^2_{\text{child}} - \chi^2_{\text{parent}})/T)$ where U(0,1) is a uniform random number on the interval [0, 1].

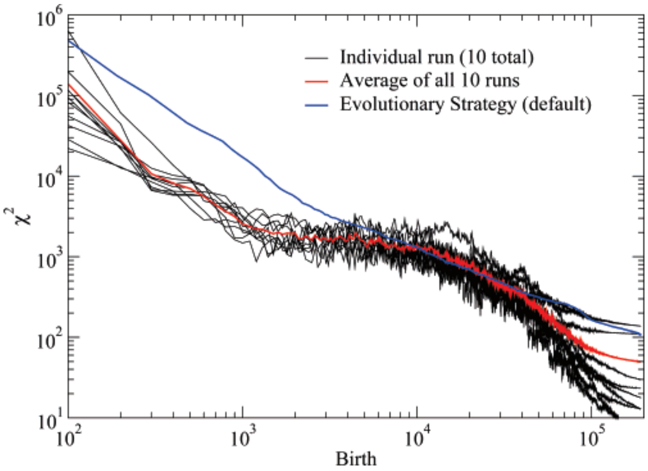


Figure 8. Simulated annealing optimizations. As in Figure 2, ten independent runs (starting from the same point) are shown, as well as their average.

As shown in Figure 8 the shape of the convergence profile in simulated annealing is substantially different from that displayed by the evolutionary strategies tested. After an initial rapid improvement, a period of slow searching occurs. The rapid feedback of simulated annealing—only considering one child per generation before choosing a new parent—may explain the advantage of SA in the first 1000 births or so. The advantage of SA toward the end of the simulation is probably related to the "1/5 rule" which allows mutation size to be adjusted on-the-fly. Interestingly, the profile of SA optimizations at very late times is still different in shape than that of any of the ES mutation size control variants, even though they are designed to have similar effects.

As the simulated temperature is lowered, the algorithm becomes trapped in a single minimum. Different annealing runs produce fitness values varying over about 1 order of magnitude, much as do the independent ES optimizations of Figure 2. The cooling schedule used here was chosen to allow the optimization to reach low temperatures, characterized by fluctuations in χ^2 much smaller than $\mathcal{O}(1)$, within the same number of function evaluations that the evolutionary strategies were allowed. There may be less variation between final fitness values when using a slower cooling schedule. Nevertheless, simulated annealing is very effective in finding a good solution.

4. Parametrization against CPMD Reference Data

Using combinations of ES options that were found to be effective in the meta-optimization study, we then ran several optimizations of the FG potential against the second training

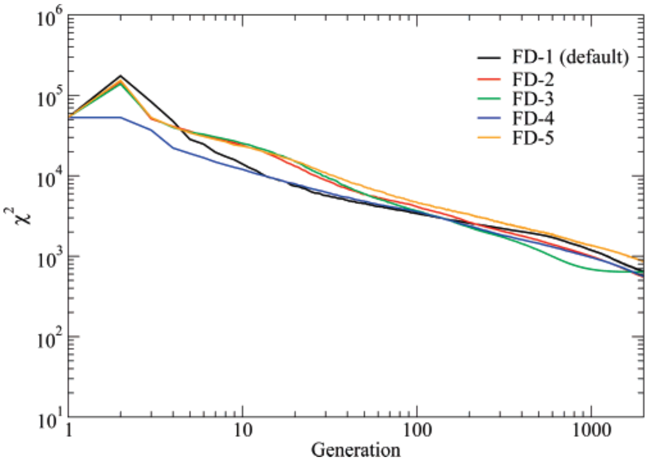


Figure 9. Fitting the FG functional form to the CPMD training set. FD-1 is the default method in the meta-optimization tests. FD-2 uses local, intermediate recombination, and other options are as in FD-1. FD-3 uses local, intermediate recombination and simple annealing mutation size control, and other options are as in FD-1. FD-4 uses local, intermediate recombination and nonoverlapping truncation plus best-ever individual selection, and other options are as in FD-1. FD-5 uses local, intermediate recombination and nonoverlapping 8-way tournament selection, with other options as in FD-1.

set, composed of DFT data. These calculations fit the FG functional form against data which it cannot perfectly reproduce, and so the minimum possible χ^2 will no longer be equal to zero. These optimizations were initialized with the original FG potential parameters as one of the parents.

These results are shown in Figure 9. FD-1 was the default method used in the meta-optimization study. FD-2 used local, intermediate recombination. FD-3 used local, intermediate recombination and simple annealing for mutation size control. FD-4 used local, intermediate recombination and semioverlapping truncation selection from the population $m + n$. FD-5 used local, intermediate recombination and 8-way tournament selection.

The FG parameters are better than almost any random guess. The use of nonoverlapping selection then creates a “spike” at the second generation in four of the five methods tested, since recombination and mutation create children with a larger χ^2 than the FG parameters while the FG potential is not carried over to the second generation.

FD-2 and FD-4 performed the best and have extremely similar profiles for the last 1000 generations of the optimization. Against this training set, the effects of recombination are observed much further into the optimization than the 250 generations usually seen during the meta-optimization study. FD-1 and FD-2 develop similar slopes after generation 1000. FD-3, using simple annealing, performs strongly until just after generation 1000, when σ became too small to make further significant improvements in fitness. Last, FD-5 lagged consistently behind the other options, showing that for this problem and the population size used even large tournament sizes may not be sufficiently elitist. Except for FD-3, all of these methods displayed optimization profiles similar to those seen in the meta-optimization study, suggesting that the approach of fitting an empirical potential to itself is a reasonable choice of a test problem for investigation of ES behavior.

The parameter sets obtained from these calculations are shown in Table 1; these are the fittest individual results from the ten independent runs using each evolutionary strategy variant. All five parametrizations are dramatically fitter (closer to the CPMD reference data) than the original FG parameters, though we should note that this does not a priori indicate that they will be more suitable for modeling a particular system or property. The obtained χ^2 values of ~500 (kJ/mol)^2 correspond to an rms deviation of 0.1 kJ/mol per atom in the energy of any given configuration relative to the reference configuration. The average hydrogen bond strength in liquid water is about 20 kJ/mol. Since hydrogen bonding is expected to dominate the energy differences between configurations, we expect that these important interactions should be described well by these parameter sets, at least to within the accuracy of the density functional theory used. The different sets vary considerably in the actual values of particular parameters, with some, such as the λs, varying over a fairly large range, while others, such as β(O−O), are very similar from one set to the next. In a few cases (γ(Si−O−Si), for example) parameters have converged to one side of their “allowed range,” which suggests that better fits could be obtained by expanding these ranges.

5. Discussion

All of the optimization profiles shown above are averaged over ten independent runs. In a typical run, for instance as shown in Figure 3, the radius of gyration R_g of the population at the endpoint had a value near to 0.03, indicating that the members of the population were all very similar to each other and that the algorithm had converged into a single minimum of the fitness function. However, the R_g value measured for the ten best solutions obtained from the ten independent runs is 1.49, approximately 2 orders of magnitude larger. Comparing the two values suggests that independent optimization runs are finding different minima of the fitness function; inspection of the actual parameter sets given in Table 1 (which is a different calculation, but with similar convergence properties) supports this. While evolutionary methods are often touted as globally convergent, it appears that for “reasonable” run conditions, performing multiple independent runs is probably a good strategy.

The number of minima, and the “shape” of the fitness function χ^2, are of interest in this regard. Given the high dimensionality of the parameter space, one might suppose that the many different solutions found in these optimizations arise from the relatively small number (128) of configurations used in the training set: the fewer conditions there are to satisfy, the more ways there should be to do so. However, this appears to not be the case. The R_g values for the ten independent optimal solutions for each of the different training set sizes of Figure 1 are all between 1.33 and 1.68, with no correlation with training set size. That is, adding additional data beyond 128 configuration energies does not bring the many locally optimal parameter sets any closer to each other. Likewise, the corresponding R_g values for the

Table 1. Feuston-Garofalini Reparametrizations by Evolutionary Strategies[a]

parameter	FD-1	FD-2	FD-3	FD-4	FD-5	FG
A (H−H), × 10^{-9} ergs	0.03103	0.02106	0.03571	0.02257	0.021513	0.0340
ρ (H−H), Å	0.2827	0.1784	0.2573	0.1786	0.2206	0.35
β (H−H), Å	1.319	1.3790	1.3526	1.3496	1.3727	2.10
a_1 (H−H), × 10^{-12} ergs	−5.335	−6.3800	−5.3370	−5.7848	−5.3192	−5.2973
b_1 (H−H), Å^{-1}	5.117	4.7664	4.7996	5.2802	5.4553	6.0
c_1 (H−H), Å	1.2663	1.2006	1.2770	1.2207	1.2542	1.51
a_2 (H−H), × 10^{-12} ergs	0.2009	0.2632	0.4197	0.2993	0.3546	0.3473
b_2 (H−H), Å^{-1}	1.8539	2.0173	1.3476	2.1513	2.2582	2.0
c_2 (H−H), Å	3.2085	3.1084	2.5569	3.0789	3.0109	2.42
A (O−H), × 10^{-9} ergs	0.3360	0.3838	0.4018	0.3882	0.3848	0.3984
ρ (O−H), Å	0.2992	0.2773	0.2695	0.2757	0.2787	0.29
β (O−H), Å	1.7270	1.7978	1.7405	1.9038	1.9026	2.26
a_1 (O−H), × 10^{-12} ergs	−2.2366	−1.2288	−1.8019	−1.7787	−1.4016	−2.0840
b_1 (O−H), Å^{-1}	10.2427	21.4197	19.0815	20.9755	17.0696	15.0
c_1 (O−H), Å	1.1064	1.1605	1.1855	1.1760	1.1541	1.05
a_2 (O−H), × 10^{-12} ergs	6.8043	7.1150	7.1936	8.4660	7.8496	7.6412
b_2 (O−H), Å^{-1}	2.8448	3.2279	3.2265	2.7840	3.0235	3.2
c_2 (O−H), Å	1.4358	1.6233	1.5092	1.5852	1.5941	1.50
a_3 (O−H), × 10^{-12} ergs	−0.8008	−1.1142	−0.8619	−0.8341	−1.0400	−0.8336
b_3 (O−H), Å^{-1}	3.8372	5.3733	4.9270	5.1868	5.1650	5.0
c_3 (O−H), Å	1.7244	1.9072	1.8161	1.9928	1.8755	2.00
A (O−O), × 10^{-9} ergs	0.6204	0.9318	0.7086	1.0126	0.6314	0.7250
ρ (O−O), Å	0.1536	0.2258	0.2316	0.1685	0.1815	0.29
β (O−O), Å	1.6597	1.7056	1.7057	1.7451	1.7893	2.34
A (Si−H), × 10^{-9} ergs	0.03488	0.04092	0.05571	0.05767	0.05520	0.0690
ρ (Si−H), Å	0.3333	0.1732	0.2241	0.1868	0.2076	0.29
β (Si−H), Å	1.7574	1.8393	1.8692	1.8520	1.9144	2.31
a_1 (Si−H), × 10^{-12} ergs	−5.9716	−5.9754	−6.2415	−6.0339	−6.3399	−4.6542
b_1 (Si−H), Å^{-1}	3.6173	3.7601	3.7488	3.7710	3.7888	6.0
c_1 (Si−H), Å	2.1270	2.1799	2.2019	2.1767	2.1761	2.20
A (Si−O), × 10^{-9} ergs	4.3049	2.0904	2.3021	2.1387	2.3477	2.9620
ρ (Si−O), Å	0.2320	0.3052	0.3041	0.3058	0.3070	0.29
β (Si−O), Å	1.2277	1.5972	1.6715	1.6305	1.7657	2.34
A (Si−Si), × 10^{-9} ergs	2.0641	2.0021	2.2312	1.7762	2.1179	1.8770
ρ (Si−Si), Å	0.3035	0.1890	0.2862	0.2197	0.1855	0.29
β (Si−Si), Å	1.1892	1.4321	1.4610	1.4137	1.5670	2.29
λ (O−Si−O), × 10^{-11} ergs	11.3068	10.1754	19.44	9.6978	19.1985	19.0
γ (O−Si−O), Å	4.1957	3.8445	3.1944	4.1697	3.9531	2.8
λ (Si−O−Si), × 10^{-11} ergs	0.4496	0.4483	0.3136	0.4447	0.4439	0.3
γ (Si−O−Si), Å	1.0005	1.0052	2.0065	1.0021	1.0067	2.0
λ (Si−O−H), × 10^{-11} ergs	4.8690	3.1015	5.1819	2.7365	3.9802	5.0
γ (Si−O−H: Si−O), Å	1.6022	1.0161	1.9427	1.0495	1.7518	2.0
γ (Si−O−H: O−H), Å	1.5203	1.7038	1.3923	1.7058	1.5326	1.2
λ (H−O−H), × 10^{-11} ergs	31.9566	25.3210	32.1643	38.3666	32.3834	35.0
γ (H−O−H), Å	1.4741	1.3718	1.4345	1.4649	1.4264	1.3
χ^2 (kJ/mol)^2	352.4	430.7	501.3	459.8	560.7	52963.0

[a] The fittest parameter sets from Figure 9 are shown as well as the original FG parametrization. Parameter names and units are as given in ref 51. Only "fitted" parameters are given in the table; other parameters (cutoffs, reference angles, and formal charges) are kept fixed at their literature values.[51]

runs of Figure 4, which vary in m and n, are all between 1.49 and 1.73 and likewise do not exhibit any trend with population parameters. It therefore appears that the many local minima in this objective function result from the potential itself and the particular definition of χ^2 used, rather than the size of the training set or other, more arbitrary parameters.

A significant feature observed in many of the optimization profiles in this study was an apparent crossover, at about 250 generations, from behavior dominated by recombination to behavior dominated by mutation. This crossover was remarkably robust to changes in the various operators involved, and therefore its appearance may be anticipated in related problems.

Since most of the computational effort is expended after the crossover, in order to more quickly locate optimized parameter sets one should make the mutation operator as efficient as possible. However, of the considerable number of mutation operators tested in this work there were no clearly superior ones, and significant further improvements may be difficult. One possible alternative could be a composite (or "memetic"[69]) optimization strategy, in which, once the ES algorithm "slows down", one switches over to a different, locally convergent, method which is good at "refining" an approximately located solution. The radius of gyration R_g introduced above is an effective signature for the ES crossover and could be monitored to trigger the change to another method. We note, in this regard, that rapidly converging methods such as conjugate-gradient optimization or Newton−Raphson root-finding are not very well suited to parameter optimization problems, since it is preferable to not have to implement derivatives of the energy with respect to the potential parameters. However, such derivatives could be efficiently estimated by using parallelized one-way finite or centered difference methods, which could provide a cost-effective route to the precise location of χ^2 minima; the effectiveness of this approach would depend on the roughness of the χ^2 function and the stability of the optimizer with respect to numerical precision. It should be noted that when applied to ligand docking, a prior study did not find local optimization to be beneficial.[44]

Based on the results of the meta-optimization study, we recommend the use of intermediate recombination operators for both the parameters $\{x_i\}$ and mutation size control variables $\{\sigma_i\}$. No substantial difference is observed between global intermediate and local intermediate recombination at long times, though at short times the global variant appears preferable. Of the mutation operators considered, the self-adaptive, independent-σ method M-1 is at least as effective as any of the others considered and lacks any "adjustable" parameters. We note that "simple annealing" is nearly as effective with one adjustable parameter (here chosen arbitrarily) and considerably simpler to implement. Finally, nonoverlapping or semioverlapping truncation methods are clearly preferred for selection, as the tournament methods appeared to not have enough selection pressure, and overlapping methods exhibited slowdowns in the later stages of optimization.

Simultaneous parametrization of all parts of a potential has the advantage of providing more uniform "quality" between different terms but greatly increases the complexity of the numerical problem to be solved. Even in fully automated parametrizations one must still provide initial estimates of the magnitude (and, likely, the allowed range) of each parameter, which requires at least some physical insight into the problem. In applications where an existing potential is to be extended, such estimates are straightforward, but for the parametrization of a new functional form or previously unstudied chemical species they may be more difficult to obtain. For very large problems, preliminary parametrization of groups of related parameters against subsets of the available reference data may also be a viable strategy.

ES methods are inherently parallelizable. While evolution of the objective (fitness) function used in this work can also be parallelized over a reasonable number of processors, the ES approach has a considerable advantage in this regard and therefore should be of particular interest when wall-clock time is a limiting factor. This suggests that ES is particularly suitable for work involving a large number of parametrizations, for instance comparisons of different functional forms, comparisons of potentials based on different reference data, or even the (common) extension of an existing potential to treat some new chemical species.

As shown in Figure 8, the efficient simulated annealing method used in this study generally outperformed the evolutionary strategies when fitting the FG potential to the FG training set. Simulated annealing can be parallelized either through distribution of configurations in the training set or by performing multiple independent runs. As discussed earlier, evolutionary strategies may spread the evaluation of groups of children across available processors. This is a significant advantage: the number of CPU cores available in modern supercomputers or clusters is increasing at a greater rate than the performance per core. We also note that the adaptive mutation algorithm in the simulated annealing optimizations may have been superior to the mutation algorithms used in the evolutionary strategy, as no equivalent to the "1/5 rule" was available for ES runs.

Finally, we note that the type of reference data used (configurational energies) and definition of the fitness function as a least-squares-like criterion are themselves arbitrary choices, and there are certainly other possibilities. Force (gradient) data could also be used in the fitness function (as in the "force-matching" studies described above,[12,13] for instance), and a "minimax" criteria could be used to define the fitness function, so that the final optimized value would limit the maximum deviation in selected quantities between the model system and the reference data. Although a wide variety of ES algorithms were considered and the potential function studied is representative of a large class of related models, the reference data used in the parametrizations described a single aqueous solution of silicate species. It is therefore possible that the ES variants selected for the parametrization against CPMD data might not transfer well to a different physical system, although one hopes that they should at least serve as a useful starting point and reference.

In summary, we have presented guidelines for the selection of ES operators and training set sizes suitable for the parametrization of empirical potentials against reference data generated using electronic-structure methods. Such parametrizations are considerably higher in dimension and complexity than the typical problems used in development of evolutionary strategies, and algorithms optimized for these different problem classes differ in nonobvious ways. The ES approach is highly parallelizable and may therefore be suited to "large" optimization problems. However, ES exhibits relatively slow convergence at later generations that may warrant changeover at late times to an alternate method which converges rapidly once a solution has been approximately located.

Acknowledgment. We acknowledge financial support from the National Science Foundation (CHE-0241005, CHE-0443501, and CHE-0626008) and the Washington University Center for Materials Innovation and NCSA (DMR030033N) and the Washington University Center for Scientific Parallel

Computing for computer resources. We also thank Rafael Salazar Tio for useful discussions.

References

(1) Gray, C. G.; Gubbins, K. E. *Theory of Molecular Fluids*; Clarendon Press: Oxford, 1984; Vol. 1, pp 27−142.

(2) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Clarendon Press: Oxford, 1987; pp 6−22.

(3) Schlick, T. *Molecular modeling and simulation: an interdisciplinary guide;* Springer-Verlag: New York, 2002; pp 225−304.

(4) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley & Sons, Ltd.: Chichester, U.K., 2007; pp 22−79.

(5) 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*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; Vol. 1.

(6) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., III *J. Phys. Chem. A* **2001**, *105*, 9396−9409.

(7) van Duin, A. C. T.; Strachan, A.; Stweman, S.; Zhang, Q.; Xu, X.; Goddard, W. A., III *J. Phys. Chem. A* **2003**, *107*, 3803- 3811.

(8) Nielson, K. D.; van Duin, A. C. T.; Oxgaard, J.; Deng, W. Q.; Goddard, W. A., III *J. Phys. Chem. A* **2005**, *109*, 493−499.

(9) Cheung, S.; Deng, W. Q.; van Duin, A. C. T.; Goddard, W. A., III *J. Phys. Chem. A* **2005**, *109*, 851−859.

(10) Han, S. S.; van Duin, A. C. T.; Goddard, W. A., III; Lee, H. M. *J. Phys. Chem. A* **2005**, *109*, 4575−4582.

(11) Han, S. S.; Kang, J. K.; Lee, H. M.; van Duin, A. C. T.; Goddard, W. A., III *J. Chem. Phys.* **2005**, *123*, 114703.

(12) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896−10913.

(13) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 6573−6586.

(14) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *Science* **2007**, *315*, 1249−1252.

(15) Faure, A.; Valiron, P.; Wernli, M.; Wiesenfeld, L.; Rist, C.; Noga, J.; Tennyson, J. *J. Chem. Phys.* **2005**, *112*, 221102.

(16) Strassner, T.; Busold, M.; Herrmann, W. A. *J. Comput. Chem.* **2002**, *23*, 282−290.

(17) Lemberg, H. L.; Stillinger, F. H. *J. Chem. Phys.* **1975**, *62*, 1677−1690.

(18) Rahman, A.; Stillinger, F. H.; Lemberg, H. L. *J. Chem. Phys.* **1975**, *63*, 5223−5230.

(19) Stillinger, F. H.; Rahman, A. *J. Chem. Phys.* **1978**, *68*, 666−670.

(20) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 77*, 2nd ed.; Cambridge University Press: Cambridge, 1992; pp 387−406.

(21) van Duin, A. C. T.; Baas, J. M. A.; van de Graaf, B. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 2881−2895.

(22) Collins, M. A. *Theor. Chem. Acc.* **2002**, *108*, 313−324.

(23) Moyano, G. E.; Collins, M. A. *J. Chem. Phys.* **2004**, *121*, 9769−9775.

(24) Netzloff, H. M. *J. Chem. Phys.* **2006**, *124*, 154104.

(25) Guo, Y.; Harding, L. B.; Wagner, A. F.; Minkoff, M.; Thompson, D. L. *J. Chem. Phys.* **2007**, *126*, 104105.

(26) Beyer, H.-G.; Schwefel, H.-P. *Nat. Comput.* **2002**, *1*, 3−52.

(27) Fogel, D. B. *IEEE Trans. Neural Networks* **1994**, *5*, 3−14.

(28) Rechenberg, I. *Cybernetic solution path of an experimental problem*; Technical Report; Royal Aircraft Establishment: Farnborough, 1965; Library Translation 1122.

(29) Schwefel, H.-P. Cybernetic evolution as a strategy for experimental research in fluid mechanics. Master’s Thesis, Technical University of Berlin, 1965.

(30) Rechenberg, I. Evolutionary strategies: optimizing technical systems with principles of biological evolution. Thesis, Technical University of Berlin, Department of Process Engineering, 1971.

(31) Rechenberg, I. *Evolutionary strategies: optimizing technical systems with principles of biological evolution*; Frommann-Holzboog Verlag: Stuttgart, 1973.

(32) Schwefel, H.-P. Evolutionary strategies and numerical optimization. Thesis, Technical University of Berlin, 1975.

(33) Fogel, L.; Owens, A.; Walsh, M. *Artificial Intelligence through Simulated Evolution;* John Wiley and Sons: 1966.

(34) Holland, J. *Adaptation in Natural and Artificial Systems;* University of Michigan Press: 1975.

(35) De Jong, K. A. *Evolutionary Computation: A Unified Approach;* MIT Press: Cambridge, MA, 2006; pp 49−69.

(36) De Jong, K. A. *Evolutionary Computation: A Unified Approach;* MIT Press: Cambridge, MA, 2006; pp 72−73.

(37) Cundari, T. R.; Deng, J.; Fu, W. *Int. J. Quantum Chem.* **2000**, *77*, 421−432.

(38) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231−236.

(39) Mohr, M.; McNamara, J. P.; Wang, H.; Rajeev, S. A.; Ge, J.; Morgado, C. A.; Hiller, I. H. *Faraday Discuss.* **2003**, *124*, 413−428.

(40) Brothers, E. N.; Merz, K. M., Jr. *J. Phys. Chem. B* **2002**, *106*, 2779−2785.

(41) Ge, Y.; Head, J. D. *Int. J. Quantum Chem.* **2003**, *95*, 617−626.

(42) Clark, D. E.; Westhead, D. R. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.

(43) Lameijer, E.-W.; Bäck, T.; Kok, J. N.; Ijzerman, A. P. *Nat. Computing* **2005**, *4*, 177−243.

(44) Thomsen, R. *BioSystems* **2003**, *72*, 57−73.

(45) Strassner, T.; Busold, M.; Radrich, H. *J. Mol. Model.* **2001**, *7*, 374−377.

(46) Wolohan, P.; Yoo, J.; Welch, M. J.; Reichert, D. E. *J. Med. Chem.* **2005**, *48*, 5561−5569.

(47) Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1219−1228.

(48) Mallik, A.; Runge, K.; Cheng, H.-P.; Dufty, J. *Mol. Sim.* **2005**, *31*, 695−703.

(49) Globus, A.; Menon, M.; Srivastava, D. *Comput. Model. Eng. Sci.* **2002**, *3*, 557−574.

(50) Feuston, B. P.; Garofalini, S. H. *J. Chem. Phys.* **1988**, *89*, 5818−5824.

(51) Feuston, B. P.; Garofalini, S. H. *J. Phys. Chem.* **1990**, *94*, 5351−5356.

(52) Bäck, T.; Hoffmeister, F.; Schwefel, H.-P. A Survey of Evolutionary Strategies. In *Proceedings of the 4th International Conference on Genetic Algorithms*; Belew, R. K., Booker, L. B., Eds.; Morgan Kaufmann: San Franscisco, CA, 1991.

(53) De Jong, K. A. *Evolutionary Computation: A Unified Approach;* MIT Press: Cambridge, MA, 2006; p 27.

(54) De Jong, K. A. *Evolutionary Computation: A Unified Approach;* MIT Press: Cambridge, MA, 2006; pp 128−130.

(55) Born, M.; Mayer, J. E. *Z. Phys.* **1932**, *75*, 1−18.

(56) Huggins, M. L.; Mayer, J. E. *J. Chem. Phys.* **1933**, *1*, 643−646.

(57) Stillinger, F. H.; Weber, T. A. *Phys. Rev. B* **1985**, *31*, 5262−5271.

(58) Rao, N. Z.; Gelb, L. D. *J. Phys. Chem. B* **2004**, *108*, 12418−12428.

(59) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474.

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

(61) Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892−7895.

(62) 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.; Tyengar, S. S.; Thomasi, 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.; Romaromi, I.; Martin, R. L.; Foz, 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*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(63) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(64) CPMD. Copyright IBM Corp 1990−2003, Copyright MPI für Festkörperforschung Stuttgart 1997−2001.

(65) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 77*, 2nd ed.; Cambridge University Press: Cambridge, 1992; pp 436−448.

(66) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. *Science* **1983**, *220*, 671−680.

(67) Alavi, A.; Alvarez, L. J.; Elliot, S. R.; McDonald, I. R. *Philos. Mag. B* **1992**, *65*, 489−500.

(68) De Jong, K. A. *Evolutionary Computation: A Unified Approach;* MIT Press: Cambridge, MA, 2006; pp 77−78.

(69) Moscato, P.; Norman, M. G. Parallel Computing and Transputer Applications. In *Chapter A "Memetic" Approach for the Traveling Salesman Problem Implementation of a Computational Ecology for Combinatorial Optimization on Message-Passing Systems*; IOS Press/CIMNE: 1992; pp 177−186.

CT700087D

JCTC Journal of Chemical Theory and Computation

General Transition-State Force Field for Cytochrome P450 Hydroxylation

Patrik Rydberg,[†] Lars Olsen,[‡] Per-Ola Norrby,[§] and Ulf Ryde*,[†]

Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, SE-221 00 Lund, Sweden, Biostructural Research Group, Faculty of Pharmaceutical Sciences, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen Ø, Denmark, and Department of Chemistry, Göteborg University, Kemigården 4, SE-412 96 Göteborg, Sweden

Received May 14, 2007

Abstract: We have developed force-field parameters for the hydrogen-abstraction transition state of aliphatic hydroxylation by cytochrome P450 using the Q2MM approach. The parametrization is based on quantum chemical (B3LYP) transition-state structures and Hessian matrices for 24 diverse substrate models (14 in the training set and 10 in the test set). The force field is intended to be applicable to any druglike molecule by the use of the general Amber force field (GAFF) for the substrates. The parameters reproduce the geometries within 0.1 Å and 1.2° for bond lengths and angles, respectively, with no significant differences between the training and test sets. The Hessian matrix is also well reproduced with a correlation coefficient of 0.99. The parametrization is performed by the ideal iterative approach of Norrby and Liljefors, which we have implemented for the Amber software.

Introduction

The cytochrome P450 enzymes (CYPs) are a superfamily of mono-oxygenases found in all types of organisms from bacteria to mammals. In the human genome, there are almost 60 genes for CYPs. They take part in the synthesis of important endogenous compounds such as steroids, prostaglandin, and fatty acids. However, they also contribute to the degradation of exogenous compounds. They affect both activation of prodrugs as well as the bioavailability and degradation of drugs. It has been estimated that the CYPs are responsible for 75% of the phase I drug metabolism.[1,2] Therefore, they have attracted much attention in pharmaceutical research.

Almost 200 crystal structures of CYPs have been published. They show a highly conserved active site, which consists of a heme group with an iron ion in the center of the porphyrin ring. In contrast to most other heme enzymes, the iron ion coordinates to the sulfur atom of a cysteine residue. This negatively charged ligand is believed to favor the formation of high-valent reactive iron intermediates of the CYPs.[3] The sixth coordination site of the iron ion, opposite to the cysteine ligand, is open to the binding of small extraneous ligands during the reaction cycle of the enzyme.

In the resting state, it is occupied by a water molecule, and the iron ion is in a low-spin Fe(III) state (cf. Figure 1). Binding of a substrate triggers the release of the water ligand, leading to a switch to the high-spin Fe(III) state, but the substrate does not directly coordinate to the iron ion. After a one-electron reduction, the Fe(II) ion binds O_2. This complex is reduced a second time, which triggers a heterolytic cleavage of the O−O bond, giving rise to a Fe(V)═O complex (formally), called compound I. This state is highly reactive and can perform many different reactions, such as hydroxylation, epoxidation, dealkylation, and N, S, and SO oxidation.[1,2]

Much effort has been devoted to the understanding of the reactivity of various CYPs. For example, the intrinsic

* Corresponding author phone: +46 − 46 2224502; fax: +46 − 46 2224543; e-mail: Ulf.Ryde@teokem.lu.se.

[†] Lund University.

[‡] University of Copenhagen.

[§] Göteborg University.

10.1021/ct700110f CCC: $37.00
Published on Web 08/01/2007

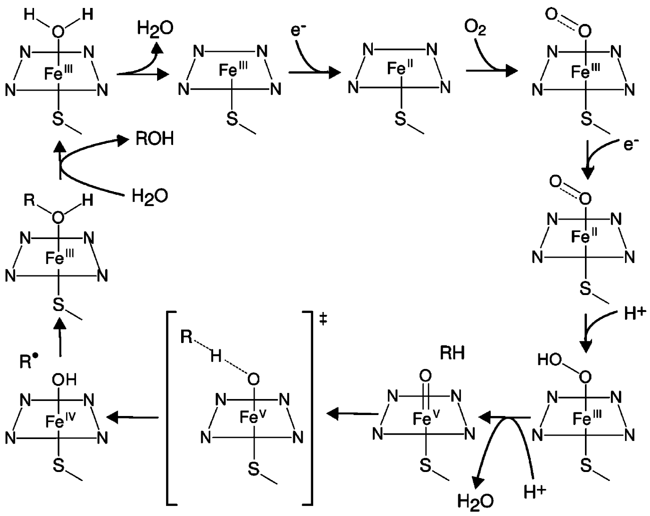


Figure 1. The CYP reaction cycle for a hydroxylation reaction, including the studied transition state and the intermediate state after it. The N quadrant represents the porphyrin ring.

reactivity of the CYP active site has been extensively studied, especially with density functional theory (DFT) and for the hydroxylation reaction.[4−6] Attempts have also been made to reproduce the DFT results with cheaper methods, allowing for the scanning of many compounds in a short time.[7−9] Moreover, the accessibility and binding of various substrates to the active site of human CYPs have been studied with pharmacophore models, docking, (3D-)QSAR, etc.[10−12]

Standard docking algorithms consider only the binding of ground-state molecules. However, for enzyme substrates, it could be more favorable to dock in the reactive transition state instead, because it must form if the substrate is to react. Such an approach would make the docking more discriminative, because the conformational space available to the substrate is more restricted in the transition state (it must bind to the oxoferryl group). In fact, it has been recently shown that docking of high-energy intermediates to enzymes can improve the prediction of the reactivity of an enzyme significantly.[13]

Unfortunately, docking of transition states is not straightforward, because standard methods of molecular mechanics cannot be directly employed (because transition states are not equilibrium states but first-order saddle points on the potential energy surface).[14] Therefore, special software is normally needed for their optimization, and available algorithms are quite time-consuming and cannot guarantee that a transition state is obtained from any starting point. Several methods are available for reproducing entire potential energy surfaces of reactions using empirical force fields.[14] Already in 1980, Warshel introduced the empirical valence bond method,[15] whereby the entire potential energy surface could be obtained by mixing of the reactant and product states, each modeled by an appropriate force field. This methodology has been used extensively, in particular for describing reactions in enzymes.[16] More recent variations on the same theme include the Rappé's reactive force field[17] and Truhlar's multiconfigurational molecular mechanics[18] methods. In an alternative approach, Goddard has recently developed a force field that allows direct bond breaking.[19] Each of these methods generates a full potential energy surface, requiring a saddle point search for locating a transition state, a task that is not easily automated and combined with a conformational search method. The SEAM method by Jensen[20] circumvents this problem by directly locating the intersection between the reactant and product force fields, a method that is very robust, even for poor starting geometries. Finally, the Q2MM method,[21] or more generally, transition-state force fields,[22] defines a new force field that treats transition-state structures as energy minima. A severe drawback with this method is that it no longer allows comparison with reactants and therefore cannot yield absolute activation energies; only relative barriers can be calculated. However, this is sufficient for investigating selectivities. Moreover, the method can be directly implemented in standard molecular mechanics software, and it is robust enough to allow full conformational searching. For these reasons we decided to proceed with the Q2MM method.

In this paper, we take the first step in this direction by developing a general transition-state force field for the hydroxylation reaction of the CYPs. A heme group has been parametrized before, both in the CYPs[23] and in other proteins (i.e., with different axial ligands),[24−28] but never a transition state. Our force field is based on 14 transition-state structures obtained at the DFT level on a diverse set of substrates.[9] In order to make the force field more general, we do not attempt to parametrize the substrate but instead take those parameters directly from the general Amber force field (GAFF), which is a general and diverse force field, designed for druglike molecules.[29] Thus, our effort is concentrated around the heme group and the reactive oxoferryl group. We evaluate the force field by comparing structures and energies with results obtained at the DFT level for a test set of 10 compounds.

Methods

Substrate Models. According to the consensus mechanism, the hydroxylation of substrates by the CYPs takes place in two steps:[4−6] First, the oxoferryl group of compound I abstracts a hydrogen atom from the substrate (Figure 1), yielding a Fe(IV)−OH intermediate and a substrate radical. In the next step, the substrate radical rebounds to the OH group, forming an iron-bound hydroxylated substrate. The DFT calculations show that the first step is rate limiting, and, therefore, we have restricted our investigation to the transition state of that step.

Our transition-state force field for the CYP hydroxylation reaction is based on DFT optimized structures of 24 small but diverse organic substrates.[9] This set was divided into a training set of 14 compounds and a test set with 10 compounds (to get a complete force field, the division of the two sets is slightly different from that in the original study[9]). These two sets are shown in Figures 2 and 3.

Quantum Mechanical Calculations. Reference data for the parametrizations were obtained with quantum mechanical calculations. We used data from our previous work,[9] which were obtained using DFT calculations with the B3LYP functional[30,31] and the 6-31G* basis set for all atoms except iron, for which we used the double-ξ basis set of Schäfer et al.[32] enhanced with a p function with the exponent 0.134915

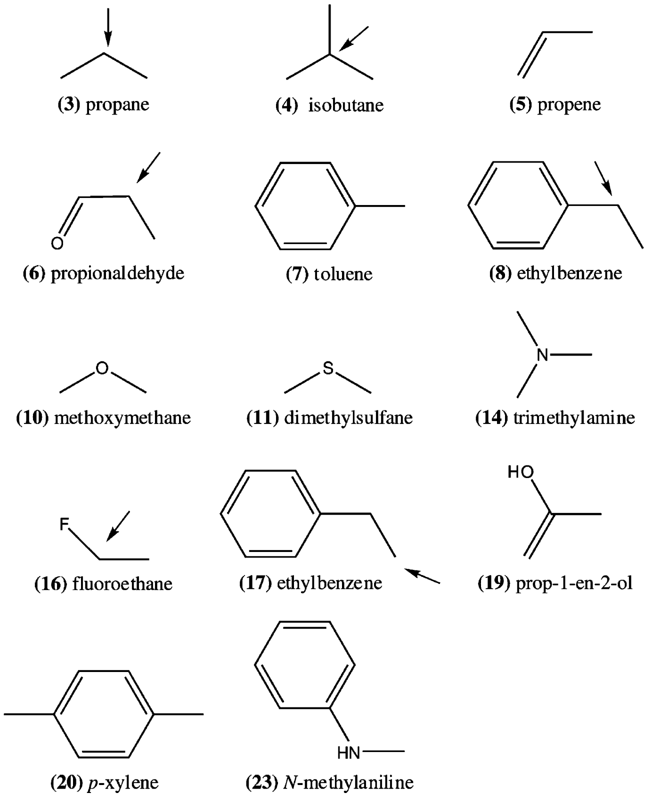

(3) propane (4) isobutane (5) propene
(6) propionaldehyde (7) toluene (8) ethylbenzene
(10) methoxymethane (11) dimethylsulfane (14) trimethylamine
(16) fluoroethane (17) ethylbenzene (19) prop-1-en-2-ol
(20) *p*-xylene (23) *N*-methylaniline

Figure 2. Substrate models used in the training set.

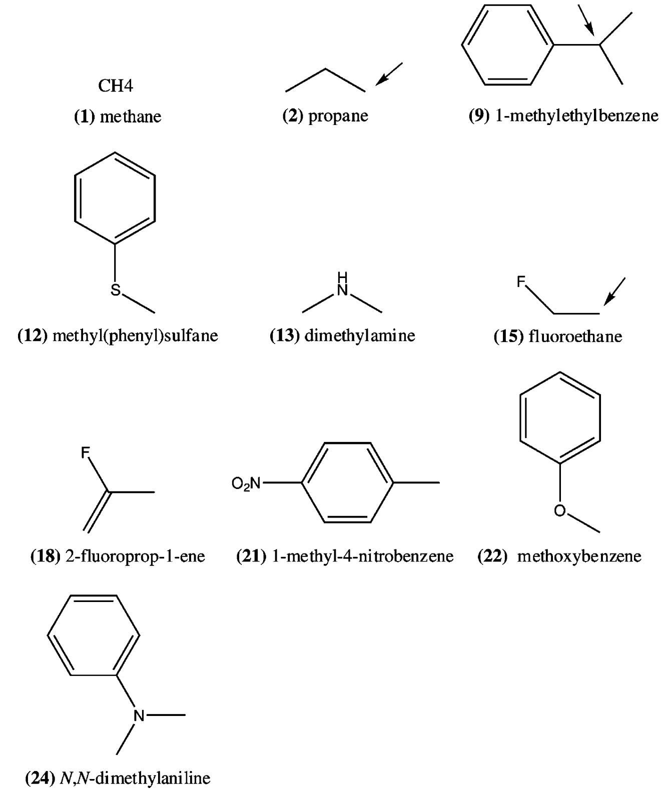

(1) methane (2) propane (9) 1-methylethylbenzene
(12) methyl(phenyl)sulfane (13) dimethylamine (15) fluoroethane
(18) 2-fluoroprop-1-ene (21) 1-methyl-4-nitrobenzene (22) methoxybenzene
(24) *N*,*N*-dimethylaniline

Figure 3. Substrate models used in the test set.

(DZP). The calculations were performed with the Gaussian03 software package.[33] The heme model was studied in the quartet state. We employed the geometries, energies, and the Hessian matrix from a frequency calculation of the transition state. All new B3LYP calculations were performed in the same way as the previous ones. The transition-state structures are insensitive to the choice of the model, theoretical method, and basis set used. Therefore, the employed structures are closely similar[9] to those obtained with other DFT methods.[5−8]

Parametrization. The Amber force field has the following functional form:

$$E_{\mathrm{MM}} = \sum_{\mathrm{bonds}} C_i(b_i - b_{i0})^2 + \sum_{\mathrm{angles}} D_i(\alpha_i - \alpha_{i0})^2 + \frac{1}{2}\sum_{\mathrm{dihedrals}}\sum_{j=1} E_{ij}(1 + \cos(j\varphi_i + \delta_{ij})) + \sum_{\mathrm{atom\_pairs}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \quad (1)$$

In this equation, the energy of a bond depends harmonically on the actual bond length (b_i; b_{i0} is the corresponding ideal bond length and C_i is the corresponding force constant). The same applies to the angles (with the actual and ideal angles α_i and α_{i0} and the force constant D_i), whereas the dihedral angles (the torsions) are assumed to be described by a cosine function with a periodicity (j) of 1, 2, 3, or 4. E_{ij} is the corresponding force constant and δ_{ij} is a phase factor, which determines the position of the minimum. Nonbonded interactions are described by Coulomb's law between partial charges on each atom (q_i) and a Lennard−Jones potential between each pair of atoms that are more than two bonds apart, using the constants A_{ij} and B_{ij}. Nonbonded interactions between atoms separated by three bonds are scaled down by a factor of 1.2 (electrostatics) or 2.0 (van der Waals), whereas those between atoms separated by one or two bonds are ignored.

To make the force field as general as possible, we decided to use the standard GAFF force field for the substrates.[29] Thereby, we avoid the need of reparametrizing the force field every time a new substrate will be studied. The atom types of the substrates were determined according to the philosophy of the GAFF force field[29] using the antechamber[34] module in Amber 8.[35] The only parameters not available in the GAFF force field were for an angle, and they were obtained according to the GAFF analogy rules (atom types c3−c2−f, force constant 66.0 kcal/mol/Å, and ideal angle 113.06°). Missing improper torsions were determined with the parmchk module of Amber.[35]

Likewise, we decided to let the carbon and hydrogen atoms of the iron ligand SCH_3^- have the same atom types as cysteine in the Amber 1999 force field.[36] Thus, we took parameters for its C−H bonds, H−C−H angle, and H−C−S−Fe dihedral from this force field.

Therefore, our parametrization of the transition state for the CYP hydroxylation is restricted to the heme group and the reactive hydrogen atom of the substrate. Nine new atom types were defined, as is shown in Figure 4. However, all parameters involving the NO and NP atom types were constrained to be identical (these two atom types are needed only to differ between the two types of N−Fe−N angles, 90 or 180°). Atom names and atom types of the full heme group with axial ligands are shown in Figures S1 and S2 in the Supporting Information.

The Lennard−Jones parameters for all atoms were taken from the Amber 1999 force field (except for HQ, for which

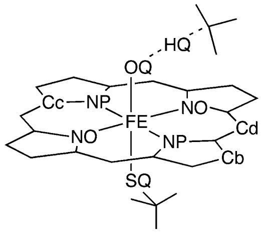


Figure 4. The new atom types.

the parameters were taken from the GAFF force field, because this atom is part of the substrate).[36] Thus, the new van der Waals parameters were taken from the following old atom types: HQ = h3, OQ = OH, NO = NP = N, SQ = S, Cb = Cc = Cd = C*, whereas those for iron were taken from the heme parameters supplied with Amber 8 (r = 1.2 Å and ϵ = 0.05 kcal/mol).[35]

Atomic charges for the isolated substrates were obtained by optimizing their geometry and calculating electrostatic potential around the molecule in points sampled with the Merz−Kollman scheme[37] using quantum mechanical calculations at the Hartree−Fock level, following the philosophy of the GAFF force field.[29] Charges were fitted to these electrostatic potentials using the RESP method,[38] as implemented in the antechamber software.[34]

For the heme group (with substrates), we instead used DFT calculations with the B3LYP functional and the DZP/6-31G* basis set (because the complicated electronic structure of the metal-containing transition states cannot be properly described at the Hartree−Fock level). Charges for all atoms were calculated using the RESP method for all 14 models in the training set. The net charge of the full model, except the substrate (which varies between the various models), was on average −0.2687 e with a mean absolute deviation of 0.05 e. The individual charges varied by up to 0.47 e, but large differences were observed only for the central iron and nitrogen atoms, which shows that they are caused mainly by the buried-charge problem of the RESP fit (atoms that are buried by other atoms in the structure have no ESP points close to them and therefore are less well-determined than more exposed atoms).[38] In fact, the well-defined hydrogen atoms varied by less than 0.05 e and the moderately buried carbon atoms varied by 0.15 e on average.

Our goal was to obtain a force field that can be combined with the GAFF force field for any substrate. Therefore, we need to have charges for all atoms in the heme group and its ligands that are independent of the substrate. We selected the model that had a net charge (excluding the substrate) closest to the mean charge of all the 14 models (−0.2687 e) and also had the smallest mean absolute deviation of all atoms compared to the average. This was the model with isobutane. We used the charges of this model for all atoms in the heme group. For the substrates, the GAFF charges were used, but a charge of +0.2679 e was added to the reactive hydrogen atom to make the full complex neutral.

In the DFT calculations,[9] the side chains of the heme group were replaced by hydrogen atoms to reduce the calculation times. In order to get parameters for a complete heme group, parameters are needed also for the side chains. It is unlikely that these groups should significantly influence the parameters around reactive center. Therefore, we simply added bonded parameters for the side chains from the GAFF force field by analogy. Charges for the side chains were determined by first running a geometry optimization of the isobutane model with the side chains added (the orientation of the heme ring was selected so that the Cys ligand and the two propionate side chains point in the same direction, as is observed in all available crystal structures[39]). During this optimization, the coordinates of the central core were fixed to those of the model without side chains, and the dihedrals of the side chains relative to the porphyrin ring were fixed to those observed in the crystal structure of human CYP 2C9[40] (to avoid that the charged propionate groups curl inward and form unphysical interactions with the heme group). With these coordinates, the charges were calculated with the RESP method as described above, keeping the charges of the core model fixed to their previous values. The resulting charges of the full heme group are shown in Table S1 in the Supporting Information. The force field can, through the parametrization from DFT data, reproduce the average oxidation state of a hydrogen abstraction transition state, but it cannot model major changes in the electronic structure, e.g., if the transition state becomes much more product- or reactantlike than in the average case, or if the oxidation state of the transition state changes significantly.

Finally, the bonded parameters (bond, angle, and dihedral terms in eq 1) for all interactions involving any of the new atom types were determined by the Q2MM method,[39] as is described in the next section. This method requires start values for all parameters. These should be as realistic as possible, to make sure that we do not end up in nonphysical local minima during the optimization. We obtained start values in the following manner: Parameters in the heme group were taken from the set of heme parameters supplied with Amber 8.[35] The parameters for the SCH_3^- group were taken from the cysteine parameters of the Amber 1999 force field.[36] Force constants of the angles including both the reacting hydrogen and carbon (atom types HQ and c3) were taken from the GAFF force field by analogy. The force constants of the iron−sulfur, iron−oxygen, and oxygen−hydrogen bonds and related angles were started from reasonable guesses.

A few parameters were excluded from the parametrization: First, all dihedrals containing an angle larger than 150° were excluded by zeroing the force constant (i.e., the dihedral A−B−C−D was excluded if either of the angles A−B−C or B−C−D was larger than 150°). These included the OQ−HQ−c3−X dihedral in the reactive bond, because the OQ−HQ−c3 angle is normally close to straight, and all dihedrals involving the two straight N−Fe−N angles of heme (atom types NO−FE−NO and NP−FE−NP). This is necessary, because the dihedral becomes undefined when one of the angles becomes straight, and close to that point, the dihedral can vary extensively for small changes in the position of the atoms. The excluded dihedral angles were selected at the start of the parametrization and were then kept fixed.

Second, the X−NO/NP−Fe−X dihedrals were also excluded by zeroing the force constants, because they ap-

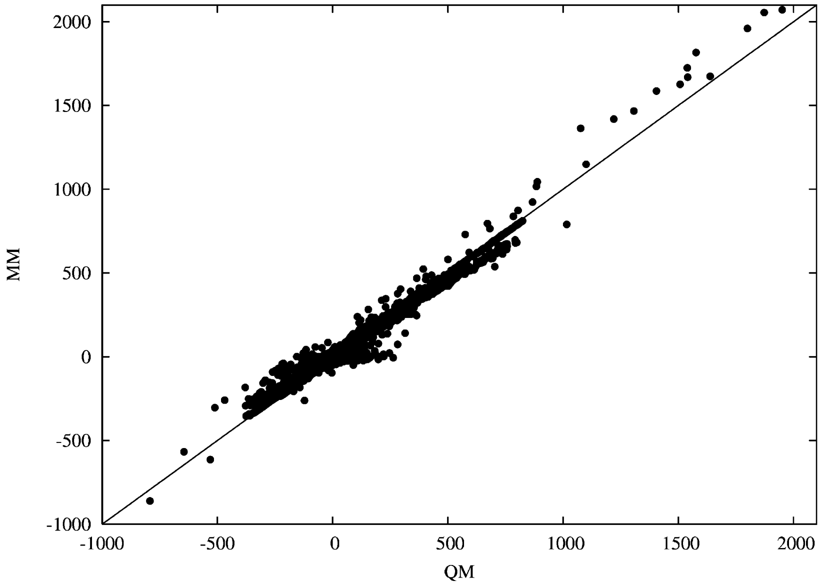


Figure 5. A comparison of the DFT and MM mass-weighted Hessian elements (in kcal mol^{-1} Å^{-2} amu^{-1}).

proached zero during the optimization and often caused instabilities.

Third, some data were excluded from the parametrization because they provided too much noise. This applies to the dihedrals N−Fe−O−H and N−Fe−S−C, because the variation in these two dihedrals is large among the 14 substrates. This would give a low force constant in a parametrization of those dihedrals, and, therefore, the optimization may randomly end up in different local rotational minima, which gives problems in the parametrization.

Q2MM Implementation with Amber. The parametrization was performed with the Q2MM approach,[21] which we have interfaced with Amber. This method tries to minimize the deviation of geometries and Hessian elements between the DFT and MM data using a penalty function that gives different weights to different kinds of data. The geometries were described as lists of all bonds, angles, and dihedral angles, rather than by absolute positions. The weight factors of the various data types were 100 Å^{-1} for bonds, 2 degree^{-1} for angles, 1 degree^{-1} for torsions, and 0.01−0.1 mol Å^2/kcal for Hessian elements (0.01 for elements involving interactions of an atom with itself, 0.02 for atoms bound to each other, 0.04 for atoms connected by two bonds, 0.1 for atoms connected by three bonds, and 0.01 for all other elements).[21] In the following, we will describe the changes made compared to the original implementation.

To generate structural data and start files for MM minimization, we read PDB files of the DFT structures into the Amber module tleap and generated coordinate and topology files. The coordinate files were used as reference structures. Using local programs, we extracted all bond lengths, angles, and dihedrals using these coordinate and topology files. Reference Hessians were taken from the Gaussian03 output files, and the negative eigenvalue was adjusted to a large positive value (1.0 au) before use, to enable us to model the transition state as a minimum.[21,41]

The structures were then minimized, either by the conjugate gradient method in the sander module or by Newton−Raphson method in the nmode module of Amber.[35] In this application, we have used nmode, because it is more robust and very seldom fails, but it is free to the user to decide what program to use. Finally, the Hessian matrix of the minimized structure was calculated by the nmode module. It was necessary to modify nmode to write out the forces and the mass-weighted Hessian matrix to a file.

The implementation of Q2MM for Amber is available from the authors on request. A detailed description of its use is available at http://www.teokem.lu.se/∼ulf/Methods/ponparm.html.

Results and Discussion

Molecular Mechanics Parameters. The final transition-state force field parameters are reported in the Supporting Information (Tables S1−S4). Figure 5 shows the relation between the DFT and MM Hessian elements. It is worth noting that there are over 180 000 data points in this figure, and the great majority of these are found along the diagonal. The correlation coefficient is 0.986. Considering that the force field is a compromise of 14 different structures, the fit is impressive and of a quality similar to previous parametrizations with the Q2MM method.[42−44] However, there is a hint of a straight line along the *x*-axis (i.e., with the MM Hessian = 0). These points represent interactions that cannot be described by the MM force field in eq 1 (e.g., trans-effects[45] or torsions across the metal), and, therefore, no parameter modifications can improve their fit.

Of the optimized parameters, two force constants for angles are quite high. These are the constants for the angles NO−FE−NP and FE−SQ−CT, which are 608 and 348 kcal mol^{-1} rad^{-2}, respectively. One could argue that these are too high and force them to have a lower value, but this gave much larger errors for the N−Fe−N and N−Fe−S angles. Therefore, we have chosen to keep them at these optimized values. Most probably, they compensate for differences in the description of nonbonded forces between MM and QM methods (remember that the Amber van der Waals parameters and the RESP charges were *not* varied in the parametrization).

Table 1. Root-Mean-Squared Deviation for All Bonds, Angles, and Dihedral Angles (in Å and deg, Respectively) between the Structures Optimized with DFT and with the Force Field[a]

training set					test set				
model	bonds	angles	dihedrals	ΔE	model	bonds	angles	dihedrals	ΔE
3	0.008	1.17	1.93	13.2	1	0.017	1.13	2.28	7.8
4	0.007	1.09	1.78	13.7	2	0.010	1.16	2.03	12.2
5	0.010	1.30	2.07	22.2	9	0.009	1.20	1.70	22.3
6	0.013	1.21	2.03	29.2	12	0.010	1.36	2.16	41.0
7	0.009	1.13	2.06	18.3	13	0.016	1.15	2.03	22.7
8	0.010	1.14	1.95	25.9	15	0.017	1.25	2.11	16.1
10	0.012	1.19	2.51	16.6	18	0.010	1.70	2.29	25.1
11	0.009	1.07	2.49	18.3	21	0.011	1.24	1.98	36.6
14	0.015	1.13	2.26	25.9	22	0.010	1.34	2.36	36.1
16	0.011	1.25	2.29	16.6	24	0.011	0.98	2.30	30.1
17	0.010	1.00	2.22	12.3					
19	0.011	1.61	2.10	36.9					
20	0.009	1.09	1.98	34.1					
23	0.011	1.03	2.22	35.4					
av	**0.010**	**1.17**	**2.14**	**22.8**		**0.012**	**1.25**	**2.12**	**25.0**

[a] Data for both the training and test sets are given. Dihedrals with zeroed force constants, any angle larger than 150°, or only consisting of substrate atoms are not included. ΔE is the energy difference (in kJ/mol) of the DFT and MM structures calculated at the B3LYP/DZP/6-31G* level. The model numbers refer to Figures 2 and 3.

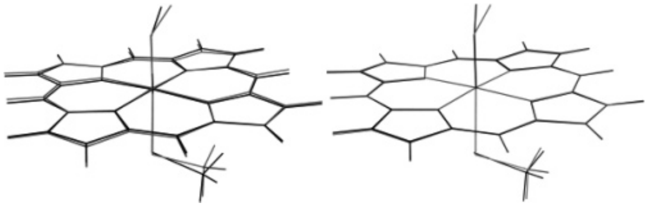


Figure 6. Overlay of the parametrized part of structures 12 and 24, which have RMSDs of 0.10 Å and 0.05 Å, respectively, for this part of the structure (the highest and lowest rmsd of all 24 structures). The DFT structures are black, and the MM structures are gray.

Reproducing Geometries. Geometries obtained after minimization with our optimized force field reproduce the DFT geometries quite well, disregarding parts of the structures that are described only by the GAFF force field (the parameters determining these parts were not optimized). When including all data (except dihedrals with an angle larger than 150°), the root-mean-squared deviation (rmsd) of bonds, angles, and dihedrals are 0.010 Å, 1.17°, 7.1° for the training set and 0.012 Å, 1.25°, 8.2° for the test set (the full data set is shown in Table 1). If we also exclude the dihedrals with a force constant of zero and those that are described by the GAFF force field, the average rmsd of the dihedrals drops to 2.1° for both the training and test sets. This shows that the parametrized part of the structures is excellently described. An overlay of the structures with the highest and lowest rmsd (of the atomic positions, disregarding the substrate) are shown in Figure 6. It can be seen that the heme ring is almost perfectly reproduced, whereas the dihedrals to the substrate and to the cysteine ligand show larger discrepancies.

The reactive center around the oxoferryl group and the reactive hydrogen atom of the substrate is the most important part of the transition-state structure, but it is also the part that is hardest to describe with the force field, because it is a transition state and the DFT structures show quite extensive variations for the bond lengths and angles around the reactive hydrogen atom. In Table 2 we illustrate how well the crucial bonds and angles around the reactive hydrogen are reproduced by the force field. It can be seen that the errors of the O−H and H−C bonds as well as the O−H−C angle are somewhat larger than the average errors for all bonds and angles (in Table 1), but they are still of the same size as the variation among the DFT structures, which is as good as one can expect to get when using many different structures in a parametrization.

Specificity versus Generality. In order to test how much the force field is deteriorated by the use of structures from many different substrates, we made a separate force field, based on a single DFT structure (model 4, isobutane). This gave RMSDs of 0.008 Å (bonds), 1.08° (angles), and 3.10° (dihedrals, 1.98° when excluding the substrate dihedrals). This is actually not significantly better than for the general force field for the bonds and angles (0.007 Å, 1.09°, and 3.37°/1.78°, cf. Table 1). Looking at the specific bonds and angles around the reaction center, we get the following deviations: Fe−O +0.003 Å, O−H +0.014 Å, H−C +0.032 Å, Fe−O−H +0.96°, and O−H−C +0.29°. These errors are similar to those in the full parametrization except for the Fe−O−H, the angle which for this system had an error of −3.35° in the full parametrization. Thus, there is no significant gain in performing a separate parametrization for each substrate.

Energies. Next, we tested if the MM optimized structures can be used to calculate DFT energies, without any further optimization. Unfortunately, this was not possible; single-point DFT energy calculations on the MM optimized structures gave 8−41 kJ/mol higher energies than those obtained from the DFT optimized structures (Table 1, column ΔE). This shows that even if the force field gives quite good geometries, one still needs to run a DFT geometry optimization to get reliable DFT energies. Interestingly, the major difference in energy comes from the substrates, which are

Table 2. Geometry around the Reaction Center Compared to DFT Data[e]

	training set				test set			
	DFT av[a]	MAD[b]	MM err[c]	MM MAD[d]	DFT av[a]	MAD[b]	MM err[c]	MM MAD[d]
Fe−O	1.743	0.010	0.009	0.011	1.745	0.011	0.007	0.012
O−H	1.238	0.025	0.006	0.026	1.219	0.040	0.024	0.046
H−C	1.316	0.017	0.012	0.022	1.333	0.027	−0.008	0.033
Fe−O−H	122.43	1.51	0.82	1.93	122.09	1.64	1.03	1.90
O−H−C	168.49	3.83	−0.84	2.87	168.69	2.99	−1.93	3.56

[a] Average of all DFT structures. [b] Mean absolute deviation of the DFT structures. [c] Average deviation of the MM optimized structures compared to the corresponding DFT structure. [d] Mean absolute deviation of the MM optimized structures compared to the corresponding DFT structure. [e] The bonds are in Å and the angles are in deg.

not parametrized: If the DFT energies were calculated only for the heme groups, with the substrate converted to methane for models 4, 10, 19, and 23, ΔE was reduced from 14−37 kJ/mol to 7−15 kJ/mol (note also that the smallest value for ΔE, 8 kJ/mol, is obtained with the smallest substrate, methane).

To make sure that there is not a significant contribution to ΔE from dispersion interactions (which are present in MM but poorly described by DFT), we also did a parametrization of model 19 (propen-2-ol), in which we also included all substrate parameters and dihedrals into the parametrization. The resulting ΔE was only 9 kJ/mol (37 kJ/mol for the general parametrization), which verifies that the substrate parameters cause the major part of ΔE.

Conclusions

In this paper, we describe an implementation of the ideal and automatic parametrization method by Norrby and Liljefors[41] for the widely used Amber software.[35] This method takes into account all interactions (bonded as well as nonbonded) in a self-consistent way during the iterative parametrization and therefore provides the best possible structure, given the functional form of the force field (eq 1). In particular, it provides appreciably better structures than methods that try to extract the bonded parameters in the force field directly from the Hessian matrix[46] (without taking into account that the Hessian elements also contain contributions from all nonbonded interactions). The method minimizes a penalty function consisting of the squared deviation of the optimized force-field structures from the reference values (in our case obtained by DFT calculations) for all bonds, angles, and dihedrals as well as the elements of the Hessian matrix, all properly weighted according to the acceptable error of each type of data.[41] The method is fully automatic, but it must be carefully checked that one does not end up in an unphysical local minimum. Of course, it is more time-consuming than simpler methods—for the present complicated application, a full optimization of all the parameters typically took about 1 week.

Using this method, we have constructed a general force field for the transition state of the hydrogen abstraction from sp^3 carbons in the cytochromes P450. Transition states cannot be treated with standard MM methods.[14] Therefore, we have used the Q2MM approach,[41] in which the transition state is treated as a normal minimum by switching the negative eigenvalue to a large positive value. Thereby, we can use a standard MM program to obtain structures of the transition state, and essentially all starting structures will converge to the transition state.

In order to obtain a force field that is as general as possible, we have used structures of transition states for 14 different substrates (Figure 2), which is an unusually large set of structures for a force-field parametrization. Still, the results are impressive: The MM optimized bond lengths and angles reproduce those obtained by DFT with an average rms deviation of only 0.01 Å and 1.2°. The dihedrals of the porphyrin ring are equally well reproduced with an average rms error of 2.1°. In most cases, our force field gives errors of the same size as the variations in the input data, which is as good as possible. However, for the dihedrals between the heme group and the substrate, the result is worse, because the 14 structures in the training set show large variations for these torsions. In real applications of the force field in proteins, this problem is less serious, because the torsions are low-energy modes and their actual values are typically dictated by the surrounding protein.

No attempt has been made to optimize the force field for the substrates, because we intended to keep the parameters as general as possible, to allow simulations of any druglike molecule. Therefore, we have used the standard GAFF force field for the substrates.[29] Of course, this leads to worse results for the substrates, but the results are not worse than in a normal use of the GAFF force field.

The use of an unusually large number of structures in the training set has ensured that a versatile force field is obtained. In fact, there is no significant difference in the performance of the force field for the training and test sets (Tables 1 and 2). We have not observed any conflicts between the various structures, except for the above-mentioned dihedrals. For example, the RESP charges for the 14 structures differed by less than 0.05 *e* for the exposed (i.e., not buried) hydrogen atoms. Most importantly, a separate parametrization for a single structure did not give any significantly improved force field.

Thus, the new force field gives excellent structures of the transition state for aliphatic hydrogen abstraction. Unfortunately, the structures are still not good enough to give accurate energies—single-point DFT calculations on the MM structures give ~24 kJ/mol higher energies than on the DFT structures. The majority of this difference comes from the substrate, which is treated with the original GAFF force field. However, even if the substrate is also parametrized, the difference is still 9 kJ/mol, showing that it is very hard to obtain accurate DFT energies from MM structures, obtained with a force field with the simple functional form in eq 1.

We see many uses of our transition-state force field. First, it can be used to rapidly obtain starting structures to DFT optimizations of transition states for other substrates. There is a great interest of predicting the reactivity of drug candidates with the CYPs, and the intrinsic reactivity of the drugs are most accurately determined by DFT methods.[9] Unfortunately, such calculations are very time-consuming (about 1 week for the small substrates in Figures 2 and 3). Good starting structures can make such calculations much faster.

Second, the force field can be used for molecular dynamics simulations of the transition state, e.g., to study how it may be stabilized by the protein. The Q2MM force field shares many properties with the empirical valence bond model potential.[16] For example, it will be possible to explore variations in the transition state through molecular dynamics or conformational searching. This will not in itself give free energies of activation, but since the transition-state cross-section is faithfully reproduced, the vibrational and conformational contributions to the free energy should be obtainable from the force field.

Third, the new force field can be used to dock various molecules into a CYP enzyme. Such docking studies will show if a drug candidate sterically fits into the active site and therefore may be a substrate of the enzyme (to be a substrate, the drug must pass the transition state). Such a transition-state docking would provide a more restrictive and therefore more discriminative test than a standard docking of the ground state of the drug candidate to an arbitrary state of the enzyme.

As with other MM methods, energies obtained with the present force field are comparable only if the models contain exactly the same bonded and nonbonded interactions. Therefore, it is in general not possible to compare the reactivity of different atoms on a substrate (i.e., the regioselectivity). However, for the special case of hydrogen atoms bound to the same carbon atoms in a substrate, the interactions are the same and the energies are comparable. Therefore, our Q2MM force field can directly be applied to study this special case of regioselectivity. Moreover, it should be possible to study differences in the regioselectivity in different CYP isoforms by studying proper energy differences. However, to compare reactions at different sites in a substrate molecule, the docking energies must be combined with estimates of the intrinsic reactivity of each site, obtained by DFT or other quantum mechanical methods.[9] Investigations along these lines are currently performed in our laboratory.[47]

Acknowledgment. This investigation has been supported by funding from the research school in pharmaceutical science (FLÄK), the Swedish Research Council, the Carlsberg Foundation, and by computer resources of LUNARC at Lund University.

Supporting Information Available: Final transition-state force field for the aliphatic hydroxylation in cytochrome P450 (charges, bond, angle, and dihedral parameters) as well as figures of the atom names and atom types used and Amber topology and parameter files of the force field. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bertz, R. J.; Granneman, G. R. *Clin. Pharmacokinet.* **1997**, *32*, 210−258.

(2) Evans, W. E.; Relling, M. V. *Science* **1999**, *286*, 487−491.

(3) Rydberg, P.; Sigfridsson, E.; Ryde, U. *J. Biol. Inorg. Chem.* **2004**, *9*, 203−223.

(4) Meunier, B.; de Visser, S. P.; Shaik, S. *Chem. Rev.* **2004**, *104*, 3947−3980.

(5) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279−2328.

(6) Hirao, H.; Kumar, D.; Thiel, W.; Shaik, S. *J. Am. Chem. Soc.* **2005**, *127*, 13007−13018.

(7) de Visser, S. P.; Kumar, D.; Cohen, S.; Shacham, R.; Shaik, S. A. *J. Am. Chem. Soc.* **2004**, *126*, 8362−8363.

(8) Park, J. Y.; Harris, D. *J. Med. Chem.* **2003**, *46*, 1645−1660.

(9) Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. *J. Med. Chem.* **2006**, *49*, 6489−6499.

(10) de Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. *J. Med. Chem.* **2005**, *48*, 2725−2755.

(11) Jones, J. P.; Korzekwa, K. R. *Methods Enzymol.* **1996**, *272*, 326−335.

(12) Vermeulen, N. P. E. *Curr. Top. Med. Chem.* **2003**, *3*, 1227−1239.

(13) Hermann, J. C.; Ghanem, E.; Li, Y.; Raushel, F. M.; Irwin, J. J.; Shoichet, B. K. *J. Am. Chem. Soc.* **2006**, *128*, 15882−15891.

(14) Jensen, F.; Norrby, P.-O. *Theor. Chem. Acc.* **2003**, *109*, 1−7.

(15) Warshel, A. *J. Am. Chem. Soc.* **1980**, *102*, 6218−6226.

(16) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523−2544.

(17) Rappé, A. K.; Pietsch, M. A.; Wiser, D. C.; Hart, J. R.; Bormann, L. M.; Skiff, W. M. *Mol. Eng.* **1997**, *7*, 385−400.

(18) Kim, Y.; Corchado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718−2735.

(19) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.

(20) Jensen, F. *J. Comput. Chem.* **1994**, *15*, 1199−1216.

(21) Norrby, P.-O. *J. Mol. Struct.* (*Theochem*) **2000**, *506*, 9−16.

(22) Eksterowicz, J. E.; Houk, K. N. *Chem. Rev.* **1993**, *93*, 2439−2461.

(23) Oda, A.; Yamaotsu, N.; Hirono, S. *J. Comput. Chem.* **2005**, *26*, 818−826.

(24) Kuczera, K.; Kuriyan, J.; Karplus, M. *J. Mol. Biol.* **1990**, *213*, 351−373.

(25) Banci, L.; Bertini, I.; Bren, K. L.; Gray, H. B.; Sompornpisut, P.; Turano, P. *Biochemistry* **1995**, *34*, 11385−11398.

(26) Autenrieth, F.; Tajkhorshid, E.; Baudry, J.; Luthey-Schulten, Z. *J. Comput. Chem.* **2004**, *25*, 1613−1622.

(27) Ma, J.-G.; Zhang, J.; Franco, R.; Jia, S.-L.; Moura, I.; Moura, J. J. G.; Kroneck, P. M. H.; Shelnutt, J. A. *Biochemistry* **1998**, *37*, 12431−12442.

(28) Park, H.; Lee, S. *J. Comput.-Aided. Mol. Des.* **2005**, *19*, 17−31.

(29) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(30) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(31) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(32) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571−2577.

(33) 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, 2003.

(34) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(35) 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, CA, 2004.

(36) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(37) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431−439.

(38) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(39) Li, H. Cytochrome P450. In *Handbook of metalloproteins*; Messerschmidt, A., Huber, R., Poulos, T., Wieghart, K., Eds.; Wiley: Chichester, pp 267−282.

(40) Yano, J. K.; Wester, M. R.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.* **2004**, *279*, 38091−38094.

(41) The Q2MM method has an incorrect response to the increase of steric strain along the reaction coordinate.[21] The problem can be somewhat alleviated by minimizing distortions along the reaction coordinate. This is accomplished by selection of a large value for the replacement eigenvalue in the Hessian modification step, in effect freezing movement along the reaction coordinate, and thus minimizing the systematic error inherent in the method.[21] However, for cases where the systematic error is expected to influence the results significantly, the Q2MM force field should only be used as a conformational search tool, with final results obtained from a method with a correct response to changes in exothermicity, like the SEAM method.[20]

(42) Norrby, P.-O.; Brandt, P.; Rein, T. *J. Org. Chem.* **1999**, *64*, 5845−5852.

(43) Norrby, P.-O.; Rasmussen, T.; Haller, J.; Strassner, T.; Houk, K. N. *J. Am. Chem. Soc.* **1999**, *121*, 10186−10192.

(44) Rasmussen, T.; Norrby, P.-O. *J. Am. Chem. Soc.* **2003**, *125*, 5130−5138.

(45) Brandt, P.; Norrby, T.; Åkermark, B.; Norrby, P.-O. *Inorg. Chem.* **1998**, *37*, 4120−4127.

(46) Seminario, J. M. *Int. J. Quantum Chem.* **1996**, *60*, 1271−1277.

(47) Rydberg, P. Theoretical studies of cytochrome P450, Ph.D. Thesis, Lund University, 2007.

CT700110F

JCTC Journal of Chemical Theory and Computation

Force Field Modeling of Amino Acid Conformational Energies

Jakub Kaminský[†,‡] and Frank Jensen*[,†]

Department of Physics and Chemistry, University of Southern Denmark, DK-5230 Odense, Denmark, and Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, 166 10 Prague, Czech Republic

Received April 2, 2007

Abstract: The conformational degrees of freedom for four amino acids in a model peptide environment have been sampled with density functional and second-order Møller-Plesset methods. Geometries have been optimized with an augmented double-ζ basis set and relative energies estimated by extrapolation of results using double, triple, and quadruple-ζ basis sets and including higher order correlation effects. In addition, the effects of vibrational zero point energies and solvation have been considered. The density functional method is unable to locate all the minima found at the MP2 level, which most likely is due to the inability for describing dispersion interactions. The use of basis sets smaller than augmented polarized double-ζ with the MP2 method may also in some cases lead to artifacts. The effects on relative energies by enlarging the basis set beyond an augmented triple-ζ and including higher order correlation beyond MP2 is small. The MP2/aug-cc-pVTZ level is recommended as a level of theory capable of an accuracy of ~1 kJ/mol for relative conformational energies. Eight different force fields are tested for reproducing the electronic structure reference data. Force fields that represent the electrostatic energy by fixed partial charges typically only account for half of the conformations, while the AMOEBA force field, which includes multipole moments and polarizability, can reproduce ~80% of the conformations in terms of geometry. This not only suggests that multipole moments and polarizability are important factors in designing new force fields but also indicates that there is still room for improvements.

Introduction

Investigating the three-dimensional structure of proteins is of prime importance for understanding their biological function. X-ray diffraction methods have in recent years provided a wealth of structural information,[1] but these methods only provide a static picture corresponding to a solid-state structure, and certain classes of proteins are difficult to investigate due to crystallization problems. NMR methods are capable of providing structural information under conditions closer resembling the natural biological environment (solution), but these methods are limited to relatively small systems.[2] Experimental information regarding the time-dependent changes in structural features can be obtained from time-resolved spectroscopy, but this can rarely provide details at the atomic level.

Theoretical simulations can in principle provide (time-dependent) atomic resolution information of proteins and other macromolecules. Predicting the three-dimensional structure of a protein from its primary amino acid sequence is still an unsolved problem due to the huge size of the phase space. Simulations of macromolecules therefore usually start from geometries derived from experiments, either directly from an X-ray structure of the actual system or by homology modeling of a closely related structure.

An essential part of a simulation is the energy function. Car−Parrinello methods take the electrons explicitly into

* Corresponding author e-mail: frj@ifk.sdu.dk.

† University of Southern Denmark.

‡ Czech Academy of Sciences.

10.1021/ct700082f CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/18/2007

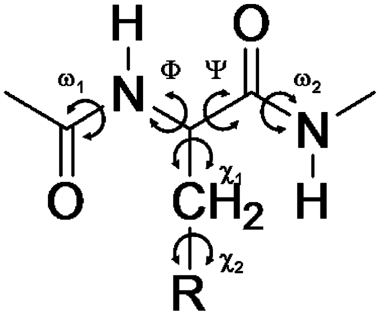
account by propagating both the nuclear and wave function parameters in time, but the computational requirements of these methods are sufficiently high that only relatively small systems can be simulated for short time spans (picoseconds).[3] Performing atomistic simulations for systems containing thousands of atoms in the nano- or microsecond time regime is only possible using parametrized energy functions.[4] Such force field methods rely on using atoms as the fundamental building block and providing the bonding information explicitly. It is evident that the quality of the energy function determines the ultimate accuracy that can be extracted from simulations.

The internal energy of a molecule is parametrized in a force field approach by a function containing constants that are fitted to reproduce experimental or quantum mechanical results.[5,6] All force fields contain the terms shown in eq 1, and additional terms may be added to improve the performance.

$$E_{\mathrm{FF}} = E_{\mathrm{str}} + E_{\mathrm{bend}} + E_{\mathrm{tors}} + E_{\mathrm{vdw}} + E_{\mathrm{el}} \qquad (1)$$

The stretch and bending terms are relatively easy to parametrize, and they have little influence on conformational degrees of freedom. The nonbonded van der Waals and electrostatic terms determine intermolecular interactions and are thus very important for calculating, e.g., binding affinities of drug molecules to enzymes. The nonbonded terms together with the torsional energy determine the internal (conformational) degrees of freedom. The existence and relative stabilities of conformations for a given molecule is thus determined by a delicate balance between these three energy terms. Being able to correctly predict structural and energetic features of conformations is an essential quantity for bringing force field methods into the predictive region of theoretical methods. Despite this, there have only been a few studies where the performance of force field for predicting structural and energetic information of conformations has been addressed.[7] Part of this calibration problem is, of course, due to the lack of reference data of sufficient accuracy.

One of the main limitations of current force field methods is the use of fixed partial charges for parametrizing the electrostatic interaction, as this neglects the known conformational dependence of the atomic charges and polarization effects.[8,9] This becomes especially problematic for polar systems, like proteins, where the electrostatic component often accounts for a large fraction of the total energy.[10] Currently research therefore focuses on improving the representation of the electrostatic energy by including multipole moments and atomic polarization.[11] These quantities are usually derived from electronic structure calculations on small model systems, but there is no clear consensus regarding which electronic structure method to use or the procedure for extracting the parameters. Lacking an objective criterion, a distinction between the different procedures will have to rely on the performance for reproducing structural and energetic feature of the target molecular systems. We here report a systematic study of the conformational space of four amino acids in a model peptide environment and evaluate the performance of different force fields for reproducing the reference data. The aim is 2-fold, to establish the necessary theoretical level for reliably determining structures and energies of amino acid conformations for use on a larger variety of systems and to subsequently use this data for developing more accurate force fields.



R = H alanine
R = OH serine
R = SH cysteine

Figure 1. Illustration of the torsional angles in the model systems.

Computational Details

All electronic structure calculations have been carried out using the Gaussian 03 program package.[12] Force field calculations have been performed with the MacroModel[13] and Tinker programs.[14] The MacroModel program allows calculations with the AMBER94, MM2*, MM3*, MMFFs, and OPLS_05 force fields which are slightly modified versions of the parent AMBER,[15] MM2,[16] MM3,[17] MMFF,[18] and OPLS[19] force fields. The Tinker program has been used for the AMBER99,[20] CHARMM27,[21] and AMOEBA[22] results. All calculations have been done using a constant dielectric constant of 1.0 for comparisons with the calculated gas-phase conformational energies. The AMOEBA force field includes distributed multipoles and polarizabilities for representing the electrostatic energy, while the others use a fixed partial charge model.

We have selected the glycine, alanine, serine, and cysteine amino acids as our trial set of systems, with side chains corresponding to H, CH_3, CH_2OH, and CH_2SH, respectively. N-Acetyl and N-methylamide functional groups were added to the C- and N-terminus to mimic the environment in longer peptide chains. The notation for the torsion angles is depicted in Figure 1. Trial structures were generated by varying the Φ and Ψ torsion angles in steps of 30°, while the χ_1 torsional angle was varied in steps of 120°. The torsion angle χ_2 was preset at a value of 180°, while the torsional angles ω_1 and ω_2 associated with the terminal amide bonds were given initial values of 180°, corresponding to trans junctions. This produces a total of 144 starting geometries for glycine and alanine and 432 for serine and cysteine. Some of these conformations are, of course, related by symmetry. We generated additional trial conformations for the serine and cysteine systems by varying the χ_2 torsional angle. All trial structures were optimized without any restraints at the MP2/6-31G(d,p) level and characterized as true minima by frequency calculations, and the unique structures were reoptimized with the aug-cc-pVDZ basis set. Although we do not claim that this procedure exhausts the conformational space, it certainly will locate the large majority of the important conformations.

Improved estimates of the relative conformational energies were obtained by single point MP2 calculations using the cc-pVDZ, cc-pVTZ, cc-pVQZ, and aug-cc-pVTZ basis sets, and method dependency was tested using CCSD(T)/cc-pVDZ calculations. Our best estimate of relative energies was obtained by separately extrapolating the Hartree−Fock[23] and MP2[24] energies to the basis set limit and adding the difference between the CCSD(T) and MP2 results with the cc-pVDZ basis set to the extrapolated MP2 results.

The same initial pool of trial structures were also optimized at the B3LYP/6-31G(d,p) level, and improved estimates of relative energies were obtained using single point calculations at the B3LYP/aug-cc-pVTZ level in order to investigate the performance of a popular density functional model. Solvent effects were estimated by single point B3LYP/6-31G(d,p) and MP2/6-31G(d,p) calculations using the PCM continuum solvent model for water.[25] Zero point energy differences were evaluated using harmonic frequencies. Force field conformations were determined by a Monte Carlo type random variation of the torsional angles followed by energy minimization, where the number of trial steps was set sufficiently high to ensure that all possible conformations are located.

In order to establish the correspondence between the force field and ab initio conformations, we have calculated root-mean-square (rms) deviations in the torsional angle space. In most cases this allows a unique connection to be made. For large rms deviations we have performed MP2/6-31G-(d,p) optimizations starting from the force field geometries in order to establish the nearest conformation.

Previous Work

Previous work on calculating conformations for peptide model systems has employed various methods based on wave function or density functional theory (DFT). A relatively large number of calculations of peptide systems have been published during the last two decades, but most of them focused on alanine or glycine systems[26−33] and only a few have considered other amino acids (e.g., serine,[34−37] cysteine,[38] asparagine,[39] proline[40]). Several of these studies have used di- or tetrapeptides as model systems, but the size and flexibility of these systems compared to simple amino acids necessitated simplification in the exploration of the conformational space. A typical approach has been to optimize a number of Φ, Ψ restrained conformers at the Hartree−Fock (HF) level of theory, followed by single point MP2 calculations. Low-energy conformations identified by this procedure is subsequently fully relaxed. Beachy et al. reported that for their test suite consisting of simple flexibile molecules such as methyl vinyl ether, the use of MP2/6-31G(d,p) geometries instead of HF/6-31G(d,p) ones changes the conformational energies by only a few tenths of a kcal/mol, and they therefore considered HF geometries as adequate.[41] Gould et al., however, showed that some minima on the potential energy surface could not be located at the HF level, while the MP2 level performed much better.[42] Beachy et al. showed for the alanine dipeptide model that canonical MP2 energies are contaminated by basis set superposition errors, which artificially lower the energies of the compact structures relative to extended ones.[43] Császár showed that inclusion of higher order (MP3, MP4, CCSD, and CCSD(T)) electron correlation in most cases makes only small contributions to the energy differences of glycine conformers and concluded that large basis set MP2 calculations should result in highly accurate energies.[44]

DFT methods have in recent years been accepted by the chemistry community as a cost-effective approach for the computing of molecular structures, vibrational frequencies, and energies of chemical reactions.[45−49] Unfortunately, most current DFT methods do not account for dispersion, and for large systems it is anticipated that intramolecular dispersion type interactions will play a significant role. DFT results for describing the conformational space are consequently likely to be of lower accuracy than MP2 results.[50]

For systems as large as tetrapeptides the number of possible starting structures for performing a systematic conformational search is prohibitive for use with electronic structure methods. Most previous work on such systems has therefore employed molecular mechanics methods for an initial evaluation of the conformational space. Beachy et al. selected 10 conformers of the alanine tetrapeptides based on AMBER* results and recomputed them with several different molecular mechanics force fields and compared the results to LMP2/cc-pVTZ//HF/6-31G(d,p) data.[53] A number of higher energy structures were deliberately chosen among the 10 structures in order to sample the conformational space representing low-lying minima. Higher energy conformers could potentially be populated due to solvation effects or by interaction with other parts of a particular protein. Gresh et al. treated Beachy's 10 alanine tetrapeptide conformations with their SIBFA force field, which includes multipoles and polarization, and evaluated relative energies by single point HF, DFT, and MP2 calculations.[51] The relative energies calculated at the LMP2/6-311G(d,p) level were reproduced by the SIBFA force field with an rms value of ∼5 kJ/mol. It was also found that the MP2 method tended to exaggerate the stabilization of folded structures, presumably due to BSSE effects.

The conformations of small linear peptides containing the Arg-Gly-Asp sequence have been studied by NMR and/or molecular dynamics simulations. These studies showed, that although linear peptides are expected to be very flexibile, the linear Arg-Gly-Asp-containing peptides could adopt mainly β-turns for Arg-Gly and Gly-Asp sequences.[52−54] The unfolding of the Ac-Ala-Ala-NHMe and Ac-Pro-Ala-NHMe sequences in water was monitored by Tobias et al.[55,56] using MD simulations with the CHARMM force field. Recently, ab intio molecular dynamic simulation techniques employing a Gaussian implementation of Kohn−Sham density functional theory was used to study the gas-phase conformational dynamics of an alanine dipeptide analog,[57] although the full range of Φ, Ψ values have yet to be sampled using this approach.

The conformational energies of dipeptides have a direct relationship with protein modeling by force fields, which has motivated several molecular mechanics studies of alanine and glycine dipeptides. Despite the fact that many force fields have been applied to a wide range of peptides, the question of their accuracy still remains. Molecular dynamics simulations of polypeptides in solution showed the existence of π-helical conformations,[58,59] but this has been reported by Feig et al. to be associated with force fields bias.[60] For systems like dipeptides or tetrapeptides, most force fields have problems reproducing quantum mechanical results.[43,61−64] We therefore feel that there is a need for improving the treatment of peptide conformations by empirical force fields. To date, optimization of empirical force fields for the protein backbone has typically been based on reproducing the relative energies of selected conformers of the alanine and glycine dipeptides.[55,65−69] This approach has the limitation that it ignores high-energy regions of the Φ, Ψ conformational space, although MacKerell et al. recently have reported an optimized CHARMM force field, where cross terms were added in order to treat the entire range of Φ, Ψ values.[70]

As indicated by the above references, a large majority of previous work has relied on force field, HF or DFT methods for selecting and optimizing conformations, and only rarely has the influence of improving the theoretical procedure by enlarging the basis set or treatment of electron correlation been addressed. The results presented in the next section provide a systematic study of the effects of different computational procedures.

Results

While HF and DFT methods have been popular for searching the conformational space due to their favorably computational requirements, it should be recognized that they lack a description of dispersion interactions. Even for a medium sized system, however, it is expected that intramolecular dispersion interactions will be important for the topology of the potential energy surface. The MP2 method is the lowest level of theory that provides a qualitative correct description of dispersion, and it is at present also the only computationally feasible method for performing a large number of structure optimization. The overall topology of the energy surface is expected to be correctly reproduced with a double-ζ type basis set, and we have selected the 6-31G(d,p) basis set for computational efficiency for the initial screening of the conformational space. The resulting conformations have been refined with the more flexible aug-cc-pVDZ basis set, which in some cases leads to different conformations collapsing to the same structure, i.e., the 6-31G(d,p) basis set may in some cases produce artificial minima.

Relative energies are sensitive to the quality of the basis set, and our reference energies are based on MP2 energies extrapolated to the basis set limit from calculations with the cc-pVDZ, cc-pVTZ, and cc-pVQZ basis sets and with an additive correction for higher order correlation based on CCSD(T)/cc-pVDZ results. From the observed changes upon improving the basis set or treatment of electron correlation, we estimate that relative energies calculated by this procedure are converged to ~1 kJ/mol. In general we do not expect such an elaborate treatment to be necessary and also report the results obtained at the MP2/aug-cc-pVTZ level, which we feel represent a suitable compromise between accuracy and computational cost.

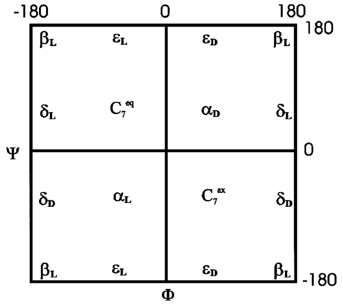


Figure 2. Connection between backbone torsional angles and conformational labeling.

Conformations for peptides are commonly discussed in terms of a Ramachandran map with Φ and Ψ as the variables, where a positive sign indicates a clockwise rotation. Most peptide residues exhibit nine unique conformations in accordance with values of their Φ and Ψ torsion angles, traditionally labeled as α_D, ϵ_D, C_7^{ax}, δ_L, β_L, δ_D, C_7^{eq}, ϵ_L, and α_L, as illustrated in Figure 2.[71] Since we in the present work only consider conformations with a trans peptide bond, the full conformational space includes two (glycine, alanine) or four (serine, cysteine) torsion angles, respectively, and the potential energy surface can therefore be described as a function of two or four variables.

Glycine. Optimization at the MP2/aug-cc-pVDZ level produced two structures, one in the C_7^{ax} and one in the β_L region, as summarized in Table 1. Both conformations were confirmed to be separate minima by frequency calculations and have been reported previously.[52,72,73] Note that C_7^{ax} and C_7^{eq} are symmetrically equivalent for glycine. Bohm et al. reported also a minimum in the α_L region on the AMBER potential energy surface, but this conformation is not a local minimum at the HF/DZP level and is not found in the present work either.[72] Our best estimate for the energy difference between the two conformations is 4 kJ/mol, but both zero point energies and solvent corrections suggest that the β_L conformer may be essentially isoenergetic under solution-phase conditions.

The B3LYP method reproduces the geometries of both minima with a root-mean-square deviation of torsion angles of only 1° relative to the MP2 structures, but the relative energy of the β_L conformer is significantly smaller. Single point MP2 calculations at the B3LYP geometries give relative energies almost identical to those at the MP2 optimized geometries, indicating that this is not a geometry effect. This is slightly at variance with work by Barone et al. where it was concluded that the B3LYP method can reproduce MP2 relative energies for free amino acids with accuracy about 1 kJ/mol,[74] but it confirms the conclusions of Beachy[43] that MP2 prefers folded structures. The zero point energy and solvent corrections determined at the B3LYP level resemble closely those at the MP2 level. Although the PCM model is expected to provide a qualitative indication of the solvent effect, it is unlikely to be able to quantitatively predict the effect of surrounding these polar systems with water, as for example hydrogen bonding is not accounted for. We therefore primarily use the PCM results

Table 1. Glycine Conformations[a]

				MP2				B3LYP			
conf	region	Φ	Ψ	A	B	ΔZPE	ΔPCM	A	ΔZPE	ΔPCM	rms
1	C_7^{ax}	82	−71	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
2	β_L	180	180	5.7	4.1	−2.8	−3.2	1.3	−2.9	−4.6	0
MAD				1.6				2.8			1

[a] Relative energies are in kJ/mol. Glycine conformations optimized at the MP2/aug-cc-pVDZ (MP2) and B3LYP/6-31G** (B3LYP) levels. Relative energies from single point calculations with the aug-cc-pVTZ basis set are labeled with A, while the best estimates obtained by basis set extrapolation and CCSD(T) corrections are labeled B. Changes in relative energies due to zero point corrections (ΔZPE) and PCM (ΔPCM) solvation corrections have been calculated at the MP2/6-31G** and B3LYP/6-31G** levels. Root-mean-square deviation (rms) of the B3LYP torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations.

Table 2. Force Field Predictions of Glycine Conformations[a]

			AMBER94		MM2*		MM3*		MMFFs		OPLS_2005		AMBER99		CHARMM27		AMOEBA	
conf	region	ref	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*
1	C_7^{ax}	0.0	9	0.0	11	0.0	8	0.0	2	0.0	21	0.0	17	0.0	1	0.0	4	0.0
									7	*0.3*	*77*	*29.9*	*54*	*12.5*			*44*	*9.0*
2	β_L	4.1	0	7.9	0	23.6	27	5.2	0	5.9	46	−3.8	0	−1.0	0	3.9	7	4.8
					61	*26.4*												
MAD			5	3.8	6	19.5	18	1.1	1	1.8	34	7.9	9	5.1	1	0.2	5	0.7

[a] Relative are energies in kJ/mol. The best estimates of the relative energies from Table 1 are labeled as ref. Root-mean-square deviation (rms) of the force field torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations not marked in italics.

to estimate solvation effects, but clearly better methods are needed for reliably calculating the effect of placing these systems in an aqueous environment. The difference in the PCM solvent corrections is sufficiently large that it is likely that the β_L conformation is the global minimum in solution.

The results of conformational searches with several force fields are shown in Table 2. For each force field is given the rms deviation of the torsional angles relative to the MP2 structure, and the relative energies calculated by the force field, which can be compared with our reference data (labeled ref in Table 2). Except for the OPLS and AMBER99 force fields, the global minimum in the C_7^{ax} region is reproduced with an acceptable accuracy, and all force fields except MM3 reproduced the β_L minimum accurately in terms of the Φ and Ψ angles. The MM2 force field predicts one additional minimum in the ϵ_D region which collapses to the β_L conformation upon MP2/6-31G(d,p) optimization, and it is therefore labeled as conformation **2** in Table 2. The MMFFs, OPLS, AMBER99, and AMOEBA force fields similarly display a third stable conformation, which upon MP2 optimization converges to the C_7^{ax} minimum, and they are listed under conformation **1** in Table 2. These minima should be considered as "artificial" minima on their respective surface. For characterization purposes, we have arranged them in the tables according to the minima to which they collapse upon MP2 optimization, but the large rms deviations in the torsional energy parameters indicate that they have substantial different geometries.

For systems with many conformations it is useful to characterize the performance in terms of a mean absolute deviation (MAD) relative to the reference data. The lack of a clear one-to-one correspondence between the force field and reference data, however, requires some decisions to be made regarding the selection of data. We have decided to classify the force field conformations into four groups: (1) "Good" minima have torsional rms values less than 40° relative to the MP2 structure. (2) "Poor" minima have torsional rms values larger than 40° but collapse to the indicated minimum upon MP2 optimization. (3) "Missing" minima indicate that there is no force field minimum that is connected to the corresponding MP2 conformation. (4) "Artificial" minima indicate that the force field minimum collapses to the indicated conformation upon MP2 optimization, but there is another force field minimum which provides a better representation of the same conformation. These artificial minima are marked with italics in the tables.

Clearly the rms value of 40° for distinguishing between "good" and "poor" minima is somewhat arbitrary. The MAD values over the conformations have only been calculated for the "good" and "poor" minima, i.e., only the best representation has been included for the conformations with multiple associated force field structures.

The ability to correctly predict geometries of local minima is only one parameter for characterizing an accurate potential energy surface. Of similar importance is the ability to predict the energetic ordering of these minima. All of the present force field can reproduce the C_7^{ax} and β_L minima, and our reference calculations place the latter 4 kJ/mol above the global minimum. This value is reasonably reproduced by the AMBER94, MM3, MMFFs, CHARMM27, and AMOEBA force fields with deviations of a few kJ/mol, while the MM2 provides a significantly higher value. The OPLS and AMBER99 force fields reverse the energetic preference of the two structures (Table 2).

Alanine. The presence of an additional methyl group in the alanine system leads to seven different minima at the MP2/aug-cc-pVDZ level, which are labeled C_7^{eq}, β_L, C_7^{ax}, δ_L, α_D, ϵ_D, and δ_D in agreement with the results obtained by Rodriguez et al.[75] The Φ and Ψ torsion angles and relative energies are presented in Table 3. The minimum in the ϵ_D region (conformation **6**) deserves a special comment. Rodriguez et al. located this structure at the HF/3-21G level as

Table 3. Alanine Conformations[a]

				MP2				B3LYP			
conf	region	Φ	Ψ	A	B	ΔZPE	ΔPCM	A	ΔZPE	ΔPCM	rms
1	C_7^{eq}	−82	76	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3
2	β_L	−161	157	6.3	6.0	−1.4	−6.0	3.4	−1.4	−4.7	5
3	C_7^{ax}	74	−54	9.5	10.1	0.7	−2.8	10.7	0.6	−3.0	3
4	α_L	−83	−10	12.8	13.5	−1.0	−9.9	12.1	−1.1	−9.9	41
5	α_D	64	30	19.8	19.3	−0.4	−13.6	23.3	−0.8	−12.0	4
6	ϵ_D	52	−130	19.7	20.5	−0.1	−2.4				
7	δ_D	−165	−38	26.7	27.1	−0.7	−12.3	27.3	−1.0	−10.0	2
MAD				0.6				1.9			10

[a] Relative energies are in kJ/mol. Alanine conformations optimized at the MP2/aug-cc-pVDZ (MP2) and B3LYP/6-31G** (B3LYP) levels. Relative energies from single point calculations with the aug-cc-pVTZ basis set are labeled with A, while the best estimates obtained by basis set extrapolation and CCSD(T) corrections are labeled B. Changes in relative energies due to zero point corrections (ΔZPE) and PCM (ΔPCM) solvation corrections have been calculated at the MP2/6-31G** and B3LYP/6-31G** levels. Root-mean-square deviation (rms) of the B3LYP torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations.

Table 4. Force Field Predictions of Alanine Conformations[a]

			AMBER94		MM2*		MM3*		MMFFs		OPLS_2005		AMBER99		CHARMM27		AMOEBA	
conf	region	ref	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*	rms	Δ*E*
1	C_7^{eq}	0.0	10	0.0	15	0.0	9	0.0	4	0.0	11	0.0	27	0.0	4	0.0	1	0.0
2	β_L	6.0	14	6.3	18	25.3	14	2.4	4	5.9	7	4.1	9	3.8	12	3.8	5	5.1
3	C_7^{ax}	10.1	9	6.2	3	6.4	3	5.6	1	7.7	4	10.4	19	9.3	10	8.6	1	10.5
4	α_L	13.5									51	12.9	42	10.8			28	11.8
5	α_D	19.3							8	19.8							1	18.7
6	ϵ_D	20.5																
7	δ_D	27.1	20	25.8			14	17.8	6	23.7	9	25.6					3	23.5
MAD			13	1.8	12	11.5	10	5.8	5	1.6	8	1.1	24	1.5	9	1.9	7	1.4

[a] Relative energies are in kJ/mol. The best estimates of the relative energies from Table 3 are labeled as ref. Root-mean-square deviation (rms) of the force field torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations not marked in italics.

well as with several force fields (AMBER, MM+, BIO+, OPLS, and ECEPP/2) but failed to find it at the semiempirical AM1 level. This conformation was not reported in the work of Yu[28] (MP2/6-311G(d,p)), Bohm[72] (HF/DZP), and MacKerell[70] (MP2/6-31G(d) and MP2/6-311++G(d,p)). We were able to optimize the structure at the HF/6-311++G(d,p) and MP2/6-311++G(d,p) levels, in addition to MP2/aug-cc-pVDZ, but it does not exist at the B3LYP level. A B3LYP/6-31G(d,p) optimization starting from the MP2 structure converged to structure **3** in the C_7^{ax} region.

The MP2/6-31G(d,p) optimized geometry for conformer **4** belongs to the δ_L region (−138°, 23°), but a reoptimization with the aug-cc-pVDZ basis set changes the value of the Ψ angle to −10°, which places the conformation in the α_L region. This value is significantly different from previously published values of ∼40°.[27,70,75] The OPLS value is 41°, while AMOEBA, AMBER99, and B3LYP/6-31G(d,p) give intermediate values (11°−23°), and the remaining force fields fail to locate this structure. This conformation corresponds to a right-handed helix in polypeptides, where the Ψ angle has a value ∼ −40°. The computational results suggest that the Ψ torsional energy surface is very flat in this region, and the exact geometry is therefore sensitive to the level of theory employed.

The B3LYP method gives a good agreement with the MP2 geometries, except for the missing conformation **6** and conformation **4**, which has an rms deviation of 41°. Compared to the reference values for the relative energies, the MP2/aug-cc-pVTZ method gives a MAD value of only 0.6 kJ/mol, while the corresponding B3LYP value is 1.9 kJ/mol. Zero point energy corrections give only minor changes in the relative conformational energies, while the PCM solvent model again preferentially stabilized the high-energy conformations and makes the C_7^{eq} and β_L conformations essentially isoenergetic.

All force fields are able to locate the three lowest energy minima in the C_7^{eq}, β_L, and C_7^{ax} regions (Table 4) with good representations of the geometries, and all except MM2, AMBER99, and CHARMM27 can also account for the highest energy conformation **7** in the δ_D region. Moreover, MMFFs locates conformation **5**, OPLS and AMBER99 locate conformation **4**, but none of the force fields finds conformation **6**. The polarizable AMOEBA force field performs well and locates six out of the seven MP2 minima, of which five have very low rms values for the torsional angles. The relative stabilities of the conformations found by the different force field reproduce fairly well the MP2 results, except that conformation **2** is predicted to be significantly destabilized by the MM2 force field.

Serine. The CH_2OH side chain in serine provides additional conformational flexibility compared to alanine, and the polarity and possibilities for hydrogen-bonding result in a more complicated potential energy surface. A conformational search at the MP2/aug-cc-pVDZ level along the lines described in the Computational Details section resulted in 39 different conformations spanning a 70 kJ/mol energy range (Table 5). The global minimum is a C_7^{eq} type conformation, and it is 13 kJ/mol below the second lowest

Table 5. Serine Conformations[a]

						MP2				B3LYP			
conf	region	Φ	Ψ	χ_1	χ_2	A	B	ΔZPE	ΔPCM	A	ΔZPE	ΔPCM	rms
1	C_7^{eq}	−82	73	55	65	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
2	C_7^{eq}	−82	68	179	−66	13.5	12.5	−1.5	−6.4	11.3	−1.8	−6.9	3
3	β_L	−157	−173	−167	79	15.2	14.8	0.1	−9.0	12.2	0.4	−10.4	2
4	β_L	−176	165	−92	51	18.2	17.0	−2.5	−5.7	11.9	−2.6	−6.7	4
5	ϵ_D	63	−138	84	−47	18.3	18.0	−2.5	−2.3	14.0	−2.6	−1.8	20
6	β_L	−157	−174	−169	172	19.1	18.4	−3.8	−8.3	15.6	−3.9	−7.6	3
7	C_7^{eq}	−85	69	−54	−179	24.0	23.0	−3.1	−7.6	22.6	−3.3	−7.0	5
8	C_7^{eq}	−85	72	−51	−68	24.2	23.5	−3.0	−11.8	21.2	−3.0	−10.4	4
9	C_7^{ax}	74	−37	84	−57	26.6	24.1	−0.6	−12.0	26.5	−1.1	−12.6	6
10	β_L	−155	177	66	−58	25.5	25.0	−3.9	−15.6	23.3	−4.4	−11.9	2
11	C_7^{eq}	−82	79	−65	51	28.3	26.2	−3.4	−14.9	26.3	−3.4	−13.0	3
12	β_L	−158	162	−175	−76	27.1	27.5	−3.7	−16.0				
13	C_7^{eq}	−104	8	53	170	27.3	27.5	−4.8	−19.2	24.5	−5.2	−17.5	8
14	ϵ_D	60	34	−166	−62	29.3	27.9	−1.8	−21.3	29.3	−2.3	−18.4	3
15	C_7^{ax}	76	−53	−52	82	31.0	29.4	−1.4	−8.0	31.0	−1.5	−7.4	6
16	α_L	−75	−20	−53	−72	31.2	31.2	−2.0	−8.0	28.3	−2.5	−7.3	4
17	C_7^{ax}	77	−50	−58	175	31.4	31.3	−3.2	−12.2	30.4	−3.3	−10.6	3
18	β_L	−161	174	71	−160	32.8	31.9	−5.6	−21.0	29.1	−6.0	−16.6	2
19	C_7^{ax}	70	−29	−161	−43	31.2	32.1	−3.6	−21.1				
20	C_7^{ax}	76	−48	−58	−76	33.4	33.3	−3.0	−16.9	31.5	−3.3	−15.6	2
21	α_L	−75	−18	−54	−178	33.0	33.9	−4.0	−17.1	33.4	−4.2	−14.3	5
22	δ_L	−150	19	−48	−52	34.1	34.1	−2.8	−21.9	31.6	−3.5	−18.5	4
23	C_7^{eq}	−81	89	−173	70	35.2	34.5	−3.7	−24.8	31.9	−4.2	−21.0	4
24	C_7^{ax}	69	−79	−178	−72	35.6	35.4	−1.6	−12.2	35.9	−2.2	−10.5	2
25	ϵ_D	59	−162	−159	−178	35.9	35.4	−2.4	−19.4	36.8	−3.0	−17.1	7
26	α_D	66	31	−55	−178	36.5	36.5	−3.7	−20.9	39.1	−4.4	−17.2	2
27	ϵ_D	57	−159	−165	70	37.4	37.0	−2.2	−23.3	39.0	−2.9	−21.6	7
28	α_D	47	52	55	60	37.3	37.9	−2.0	−23.2	42.6	−3.0	−21.1	4
29	ϵ_L	−68	161	59	−176	40.0	39.4	−4.7	−31.1				
30	δ_D	171	−37	−88	61	40.1	39.9	−2.2	−19.4	36.5	−2.3	−20.0	11
31	δ_D	−159	−69	57	−173	39.7	40.0	−4.4	−21.6	35.7	−4.8	−15.8	9
32	α_D	65	30	−56	−90	40.8	40.6	−3.8	−25.5				
33	α_D	66	26	−55	86	40.9	41.2	−5.1	−23.9	42.4	−5.6	−21.4	3
34	ϵ_D	49	−134	−58	56	42.0	41.3	−2.9	−15.2	44.5	−3.7	−11.9	4
35	δ_D	−167	−26	−172	−41	49.5	49.7	−2.9	−18.0	49.9	−3.3	−18.1	7
36	C_7^{ax}	58	−29	67	173	52.7	53.4	−3.9	−21.3	49.8	−4.3	−19.9	1
37	ϵ_D	37	−120	65	−175	53.2	53.8	−3.9	−27.1	55.2	−4.5	−24.3	4
38	α_L	−65	−37	−175	−172	60.4	61.3	−6.1	−39.7				
39	δ_D	−126	−69	−63	−176	69.9	69.1	−7.9	−39.3				
MAD						**0.6**				**2.2**			**5**

[a] Relative energies are in kJ/mol. Serine conformations optimized at the MP2/aug-cc-pVDZ (MP2) and B3LYP/6-31G** (B3LYP) levels. Relative energies from single point calculations with the aug-cc-pVTZ basis set are labeled with A, while the best estimates obtained by basis set extrapolation and CCSD(T) corrections are labeled B. Changes in relative energies due to zero point corrections (ΔZPE) and PCM (ΔPCM) solvation corrections have been calculated at the MP2/6-31G** and B3LYP/6-31G** levels. Root-mean-square deviation (rms) of the B3LYP torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations.

conformation. There are six conformations within 20 kJ/mol of the global minimum and 15 conformations within a 30 kJ/mol window. The differences in zero point energies are small, but the estimates of solvent effects indicate that many of these higher energy conformations may be energetically accessible in solution. The MP2/aug-cc-pVTZ method again provides a good correlation with the accurate reference energies, with a MAD value of 0.6 kJ/mol.

A corresponding conformational search at the B3LYP level produced 33 conformations with the large majority being in good agreement with the MP2 geometries. It is significant, however, that six conformations located at the MP2 level do not exist on the B3LYP energy surface.[76] Most of these missing conformations are relatively high in energy, although solvation potentially could bring some of these down in energy. The MAD for the relative energies of the 33 conformations located at the B3LYP level is 2.2 kJ/mol.

Table 6 shows that the various force fields perform erratically for locating conformations, producing 18−41 different conformation compared to the 39 found at the MP2 level. Not only are there several conformations which do not exist on the force field energy surfaces, but there are also many artificial minima. The MMFFs, for example, not only have a good description of the global minimum, with an rms deviation of only 2° compared to the MP2 geometry, but also have another minimum 10.4 kJ/mol higher in energy

Table 6. Force Field Predictions of Serine Conformations[a]

			AMBER94		MM2*		MM3*		MMFFs		OPLS_2005		AMBER99		CHARMM27		AMOEBA	
conf	region	ref	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE
1	C_7^{eq}	0.0	5	0.0	10	0.0	5	0.0	2	0.0	6	0.0	13	0.0	3	0.0	4	0.0
							6	*1.5*	*50*	*10.4*			*45*	*19.5*			*45*	*26.1*
2	C_7^{eq}	12.5	8	13.3	14	5.7	5	7.5	8	24.8	5	17.3	17	10.9	5	9.9	7	18.4
									36	*35.7*	*32*	*27.2*	*36*	*22.0*			*36*	*31.3*
									55	*32.4*							*58*	*37.8*
3	β_L	14.8	7	14.4	58	50.8			9	17.0	1	10.7	9	10.8	9	3.0	5	11.8
					72	*50.8*							*66*	*40.5*			*70*	*53.3*
4	β_L	17.0	12	4.6	40	39.0	11	7.0	3	5.6	8	14.2	11	−0.6	7	10.7	5	22.7
					45	*41.2*												
5	ϵ_D	18.0	46	19.5					21	15.7	21	16.2	37	16.4	17	17.2	6	30.8
													64	*23.7*	*43*	*19.6*	*34*	*30.7*
																	65	32.0
6	β_L	18.4	6	13.7	8	27.9	12	−3.9	7	16.4	5	13.3	7	9.8	4	−0.1	9	16.6
									64	*61.1*			*62*	*48.1*				
7	C_7^{eq}	23.0	28	5.9	9	13.1	6	9.7	9	11.4	13	20.0	23	4.5	9	9.8	5	19.2
													46	*33.9*			*68*	*29.1*
8	C_7^{eq}	23.5	29	7.3	13	19.8			9	13.4	13	14.6	23	6.1	6	4.0	5	17.2
9	C_7^{ax}	24.1	13	12.3	11	15.8	6	17.6	10	22.0	8	21.2	14	13.8			10	40.9
																	62	51.8
10	β_L	25.0	2	24.8	11	43.7	16	20.9	9	22.5	5	18.8	3	20.9	44	14.8	22	29.8
			55	*27.8*			*51*	*5.0*	*49*	*22.4*	*34*	*22.1*	*71*	*25.6*			*54*	*34.5*
11	C_7^{eq}	26.2			16	27.2			5	16.2	5	23.5					16	24.9
12	β_L	27.5									6	20.4					10	21.7
13	C_7^{eq}	27.5							39	30.5	39	30.7						
14	ϵ_D	27.9			7	20.1			10	30.3					11	18.5	7	26.7
15	C_7^{ax}	29.4	9	17.2	7	22.9	5	14.0	7	19.7	8	28.0	14	18.5			5	31.3
							14	*17.4*										
16	α_L	31.2																
17	C_7^{ax}	31.3	19	13.4	4	20.3	7	14.8	4	18.9	8	33.7	19	14.5	62	31.6	5	32.7
18	β_L	31.9	6	30.1	20	41.7	17	22.5	9	30.9	7	32.3	4	26.3				
19	C_7^{ax}	32.1	11	16.3	60	30.7	9	13.6	13	29.2	7	25.7	11	17.5	12	22.7	58	40.8
											53	*34.7*						
20	C_7^{ax}	33.3	19	19.5					5	26.5	8	33.8	19	20.7	11	23.5	3	32.1
															52	*23.3*		
21	α_L	33.9															4	33.9
22	δ_L	34.1	17	20.2					50	19.4			21	11.6	7	21.4		
23	C_7^{eq}	34.5	7	37.5			10	27.3									7	37.3
24	C_7^{ax}	35.4	52	30.1	53	22.7	50	18.7			53	38.1	53	34.4			54	43.6
25	ϵ_D	35.4	7	33.8	22	35.0			15	30.1	21	29.1	7	39.6	7	18.7	6	29.9
26	α_D	36.5							4	22.3							8	36.2
									36	35.7								
27	ϵ_D	37.0							51	26.4	21	30.9			11	13.3	5	29.4
28		37.9			18	31.4			21	32.9					11	28.9		
29	ϵ_L	39.4																
30	δ_D	39.9							5	26.1	13	41.7	31	24.9	35	44.3	45	65.1
																	80	42.9
																	63	53.3
31	δ_D	40.0	5	26.0			17	4.8	7	25.0	5	29.7	15	21.2			7	37.6
32	α_D	40.6															4	38.9
33	α_D	41.2															7	36.3
34	ϵ_D	41.3															13	40.7
35	δ_D	49.7					14	33.7	8	55.8								
36	C_7^{ax}	53.4	7	39.1	19	33.2			11	50.3			7	40.3				
37	ϵ_D	53.8													25	44.4	6	59.8
38	α_L	61.3															9	65.3
39	δ_D	69.1																
MAD			15	8.9	21	10.7	13	14.0	13	7.4	13	4.1	17	10.9	16	11.1	11	4.2

[a] Relative energies are in kJ/mol. The best estimates of the relative energies from Table 5 are labeled as ref. Root-mean-square deviation (rms) of the force field torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations not marked in italics.

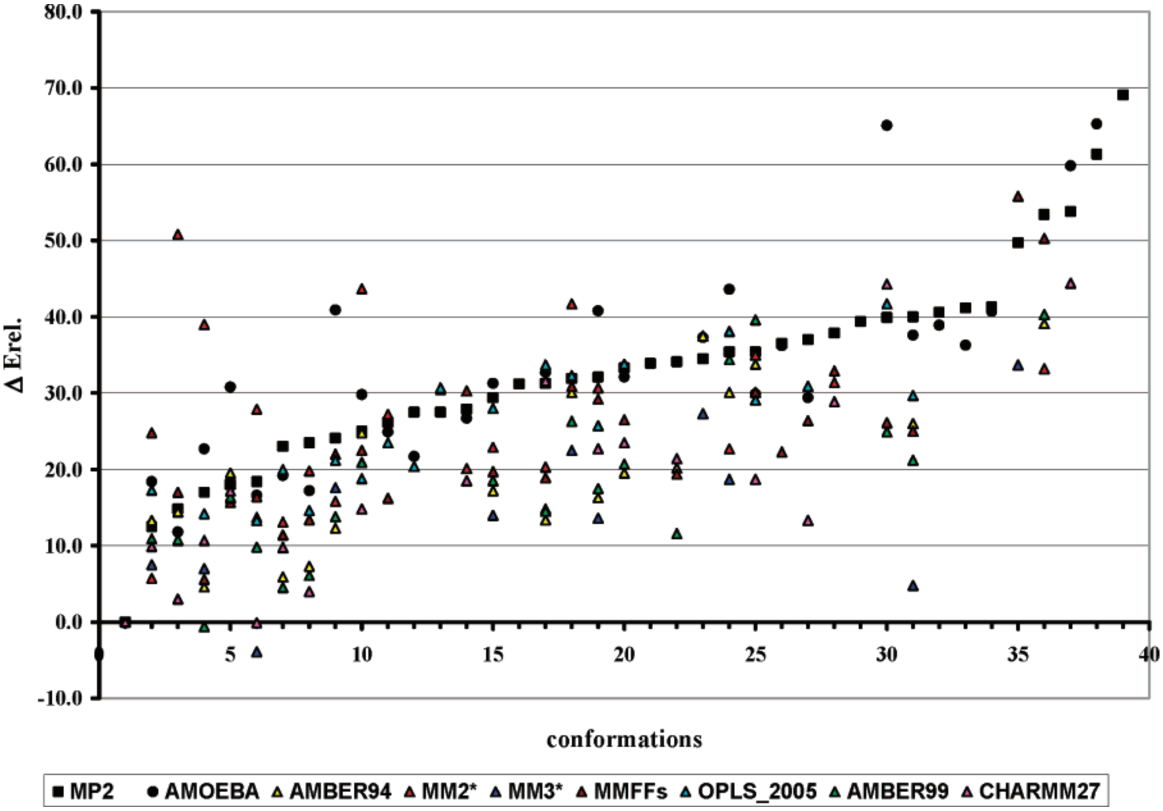


Figure 3. Grapical representation of the relative energies for the serine conformations in Table 6.

with an rms deviation of 50° which converges to the global minimum upon MP2 minimization. Using an rms deviation of 40° as a criterion of a satisfactory representation of the geometries, the MM2, MM3, and CHARMM27 force fields have a similar performance and account satisfactorily for ~15 conformations, the AMBER94, AMBER99, and OPLS reproduce ~20 conformations, while the MMFFs force field gives reasonable representations for 25 of the 39 conformations. The AMOEBA again performs better than the other force fields, being able to locate 27 of the 39 MP2 minima with rms deviation lower than 40°, but also finds 11 artificial minima. The difference in performance is also indicated by the MAD rms values, where the MMFFs, OPLS, and AMOEBA have a value of ~12°, while the other force fields have values of ~17°. The energetic ordering of the force field conformations is somewhat more erratic. The AMOEBA and OPLS force fields perform best and reproduce the reference energies with MAD values of ~4.0 kJ/mol, while the other force fields have MAD values of ~10 kJ/mol. Figure 3 shows a graphical representation of the relative energies of the different force fields as a function of the conformation numbering in Table 6.

Cysteine. The CH_2SH side chain in cysteine has the same torsional degrees of freedom as serine, but the SH group is more polarizable and has less hydrogen bonding ability than the OH group in serine. An MP2 conformational search located 47 conformations of which 14 are within 20 kJ/mol of the global minimum and 28 are within a 30 kJ/mol energy window (Table 7). Estimates of the solvation effect again suggest that many of these will be substantially stabilized in solution, perhaps to the point where the global minimum will change. The computationally less demanding MP2/aug-cc-pVTZ level has a MAD value of 0.7 kJ/mol compared to the reference energies.

The B3LYP method fails to account for six of the 47 conformations, again mostly in the (gas phase) higher energy region. With a couple of exceptions, the geometries of the 41 conformations are in good agreement with the MP2 values. Most of the differences between the B3LYP and MP2 results are in the L-regions, i.e., β_L, α_L, δ_L, and C_7^{eq}. The B3LYP relative energies are in fair agreement (MAD = 2.5 kJ/mol) with the reference values. We note that Zamora and Bombaraso[77,78] reported 47 local minima located by the HF/3-21G model and 42 minima at the B3LYP/6-31G level for this system.

The force field results in Table 8 again show erratic performance. The best force field by far is the AMOEBA, which provides a fair description of 39 of the 47 conformations but also has five artificial minima. The MMFFs is again the best of the fixed charge force fields and accounts for 25 conformations, while the other force fields reproduce between 15 and 24 of the 47 minima. The MAD over the rms in torsion angles is ~20° for the AMBER94, AMBER99, MM2, and MM3 force fields, while the other force fields give values in the range 9−14°. The MAD for the energetic ordering of the conformations is largest for the MM2 and MM3 force fields with values of 10−14 kJ/mol, while the remaining force fields provide values of ~5 kJ/mol. Figure 4 shows a graphical representation of the relative energies of the different force fields as a function of the conformation numbering in Table 8.

Table 7. Cysteine Conformations[a]

						MP2				B3LYP			
conf	region	Φ	Ψ	χ_1	χ_2	A	B	ΔZPE	ΔPCM	A	ΔZPE	ΔPCM	rms
1	C_7^{eq}	−82	64	51	65	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
2	β_L	−161	173	−161	69	8.8	7.6	−1.1	−9.9	7.9	−2.4	−6.7	4
3	C_7^{eq}	−82	80	−174	−71	9.8	8.9	−0.1	−16.7	6.4	−0.8	−15.6	2
4	C_7^{eq}	−81	64	55	−119	11.7	10.8	0.1	−13.1	10.4	−0.7	−10.7	2
5	δ_L	−128	22	59	71	11.6	11.8	−0.5	−15.9	9.9	−1.1	−15.1	3
6	C_7^{eq}	−86	76	−52	−56	13.5	12.1	−0.8	−15.6	9.9	−1.9	−14.3	2
7	β_L	−164	149	−174	−80	13.8	12.7	−1.6	−23.1	10.8	−2.3	−16.2	5
8	C_7^{eq}	−86	73	−57	−179	16.1	14.8	−1.0	−18.1	12.5	−1.9	−17.1	2
9	C_7^{eq}	−83	77	−68	52	16.8	15.6	−1.0	−19.9	13.7	−2.0	−18.4	4
10	C_7^{eq}	−82	85	−174	78	17.4	15.8	−1.0	−25.1	17.0	−2.0	−21.5	5
11	β_L	−158	179	68	−59	16.5	16.2	−2.2	−24.9	12.9	−3.1	−20.9	3
12	C_7^{ax}	78	−52	−54	81	17.7	17.7	−0.4	−19.5	16.6	−1.0	−13.8	3
13	β_L	−171	157	54	54	18.1	17.7	−1.7	−25.4	17.4	−2.3	−21.7	5
14	C_7^{ax}	79	−51	−56	−70	18.8	18.3	0.0	−21.8	16.6	−1.1	−17.5	4
15	δ_L	−144	23	58	−95	20.8	20.4	−1.9	−27.3	17.2	−2.4	−24.1	6
16	β_L	−140	148	−54	−32	22.7	21.9	−1.4	−24.8	22.0	−1.9	−22.4	5
17	α_L	−74	−22	−55	−60	22.1	22.5	−1.3	−22.4	21.4	−2.1	−18.0	32
18	C_7^{ax}	78	−50	−52	−162	23.0	23.0	−1.0	−22.4	24.5	−1.0	−18.3	19
19	C_7^{ax}	72	−74	−174	−68	23.8	23.5	0.4	−19.7				
20	β_L	−167	165	66	−166	24.0	23.6	−0.8	−31.6	27.2	−2.6	−27.8	2
21	C_7^{ax}	73	−68	−179	59	25.8	25.0	−1.2	−27.3	28.7	−1.3	−22.7	3
22	α_D	67	29	−57	−75	26.1	25.5	−0.6	−27.3	25.1	−2.4	−21.9	2
23	α_D	68	26	−57	79	25.9	25.6	−1.7	−27.6	22.4	−2.4	−20.1	3
24	α_L	−78	−17	−61	165	27.5	27.4	−0.7	−27.1				
25	β_L	−138	146	−63	173	30.0	28.4	−0.5	−34.2	33.2	−3.2	−31.1	3
26	δ_D	−176	−44	47	−77	27.8	28.8	−0.8	−17.1				
27	α_D	70	26	−51	−162	29.5	29.5	−2.2	−28.6	31.5	−2.0	−22.7	6
28	α_D	60	39	−157	−67	30.1	29.9	−0.9	−26.8	25.3	−2.0	−26.3	4
29	C_7^{ax}	74	−64	−173	−178	31.4	30.9	−0.5	−28.0	34.8	−1.4	−23.8	5
30	ϵ_D	58	−161	−158	56	32.4	31.0	−1.2	−20.2	32.1	−2.2	−12.5	5
31	δ_D	−164	−34	60	87	31.7	31.9	−1.4	−23.4	30.3	−2.4	−16.6	2
32	C_7^{ax}	71	−34	98	−65	33.6	32.9	−2.7	−20.2	34.5	−0.6	−15.6	9
33	α_L	−74	−23	−165	−42	32.7	33.0	−1.3	−27.0				
34	ϵ_D	58	−159	−151	−166	35.1	33.8	0.0	−19.1	37.5	−2.3	−10.9	4
35	ϵ_D	40	−134	72	−19	34.6	35.3	−2.1	−14.6				
36	α_D	44	52	51	59	35.4	35.6	0.2	−26.4	42.6	−1.8	−18.6	3
37	C_7^{ax}	52	−26	58	55	39.2	39.5	−2.1	−18.7	38.3	−1.5	−15.9	3
38	δ_D	−154	−69	175	60	40.9	40.1	0.1	−35.9	43.5	−3.9	−30.4	5
39	α_D	58	42	−164	108	41.4	40.3	−0.6	−40.1	38.2	−2.5	−34.8	2
40	δ_D	−163	−44	−173	−48	40.8	40.6	−0.1	−29.4	47.2	−2.4	−26.4	4
41	ϵ_D	42	−129	62	176	41.6	41.0	−1.7	−25.9	41.3	−1.3	−12.9	5
42	α_L	−70	−30	−180	165	44.7	44.6	−0.7	−42.3				
43	δ_D	−158	−58	175	171	47.3	46.8	−2.8	−36.7	46.1	−3.0	−32.2	7
44	α_D	51	43	51	−161	47.6	46.9	0.0	−30.7	54.0	−3.7	−26.7	6
45	δ_D	−134	−68	−55	−30	47.6	48.8	−1.1	−36.2	46.4	−2.1	−28.1	27
46	δ_D	−173	−41	−118	64	50.0	49.0	−0.5	−27.9				
47	δ_D	−135	−69	−64	175	55.8	54.3	−0.8	−39.6	51.0	−3.8	−34.2	4
MAD						0.7				2.5			6

[a] Relative energies are in kJ/mol. Cysteine conformations optimized at the MP2/aug-cc-pVDZ (MP2) and B3LYP/6-31G** (B3LYP) levels. Relative energies from single point calculations with the aug-cc-pVTZ basis set are labeled with A, while the best estimates obtained by basis set extrapolation and CCSD(T) corrections are labeled B. Changes in relative energies due to zero point corrections (ΔZPE) and PCM (ΔPCM) solvation corrections have been calculated at the MP2/6-31G** and B3LYP/6-31G** levels. Root-mean-square deviation (rms) of the B3LYP torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations.

Discussion

The present results for four amino acid systems show that the B3LYP method cannot locate all minima found at the MP2 level. One would expect similar problems for density functional methods in general, as they lack a proper description of dispersion interactions. The missing conformations tend to be among the higher energy ones in the gas phase, but whether this is a general trend will have to await results for a larger selection of systems. For the minima that actually exist on the B3LYP surface, the geometries and

Table 8. Force Field Predictions of Cysteine Conformations[a]

			AMBER94		MM2*		MM3*		MMFFs		OPLS_2005		AMBER99		CHARMM27		AMOEBA	
conf	region	ref	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE	rms	ΔE
1	C_7^{eq}	0.0	7	0.0	11	0.0	10	0.0	4	0.0	5	0.0	14	0.0	2	0.0	4	0.0
2	β_L	7.6	53	11.2	57	26.3	59	−3.0	4	14.6	5	7.1	4	10.0	4	8.1	3	10.6
													53	*9.2*	*53*	*7.8*		
															72	*28.2*		
3	C_7^{eq}	8.9	8	8.4	11	4.6	4	1.5	3	15.9	7	12.1	18	10.1	5	5.3	4	7.5
			51	*18.1*	*51*	*10.0*	*53*	*5.2*					*52*	*23.0*	*52*	*12.9*		
4	C_7^{eq}	10.8			15	13.0	18	13.2	19	17.3	21	17.6	42	8.0	17	20.5	8	17.1
					37	*5.4*			*37*	*9.8*	*36*	*9.8*			*36*	*10.5*		
5	δ_L	11.8							53	16.8	6	10.9	3	9.7				
											53	16.0						
6	C_7^{eq}	12.1	10	19.1	13	0.2	9	−2.0	7	3.6	11	12.8	23	2.8	7	2.8	4	9.5
7	β_L	12.7	10	19.1					3	21.3	8	11.2	14	16.0	10	11.8	4	13.0
									52	*28.2*			*61*	*21.1*			*40*	*22.0*
8	C_7^{eq}	14.8	27	1.8	10	0.8	5	−2.7			10	16.1	21	2.0	4	2.2	4	13.1
9	C_7^{eq}	15.6	6	9.4	15	4.8	36	5.5					16	11.0	4	6.9	2	16.4
																	61	21.7
10	C_7^{eq}	15.8	10	17.4			10	7.5	9	23.6			16	27.1	8	15.5	5	19.6
																	41	22.0
11	β_L	16.2	4	18.7	7	28.2			8	15.3	9	11.4	3	17.2	6	12.4		
											61	*37.3*	*87*	*22.1*				
12	C_7^{ax}	17.7	9	13.1	3	6.3	5	1.1	6	10.0	9	23.1	15	16.4	10	12.9	3	23.9
					64	*39.9*	*85*	*31.4*	*68*	*38.2*								
13	β_L	17.7	15	18.6	30	29.2	11	4.5	6	17.0	12	11.9	13	15.9			11	17.0
14	C_7^{ax}	18.3					86	42.3					18	16.4	9	16.1	5	22.5
15	δ_L	20.4	61	6.6			13	10.4					46	12.2			11	25.6
																	47	23.9
16	β_L	21.9	41	12.9					44	16.6			42	10.9	35	9.7	4	21.8
17	α_L	22.5											34	15.1			8	22.7
18	C_7^{ax}	23.0	20	9.3			11	0.8	70	46.8	69	55.0	54	12.4	14	13.2	13	26.5
19	C_7^{ax}	23.5	26	14.6	10	7.6	12	2.5	16	20.4	19	16.9	27	17.8	7	14.8	13	17.4
20	β_L	23.6							6	22.1	10	17.8	10	19.4	11	15.0	9	25.3
21	C_7^{ax}	25.0	85	33.1			8	8.1	9	26.4	12	22.4	13	28.5	6	19.5	82	42.5
									82	*46.0*	*78*	*45.3*	*85*	*36.4*	*82*	*35.2*		
22	α_D	25.5															6	26.6
23	α_D	25.6							2	17.0							4	26.7
24	α_L	27.4															7	28.2
25	β_L	28.4	9	28.2									5	27.4				
26	δ_D	28.8	20	26.8	5	31.6			2	23.9	11	22.7	10	24.5			7	28.0
									21	*27.2*								
27	α_D	29.5															14	30.1
28	α_D	29.9							10	27.4							4	24.5
									52	*39.1*								
29	C_7^{ax}	30.9	6	20.9			6	5.6					55	27.8	6	18.2	5	26.1
30	ϵ_D	31.0	9	38.2							18	28.9	9	45.7			6	26.2
																	74	36.3
31	δ_D	31.9					8	19.5	6	29.4					15	30.3	5	28.6
							48	*14.2*	*50*	*26.0*								
32	C_7^{ax}	32.9	17	19.1	12	21.5			14	31.2	12	27.2	18	22.4	20	23.7	6	35.5
33	α_L	33.0															6	28.1
34	ϵ_D	33.8	9	33.8	49	43.9			16	34.9	51	37.4	9	41.3			7	28.4
35	ϵ_D	35.3															6	30.2
36	α_D	35.6							18	41.2							6	33.8
37	C_7^{ax}	39.5			15	20.4	15	21.9			12	38.5			16	31.4	7	32.9
38	δ_D	40.1	8	35.0	7	36.2	11	18.1	13	43.7	7	37.4	10	34.6				
39	α_D	40.3															18	40.4
40	δ_D	40.6			11	40.5			6	43.0	2	39.2						
41	ϵ_D	41.0															3	41.3
42	α_L	44.6															7	40.6
43	δ_D	46.8							9	50.1	5	44.3	4	39.6			3	44.6
44	α_D	46.9															5	54.7
45	δ_D	48.8					15	30.1	13	42.4							8	42.1
							48	32.3										
46	δ_D	49.0							11	39.7	11	46.7						
47	δ_D	54.3																
MAD			20	6.3	17	10.0	18	13.9	14	5.4	14	4.6	19	5.4	10	6.6	9	3.1

[a] Relative energies are in kJ/mol. The best estimates of the relative energies from Table 7 are labeled as ref. Root-mean-square deviation (rms) of the force field torsional angles relative to the MP2 values in deg. MAD indicates the mean absolute deviation over the conformations not marked in italics.

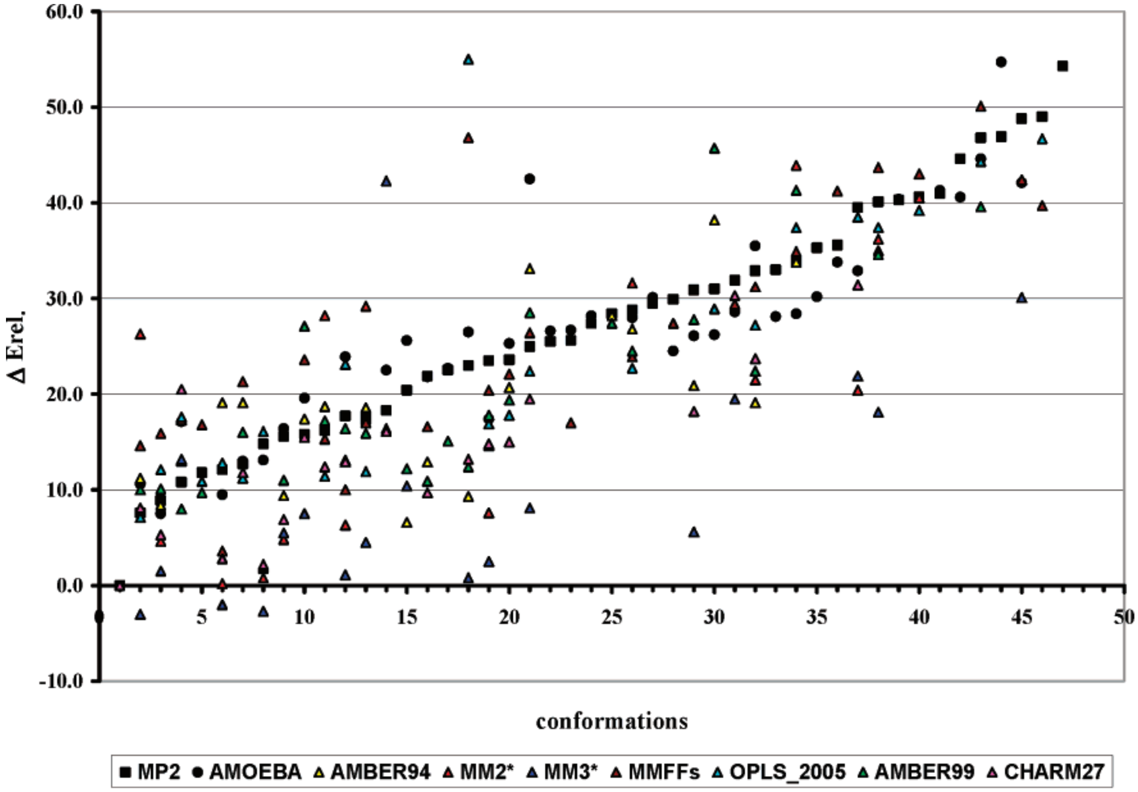


Figure 4. Grapical representation of the relative energies for the cysteine conformations in Table 8.

relative energies are mostly in good agreement with the MP2 results, although there certainly are exceptions. Inclusion of zero point energies only makes small changes in relative energies, typically a few kJ/mol. Estimates of the solvation effect with the PCM model indicate that the energetic ordering in solution may be substantially different than in the gas phase. For these polar systems with hydrogen-bonding capabilities, however, it is questionable whether a continuum solvation model can satisfactorily account for solvation.

Figure 5 shows a graphical representation of the performance of eight different force fields and the B3LYP model against the reference data consisting of all 95 conformations for the four systems. The MM2, MM3, CHARMM27, AMBER94, AMBER99, and OPLS force fields perform almost at par but are only able to satisfactorily reproduce the geometries of approximately half of the conformations. The MMFFs performs somewhat better and is the best of the traditional fixed charge force fields. The more recent AMOEBA force field, which includes multipole moments and polarizabilities, represents a significant improvement and performs almost as well as the B3LYP method. Nevertheless, the AMOEBA force field only accounts for ~80% of the conformations and, in addition, has ~20% artificial minima which are not present on the MP2 energy surface.

Our reference energies are derived from basis set extrapolated MP2 results combined with an additive correction for higher order correlation effects. Compared to these results, the MP2/aug-cc-pVTZ level provides a MAD for relative energies of only 0.7 kJ/mol, indicating that this level of theory is a good compromise between accuracy and computational efficiency and capable of providing relative conformational energies accurate to ~1 kJ/mol.

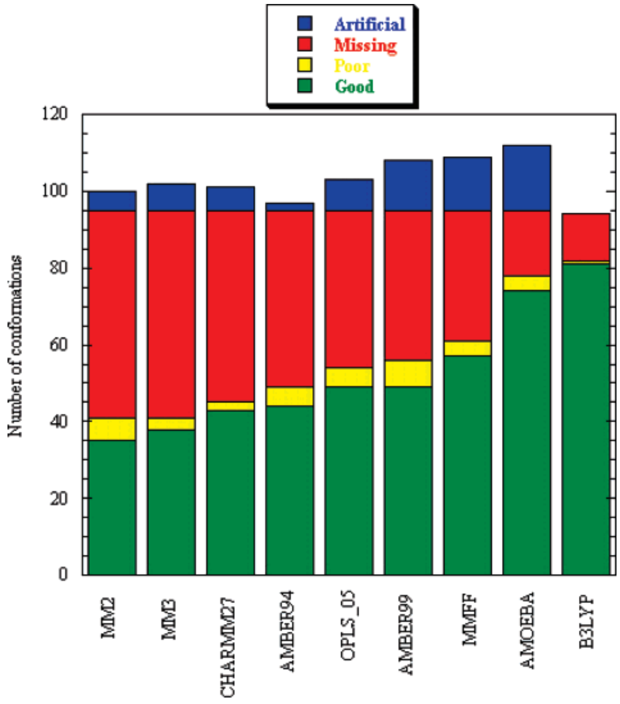


Figure 5. Performance of different force fields and the B3LYP method for reproducing MP2 conformations for all four systems. See the text for the definition of Good, Poor, Missing, and Artificial conformations.

Figure 6 shows the MAD values for the rms torsional angles and relative energies for all four systems for the “good” and “poor” conformations in Figure 5. The MM2

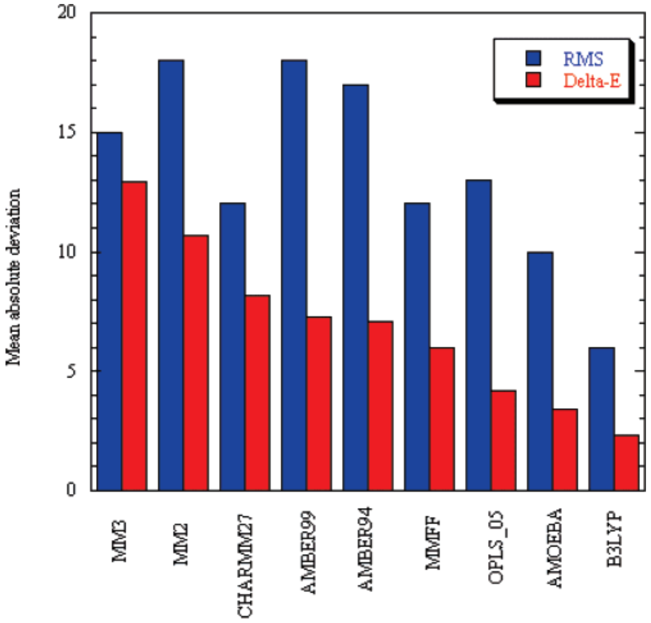


Figure 6. Mean absolute deviations for root-mean-square (rms) differences in the torsional angles and relative energies (Delta-E) of different force fields and the B3LYP method for all conformations over all four systems. rms values in degrees and Delta-E values in kJ/mol.

and MM3 force fields provide the poorest performance, with the CHARMM27, AMBER99, and AMBER94 being slightly better for reproducing relative energies. The MMFFs and OPLS force fields are the best of the fixed partial charge methods, with the former performing best for geometries and the latter performing slightly better for relative energies. The polarizable AMOEBA force field provides the best results with a typical rms error in the torsional angles of 10° and a deviation of 3.4 kJ/mol for relative energies. Nevertheless, the inability of this force field to describe ~20% of the conformations, the presence of artificial minima, and a mean rms deviation for the torsional angles of 10° for those conformations that can be described indicate further room for improvement.

The longer term goal is to use data of the present type for developing new force fields with improved capabilities for reproducing the potential energy surface of, e.g., proteins. For this purpose a decision must be made regarding which reference data to use. Including both solvation and zero point energies in the reference data produces a force field where these effects are absorbed in the parameters. Alternatively, the force field can be parametrized against data without solvent and/or zero point energies, and these effects can then be calculated explicitly within the force field model, for example using the GB/SA solvent model[79] or by explicit solvation. The use of continuum solvent models, however, is only expected to produce qualitative solvent effects, as for example hydrogen bonding is neglected in these models, and using explicit solvation with accurate electronic structure methods is computationally expensive. Zero point energies, on the other hand, are cumbersome to calculate within a force field environment. A viable strategy could be to parametrize the force field for reproducing gas-phase results including zero point energies and subsequently parametrize the interaction with explicit solvent. Since zero point energies are a relatively minor correction, the parametrization could also be done using just electronic energies.

The present work has focused on locating stable conformations, i.e., minima on the potential energy surface for a few simple systems. For parametrization of force field it is necessary to extend the present work to more diverse systems, i.e., other amino acids and larger peptide models. For modeling the dynamics it is also of importance to be able to describe the energetics of interconversion between conformations, i.e., reproduce geometries and stabilities of transition structures connecting minima. Such extensions will be considered at a later date.

Acknowledgment. This work was supported by grants from the Danish Natural Science Research Council, the Danish Center for Scientific Computing, an EU Marie Curie training grant, and the Grant Agency of the Academy of Sciences of the Czech Republic (IAA400550702).

Supporting Information Available: MP2/aug-cc-pVDZ optimized geometries and associated energies for all 95 conformations. The material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) http://www.rcsb.org (accessed April 1, 2007).

(2) Fernandez, C.; Wider, G. *Mod. Mag. Res*. **2006**, *1*, 483. Candel, A. M.; Conejero-Lara, F.; Martinez, J. C.; van Nuland, N. A. J.; Bruix, M. *FEBS Lett.* **2007**, *581*, 687. Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. J. *Biomol. NMR* **2007**, *37*, 117.

(3) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471. Tse, J. S. *Ann. Rev. Phys. Chem.* **2002**, *53*, 249.

(4) Schlick, T. *Molecular modeling and simulation*; Springer: 2002.

(5) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; Caserio, M. S., Ed.; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(6) Jensen, F. In *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, England, 2006; Chapter 2, pp 22−77.

(7) Gundertofte, K.; Liljefors, T.; Norrby, P. O.; Pettersson, I. *J. Comput. Chem.* **1996**, *17*, 429. Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III *J. Am. Chem. Soc*. **2004**, *126*, 698. Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; Mackerell, A. D., Jr. *Biophys. J.* **2006**, *90*, L36. Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem*. **2006**, *27*, 781.

(8) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253. Koch, U.; Stone, A. J. *J. Chem. Soc., Faraday Trans*. **1996**, *92*, 1701.

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

(10) Rasmussen, T. D.; Jensen, F. *Mol. Simul.* **2004**, *30*, 801. Mackerell, A. D., Jr. *J. Comput. Chem*. **2004**, *25*, 1584. Mackerell, A. D., Jr. *Ann. Rep. Comp. Chem*. **2005**, *1*, 91.

(11) Patel, S.; Brooks, C. L., III *J. Comput. Chem.* **2006**, *25*, 1. Patel, S.; MacKerell, A. D., Jr.; Brooks, C. L., III *J. Comput. Chem.* **2006**, *25*, 1504. Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theor. Comput.* **2005**, *1*, 694.

(12) 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.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(13) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440.

(14) Ponder, J. W. *Tinker*, *Version 4.2*; Biochemistry & Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 2004.

(15) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(16) Alinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(17) Allinger, N. L.; Yan, L. *J. Am. Chem. Soc.* **1993**, *115*, 11918.

(18) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.

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

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

(21) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.

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

(23) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330.

(24) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

(25) Amovilli, C.; Barone, V.; Cammi, R.; Cances, E.; Cossi, M.; Mennuci, B.; Pomelli, C. S.; Tomasi, J. *Adv. Quantum Chem.* **1999**, *32*, 227.

(26) Moon, S.; Case, D. A. *J. Comput. Chem.* **2006**, *27*, 825.

(27) Hudáky, I.; Hudáky, P.; Perczel, A. *J. Comput. Chem* **2004**, *25*, 1522.

(28) Yu, C.-H.; Norman, M. A.; Schäfer, L.; Ramek, M.; Peeters, A.; van Alsenoy, C. *J. Mol. Struct.* **2001**, *567*, 361.

(29) Černohorský, M.; Vaultier, M.; Koča, J. *J. Mol. Struct.* **1999**, *489*, 213.

(30) Brijbassi, S. U.; Sahai, M. A.; Setiadi, D. H.; Chass, G. A.; Penke, B.; Csizmadia, I. G. *J. Mol. Struct.* **2003**, *666*, 291.

(31) Adamo, C.; Dillet, V.; Barone, V. *Chem. Phys. Lett.* **1996**, *263*, 113.

(32) Chakraborty, D.; Manogaran, S. *J. Phys. Chem. A* **1997**, *101*, 6964.

(33) Kaschner, R.; Hohl, D. *J. Phys. Chem. A* **1998**, *102*, 5111.

(34) Gronert, S.; O'Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, *117*, 2071.

(35) Jarmelo, S.; Lapinski, L.; Nowak, M. J.; Carey, P. R.; Fausto, R. *J. Phys. Chem. A* **2005**, *109*, 5689.

(36) Gong, X.; Zhou, Z.; Du, D.; Dong, X.; Liu, S. *Int. J. Quantum Chem.* **2005**, *103*, 105.

(37) Bogar, F.; Szekeres, Z.; Bartha, F.; Ladik, J. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2965.

(38) Fernandez-Ramos, A.; Cabaleiro-Lago, E.; Hermida-Ramón, J. M.; Martínez-Núnez, E.; Pena-Gallego, A. *J. Mol. Struct.* **2000**, *498*, 191.

(39) Aleman, C.; Puiggali, J. *J. Phys. Chem. B* **1997**, *101*, 3441.

(40) Antohi, O.; Naider, F.; Sapse, A.-M. *J. Mol. Struct.* **1996**, *360*, 99.

(41) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481.

(42) Gould, I. R.; Cornell, W. D.; Hillier, I. H. *J. Am. Chem. Soc.* **1994**, *116*, 9250.

(43) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908.

(44) Czaszar, A. G. *J. Phys. Chem.* **1996**, *100*, 3541.

(45) Gang, Z.; Chunsheng, D.; Zhengyu, Z.; Qunyan, W.; Jinfeng, L. *J. Mol. Struct.* **2005**, *765*, 143.

(46) Korter, T. M.; Balu, R.; Campbell, M. B.; Beard, M. C.; Gregurick, S. K.; Heilweil, E. J. *Chem. Phys. Lett.* **2006**, *418*, 65.

(47) Broda, M. A.; Siodłak, D.; Rzeszotarska, B. *J. Pept. Sci.* **2005**, *11*, 546.

(48) Linder, R.; Nispel, M.; Häber, T.; Kleinermanns, K. *Chem. Phys. Lett.* **2005**, *409*, 260.

(49) Gerlach, A.; Unterberg, C.; Fricke, H.; Gerhards, M. *Mol. Phys.* **2005**, *103*, 1521.

(50) Toroz, D.; van Mourik, T. *Mol. Phys.* **2006**, *104*, 559. van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8. Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442*, 42.

(51) Gresh, N.; Kafafi, S. A.; Truchon, J.-F.; Salahub, D. R. *J. Comput. Chem.* **2004**, *25*, 823.

(52) Pieshoff, C. D.; Ali, F. E.; Bean, J. W.; Calvo, R.; D'Ambrosio, C. A.; Eggleston, D. S.; Hwang, S. M.; Kline, T. P.; Koster, P. F.; Nichols, A.; Powers, D.; Romoff, T.; Samanen, J. M.; Stadel, J.; Vasko, J. A.; Kopple, K. D. *J. Med. Chem.* **1992**, *35*, 3962.

(53) Reed, J.; Hull, W. E.; von der Lieth, C.-W.; Kübler, D.; Suhai, S.; Kinyel, V. *Eur. J. Biochem.* **1988**, *178*, 141.

(54) Johnson, W. C., Jr.; Pagano, T. G.; Basson, T. C.; Madri, J. A.; Gooley, P.; Armitage, I. M. *Biochemistry* **1993**, *32*, 268.

(55) Tobias, D. J.; Brooks, C. L., III *J. Phys. Chem.* **1992**, *96*, 3864.

(56) Tobias, D. J.; Sneddon, S. F.; Brooks, C. L., III *J. Mol. Biol.* **1990**, *216*, 783.

(57) Wei, D.; Guo, H.; Salahub, D. R. *Phys. Rev. E* **2001**, *64*, 011907.

(58) Shirley, W. A.; Brooks, C. L., III *Proteins* **1997**, *28*, 59.

(59) Sajot, N.; Garnier, N.; Genest, M. *Theor. Chem. Acc.* **1999**, *101*, 67.

(60) Feig, M.; MacKerell, A. D., Jr.; Brooks, C. L., III *J. Phys. Chem. B* **2002**, *107*, 2831.

(61) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(62) Park, H. S.; Kim, C.; Kang, Y. K. *Biopolymers* **2002**, *63*, 298.

(63) Ono, S.; Nakajima, N.; Higo, J.; Nakamura, H. *J. Comput. Chem.* **2000**, *21*, 748.

(64) Bashford, D.; Case, D. A.; Choi, C.; Gippert, G. P. *J. Am. Chem. Soc.* **1997**, *119*, 4964.

(65) Gresh, N.; Tiraboschi, G.; Salahub, D. R. *Biopolymers* **1998**, *45*, 405.

(66) Gould, I. R.; Kollman, P. A. *J. Phys. Chem.* **1992**, *96*, 9255.

(67) Möhle, K.; Hofmann, H.-J. *J. Mol. Struct.* **1995**, *339*, 57.

(68) Hermida-Ramón, J. M.; Brdarski, S.; Karlström, G.; Berg, U. *J. Comput. Chem.* **2003**, *24*, 161.

(69) Yang, Z. Z.; Zhang, Q. *J. Comput. Chem.* **2006**, *27*, 1.

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

(71) Some authors chose a different labeling where β_L is equivalent to C5, α_D is equivalent to α_L, and α_L is equivalent to α_R.

(72) Bohm, H. J.; Brode, S. *J. Am. Chem. Soc.* **1991**, *113*, 7129.

(73) Iwaoka, M.; Okada, M.; Tomoda, S. *Theochem* **2002**, *586*, 111.

(74) Barone, V.; Adamo, C.; Lelj, F. *J. Chem. Phys.* **1995**, *102*, 364.

(75) Rodriguez, M. A.; Baldoni, H. A.; Suvire, F.; Vásquez, R. N.; Zamarbide, G.; Enriz, R. D.; Farkas, O.; Perczel, A.; McAllister, M. A.; Torday, L.; Papp, J. G.; Csizmadia, I. G. *J. Mol. Struct.* **1998**, *455*, 275.

(76) Perczel reported the presence of 44 local minima at the HF/3-21G level for the HCO-Ser-NH_2 system: Perczel, A.; Farkas, Ö.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1996**, *118*, 7809.

(77) Zamora, M. A.; Baldoni, H. A.; Bombaraso, J. A.; Mak, M. L.; Perczel, A.; Farkas, O.; Enriz, R. D. *J. Mol. Struct.* **2001**, *540*, 271.

(78) Bombaraso, J. A.; Zamora, M. A.; Baldoni, H. A.; Enriz, R. D. *J. Phys. Chem. A* **2005**, *109*, 874.

(79) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

CT700082F

JCTC Journal of Chemical Theory and Computation

Atomistic Force Field for Azobenzene Compounds Adapted for QM/MM Simulations with Applications to Liquids and Liquid Crystals

Marcus Böckmann,[†] Christine Peter,*[,‡] Luigi Delle Site,[‡] Nikos L. Doltsinis,*[,†] Kurt Kremer,[‡] and Dominik Marx[†]

Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany, and Max-Planck Institut für Polymerforschung, Ackermannweg 10, 55128 Mainz, Germany

Received March 22, 2007

Abstract: An atomistic force field has been adapted for use in molecular dynamics simulations of molecular materials that contain azobenzene (AB) functional groups. Force field parameters for bonded interactions and partial charges in the AB unit have been derived from ab initio molecular dynamics reference calculations. First applications of the new force field to liquid *trans*- and *cis*-AB are presented, both using a purely classical approach (MM) and a hybrid quantum-classical (QM/MM) simulation scheme. Detailed structural analysis confirms that QM/MM and purely MM simulations yield results that are in good agreement with each other. The force field of the AB core has been extended to include aliphatic chains that are attached via ether bridges to the two AB benzene rings. This allows for studying temperature induced phase transitions in the liquid−crystalline 8AB8 system. Using replica exchange techniques the new force field has successfully reproduced the smectic to isotropic-phase transition.

1. Introduction and Motivation

Azobenzene (AB) has become one of the most widely studied photoactive compounds in physics and chemistry[1−5] mainly because it is possible to reversibly switch between the cis and trans isomers by photoexcitation (see Figure 1a). Although the structure of the electronically excited molecule as well as the mechanism of this photoisomerization still remains an unsolved problem,[6−16] AB is frequently used in practice as an optical switch, for instance to fold/unfold peptides,[17,18] in optomechanical cycles and molecular machines,[19,20] and in developing optically active materials.[1,5,21−29] Most of these applications exploit the large difference of approximately 2.4 Å in the end-to-end distance of the extended trans isomer with respect to the more compact cis isomer. The difference in molecular shape results in crucially different properties. In the field of liquid crystals, for instance, the stiff, rod-shaped trans isomer of AB can function as a mesogen, occurring in numerous liquid−crystalline compounds, while the nonmesogenic cis isomer is not able to induce any long-range order.[24,30] Hence, it is

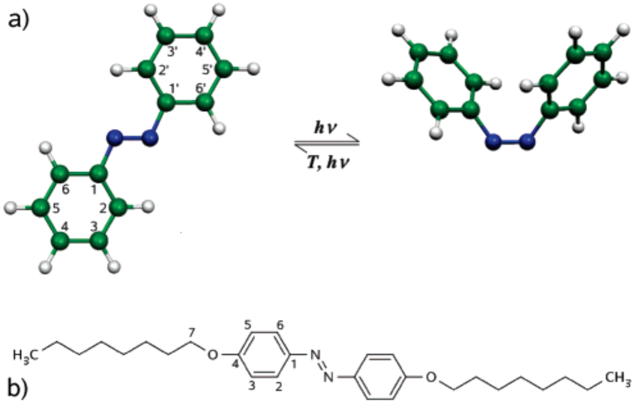


Figure 1. (a) Structure and atom numbering scheme of *trans*- and *cis*-azobenzene in the left and right panels, respectively. (b) Chemical structure of 4,4′-dioctyloxyazobenzene (8AB8).

* Corresponding author e-mail: peter@mpip-mainz.mpg.de (C.P.) and nikos.doltsinis@theochem.rub.de (N.L.D.).

[†] Ruhr-Universität Bochum.

[‡] Max-Planck Institut für Polymerforschung.

10.1021/ct7000733 CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/17/2007

possible to design AB-containing materials which form photoswitchable liquid crystals.[24,30]

Theoretical modeling of such optically active materials presents a challenge since they are characterized by phenomena that take place on largely different length and time scales: the ultrafast process of photoisomerization, occurring on the (sub-)picosecond time scale, requires the use of state-of-the-art ab initio simulation techniques eventually even going beyond the Born−Oppenheimer approximation, while the macroscopic changes induced by the photoswitching of the AB chromophore, such as phase transitions in liquid crystals, calls for simulation techniques reaching at least into the mesoscopic realm. In the ‘hierarchy’ of computational methods, going from the atomistic (quantum mechanical: QM) via the microscopic (molecular mechanics: MM) to the mesoscopic (coarse grained: CG) level implies that ab initio molecular dynamics is followed by classical force field molecular dynamics and coarse grained dynamics. A valuable step in between such ‘all-QM’ and ‘all-MM’ techniques is the well-known hybrid quantum-classical QM/MM approach, in which only a chemically active subsystem is treated quantum-mechanically (QM), while the remaining part is described by molecular mechanics (MM) employing classical force fields. Unfortunately there is no appropriate force field available that allows for a sound description of the azobenzene functional unit in a macroscopic context.

The strategy we followed in this work was to use a well tested standard force field with a broad range of applications (in this case the GROMOS force field, see below) and to restrict the reparametrization to the photoswitchable azo functional unit, thus retaining the advantages of the well-established standard force field (especially the nonbonded parameters and small charge groups). This implies that we did not aim to develop an optimized (and hence specialized) force field for liquid azobenzene or 8AB8. However, by the same token this means that we retain the universal applicability of the (GROMOS) standard force field. In particular, the standard nonbonded parameters derived from thermodynamic data for various aliphatic and aromatic groups have not been changed in order to ensure this transferability. Furthermore, keeping in mind the aim to treat the photoswitch fully quantum mechanically in a QM/MM framework, its parametrization has been done with reference to the particular QM treatment (in this case the PBE density functional, see below) to be used therein. This allows for a most seemless change between a force field (MM) description and an electronic structure based (QM) description of the photoswitch. These important features of our approach may, for instance, be exploited in future studies of photoswitchable biomolecules.[31]

In the present paper, we attempt to consistently link the most fundamental three hierarchical levels for the specific case of AB itself as well as for AB-containing materials: from QM to QM/MM to MM. The chosen strategy is as follows.

(1) We first carry out finite-temperature ab initio molecular dynamics simulations (all-QM) of *cis*- and *trans*-AB in the gas phase, which will serve as a reference for the MM simulations.

(2) Classical force field simulations (all-MM) of gas-phase AB are performed. Force field parameters for bonded interactions and partial charges in the AB unit are adapted so as to yield maximum agreement between all-QM and all-MM results. In line with our philosophy outlined above, parameters for nonbonded interactions are taken from a well-tested standard force field so as to retain its applicability in a more general context.

(3) The obtained force field is applied to the study of liquid AB using two types of simulation techniques, all-MM and hybrid QM/MM. A detailed comparison in terms of structural parameters is performed.

(4) An extended force field is employed to perform all-MM simulations of an order−disorder-phase transition in the 4,4′-dioctyloxyazobenzene (8AB8)[32] liquid crystal (see Figure 1b). This serves to validate the extended force field by direct comparison with experimental observations.

The above procedure represents the foundation of a genuine multilevel approach to studying optically active materials. In such a scheme, the photoinduced events involving the AB chromophore are treated fully quantum-mechanically, ultimately including even non-Born−Oppenheimer effects, whereas the complex condensed-phase environment, expected to have a considerable impact on the photoisomerization mechanism and efficiency, is described in a QM/MM approximation. Phenomena occurring on much longer (nanosecond) time scales as a result of the initial photoswitching and its picosecond relaxation are modeled using classical MM simulation techniques. It is envisaged to extend further the range of applicability of the present “QM to QM/MM to MM” multiscale ansatz in space and time by introducing both nonadiabatic effects and an additional CG level in the hierarchy which no longer carries atomic resolution but relies on coarse grained interaction potentials. Ultimately, this “QM to nonadiabatic−QM/MM to MM to CG” approach will provide a genuine link between quantum-mechanical events at the level of electrons and nuclei up to the level of macroscopic properties of photo-(re)active materials.

2. Computational Details

2.1. Ab Initio Molecular Dynamics Reference Simulations. Gas-phase Car−Parrinello ab initio molecular dynamics simulations[33,34] have been performed using the plane wave density functional theory code CPMD.[34,35] We employed the PBE exchange and correlation functional[36,37] and truncated the plane wave expansion of the Kohn−Sham orbitals at 70 Ry and used normconserving pseudopotentials of the Goedecker type for the core electrons. A fictitious electron mass of 800 au and a time step of 5 au (0.12 fs) were chosen replacing as usual the hydrogen mass by the deuterium mass in order to increase the computational efficiency.[34] The size of the periodic simulation cell was 20 × 13 × 8 Å^3 and 17 × 12.5 × 12.5 Å^3 for the isolated *trans*- and *cis*-AB molecules, respectively. The systems were pre-equilibrated for 10 ps at 300 K using a separate Nosé−Hoover chain thermostat[38] for each degree of freedom which ensures very efficient thermalization and energy equipartitioning of even stiff intramolecular rovibrational modes.[39] Based on such initial conditions, the systems were left to evolve microcanonically without the use of any thermostats during the final 12 ps production runs.

In order to determine the barrier height for the torsional cis−trans ground-state isomerization about the N═N bond, which is important for the parametrization of the associated classical force constant, a minimum energy path has been calculated along the CNNC dihedral angle from 0° and 180° in steps of 10°. In addition, ab initio molecular dynamics simulations of methyl phenyl ether ($H_3C{-}O{-}C_6H_5$) were carried out at 300 K. This serves as a model system for the ether linkage occurring in AB containing molecules such as 8AB8 (see Figure 1b). A cubic unit cell of length 15 Å was used, whereas the remaining simulation parameters were identical to the AB simulations. After equilibration a production run of 2.4 ps length was performed. In order to determine partial charges, restrained electrostatic potential (RESP) fits[40] were carried out using the CPMD program for constrained optimized geometries at fixed $C^{(5)}C^{(4)}OC^{(7)}$ dihedral angles (see Figure 1 for nomenclature) from 0° to 90° in steps of 10°. The point charges determined for the classical force field have been calculated by Boltzmann averaging over this dihedral angle.

2.2. Classical Gas-Phase Simulations and Force Field Parametrization. A force field suitable for simulations of *trans*- and *cis*-AB was derived starting from the GROMOS 45a3 force field.[41] Using the GROMOS program package,[42,43] we performed MM runs of the isolated trans and cis conformers in the gas phase for 25 ps (after 25 ps of equilibration) using a time step of 1 fs. The temperature was kept constant at 300 K using the Berendsen thermostat[44] with a coupling constant of 0.1 ps. Nonbonded interactions were calculated using a cutoff of 14 Å (updated every 10 steps) (i.e., all nonbonded interactions were taken into account). The parametrization calculations were performed without the use of any bond constraints. Nonstandard force field parameters (i.e., point charges and bonded parameters in the azo group) were determined to achieve maximum compatibility between the QM and MM descriptions. For the bonded parameters this requires that the distributions of bond lengths, bond angles, and dihedral angles obtained from the QM reference simulations are reproduced by the MM force field.

The point charges have been assigned according to the RESP atomic charges as described in section 2.1 from the reference QM calculations on *trans*- and *cis*-AB. In analogy to the force field parametrization procedure for AB, we have developed an extended force field suitable for the 8AB8 compound (see Figure 1b). The required force field parameters for the $C^{(4)}{-}O{-}C^{(7)}$ linkage were determined by matching MM and QM simulation data for the methyl phenyl ether ($H_3C{-}O{-}C_6H_5$) model system introduced in section 2.1.

2.3. Condensed-Phase Classical Simulations. The purely classical simulations of liquid AB and liquid−crystalline 8AB8 were performed with the Gromacs simulation package.[45] In the case of liquid AB, our simulation box contained 343 such molecules either all in cis or all in trans conformation. The simulations were carried out using a time step of 0.5 fs at a temperature of 400 K and a pressure of 1 bar. A weak coupling Berendsen algorithm[44] was used to control both temperature (coupling constant: 0.1 ps) and pressure (coupling constant: 5.0 ps, isothermal compressibility: 1.0×10^{-5} bar^{-1}). Electrostatic interactions were computed using the Particle Mesh Ewald method,[46,47] Lennard-Jones interactions were computed using a cutoff of 20 Å (with pair list updates every 10 steps), and no bond constraints were applied (these settings were also used for the classical part of the QM/MM calculations described below). For the sake of comparison, we also carried out simulations of a single AB molecule where stochastic temperature coupling was used with a friction coefficient of 10 ps^{-1}.

The liquid−crystalline systems contained 1296 8AB8 molecules (four layers of 18 × 18 molecules) prepared in smectic order. Unless stated otherwise, the same computational settings were used as for liquid AB. Here, all chemical bonds were constrained using the LINCS algorithm[48] enabling us to use a time step of 2 fs, and for Lennard-Jones interactions a twin-range cutoff of 10 and 14 Å (updated every 5 steps) was applied.[42] Anisotropic pressure coupling was used, allowing the three box directions to fluctuate independently but keeping the box orthorhombic. After energy minimization and a very short (5 ps) initial relaxation at 300 K with a 0.5 fs time step, the system was equilibrated at 400 K for 800 ps. After equilibration, replica exchange simulations[49] were carried out with seven replicas in a temperature range from 460 to 475 K. This range had been determined by previous simulations at various temperatures between 400 and 500 K such that it covers temperatures where the system remains in the initial layered setup up to temperatures where a transition from the ordered setup to the isotropic phase is observed; the number of replicas is limited due to the required simulation lengths to observe phase changes in the chosen system. Each replica evolved independently, and, after every 4000 MD-steps (∼8 ps), an exchange of pairs of neighboring (in temperature space) replicas was attempted.[49]

For analysis of the liquid−crystalline system, the order parameter S of the system was evaluated, with S being the largest eigenvalue of following tensor

$$Q_{\alpha\beta} = \sum_{j=1}^{N} \left(\frac{3}{2}u_{\mathrm{j}\alpha}u_{j\beta} - \frac{1}{2}\delta_{\alpha\beta}\right) \quad \alpha, \beta = x, y, z \qquad (1)$$

where the sum is over all N molecules, the unit vector u_j points along the axis of one molecule or mesogenic unit j (in our case this axis was defined by the positions of the two carbon atoms $C^{(4)}$ and $C^{(4')}$, see Figure 1), and $\delta_{\alpha\beta}$ is the Kronecker delta. For nematic and smectic phases this eigenvalue is significantly larger than the other two indicating the degree of alignment along the direction of the corresponding eigenvector, the so-called director. Thus, the appearance of liquid−crystalline phases can be identified through the order parameter; a distinction of smectic and nematic phase is not possible however. For this, the 3D arrangement of the molecules needs to be analyzed. The smectic phase is characterized by a 2D arrangement of the molecules in (fluid) layers, whereas the nematic phase only exhibits orientational order along the director but no positional order.

2.4. QM/MM Simulations. The hybrid QM/MM simulations of liquid AB were carried out using the CPMD/GROMOS interface[50] developed by Laio et al. that couples the Car−Parrinello molecular dynamics package CPMD[34,35] to the GROMOS molecular dynamics package[42,43] (see ref 51 for a review). Within this coupling scheme a single QM AB molecule was immersed in a liquid consisting of MM AB molecules. The electronic structure treatment of the QM subsystem is identical to that described in section 2.1 for isolated AB molecules, while the classical force field (see sections 2.2 and 2.3) is used to describe the MM subsystem.

Starting from an equilibrated configuration obtained from a purely classical (MM) simulation of 343 *trans*-AB molecules in a cubic box of length 45.2 Å (the average from the constant pressure MM simulation described in section 2.3), a hybrid QM/MM simulation was performed at 400 K using Nosé−Hoover chains[38] for both ions and electrons and a time step of 5 au (0.12 fs). The length of the production run was about 3 ps. The center-of-mass translation and overall rotation was removed every 10 time steps.

A single AB molecule was treated quantum-mechanically in a cubic box of length 18 Å for both *trans*- and *cis*-AB. QM atoms carry the same nonbonded parameters as the MM atoms. The cutoff for the explicit interaction of the remaining MM atoms with the QM charge density, i.e., the nearest-neighbor cutoff, between the charge group centers was set to 20.0 au (10.6 Å), and the nearest-neighbor list was updated every 10 time steps. Charge groups at distances greater than 20.0 au interact with the QM subsystem through a multipole expansion of the QM charge density. In order to prevent the QM subsystem from artificial cooling as often observed in QM/MM calculations we applied two separate Nosé−Hoover chains[38] to the QM and MM subsystems.

3. Results and Discussion

3.1. Ab Initio Simulations of Azobenzene in the Gas Phase. As a basis for the AB force field parametrization, we performed two separate ground-state Car−Parrinello simulations of an isolated *trans*- and *cis*-AB molecule at 300 K. In the following discussion, the focus is on internal coordinates involving the azo group (see Figure 1 for structure and numbering), that is, the NN′ and $C^{(1)}N$ bond lengths (the latter being equivalent to the $C^{(1')}N'$ bond length due to the molecular symmetry), the $C^{(1)}NN'$ (and also $C^{(1')}N'N$) bond angle, the $C^{(1)}NN'C^{(1')}$ dihedral angle, and the rotation of the two aromatic ring systems described by the dihedral angles $\Phi_1 = \angle C^{(2)}C^{(1)}NN'$ and $\Phi_2 = \angle C^{(2')}C^{(1')}N'N$. As a measure for pyramidalization at $C^{(1)}$ we chose the angle between the $C^{(1)}N$ bond and the $C^{(6)}C^{(1)}C^{(2)}$ plane which is the so-called Wilson angle.[52] In the following we omit the atomic indices and refer to the CN bond length, the CNN bond angle, and the CNNC dihedral angle for simplicity.

Figure 2 depicts the time evolution of the CNNC dihedral, the ring rotational angles Φ_1 and Φ_2, and the pyramidalization at $C^{(1)}$, in the case of the *trans*-AB conformer. The upper panel of Figure 2 illustrates that the molecule is planar on average, the CNNC dihedral exhibiting rather regular oscillations about 180° with an amplitude of roughly 20°. The middle panel of Figure 2 demonstrates that the rotational motion of the phenyl rings is concerted for *trans*-AB. Each individual phenyl ring rotates up to $\Phi_{1,2} \approx \pm 30°$, while maintaining $\Phi_1 \approx \Phi_2$ throughout. In other words, given the definition of Φ_1 and Φ_2, the two phenyl rings are found to rotate against each other ('out-of-phase'), resulting in an overall (approximate) C_2 symmetry of *trans*-AB.

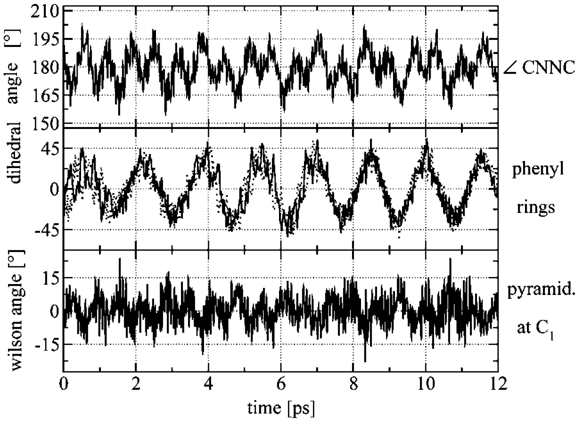


Figure 2. Time evolution of structural parameters from an ab initio molecular dynamics simulation of isolated *trans*-AB at 300 K. Top panel: dihedral angle CNNC. Middle panel: rotation angles Φ_1 and Φ_2 for the two phenyl rings (solid and dotted lines). Bottom panel: pyramidalization angle (see text) at the carbon atom $C^{(1)}$.

If the rotation were 'in-phase', i.e. $\Phi_1 = -\Phi_2$, the molecular symmetry would be C_i (assuming planarity and equal bond lengths in both aromatic rings). The question of whether vibrational motion in AB leads to C_2 or C_i symmetry has been controversially debated in the literature.[53,54] Our current results are in line with the experimental findings from gas-phase electron diffraction studies by Adamson et al.[53] suggesting C_2 symmetry. As can be seen from the bottom panel of Figure 2, there is no pyramidalization at $C^{(1)}$ on average, but fluctuations of ≈±15° are observed; this observation holds for $C^{(1')}$, too.

It is revealing to compare the dynamics of *trans*-AB (Figure 2) to that of its *cis*-AB analogue shown in Figure 3. In the case of *cis*-AB, the central CNNC unit is nonplanar on average, the CNNC dihedral being $-10 \pm 20°$ (Figure 3, upper panel). Furthermore, the CNNC dihedral oscillations are seen to have a much higher frequency compared to *trans*-AB. In addition, in order to minimize steric interactions, the phenyl rings are rotated by $\Phi_{1,2} = -52 \pm 17°$ as seen from the middle panel. The events where Φ_1 gets close to zero while Φ_2 is close to −90° can be viewed as attempted ring flips. In contrast to the trans conformer, a substantial amount (≈8°) of pyramidalization is observed for the $C^{(1)}$ carbon atom connected to the nitrogens (Figure 3, bottom panel); the same amount of pyramidalization is found for $C^{(1')}$.

The average values for selected structural parameters from these ab initio molecular dynamics gas-phase runs are collected in Table 1 together with our results from structure optimizations as well as published quantum chemical and experimental data. As expected, for the (nonplanar) cis conformation the NN′ bond distance is shortened, and the CN bond distance is elongated compared to the trans conformation. The steric hindrance of the aromatic ring

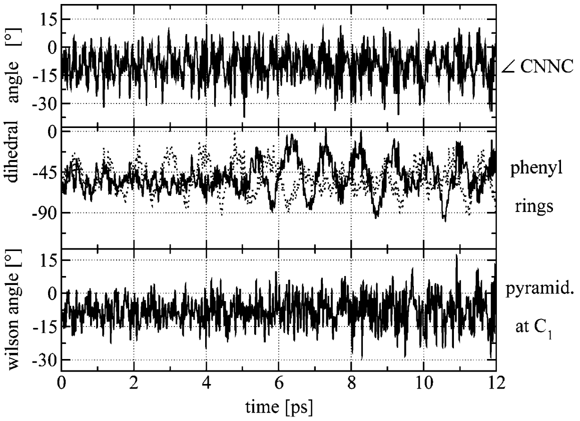


Figure 3. Time evolution of structural parameters from an ab initio molecular dynamics simulation of isolated *cis*-AB at 300 K. Top panel: dihedral angle CNNC. Middle panel: rotation angle Φ_1 and Φ_2 for the two phenyl rings (solid and dotted lines). Bottom panel: pyramidalization angle (see text) at the carbon atom $C^{(1)}$.

systems in the cis conformation results in a CNNC dihedral angle of 9.6° (optimized structure) and, in addition, a significantly (by ≈5−10°) increased CNN bond angle. Comparing the average structures to the optimized structures it is interesting to note that our simulations suggest a shortening of the NN bond and a lengthening of the CN bond at finite temperature in the trans case, while the opposite trend is observed for the cis isomer.

Importantly, the overall agreement of our structures and relative energetics for the optimized structures in comparison to experimental data and quantum chemical calculations is good. Thus, the electronic structure approach chosen and the generated ab initio trajectories can be considered as accurate reference data in order to parametrize a classical force field.

3.2. Force Field Parametrization. Using the GROMOS 45a3 force field as a starting point to derive a new force field suitable for AB, we adjusted only the bonded parameters and the charges of the azo group while keeping the original values for the remaining parameters. The parametrization was carried out in such a way as to achieve maximum agreement between the dynamical distributions obtained from force field and ab initio molecular dynamics simulations in the gas phase at 300 K concerning the relevant bond lengths, bond angles, and dihedral angles. This will ensure maximum compatibility between the QM and MM descriptions—so that switching adaptively between the quantum-mechanical (atomistic) and classical (microscopic) description for a given AB unit is as smooth as possible in future adaptive applications. Table 2 summarizes our results for the nonstandard parameters. In all cases, the values for the force constants are adapted to reproduce the widths of the distributions of bond lengths, angle, and dihedrals, starting out from the force field's standard values for chemically similar internal coordinates. In this spirit, we also decided to use the same force constants and point charges for the cis and trans isomers; only the equilibrium reference values for the bonded potentials differ.

The point charges for the atoms $C^{(1)}$ and N were adapted from average RESP charges[55] computed along the ab initio molecular dynamics reference trajectories. Since the values obtained for the aromatic ring atoms $C^{(2)}-C^{(6)}$ and the hydrogen atoms were close to their standard force field values, we decided to use the standard values and thus to take advantage of the resulting small charge groups. Note that in contrast to the GROMOS convention but in line with the AMBER convention[56] and in order to distribute the forces evenly over all contributing atoms we define explicitly *all* dihedral angles involving the two CN bonds, i.e., the dihedral angles $C^{(6)}C^{(1)}NN'$, $C^{(2)}C^{(1)}NN'$, $C^{(6')}C^{(1')}N'N$, and $C^{(2')}C^{(1')}N'N$, which consequently results in smaller force constants per dihedral compared to standard force field values.

The resulting distributions obtained with our new, separate cis and trans force fields are presented in Figure 4 together with the ab initio molecular dynamics results. Note that using different parameters for *trans*- and *cis*-AB, the distributions obtained from the ab initio molecular dynamics runs can be reproduced very accurately. However, it is our aim here to derive a single, unified force field which can be applied to study, for instance, a mixture of *trans*- and *cis*-AB molecules in the condensed phase. In addition, such a force field does not need to be modified during a simulation once a photoinduced cis/trans conformational change (or vice versa) has occurred in future nonadiabatic QM/MM simulations. Therefore, in the applications described below, we employ the average values given in the last column of Table 2.

The most striking deviation between the force field and reference data occurs for the dihedral angle ∠CNNC (fourth row in Figure 4). Note that the force constant associated with this degree of freedom not only influences the broadness of the thermal distribution observed in a molecular dynamics run but also, more importantly, governs the barrier height for the thermal trans/cis isomerization in the ground state. Based on recent ab initio calculations[8] this barrier is predicted to be ≈160 kJ/mol (see Table 1), and the new force field yields a barrier along the torsional reaction coordinate of ≈140 kJ/mol. Increasing the barrier in the force field would yield to an even narrower distribution function for the dihedral CNNC angles. On the other hand, the presently parametrized barrier height is sufficiently large to prevent thermally induced cis/trans isomerizations in the simulations of bulk AB at 400 K (see section 3.3), which should not occur on the time scale accessible to classical simulation.

In Table 1 the resulting data for the optimized structures (see 'MM (trans)' and 'MM (cis)' entries) as well as those for the unified, average force field (see 'MM (average)' entries) are collected together with our static and dynamic QM reference results (see 'DFT/PBE' and 'DFT/PBE (300 K)' entries, respectively) and compared to the literature. The data again underline the good quality of our parameter set for *trans*- and *cis*-AB, respectively. However, it becomes obvious that using the average values (for the sake of methodological consistency and ease) instead of the separate cis and trans parameters leads necessarily to some systematic deviations concerning the bond angle and, more pronounced, the bond lengths.

Based on this force field for the AB chromophore we have extended the set of force field parameters to be able to study materials containing this photoswitch such as 8AB8 introduced in Figure 1 where aliphatic side chains are attached

Table 1. Comparison of Structures and Relative Energies of an Isolated Azobenzene Molecule[a]

	method	E_{rel}	r_{NN}	r_{CN}	∠CNN	∠CNNC
			trans-AB			
literature:	CASSCF(14,12)/6-31G* [8]	0.0	1.243	1.422	115.1	180.0
	MP2/cc-pVTZ[60]	-	1.268	1.417	113.7	180.0
	DFT/BP86/TZVP[60]	-	1.267	1.420	114.8	180.0
	semiemp AM1 (mod)[12]	0.0	1.239	-	117.5	180.0
	exp. (X-ray)[61]	-	1.247	1.428	114.1	180.0
	exp. (electr diffr)[62]	-	1.25	1.43	114.1	180.0
this work:	DFT/PBE	0.0	1.270	1.426	114.6	180.0
	DFT/PBE (300 K)	-	1.263	1.450	114.5	180.0
	MM(trans)	-	1.269	1.428	115.8	180.0
	MM(average)	-	1.261	1.435	118.0	180.0
			cis-AB			
literature:	CASSCF(14,12)/6-31G* [8]	68.5	1.242	1.435	122.9	4.2
	CASPT2(14,12)/6-31G* [8]	50.2	1.242	1.435	122.9	4.2
	MP2/cc-pVTZ[60]	-	1.261	1.432	120.8	7.3
	DFT/BP86/TZVP[60]	-	1.255	1.437	124.1	11.4
	semiemp AM1 (mod)[12]	38.6	1.221	-	124.3	4.1
	exp. (X-ray)[63]	-	1.253	1.449	121.9	8.0
	exp. (ΔH)[64]	56.0	-	-	-	-
	exp. (ΔH)[53]	50.3	-	-	-	-
this work:	DFT/PBE	58.3	1.261	1.442	123.6	9.6
	DFT/PBE (300 K)	-	1.277	1.430	124.5	7.5
	MM(cis)	-	1.260	1.444	125.1	5.7
	MM(average)	-	1.266	1.435	122.6	5.4
			AB Transition State			
literature:	CASSCF(14,12)/6-31G* [8]	173.7	1.304	1.370	122.2	85.3
	CASPT2(14,12)/6-31G* [8]	159.2	1.304	1.370	122.2	85.3
	semiemp AM1 (mod)[12]	193.0	1.244	-	129.5	90.0
this work:	DFT/PBE	169.9	1.290	1.364	125.2	89.4
	MM(average)	137.6	1.269	1.427	115.2	90.0

[a] Relative energies, E_{rel}, are given in kJ/mol, bond lengths in Å, and angles in degrees. The DGT/PBE (300 K) data are average values obtained from the ab initio molecular dynamics trajectories (see text). The MM data are from optimized geometries using separate force fields for cis and trans and the unified *average* force field, respectively.

Table 2. Nonstandard Force Field Parameters Derived for Azobenzene[a]

entity		force constant	reference value		
			trans	cis	average
bonds:	NN	1.40×10^3 kJ/(mol Å^4)	1.270 Å	1.255 Å	1.2625 Å
	CN	0.72×10^3 kJ/(mol Å^4)	1.425 Å	1.440 Å	1.4325 Å
angles:	CNN	650.0 kJ/mol	114.0°	119.0°	116.5°
	CCN	560.0 kJ/mol	120.0°	120.0°	120.0°
dihedrals:	CNNC	70.0 kJ/mol	180°	180°	180°
	CCNN	6.0 kJ/mol	180°	180°	180°
	XCCX	40.0 kJ/mol	180°	180°	180°
point charges:	N		−0.200 *e*	−0.200 *e*	−0.200 *e*
	$C^{(1)}$		0.200 *e*	0.200 *e*	0.200 *e*
	$C^{(2)}-C^{(6)}$		−0.100 *e*	−0.100 *e*	−0.100 *e*
	H		0.100 *e*	0.100 *e*	0.100 *e*

[a] For azobenzene structure and atomic numbering scheme see Figure 1a (X denotes *any* atom).

to the phenyl rings of AB via ether bridges. We used methyl phenyl ether ($H_3C-O-C_6H_5$) as a model system to derive the force field parameters necessary to describe the $C^{(4)}-O-C^{(7)}$ link unit (see Figure 1 for atomic numbering scheme), whereas the remainder of these side chains will be treated using standard force field parameters.

As for AB itself a Car−Parrinello run at 300 K was performed with $HH_3C-O-C_6H_5$ as QM reference with the aim to parametrize those internal coordinates that involve the oxygen atom of the ether group. In order to take into account dynamical fluctuations of the $C^{(5)}C^{(4)}OC^{(7)}$ dihedral angle, RESP charges for $C^{(4)}$, O, and $C^{(7)}$ were calculated for different angles between 0° and 90° in steps of 10°. The resulting charge of the methyl group is taken as the charge of the alkyl carbon atom $C^{(7)}$ (united atom approach), and the resulting charge of the aryl carbon atom $C^{(4)}$ is adjusted so as to yield a neutral $C^{(4)}-O-C^{(7)}$ unit. Force field point charges were then obtained by Boltzmann averaging over

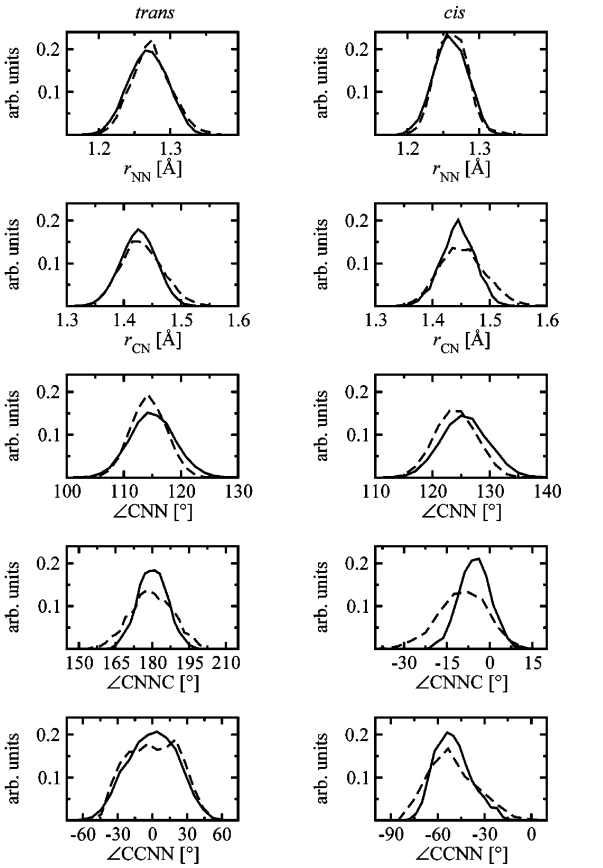


Figure 4. Structural analysis for classical force field (solid lines) and ab initio (dashed lines) molecular dynamics simulations of *trans*- (left panels) and *cis*- (right panels) AB at 300 K in terms of distribution functions for selected internal coordinates. The classical distributions were obtained from two separate force field parametrizations for cis and trans.

Table 3. Extension of Azobenzene Force Field To Include Ether Linkage $C^{(4)}-O-C^{(7)}$ [a]

entity		force constant	ref value
bonds:	$OC^{(7)}$	8.18×10^2 kJ/(mol Å^4)	1.430 Å
	$C^{(4)}O$	1.02×10^3 kJ/(mol Å^4)	1.360 Å
angle:	$C^{(4)}OC^{(7)}$	620.0 kJ/mol	116.0°
dihedral:	$C^{(3,5)}C^{(4)}OC^{(7)}$	6.0 kJ/mol	180°
point charges:	O		−0.332 *e*
	$C^{(7)}$		0.178 *e*
	$C^{(4)}$		0.154 *e*

[a] See Figure 1 for structure and atomic numbering scheme.

the torsion−angle dependent RESP charges. The resulting parameters for the ether linkage are collected in Table 3.

3.3. Classical Simulation of Liquid Azobenzene. Parametrizations of the azo group had been carried out at 300 K having in mind future applications of the force field at ambient conditions. The present applications to the study of liquid AB (with a melting point of 341 K) and liquid−crystalline AB-containing compounds (with phase transition temperatures of 8AB8 between 372 and 385 K) required testing the validity of the classical force field at an increased temperature of 400 K.

In the following we apply the new force field to study liquid AB at 400 K and analyze separately the influence of the liquid environment on the structural properties of the cis and trans conformers of AB by comparison with gas-phase simulations at 400 K. As a measure for the extension of the AB unit serves the distribution of the distance between the geometric centers of the two phenyl rings as shown in Figure 5a,b. For *trans*-AB the distributions of the single molecule and the liquid phase are indistinguishable (solid line in Figure 5a), whereas the conformations of the cis isomer are slightly more affected by the bulk environment (see Figure 5b). In the liquid phase the *cis*-AB unit is slightly stretched out compared to the vacuum simulations. Similar observations can be made when analyzing the out-of-plane motions of the phenyl rings by monitoring the distribution functions of the dihedral angle between the normal vectors of the two phenyl rings (data not shown) and of the CCNN dihedral angle (see Figure 5c,d). The conformations in *trans*-AB are not affected by the liquid environment (solid line in Figure 5c), whereas in the case of the cis conformer the amplitude of the ring motion (compared to a planar structure) is slightly larger in the isolated molecule than in the bulk liquid at the same temperature (see Figure 5d). Additionally, it was found that the distribution functions of the CNNC dihedral angle are not affected by the liquid environment—neither in the case of the trans nor in the case of the cis isomer (data not shown), and it was tested that the system does not undergo thermal cis/trans isomerization, which is a rare event that indeed should not occur on the time scale presently accessible by such classical molecular dynamics simulations.

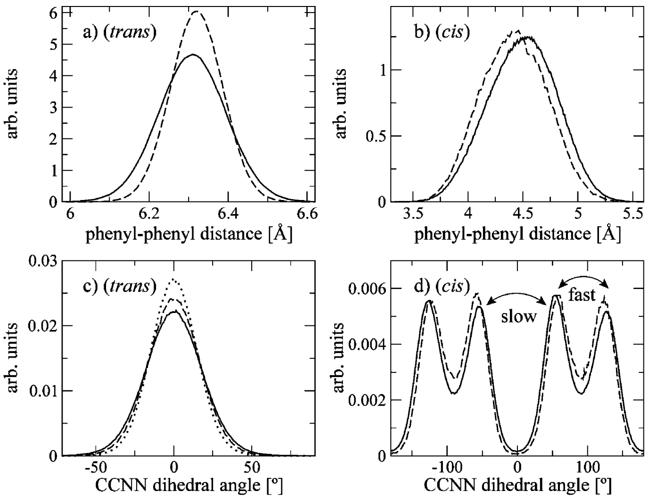


Figure 5. Distributions of structural parameters of the AB unit under various conditions at 400 K. Panels a and b: distance of the geometrical centers of the two phenyl rings. Panels c and d: CCNN dihedral angle. Panel a: *trans*-AB in liquid phase as well as in vacuum − solid line, *trans*-8AB8 in liquid phase (both isotropic and smectic) and in vacuum − dashed line; panel b: *cis*-AB in liquid phase − solid line, *cis*-AB in vacuum − dashed line; panel c: *trans*-AB in liquid phase (equivalent to *trans*-AB in vacuum and *trans*-8AB8 in vacuum) − solid line, *trans*-8AB8 in isotropic liquid − dashed line, *trans*-8AB8 in anisotropic (smectic) phase − dotted line; panel d: *cis*-AB in liquid phase − solid line, *cis*-AB in vacuum − dashed line.

Figure 5d shows that the distribution of the CCNN dihedral angle in *cis*-AB has four chemically equivalent maxima

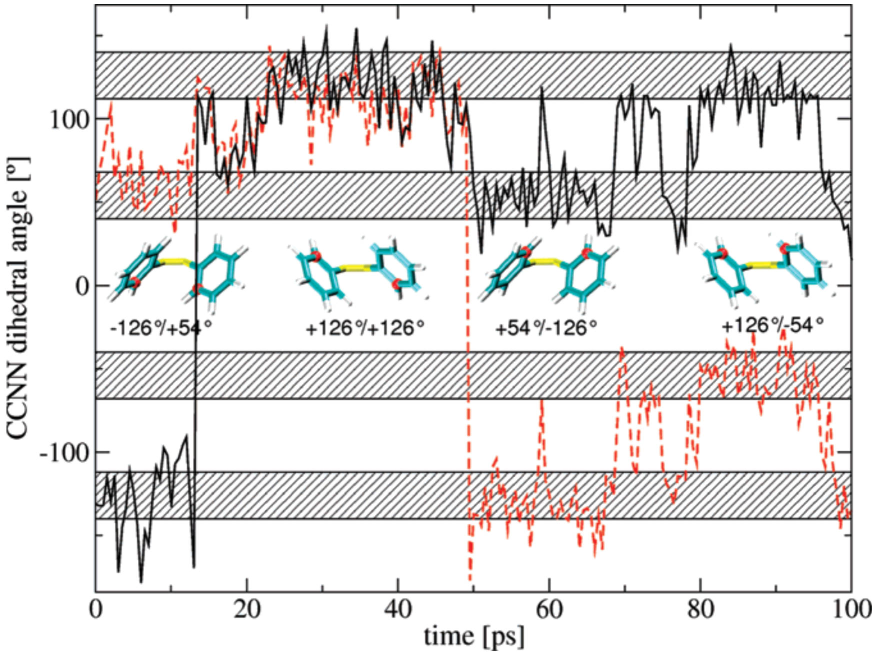


Figure 6. Time evolution of the CCNN dihedral angles of a *cis*-AB molecule in liquid environment at 400 K. The striped bars indicate the regions (±14° around the maxima of the distributions at ±54° and ±126°, see Figure 5d) used to count the transitions between the states (see text). Snapshots of typical conformations are included with the $C^{(2)}$ and $C^{(2')}$ carbon atoms that are used to define the CCNN dihedral angles marked in red.

around ±54° and ±126° and consequently two types of transitions between these states. As indicated in the figure, there is one "fast" type of transition, where the phenyl ring is intermediately standing perpendicular to the plane spanned by the $C^{(1)}$ (or the $C^{(1')}$) carbon and the two N atoms, and one "slow" type of transition, where the ring is intermediately in-plane with the $C^{(1)}$ and the two N atoms (to avoid steric hindrance in this planar conformation during the "slow" transition, the second phenyl ring has to "make way" by adopting a conformation perpendicular to the plane). These transitions are observed in the classical simulations of *cis*-AB, whereas the time scale of QM simulations of a few ps are too short to sample these transitions systematically. Figure 6 shows one example of such a process by monitoring the dynamics of the CCNN dihedral angles of one AB unit in a simulation of liquid *cis*-AB at 400 K, where both types of transitions are observed. In addition, snapshots of representative *cis*-AB conformations are shown to illustrate the conformational changes during the transitions. In order to get a rough estimate for the time scale of these ring flips the transitions of both types are counted for all CCNN dihedrals in a simulation of 343 *cis*-AB molecules at 400 K. Since the separation between the states, in particular between the states involved in the "fast" transitions, is ambiguous, narrow regions (±14°) around the maxima of the distributions at ±54° and ±126° were defined (as marked in Figure 6), and only transitions between these regions were counted. This results in transition times of approximately 20 ps for the "fast" and 200 ps for the "slow" ring flips. By Boltzmann inverting the dihedral distribution in Figure 5d, one obtains an effective barrier for the "fast" transition of the order of about 3 kJ/mol ($\approx 1\ k_BT$, where k_B is the Boltzmann constant) and for the "slow" transition a barrier of about 12 kJ/mol ($\approx 4\ k_BT$), which approximately reproduces the relative magnitude of the two transition rates extracted from the dynamics.

3.4. Hybrid QM/MM Simulations of Bulk Liquid Azobenzene. Using the hybrid CPMD/GROMOS interface[50,51] we performed two QM/MM Car−Parrinello simulations of liquid AB consisting of 343 *cis*- and *trans*-AB molecules each as described in more technical detail in section 2.4. Therein, only a single AB molecule is treated quantum-mechanically (QM AB) and interacts via the interface with the remaining 342 AB molecules which themselves interact with each other according to the molecular mechanics force field (MM AB) generated in section 3.2.

For liquid *trans*-AB, Figure 7 shows the dihedral CNNC angle, the rotation angles Φ_1 and Φ_2 of the two phenyl rings, and the pyramidalization at the $C^{(1)}$ and $C^{(1')}$ carbon atoms of the QM AB molecule (compare to Figure 2 for the corresponding gas-phase results). In contrast to the gas phase, the two aromatic ring systems now move in a C_i-like fashion ($\Phi_1 \approx -\Phi_2$) in the condensed phase, and thus the relative orientation of the two aromatic ring systems is much more conserved in the bulk liquid (compare to Figure 7, center panel). This suggests that in the liquid phase the aromatic ring systems are more or less "pinned' by the neighboring molecules, and the change in the dihedral angles Φ_1 and Φ_2 is mainly due to the mobility of the central nitrogen atoms.

In order to prove that the division of the QM/MM system into 'near' and 'far' coupling regions (see section 2.4) produces a homogeneous description of the liquid phase, we analyzed the structural distributions in these different regions.

Note that the 'near' ABs (center-of-mass distance < 20 au) interact directly with the charge density of the quantum AB molecule, while the 'far' ABs (center-of-mass distance > 30 au) interact through the multipole expansion.

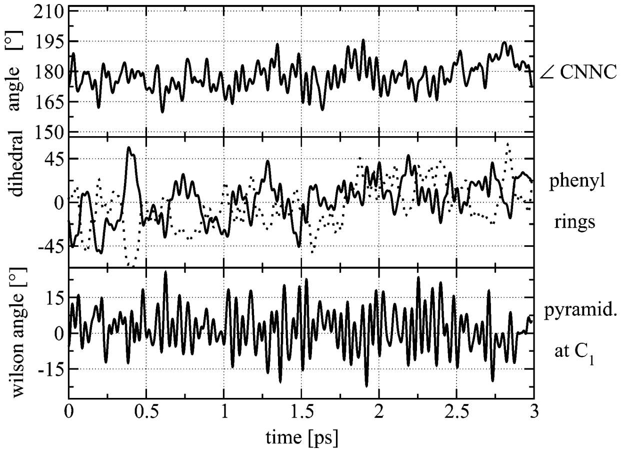


Figure 7. Time evolution of structural parameters of the QM AB unit in bulk liquid *trans*-AB from a QM/MM simulation at 400 K. Top panel: dihedral angle CNNC. Center panel: rotation angle Φ_1 and Φ_2 for the two aromatic rings (solid and dotted lines). Bottom panel: pyramidalization angle (see text) at the carbon atom $C^{(1)}$.

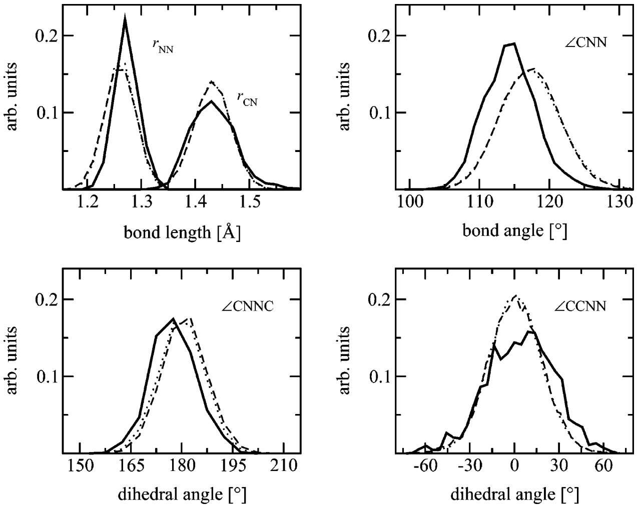


Figure 8. Distributions of selected internal coordinates from a QM/MM simulation of bulk liquid *trans*-AB at 400 K: QM AB (solid lines), near (center-of-mass distance $R < 20$ au) MM molecules (dashed lines), far ($R > 30$ au) MM molecules (dotted lines).

The resulting distribution functions for the QM AB, the 'near' ABs, and the 'far' ABs of the structural parameters of the CNNC group are given in Figure 8. For the 'near' and 'far' MM ABs, the curves are almost identical (dashed and dotted lines, respectively). This demonstrates that the QM/MM coupling scheme applied indeed results in homogeneous properties of the liquid azobenzene system.

In addition, the distributions for the QM AB molecule show the same broadness, but they appear to be slightly but systematically shifted as made most evident by the CNN bond angle distribution. This shift is due to the fact that in the force field description we use for reasons explained earlier the average reference values from Table 2 (see Figure 4) instead of the specific parametrizations for the *cis*- and *trans*-AB conformers which are also available. We would like to point out, however, that the statistical uncertainty in the distributions for the QM molecule is of course much larger than for the ensemble-averaged MM distributions, which must be considered when comparing the QM distribution to the two MM distributions in Figure 4.

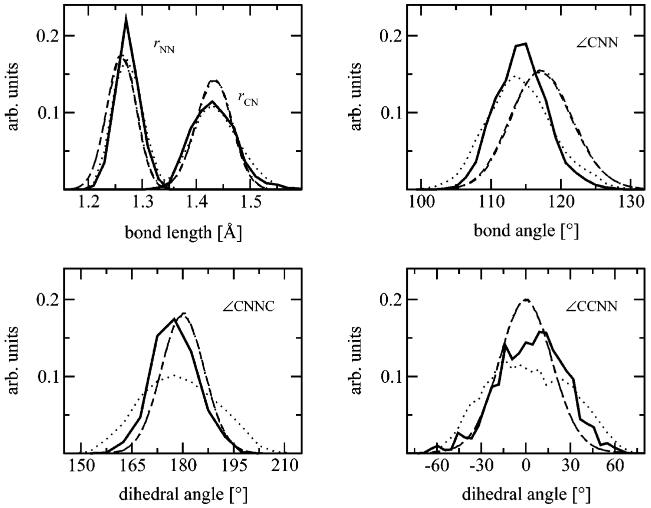


Figure 9. Comparison of selected internal coordinates from the QM/MM simulation (QM molecule, solid lines) of bulk liquid *trans*-AB at 400 K to all-MM bulk (dashed lines), QM gas-phase (dotted lines), and MM gas-phase (dashed-dotted lines) simulations.

A comparison of the distribution functions of structural parameters for the QM AB molecule in the QM/MM simulation with purely classical simulation data of all-MM liquid AB is shown in Figure 9. The difference between the QM/MM and the all-MM data is comparable to the difference between the QM and MM subsystems within the QM/MM simulation itself (see Figure 8). Furthermore, Figure 9 shows a comparison of bulk distribution functions with those in the gas phase. In particular the CNNC dihedral angle distribution is seen to be much more confined in the liquid-state environment.

The time evolution of different structural variables from the QM/MM simulation of liquid *cis*-AB can be analyzed with the help of Figure 10 (to be compared with Figure 7 for the situation with only *trans*-AB molecules in the liquid). The top panel depicts the rotational motion about the N=N double bond; the corresponding CNNC dihedral angle is found to fluctuate about 12° by ±15° flipping twice into a symmetrically equivalent position with a negative value of the CNNC dihedral angle at about 0.9 and 1.7 ps. As such a flip is not observed for the gas-phase trajectory (see Figure 3), this may be due to solvation effects of the neighboring MM AB molecules. The rotational motion of the two phenyl rings is seen to be much more hindered (see Figure 10: middle panel) than in the gas phase (see Figure 3), since the CCNN dihedral angle is only about 45° and the fluctuations are less pronounced. At about 1.7 ps and coinciding with the second flip of the CNNC dihedral angle (see above), the relative orientation of the two phenyl rings (see Figure 3: middle panel, solid line) approaches 90°, with one ring roughly perpendicular and the other one roughly parallel to an idealized CNNC plane. This may be viewed as an (unsuccessful) ring flip scenario similar to those described

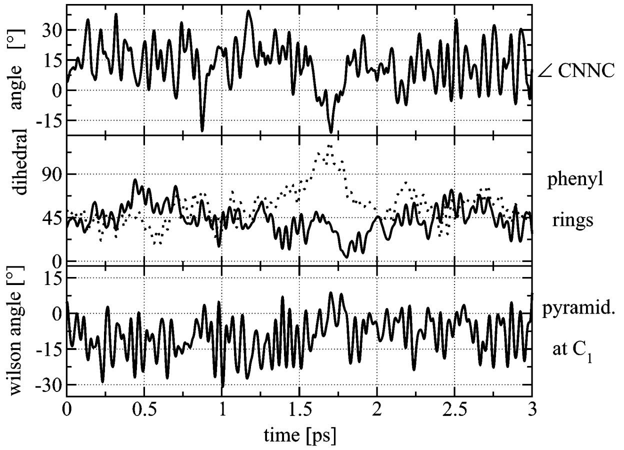


Figure 10. Time evolution of structural parameters of the QM AB unit in bulk liquid *cis*-AB from a QM/MM simulation at 400 K. Top panel: dihedral angle CNNC. Center panel: rotation angle Φ_1 and Φ_2 for the two aromatic rings (solid and dotted lines). Bottom panel: pyramidalization angle (see text) at the carbon atom $C^{(1)}$.

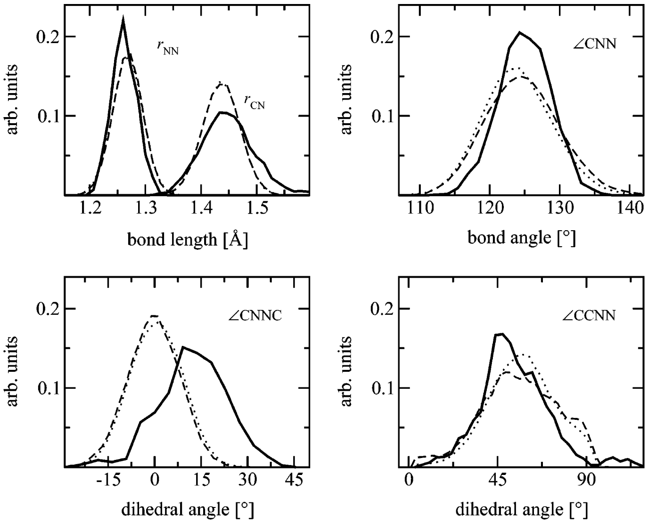


Figure 11. Distributions of selected internal coordinates from a QM/MM simulation of bulk liquid *cis*-AB at 400 K: QM AB (solid lines), near (center-of-mass distance R < 20 au) MM molecules (dashed lines), far (R > 30 au) MM molecules (dotted lines).

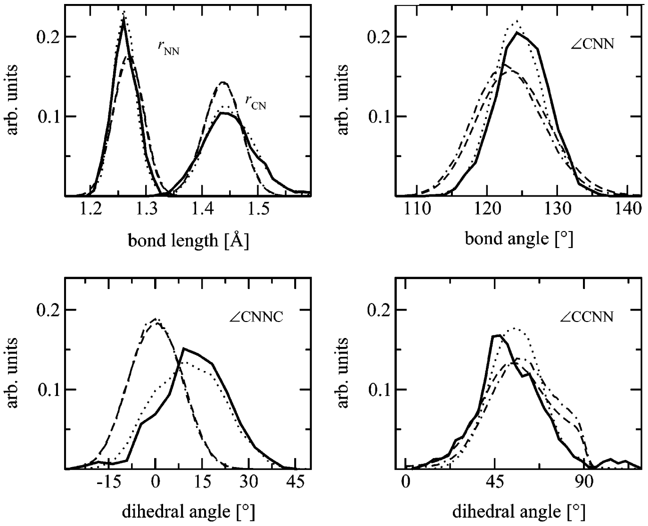


Figure 12. Comparison of selected internal coordinates from the QM/MM simulation (QM molecule, solid lines) of bulk liquid *cis*-AB at 400 K to all-MM bulk (dashed lines), QM gas-phase (dotted lines), and MM gas-phase (dashed-dotted lines) simulations.

for the all-MM calculation (see section 3.3). The bottom panel of Figure 10 shows that the average pyramidalization of the carbon atom $C^{(1)}$ ($C^{(1')}$ behaves analogously) is as pronounced as in the gas phase (see Figure 3), however exhibiting even larger fluctuations.

Like in the *trans*-AB case, Figure 11 presents histograms of structural parameters from the *cis* QM/MM simulation divided into contributions from the QM molecule itself as well as the near (center-of-mass distance R < 20 au) and far (R > 30 au) MM solvent molecules. As for the liquid *trans*-AB, there is little difference between near and far MM AB molecules, and the distributions for the QM AB molecule again show the same broadness. In case of the CNNC dihedral angle, there is a significant shift due to the fact that the reference value of the force field corresponds to 0°. The CCNN dihedral angle distributions for the different regions are in good agreement with each other. For the ease of comparison the four peaks visible in Figure 5d have been mapped onto the ±54° region.

Figure 12 compares the structural distribution functions from the liquid *cis*-AB QM/MM simulation to all-MM simulations of liquid AB, on the one hand, and to gas-phase all-QM and all-MM data, on the other hand. In particular, when comparing all-QM gas-phase data to the QM subsystem in QM/MM simulations, one notices a slight smearing out of the histograms in solution. Deviations between the liquid-phase QM/MM and MM distributions are similar to those observed in Figure 11 between the QM and MM subsystems in the QM/MM simulation.

3.5. Classical Simulations of the 8AB8 Liquid Crystal. In this section, we present the application of the azobenzene force field to the liquid−crystalline-phase behavior of 8AB8 (sketched in Figure 1b). This AB containing liquid crystal exhibits in experiments a crystalline → nematic-phase transition at 372 K followed by a nematic → isotropic-phase transition at 385 K; in addition there is a monotropic nematic → smectic C-phase transition at 368 K upon cooling. The main intention here is to explore the phase behavior of 8AB8 using classical atomistic (MM) simulations. The stimulation for doing so is 2-fold. First of all, we want to lay the foundation for future QM/MM studies on the photoisomerization of azobenzene compounds in bulk liquid and the anisotropic liquid−crystalline environments and, second, for the development of a coarse grained model of 8AB8 to investigate the photoinduced-phase transition itself. It should be noted that this section does not primarily serve as a test of the quality of the force field as explained in the Introduction.

A system of 1296 8AB8 molecules was setup initially in a smectic arrangement (i.e., the starting structure consisted of four 8AB8 layers of 324 molecules each in an orthorhombic box). In order to investigate the phase behavior of the system, replica exchange simulations in a temperature

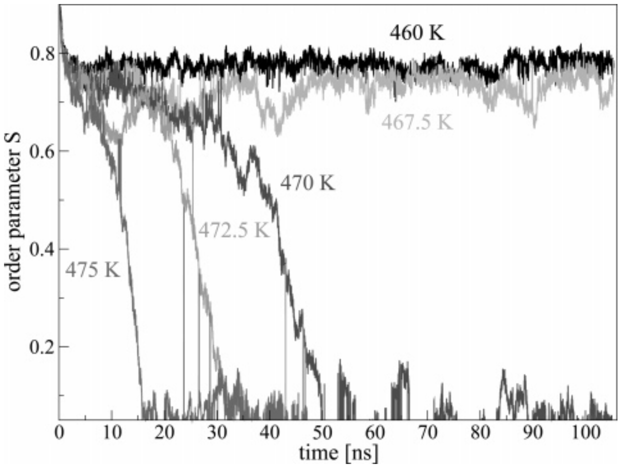


Figure 13. Order parameter in the replica exchange simulations of 8AB8 (460−475 K). Each gray shade corresponds to one replica temperature as indicated in the graph (replicas at 462.5 and 465 K not displayed, they behave like the replica at 460 K). Exchanges between two replicas lead to the discontinuities in the order parameter due to the exchange of conformations. (The replicas at 475 and 472.5 K were discontinued after 48 ns of simulation time.)

range between 460 and 475 K were carried out as outlined in section 2.3. The temperature range had been determined beforehand by simulating the system at a number of different temperatures ranging from temperatures where the systems remains stable in the layered/ordered setup up to temperatures where the initial order quickly decays and one obtains an isotropic phase. The temperature shift of this range compared to the experimentally observed transition temperatures will be further addressed below. Figure 13 shows the order parameter eq 1 in the replica exchange simulations of 8AB8 in the temperature range from 460 to 475 K. Exchanges between replicas at different temperatures are observed which is visible as discontinuities in the order parameter due to the exchange of conformations. One nicely sees the loss of the original order in the replicas from 475 K down to 470 K (note that the replicas at 475 and 472.5 K were discontinued after 48 ns of simulation time). The remaining replicas below 470 K remain stable with an order parameter of ≈0.75.

Snapshots from these replica exchange simulations are shown in Figure 14. The left snapshot is taken from the replica at 470 K at a simulation time of 37 ns (order parameter $S \approx 0.6$), thus showing a structure during the transition from the layered setup to the isotropic state. The right snapshot is taken from the replica at 465 K at a simulation time of 86 ns ($S \approx 0.75$), showing a representative structure for the systems that remain stable in an ordered phase (the inital layered setup is still very well recognizable in this snapshot). In the left snapshot, which shows the system during the transition from smectic ordering to the isotropic state, one sees that at least partially the system does maintain some degree of order/alignment while the smectic layers are dissolving. The right snapshot shows the system at a lower temperature, and we can observe that, while the alignment along the *z*-direction is maintained, the layers start to dissolve, i.e., molecules move in the *z*-direction while they remain oriented. We also observe realignment of molecules that had been lying between the layers perpendicular to the director.

The experimentally observed phases and transition temperatures cannot be fully reproduced with these classical atomistic simulations. The nematic phase (which would experimentally be expected between 372 and 385 K) is not observed, but the layered setup is stable up to a much higher temperature until it dissolves rapidly and the system goes directly to the isotropic phase (even though transient nematic-like structures are observed as well as molecules that leave the smectic layers and yet stay oriented). This phase behavior most likely has several reasons. First of all, the temperature range in which the nematic phase is observed experimentally for this compound is rather narrow (≈10 K), which is a challenge for computer simulations where overheating and finite size effects may play an important role and possibly lead to an overstabilization of the smectic layers. For example, an artificial stabilization of the smectic layers due to the anisotropic pressure scaling, which allows the aspect ratio of the sides of the orthorhombic box to adjust to the layered structure, cannot be ruled out and needs to be further investigated in the future. In addition, rather small variations in the classical force field may have large effects on the transition temperatures. Different factors can play a role in this context. For one, the intramolecular degrees of freedom determine the flexibility/stiffness of the individual molecules which plays an important role in the liquid−crystalline-phase behavior. In the first tests when the force field was parametrized we noticed that changes in the molecular flexibility have an effect on the thermal stability of the smectic phase (for instance we found that the transition temperature decreased after the azo group and the ether linkage between the azobenzene unit and the tails were reparametrized based on the all-QM simulations compared to first guesses made based on standard force field parameters for similar groups). A second important aspect concerning the force field are the intermolecular/nonbonded interactions. For 8AB8 the balance of aromatic/aromatic (azobenzene core), aliphatic/aliphatic (alkoxy tails) and aromatic/aliphatic interactions is particularly relevant (this menas that the liquid densities of the alkyl and aromatic fragments as well as the mixing properties of alkyl and aromatic groups have to be correctly reproduced). It should be noted that we are aware that there are procedures to optimize parameters for classical atomistic force fields for liquid−crystalline systems;[57−59] however, it was not our aim to provide an optimized classical force field for 8AB8 liquid crystals but to apply an existing force field (the GROMOS force field which is parametrized on thermodynamic properties in the liquid state) and to supplement the nonstandard parameters for the azo group. One key objective in the parametrization was the suitability of the classical model for further use in QM/MM simulations in order to study the influence of bulk liquid and anisotropic liquid−crystalline environments on the azobenzene photoisomerization. We also note that our set of bonded parameters for the azo group could be combined with a different nonbonded force field without having to reparametrize the AB unit itself.

We hope to get a better understanding of the phase behavior of 8AB8 in future multiscale simulations using

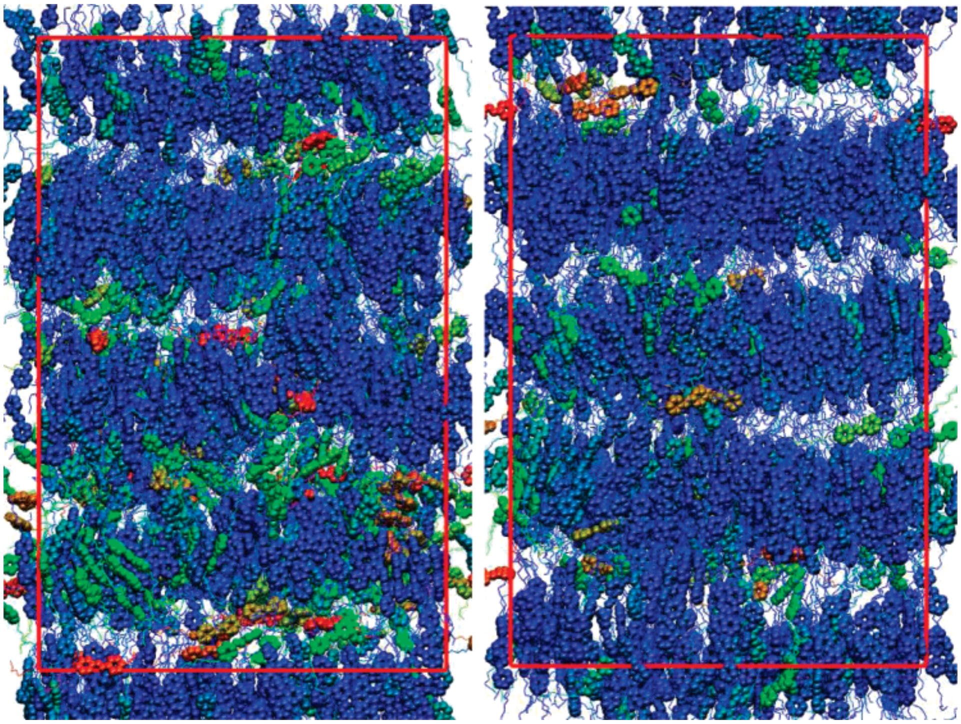


Figure 14. Snapshots from replica exchange simulations. The left snapshot stems from the replica at 470 K at a simulation time of 37 ns ($S \approx 0.6$), and the right snapshot stems from the replica at 465 K at a simulation time of 86 ns ($S \approx 0.75$). The red frame indicates the size of the periodic box. Spherical beads: azobenzene atoms, lines: alkoxy tails. The color shows the alignment of each molecule with respect to the director (from blue to red; blue: aligned along the director, i.e., along the *z*-direction; red: perpendicular to the director).

classical atomistic and coarse grained simulations. First simulations using a coarse grained (CG) model that was built on the classical atomistic one did indeed reproduce the nematic phase (a publication on the development of the CG model is in preparation). This CG model can be utilized to investigate systematically the phase behavior of 8AB8 and (after reintroduction of atomistic details into the CG coordinates, so-called inverse mapping) to obtain equilibrated atomistic coordinates. Since CG simulations lead to a significant speedup, we can also systematically study the influence of box size and shape on the relative stabilities of nematic and smectic phases and the resulting finite size effects—a comparison which due to the simulation time limitations would not be possible with all-atom simulations alone.

In order to characterize the ordered/smectic phase that is observed in the simulations we reorder the replicas to obtain trajectories which are continuous in conformation space which implies necessarily fluctuating temperatures. For these reordered replicas one can analyze diffusion constants in the *xy* planes and perpendicular to the layers along the *z*-direction. It is observed that in the case of the replicas that remain in the layered arrangement the in-plane diffusion is considerably faster than the diffusion perpendicular to the planes—characterizing the observed phase indeed as smectic and not as frozen.

In the following we study the influence of the anisotropic liquid−crystalline environment on the conformational distributions of the *trans*-AB functional units within the 8AB8 molecules. In order to compare with the conformations in pure liquid AB as discussed in the previous sections, the liquid−crystalline system was quenched down to 400 K, both starting from a disordered/isotropic structure and from a layered/smectic structure. The resulting distributions are added to Figure 5a,c, and the distributions obtained from a vacuum simulation of 8AB8 were added as well for the sake of comparison. For all three simulations of 8AB8, the distribution of the phenyl−phenyl distance is narrower than in both gas-phase and liquid-phase *trans*-AB (see Figure 5a). This shows that the aliphatic tails have a sort of ‘pulling influence’ on the conformational equilibria of the AB core unit, again an effect which is potentially important for photoisomerization in condensed-phase environments such as liquid−crystalline materials. Figure 5c shows that the distribution functions of the CCNN dihedral angle, which measures the out-of-plane motions of the phenyl rings, in a single *trans*-8AB8 chain do not differ from *trans*-AB, whereas they are clearly affected in the two condensed-phase simulations of 8AB8. The distributions are narrower in the isotropic phase and even more so in the smectic phase, i.e., the oriented molecules are kept more planar due to the anisotropic environment.

Overall, our condensed-phase simulations show that there are various factors that very clearly influence the conformations of the AB photoactive functional unit compared to the noninteracting reference situation in the gas phase.

In particular, intramolecular effects due to the ‘pulling’ of the tails are observed in 8AB8, which are likely to play an important role in AB-containing photo(re)active materials in more general terms. In addition, intermolecular effects exist due to the bulk environment that are most pronounced in the case of an anisotropic liquid−crystalline environment.

These results suggest that the condensed-phase environment may induce changes to the mechanism and efficiency of the photoisomerization which needs to be investigated in the future.

4. Conclusions and Outlook

In a long-term effort to being able to simulate photo(re)-active materials containing covalently bonded azobenzene (AB) functional units as photoswitches, we have extended and adapted a classical force field for AB based on ab initio molecular dynamics reference simulations and the GROMOS 45a3 force field. Using this novel force field we have presented a first application to the study of AB in the bulk liquid state. To this end, not only a standard purely classical molecular−mechanics (MM) setup but also a hybrid quantum-classical QM/MM approach, where only a single AB molecule has been treated quantum-mechanically (QM), has been used to conduct molecular dynamics simulations. Based on various observables we have verified that both levels of simulation yield conformational distributions that are consistent with each other. This agreement was an important initial goal of the effective parametrization in order to allow for smooth transitions between classical and quantum mechanical description of AB units in future adaptive multiscale simulations. This will be necessary to investigate the influence of bulk liquid and anisotropic liquid−crystalline environments on the photoisomerization mechanism of the AB functional group.

Furthermore, the AB force field has been extended to include aliphatic hydrocarbon chains that are connected via ether links to the phenyl rings. In particular, the extended azobenzene force field has been employed to study the liquid−crystalline azobenzene compound 8AB8. Using this force field the phase behavior of the liquid−crystalline AB compound 8AB8 has been studied by classical simulations using replica exchange techniques. The structures observed in these simulations give confidence that the experimental phases of 8AB8 can be described by the generated force field. However, in order to simulate time and length scales required to properly cover such phase transition processes and to systematically investigate the phase behavior of 8AB8, a coarser description of the system is necessary. The development of such a coarse grained model for 8AB8 based on the atomistic force field is on the way.

Apart from the above aspects of force field generation and validation we could also show that the liquid environment, and even more so the anisotropic environment in a liquid crystal, influences the conformational distributions of the *trans*-and *cis*-AB units as compared to the gas-phase reference at the same temperature. In particular, AB units in the cis arrangement are found to be much more affected by condensed-phase effects in comparison to the trans conformer. It is therefore essential for the study of photoisomerizations of AB chromophores that are covalently embedded in materials to properly take into account environmental condensed-phase effects. Work along these directions is currently in progress.

Acknowledgment. We are grateful to the Volkswagen Stiftung for supporting our joint project “Adaptive Multiscale Simulation: Connecting the Quantum to the Mesoscopic Level” within the framework of the program “New Conceptual Approaches to Modeling and Simulation of Complex Systems − Computer Simulation of Molecular and Cellular Biosystems as well as Complex Soft Matter”. The simulations were performed using resources from NIC Jülich, Bovilab@RUB, Rechnerverbund−NRW, and Rechenzentrum Garching (MPG). C.P. wants to thank Berk Hess for providing the fast replica exchange code and Berk Hess and Nico van der Vegt for many stimulating discussions.

References

(1) *Photoreactive organic thin films;* Sekkat, Z., Knoll, W., Eds.; Academic Press: San Diego, CA, 2002.

(2) *Molecular Switches;* Feringa, B. L., Ed.; Wiley-VCH: Weinheim, 2001.

(3) Stolow, A. *Ann. Rev. Phys. Chem.* **2003**, *54*, 89.

(4) Yesodha, S. K.; Pillai, C. K. S.; Tsutsumi, N. *Prog. Polym. Sci.* **2004**, *29*, 45.

(5) Yager, K. G.; Barrett, C. J. *J. Photochem. Photobiol., A* **2006**, *182*, 250.

(6) Nägele, T.; Hoche, R.; Zinth, W.; Wachtveitl, J. *Chem. Phys. Lett.* **1997**, *272*, 489.

(7) Fujino, T.; Arzhantsev, S. Y.; Tahara, T. *J. Phys. Chem. A* **2001**, *105*, 8123.

(8) Cembra, A.; Bernardi, F.; Garavelli, M.; Gagliardi, L.; Orlandi, G. *J. Am. Chem. Soc.* **2004**, *126*, 3234.

(9) Cattaneo, P.; Persico, M. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4739.

(10) Ishikawa, T.; Noro, T.; Shoda, T. *J. Chem. Phys.* **2001**, *115*, 7503.

(11) Diau, W.-G. *J. Phys. Chem. A* **2004**, *108*, 950.

(12) Ciminelli, C.; Granucci, G.; Persico, M. *Chem. Eur. J.* **2004**, *10*, 2327.

(13) Schultz, T.; Quenneville, J.; Levine, B.; Toniolo, A.; Martinez, T. J.; Lochbrunner, S.; Schmitt, M.; Shaffer, J. P.; Zgierski, M. Z.; Stolow, A. *J. Am. Chem. Soc.* **2003**, *125*, 8098−8099.

(14) Tiago, M. L.; Ismail-Beigi, S.; Louie, S. G. *J. Chem. Phys.* **2005**, *122*, 094311.

(15) Toniolo, A.; Ciminelli, C.; Persico, M.; Martinez, T. *J. Chem. Phys.* **2005**, *123*, 234308.

(16) Crecca, C. R.; Roitberg, A. E. *J. Phys. Chem. A* **2006**, *110*, 8188−8203.

(17) Spörlein, S.; Carstens, H.; Satzger, H.; Renner, C.; Behrendt, R.; Moroder, L.; Tavan, P.; Zinth, W.; Wachtveitl, J. *Proc. Natl. Acad. Sci.* **2002**, *99*, 7998.

(18) Wachtveitl, J.; Spörlein, S.; Satzger, H.; Fonrobert, B.; Renner, C.; Behrendt, R.; Oesterhelt, D.; Moroder, L.; Zinth, W. *Biophys. J.* **2004**, *86*, 2350.

(19) Browne, W. R.; Feringa, B. L. *Nat. Nanotechnol.* **2006**, *1*, 25.

(20) Hugel, T.; Holland, N. B.; Cattani, A.; Moroder, L.; Seitz, M.; Gaub, H. E. *Science* **2002**, *296*, 1103.

(21) Kumar, G. S.; Neckers, D. C. *Chem. Rev.* **1989**, *89*, 1915−1925.

(22) Liu, Z. F.; Hashimoto, K.; Fujishima, A. *Nature* **1990**, *347*, 658.

(23) Sekkat, Z.; Dumont, M. *Appl. Phys. B* **1992**, *54*, 486.

(24) Ikeda, T.; Tsutsumi, O. *Science* **1995**, *268*, 1873.

(25) Hagen, R.; Bieringer, T. *Adv. Mater.* **2001**, *13*, 1805.

(26) Natansohn, A.; Rochon, P. *Chem. Rev.* **2002**, *102*, 4139−4175.

(27) Shibaev, V.; Bobrovsky, A.; Boiko, N. *Prog. Polym. Sci.* **2003**, *28*, 729−836.

(28) Yu, Y.; Nakano, M.; Ikeda, T. *Nature* **2003**, *425*, 145.

(29) Banerjee, I.; Yu, L.; Matsui, H. *J. Am. Chem. Soc.* **2003**, *125*, 9542.

(30) Tsutsumi, O.; Shiono, T.; Ikeda, T.; Galli, G. *J. Phys. Chem. B* **1997**, *101*, 1332−1337.

(31) Pieroni, O.; Fissi, A.; Angelini, N.; Lenci, F. *Acc. Chem. Res.* **2001**, *34*, 9.

(32) de Jeu, W. H. *J. Phys.* **1977**, *38*, 1265−1273.

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

(34) Marx, D.; Hutter, J. Ab Initio Molecular Dynamics: Theory and Implementation. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; NIC: Jülich, 2000. www.theochem.rub.de/go/cprev.html (accessed June 2007).

(35) Hutter, J. et al. Car−Parrinello Molecular Dynamics: An *Ab Initio* Electronic Structure and Molecular Dynamics Program. www.cpmd.org (accessed June 2007).

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

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

(38) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.

(39) Fukui, K.; Cline, J. I.; Frederick, J. H. *J. Chem. Phys.* **1997**, *107*, 4551−4563.

(40) Cox, S. R.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

(41) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205−1218.

(42) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Vdf Hochschulverlag AG an der ETH Zürich: Zürich, 1996.

(43) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596−3607.

(44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(45) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.

(46) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(47) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(48) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(49) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(50) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(51) Carloni, P.; Rothlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455−464.

(52) Wilson, E. B., Jr.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill: New York, 1955.

(53) Adamson, A. W.; Vogler, A.; Kunkely, H.; Wachter, R. *J. Am. Chem. Soc.* **1978**, *100*, 1300.

(54) Tsuji, T.; Takashima, H.; Takeuchi, H.; Egawa, T.; Konaka, S. *J. Phys. Chem. A* **2001**, *105*, 9347.

(55) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(56) 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, CA, 2002.

(57) Cheung, D. L.; Clark, S. J.; Wilson, M. R. *Phys. Rev. E* **2002**, *65*, 051709.

(58) Wilson, M. R. *Int. Rev. Phys. Chem.* **2005**, *24*, 421−455.

(59) Bizzarri, M.; Cacelli, I.; Prampolini, G.; Tani, A. *J. Phys. Chem. A* **2004**, *108*, 10336−10341.

(60) Fliegl, H.; Köhn, A.; Hättig, C.; Ahlrichs, R. *J. Am. Chem. Soc.* **2003**, *125*, 9821.

(61) Bouwstra, J. A.; Schouten, A.; Kroon, J. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **1983**, *39*, 1121.

(62) Traetterberg, M.; Hilmo, I.; Hagen, K. *J. Mol. Struct.* **1977**, *39*, 231.

(63) Mostad, A.; Romming, C. *Acta Chem. Scand.* **1971**, *25*, 3561.

(64) Schulze, F. W.; Petrick, H.-J.; Cammenga, H. K.; Klinge, H. *Z. Phys. Chem. (Munich)* **1977**, *107*, 1.

CT7000733

JCTC Journal of Chemical Theory and Computation

Parametrization and Validation of Intramolecular Force Fields Derived from DFT Calculations

Ivo Cacelli* and Giacomo Prampolini

Dipartimento di Chimica e Chimica Industriale, Università di Pisa, via Risorgimento 35, I-56126 Pisa, Italy

Received May 15, 2007

Abstract: The energy and its first and second geometrical derivatives obtained by DFT calculations for a number of conformations of a single molecule are used to parametrize intramolecular force fields, suitable for computer simulations. A systematic procedure is proposed to adequately treat either fully atomistic or more simplified force fields, as within the united atom approach or other coarse grained models. The proposed method is tested and validated by performing molecular dynamics simulations on several different molecules, comparing the results with literature force fields and relevant experimental data. Particular emphasis is given to the united atom approach for flexible molecules characterized by “soft” torsional potentials which are known to retain a high degree of chemical specificity.

1. Introduction

Thanks to the massive increase of computational resources of the past two decades, the study and design of novel materials possessing the desired physical, chemical, and biological requirements can be significantly aided by a detailed investigation on structure and interactions at molecular scale. In this theoretical approach to material science, computer simulations methods[1,2] such as Monte Carlo (MC) and Molecular Dynamics (MD) have been employed successfully in many complex systems as, for instance, clays,[3] nanocomposites,[4] polymers,[5−7] liquid crystals,[8] macromolecules,[9] or biological membranes.[10] In MC or MD simulations, all the information on the molecular framework and interactions are encrypted in the adopted force field (FF), which can be considered as the link between the microscopic description and the resulting structural and dynamic macroscopic properties.

In computer simulation studies of advanced materials two main problems arise, due to the complexity of the intra- and intermolecular forces. First, notwithstanding the usually large dimensions of the forming molecules, small differences in the specific molecular structure may produce impressive variations in the resulting material properties due to the delicate interplay between energetic and entropic effects. This implies that caution must be used in extending the straightforward adoption of the most widely employed FFs[11−17] to target molecules. Second, the wide range of length and time scales that characterizes the dynamics of complex systems impose some limitations on the complexity of the FF that can be adopted. Consequently, increasing attention has been devoted to the construction of united atom (UA) or coarse grained (CG) models,[6,10,18−20] capable of capturing the main physics of the problem while reducing the computational expense with respect to fully atomistic (FA) descriptions. These two issues (transferability problems and computational advantages of simplified FFs) convinced us to work toward a route that could allow the parametrization of FFs specifically suited for a definite target molecule rather than performing ad hoc empirical corrections on potential parameters transferred from the general literature FFs.

To make up for the lack of specificity in the standard FF description of the molecular interactions, our group has recently proposed the Fragmentation Reconstruction Method (FRM),[21,22] for the calculation of the intermolecular potential energy surface (PES) of large molecules from ab initio data only. Once the two-body potential has been computed for a number of dimer arrangements, the parametrization of an intermolecular FF of the desired complexity can be performed by a least-square fitting. The obtained parameters, employed in simulations, can then be validated through the

* Corresponding author e-mail: ivo@dcci.unipi.it.

10.1021/ct700113h CCC: $37.00
Published on Web 08/21/2007

comparison of the resulting macroscopic properties with the relevant experimental measures. Up to now, the whole approach has been successfully tested in the field of liquid crystals for both UA[23−26] and CG[27] models of some mesogenic molecules.

With regards to the intramolecular part of the potential, a great variety of internal FFs has been proposed in the last two decades,[11,12,14,17,28−30] and many different sets of parameters can be found in the literature. Among these, it may be useful to distinguish between class I and class II FFs (see e.g. ref 29), on the basis of the complexity of their analytical expressions: class I FFs[11,14,17] are diagonal in a given set of internal coordinates and only contain harmonic terms for bond stretching and angle bending, while the more recent class II FFs[12,28,29] usually include anharmonic terms and off-diagonal couplings. Broadly speaking, while the former are designed for simulations of large systems in condensed phase, the latter have been mostly used for the study of structures and energetics in the gas phase.[31]

The attempt to extend literature intramolecular parameters to a well-defined target molecule requires a distinction between “hard” and “soft” internal degrees of freedom. The formers, such as bond stretching and bond angle bending, are usually less affected by the molecular environment, and the transferability of literature FFs is expected to be satisfactory for most purposes. Nonetheless, many cases have been reported where the intramolecular description achieved by standard class I FFs was found not to be adequate as, for instance, the reproduction of spectroscopic properties[32,33] or in the presence of nonstandard structures as heteroaromatic rings.[34] The situation is even more complex for flexible molecules, where flat energy profiles and small energy barriers are found for “soft” internal degrees of freedom, generally dihedral angles. This causes different conformations to be populated even at low temperatures with a marked sensitivity to both intra- and intermolecular environment. An accurate description of these torsional curves is therefore necessary, and a reparametrization of some specific torsions has often been performed[20,24,35−38] to correct the standard literature FFs. More importantly, when dealing with large molecules, the necessity of adopting novel UA or CG models calls for new parametrizations of those intramolecular potentials which describe the forces between the nonatomistic sites.[27,39] To our knowledge, no attempt has yet been made to propose a systematic approach capable of providing a set of intramolecular parameters for newly designed CG models.

In this context, the method here proposed is intended to provide a complete parametrization of intramolecular FFs of a target molecule, at various degrees of complexity, on the basis of quantum mechanical (QM) calculations only. For these reasons the computed FF will be labeled QMD, i.e., quantum mechanics derived. The implementation of this method should provide both spectroscopic quality FFs or less complex intramolecular parametrizations suitable for simulations of condensed phases.

Similar QM-derived approaches have already been proposed by other groups[28,30,40] who employed QM ab initio data to parametrize intramolecular FFs at several levels of complexity. However, these studies aim at providing spectroscopic quality class II FFs, and none of these considers intramolecular parametrizations in the UA approach. It should be pointed out that, owing to the fictitious molecule inherent to the UA approach, a UA-FF cannot be rigorously derived from QM data. Nevertheless, as discussed in this paper, suitable approximations can still be made in order to judiciously use the QM information in UA-FF parametrization. Once the model to be adopted for the target molecule has been chosen, the present method should thus provide the best values of the parameters entering the assigned functions of internal coordinates. In fact, as previously stressed, another peculiarity of the proposed approach with respect to other similar methods as, for example, QMFF[29] is to privilege FF specificity rather than FF transferability. In addition to the above-mentioned reasons, this strategy is also advisable for the increase of computational resources, which allow nowadays to perform accurate QM calculations even for large systems, as for instance liquid crystal forming molecules,[23,24,38] thus both providing a QM database to calibrate specific intramolecular FFs and reducing the need of invoking transferability. A different situation is encountered when treating very large biological systems, where the dimension of the involved molecules rules out QM calculations, and transferable parameters are strongly needed to perform computer simulation studies.

The paper is organized as follows: in section 2 the theory of the method is described, and the computational details of the employed techniques are given in section 3. The first part of section 4 is devoted to the validation of the method for “hard” internal coordinates (IC). The intramolecular FFs, computed for several stiff heteroaromatic molecules, are employed in MD-FA simulations, and some of the resulting thermodynamic and structure properties are compared with both theoretical and experimental data. In the second part of section 4, the capability of the proposed approach to yield intramolecular FF for large, flexible molecules in the UA approach is tested. Finally, main conclusions are drawn in section 5.

2. Theory

The QMD-FF is modeled on the basis of QM results, namely energies, energy gradient, and Hessian matrix, for a number of molecular geometries. To make the formulas easier to be understood, the following notation will be adopted for the summation indices and symbols: i and j are used for the Cartesian coordinates (CCs) x or mass weighted Cartesian coordinates ($1 \div 3N$), μ and ν indicate the redundant internal coordinates[41,42] (RICs) q ($1 \div N_{\mathrm{RIC}}$), K and L run over the normal coordinates (NCs) Q ($1 \div 3N - 6$) ($3N - 5$ for linear molecules), g runs over the considered molecular geometries ($0 \div N_g$), a and b indicate the functions f used to represent the empirical FF and/or the number of linear parameters of the FF ($1 \div N_{\mathrm{func}}$), and s and t run over the quantities to be represented by the FF (energies, energy gradients and Hessian) for the considered geometries ($1 \div N_{\mathrm{points}}$).

The QMD-FF, to be used in molecular dynamics or molecular mechanics, is expressed through a linear combination of functions f_a of a set of RICs

$$V(q) = \sum_{a=1}^{N_{\text{func}}} p_a f_a(q) \tag{1}$$

where the q symbol collects all RICs. The functions may conveniently be expressed in terms of displacements with respect to a given reference geometrical conformation identified by the vector q^0

$$\Delta q_\mu = q_\mu - q_\mu{}^0 \tag{2}$$

Usually the RICs consist of all bond stretches, angle bendings, and dihedral torsions that can be obtained from a given connectivity criteria referred to the reference conformation. The inversion coordinate[43] can be included for atoms bonded to three other atoms. Nonbonded intramolecular interactions can also be added in order to make the FF more accurate. In usual FFs the number of RICs exceeds $3N-6$, and therefore they form a redundant set of coordinates. Although eq 1 has been written in a general form, each function f_a only depends on one or two RICs, as reported in detail later on (eqs 36−41).

2.1. Internal Coordinates Transformations. Since the Hessian and gradients are computed in CCs, whereas the FF is usually expressed through RICs, some coordinate transformation is required. For infinitesimal displacements with respect to a given geometrical conformation, the RICs are related to the nuclear CCs x through a noninvertible transformation

$$\delta q = B\delta x \tag{3}$$

where δq and δx are column vectors. The Wilson rectangular B matrix

$$B_{\mu i} = \left(\frac{\partial q_\mu}{\partial x_i}\right) \tag{4}$$

is related to the geometry the displacements are referred to and can be accurately computed both in analytical[44] and numerical ways.

The normal coordinates are computed from the Hessian matrix in CCs

$$H_{ij} = \left(\frac{\partial^2 E}{\partial x_i \partial x_j}\right) = E''_{ij} \tag{5}$$

obtained by a QM calculation at a given geometry. H is transformed to the mass weighted CCs form and diagonalized by a unitary matrix C

$$M^{-1/2} H M^{-1/2} C = C\Lambda \tag{6}$$

The matrix M is diagonal and for each CC contains the mass m of the related atom. The columns of the C matrix are the linear combinations of the mass weighted CCs that correspond to the NCs displacements

$$\delta Q_K = \sum_{i=1}^{3N} \sqrt{m_i} C_{iK} \delta x_i \tag{7}$$

or in matrix form

$$\delta Q = \tilde{C} M^{1/2} \delta x \tag{8}$$

where δQ and δx are column vectors. In the case the geometry corresponds to an absolute or local energy minimum, $3N - 6$ eigenvalues Λ_K are positive and refer to vibrations, whereas the 3 translational and 3 rotational modes are identified by zero eigenvalues. In other cases negative eigenvalues can occur, and these do not correspond to vibrational modes. If all the NCs are retained, the transformation of eq 7 is fully invertible

$$\delta x = M^{-1/2} C \delta Q \tag{9}$$

The relation between the RICs and the NCs can be easily obtained exploiting the completeness of the CCs basis set. Using eqs 3 and 9

$$\delta q = B M^{-1/2} C \delta Q = T \delta Q \tag{10}$$

where the T matrix is defined as

$$T_{\mu K} = \left(\frac{\partial q_\mu}{\partial Q_K}\right) \tag{11}$$

Thus the RICs may be expressed in terms of the NCs, and the inclusion or not of the rotational and translational modes is uninfluential since they leave the RICs unchanged.

2.2. The Optimal Parameters of the Force Field. The best parameters for the QMD-FF in order to represent the internal molecular motion are obtained by minimizing the following merit function, written as a sum over the considered molecular geometries

$$I = \sum_{g=0}^{N_g} I_g \tag{12}$$

where

$$I_g = W_g[(E_g - E_0) - V_g]^2 + \sum_{K=1}^{3N-6} \frac{W'_{Kg}}{3N-6}[E'_{Kg} - V'_{Kg}]^2 + \sum_{K \le L}^{3N-6} \frac{2W''_{KLg}}{(3N-6)(3N-5)}[E''_{KLg} - V''_{KLg}]^2 \tag{13}$$

The indices K and L (capital letters) run over the normal coordinates and include all the modes except for the rotational and translational ones. E_g is the total energy obtained by a QM calculation, and E_0 is the same at the reference geometry ($g = 0$). E'_{Kg} (E''_{KLg}) is the energy gradient (Hessian) at a given geometry with respect to the NC evaluated at the same geometry. V, V', and V'' are the corresponding quantities calculated by the FF in eq 1. The constants W, W', and W'' weight the several terms at each geometry and can be chosen in order to drive the results depending on the circumstances. The energy, gradient, and Hessian terms are normalized in order to account for the different number of terms and to make the weights independent from the number of atoms in the molecule.

To compute the energy derivatives entering the merit function (13) we have to perform some transformations since no derivative is originally expressed with respect to the NCs. Indeed standard quantum chemistry programs provide derivatives E' and E'' with respect to CCs. Using the above relations and exploiting the completeness of the CCs, the transformation is simple

$$E'_K = \left(\frac{\partial E}{\partial Q_K}\right) = \sum_{i=1}^{3N}\left(\frac{\partial E}{\partial x_i}\right)\left(\frac{\partial x_i}{\partial Q_K}\right) = \sum_{i=1}^{3N} E'_i\, m_i^{-1/2}\, C_{iK} \quad (14)$$

or in matrix form

$$[E']_{\rm NC} = \tilde{C}M^{-1/2}[E']_{\rm CC} \quad (15)$$

where the square parentheses indicate column vectors of energy gradients computed with respect to the NCs and the CCs. The QMD-FF energy gradients at a given geometry

$$V'_K = \sum_{a=1}^{N_{\rm func}} p_a\left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{a=1}^{N_{\rm func}} p_a\, f'_{aK} \quad (16)$$

can be conveniently computed using the derivatives of the basis function with respect to the RICs, that is

$$\left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{\rm RIC}}\left(\frac{\partial f_a}{\partial q_\mu}\right)\left(\frac{\partial q_\mu}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{\rm RIC}}\sum_{i=1}^{3N}\left(\frac{\partial f_a}{\partial q_\mu}\right)T_{\mu K} \quad (17)$$

or in matrix form

$$[f'_a]_{\rm NC} = \tilde{T}[f'_a]_{\rm RIC} \quad (18)$$

The Hessian matrix of the QM calculation in NCs

$$E''_{KL} = \left(\frac{\partial^2 E}{\partial Q_K \partial Q_L}\right) \quad (19)$$

is obtained from the Hessian matrix in the CC basis according to

$$[E'']_{\rm NC} = \tilde{C}M^{-1/2}[E'']_{\rm CC}M^{-1/2}C \quad (20)$$

The second derivatives of the FF are a bit more complicated since they involve derivatives of the B matrix and are conveniently expressed in explicit form

$$\left(\frac{\partial^2 f_a}{\partial Q_K \partial Q_L}\right) = \sum_{\mu\nu=1}^{N_{\rm RIC}} T_{\mu K}\left(\frac{\partial^2 f_a}{\partial q_\mu \partial q_\nu}\right)T_{\nu L} + \sum_{\mu\nu=1}^{N_{\rm RIC}} T_{\mu K}\left(\frac{\partial f_a}{\partial q_\nu}\right)\left(\frac{\partial T_{\nu L}}{\partial q_\mu}\right) \quad (21)$$

As shown in eq 1, the QMD-FF is linear in the p parameters, thus the least-squares minimization of functional (13) can be written as

$$\sum_a^{N_{\rm func}}\sum_s^{N_{\rm point}} \alpha_{bs}W_s\alpha_{as}p_a = \sum_s^{N_{\rm point}} \alpha_{bs}W_s\beta_s \quad (22)$$

where the index s runs over the collections $[g]$, $[Kg]$, and $[KLg]$ defined in eq 13 for energy, gradient, and Hessian, respectively. Following this notation the matrix α and the vector β are defined as

$$\alpha_{as} = f_{as} \text{ or } f'_{as} \text{ or } f''_{as}; \quad \beta_s = E_s \text{ or } E'_s \text{ or } E''_s$$

and

$$W_s = W_s \text{ or } \frac{W'_s}{3N-6} \text{ or } \frac{W''_s}{(3N-6)(3N-5)}$$

where f's are the functions of eq 1, E, E', and E'' are the functions of the QM data, and W, W', and W'' are the weights of the merit function (13). Thus, defining

$$A = \alpha W\tilde{\alpha}$$

$$b = \alpha W\beta$$

one has to solve a standard linear equation in the form

$$Ap = b \quad (23)$$

where A is a symmetric matrix.

In usual FF it is convenient for practical purposes to employ functions of the RIC that will be in general redundant over the considered points. The scalar product between the FF functions is defined as

$$f_a \cdot f_b = \sum_{s=1}^{N_{\rm point}} W_s\ f_{as}\ f_{bs} \quad (24)$$

and the redundancy strongly depends on the number and type of points included in the fitting. However in general the f set might not be linearly independent. This leads to a singular A matrix, and the direct inversion method cannot be used to solve the linear system (23). On the contrary, the Singular Value Decomposition method[40,45] adapted to symmetric matrices is adequate and provides a stable solution of the linear system.

2.3. United Atom Theory. In many molecular simulations a group of atoms whose individual behavior is considered not to be crucial for the properties to be investigated can be grouped in a single interaction site. This approach, henceforth named United Atom (UA), allows saving computational time and simultaneously removes some high-frequency vibrational modes which can limit the integration time step in MD simulations. The most common example concerns aliphatic chains where each CH_2 group is treated as a single interaction site (C_2) with FF parameters accounting for the effect of the hydrogen atoms both in the nonbonded interactions and electrostatic charge. Despite recent work that has been done for some torsional potentials,[20] usually the intramolecular FF parameters of "hard" IC are not changed in the UA approach, thus the parameters driving the C_2-C_2-C_2 stretching and bending motion in the aliphatic chains are the same as those commonly employed in the FA description.

In the UA approximation the involved atoms are considered to move as a single point with the consequence that the translational movements with respect to the rest of the molecule can be somehow taken into account, but the relative rotational movements are irreparably lost. In other words a three-dimensional object described by 6 coordinates is transformed into a single point described by 3 coordinates. Even in the (nonrealistic) hypothesis that there exists some local vibrational modes much faster than those involving the atoms close to the UA, this approximation affects the motion of the neighboring atoms. Thus the remaining vibrational frequencies are altered by the UA approach, and it is convenient focusing on the representation of the intramolecular potential energy rather than on the vibrational analysis.

In this paper the UA atom approach, consistently with the previous FA approach, is treated on the basis of ab initio calculation of energies, gradients, and Hessian. The main problem is concerned with the transformation of the gradient vector and Hessian matrix in eq 13 in case the number of effective atoms is less than than the number of true atoms in the molecule. Let us consider for simplicity the case of a single UA in which N_{UA} atoms are grouped together. We use the indices μ and ν for the Cartesian coordinates referred to as the atoms involved in the UA and the indices a and b for those of the remaining atoms not involved in the UA (in this section we are forced to change the previous notation). For simplicity we suppose that only one atom in the UA group is linked to the unaltered atoms. The first-order energy expansion around a given geometry is

$$E^{(1)} = \sum_{a} \sum_{s}^{x,y,z} E'_{as}\delta t_{as} + \sum_{\mu} \sum_{s}^{x,y,z} E'_{\mu s}\, \delta t_{\mu s} \tag{25}$$

where t_{as} represents the sth component of the CC of the ath atom. The new gradient vector of the united atom U for a given geometry is transformed according to the simple expression

$$E'_{\mathrm{U}s} = \sum_{\mu} E'_{\mu s} \quad (s = x,y,z) \tag{26}$$

where E'_{U} represents the energy gradient with respect to the UA displacements. This expression is consistent with the hypothesis that the UA represents a set of internally frozen atoms: $\delta t_{\mathrm{U}s} = \delta t_{\mu s}$ ($\mu = 1...N_{\mathrm{UA}}$) and holds for simultaneous translations but not for rotations of the grouped atoms.

The second-order energy is

$$E^{(2)} = \frac{1}{2} \sum_{ab} \sum_{sr}^{x,y,z} E''_{as,br}\delta t_{as}\delta t_{br} + \frac{1}{2} \sum_{\mu\nu} \sum_{sr}^{x,y,z} E''_{\mu s,\nu r}\delta t_{\mu s}\delta t_{\nu r} + \sum_{a\mu} \sum_{sr}^{x,y,z} E''_{as,\mu r}\delta t_{as}\delta t_{\nu r} \tag{27}$$

Defining the UA Hessian matrix as

$$E''_{\mathrm{U}s,\mathrm{U}r} = \sum_{\mu\nu} E''_{\mu s,\nu r} \tag{28}$$

$$E''_{as,\mathrm{U}r} = \sum_{\mu} E''_{as,\mu r} \tag{29}$$

the energy expression becomes

$$E^{(2)} = \frac{1}{2} \sum_{ab} \sum_{sr} E''_{as,br}\delta t_{as}\delta t_{br} + \frac{1}{2} \sum_{sr} E''_{\mathrm{U}s,\mathrm{U}r}\delta t_{\mathrm{U}s}\delta t_{\mathrm{U}r} + \sum_{a} \sum_{sr} E''_{as,\mathrm{U}r}\delta t_{as}\delta t_{\mathrm{U}r} = \frac{1}{2}\, \tilde{\delta t} E'' \delta t \tag{30}$$

It is easy to verify that such a transformation of the Hessian matrix will preserve the three null eigenvalues due to translations, whereas the rotational modes of a molecule with UA included may lead to small (unphysical) energy contributions with the further undesirable consequence of small mixing between rotational and vibrational modes.

The two other quantities of the UA to be defined are the mass and the relative position. For the UAs considered in this paper (methylene and methyl groups) the mass was taken as the sum of the involved atoms. In the case where only one atom of the grouped atoms forms bonds with the rest of molecule, the natural choice for the position seems to make the UA coincident with that atom. However, other choices are possible; for example, the UA may be placed in the center of mass of the grouped atoms at the equilibrium geometry and/or its mass may be chosen in order to preserve the original inertia moments. Taking as criteria the magnitude of the rotational eigenvalues and the perturbation of the vibrational modes, these attempts do not lead to any improvement and were rejected. With the original choice the rotational eigenvalues at the equilibrium geometry are found to be much lower than the low-frequency vibrational modes, and the contamination is very small.

In summary the UA approach preserves some of the original atom−atom interactions contained in the Hessian matrix and leads to a useful simplification of the intramolecular energy hypersurface but does not allow conserving the rigorous implementation of the all-atom force field presented in this paper.

2.4. MD Model Force Field. The FF employed in MD simulations has the following expression: The intermolecular

$$E_{\mathrm{tot}} = E_{\mathrm{inter}} + E_{\mathrm{intra}} \tag{31}$$

part, E_{inter}, is computed as

$$E_{\mathrm{inter}} = E_{\mathrm{LJ}} + E_{\mathrm{Coul}} \tag{32}$$

where the long-range electrostatic term is

$$E_{\mathrm{Coul}} = \sum_{i=1}^{N_{\mathrm{sites}}} \sum_{j=1}^{N_{\mathrm{sites}}} \frac{q_i q_j}{r_{ij}} \tag{33}$$

and a Lennard-Jones term has been employed for the short-range part, i.e.

$$E_{\mathrm{LJ}} = \sum_{i=1}^{N_{\mathrm{sites}}} \sum_{j=1}^{N_{\mathrm{sites}}} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{34}$$

where i and j belong to different molecules, and N_{sites} is the total number of interacting sites. The intermolecular parameters q_{ij}, σ_{ij}, and ϵ_{ij} were taken for all molecules from the OPLS[15,16] literature force field.

The intramolecular part of the QMD-FF is expressed as a sum of different terms, namely

$$E_{\mathrm{intra}} = V(q) = E_{\mathrm{stretch}} + E_{\mathrm{bend}} + E_{\mathrm{Rtors}} + E_{\mathrm{Ftors}} + E_{\mathrm{Coupl}} \tag{35}$$

The first three terms count for the “hard” IC, i.e., bond stretchings, angle bendings, and stiff angle dihedrals (Rdihedrals), as those that drive the planarity of aromatic rings and are expressed with harmonic potentials:

$$E_{\mathrm{stretch}} = \frac{1}{2}\sum_{\mu}^{N_{\mathrm{bonds}}} k_{\mu}^{s}(r_{\mu} - r_{\mu}{}^{0})^{2} \tag{36}$$

$$E_{\mathrm{bend}} = \frac{1}{2}\sum_{\mu}^{N_{\mathrm{angles}}} k_{\mu}^{b}(\theta_{\mu} - \theta_{\mu}{}^{0})^{2} \tag{37}$$

$$E_{\mathrm{Rtors}} = \frac{1}{2}\sum_{\mu}^{N_{\mathrm{Rdihedrals}}} k_{\mu}^{t}(\phi_{\mu} - \phi_{\mu}{}^{0})^{2} \tag{38}$$

Conversely, the model functions employed for more flexible dihedrals (Fdihedrals) are the sums of periodic functions, namely

$$E_{\mathrm{Ftors}} = \sum_{\mu}^{N_{\mathrm{Fdihedrals}}} \sum_{j=1}^{N_{\cos_{\mu}}} k_{j\mu}^{d}[1 + \cos(n_{j}^{\mu}\phi_{\mu} - \gamma_{j}^{\mu})] \tag{39}$$

where $N_{\cos_{\mu}}$ is the number of cosine functions employed to describe the potential of the ϕ_{μ} dihedral. It is worth noticing that eqs 36−39 can be easily expressed in the formalism of eq 1 by setting $q_{\mu} = r_{\mu}$, θ_{μ}, and ϕ_{μ}, respectively. Following the notation introduced in eq 1, the last term of eq 35 can be written as

$$E_{\mathrm{Coupl}} = \sum_{i}^{N_{\mathrm{Coupl}}} V_{i}(q_{\mu}, q_{\nu}) \tag{40}$$

and may contain specific cross terms between the ICs q_{μ} and q_{ν}. The presence or absence of these off-diagonal coupling terms, which for instance may be of the form proposed in ref 29, discriminates between QMD-FFs of class II or class I. In this paper only couplings between soft dihedrals have been tested, which take the following expression

$$V_{i}(\phi_{\mu}, \phi_{\nu}) = \sum_{j=1}^{N_{\sin_{\mu}}{}^{i}} \sum_{k=1}^{N_{\sin_{\nu}}{}^{i}} k_{ijk}^{tc} \sin(n_{j}^{i}\phi_{\mu} - \gamma_{j}^{i}) \sin(m_{k}^{i}\phi_{\nu} - \gamma_{k}^{i}) \tag{41}$$

3. Computational Details

3.1. DFT Calculations. In all QM calculations the well tested density functional B3LYP method[46] with a correlation consistent basis set cc-pvDz was employed. For all tested molecules, the absolute energy minimum was obtained by a complete geometry optimization. Vibrational frequencies, gradients, and a Hessian matrix were computed from this optimized conformation. Torsional energy profiles for flexible molecules are obtained by performing calculations of the optimized energy without any restriction but the investigated “soft” IC, which was increased in a stepwise manner. All calculations were performed with the GAUSSIAN 03 package.[47]

3.2. Optimization of the FF Parameters. The program to compute the FF parameters through a fitting of the QM data was coded (in Fortran language) by the authors. It is coupled with the Gaussian 03 package for the input QM data, whereas the FF functions are read from a Moscito[48] input file. The output is again a Moscito input file which contains the optimized intramolecular input parameters. This program was named JOYCE in honor of the great Irish writer who spent many years in Italy. This program is free and can be obtained from the authors upon request.

3.3. Simulations. All MD simulations are carried out with a parallel version of the Moscito3.9[48] package. The equilibration runs are performed in the NPT ensemble on systems of 125 molecules at 298 K and 1 atm for at least 1 ns, keeping temperature and pressure constant using the weak coupling scheme of Berendsen et al.[49] The short-range intermolecular interactions are truncated at $R_c = 10$ Å, employing standard corrections for energy and virial.[1] Charge−charge long-range interactions are treated with the particle mesh Ewald (PME) method,[50,51] using a convergence parameter α of $5.36/2R_c$ and a fourth-order spline interpolation. In all cases the time step is set to 0.1 fs, since bond stretching is explicitly considered. An exception is made for the butoxybenzene molecule, where the bond lengths are kept fixed at their equilibrium value using the SHAKE algorithm[52] allowing a time step of 1 fs. After equilibration, NVE trajectories of 100 ps were produced in the NVE ensemble and used for the calculation of the velocity autocorrelation function

$$Z(t) = \langle v(t) \cdot v(0) \rangle \tag{42}$$

The latter is employed to obtain the power spectra as

$$P(\omega) = \frac{6}{\pi Z(0)} \int_{0}^{\infty} Z(t)\cos(\omega t)\mathrm{d}t \tag{43}$$

4. Results

4.1. Rigid Molecules. The first test was performed on four heterocyclic aromatic molecules, namely pyridine, furan, oxazole, and isoxazole. The Hessian matrix of each molecule, computed after DFT complete geometry optimization, was used to parametrize the intramolecular QMD-FF according to eq 13. Two different weighting factors W''_{KL} of 10^4 and 0.8×10^4 were used for the diagonal ($K = L$) and off-diagonal ($K \neq L$), respectively. Since all chosen molecules show rather stiff ICs we employed only the harmonic terms (36−38) in eq 35. It may be worth noticing that with this choice only one QM calculation (a complete optimization with frequencies) is needed to construct the QM database necessary for the intramolecular parametrization. A FA description was adopted as shown in Figure 1, and no restriction was imposed on the fitting parameters except those dictated by symmetry. All parameters were obtained with a standard deviation of $1.2-1.4 \times 10^{-2}$ kJ/mol and are reported in Tables 1−3, for stretching, bending, and torsions, respectively. For comparison, AMBER[14,17,53,54] parameters are also reported in the same tables.

By looking at Table 1, one can see that QMD and literature[53,54] stretching constants k^s are rather similar. Exceptions are those for the C−N bond in pyridine and the C−C in furan, which are found smaller by 25−30% in the QMD-FF.

For the bending constants k^b, reported in Table 2, the QMD parameters appear to retain a higher level of chemical specificity. Indeed, the QMD values indicate a marked difference between different bending motions, as for instance those regarding the C_N-N-C_N ($k^b = 569$ kJ/mol rad^{-2}) and the $N-C_N-C$ ($k^b = 869$ kJ/mol rad^{-2}) triplets in pyridine, which is not accounted for in the literature FF. Also in the

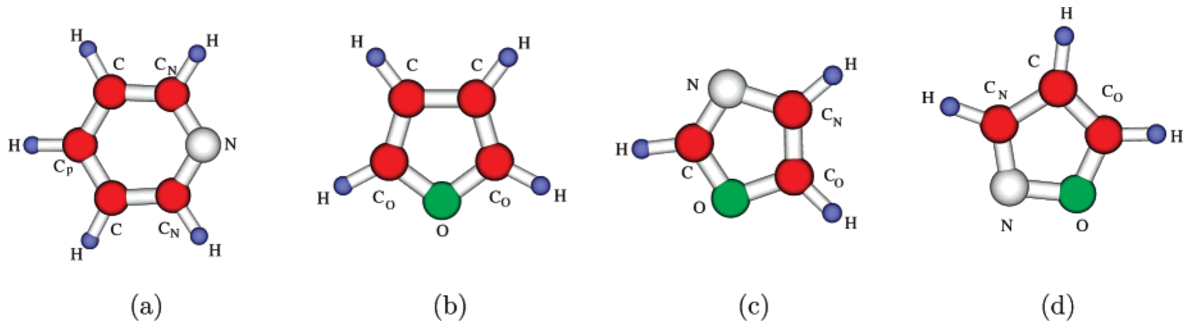


Figure 1. Test molecules in the FA models: (a) pyridine, (b) furan, (c) oxazole, and (d) isoxazole.

Table 1. QMD Fitted and AMBER[53,54] Parameters for Bond Stretching Potential of Eq 36[a]

pyridine				furan			
	k^s (kJ/mol Å^{-2})				k^s (kJ/mol Å^{-2})		
IC	QMD	AMBER	r_0 (Å)	IC	QMD	AMBER	r_0 (Å)
C_N−N	3036	4041	1.34	C_O−C	4133	4569	1.36
C_N−C	3132	3924	1.40	C−C	2793	3924	1.44
C−C $_p$	3398	3924	1.40	C_O−O	2921	2845	1.36
C*−H	3301	3071	1.09	C*−H	3497	3071	1.09

oxazole				isoxazole			
	k^s (kJ/mol Å^{-2})				k^s (kJ/mol Å^{-2})		
IC	QMD	AMBER	r_0 (Å)	IC	QMD	AMBER	r_0 (Å)
O−C_O	2836	2845	1.37	O−C_O	3147	2845	1.34
C_O−C_N	4067	4351	1.36	C_O−C	4090	4569	1.36
C_N−N	2904	3430	1.39	C−C_N	2789	3923	1.43
N−C	4433	4083	1.30	C_N−N	4028	3431	1.31
O−C	2535	3866	1.36	O−N	1884	3866	1.39
C*−H	3515	3071	1.09	C*−H	3485	3071	1.09

[a] Symbols refer to Figure 1; C* indicates every carbon atom in the ring. All the C*−H stretchings were constrained to the same value, since they resulted spontaneously very similar from the fitting procedure.

Table 2. QMD Fitted and AMBER[53,54] Parameters for Angle Bending Potential of Eq 37[a]

pyridine				furan			
	k^b (kJ/mol rad^{-2})		θ_0		k^b (kJ/mol rad^{-2})		θ_0
IC	QMD	AMBER	(deg)	IC	QMD	AMBER	(deg)
N−C_N−C	869	586	124	C_O−C−C	708	586	106
C−C_p−C	637	527	120	C−C_O−O	641	586	111
C_N−N−C_N	569	586	117	C_O−O−C_O	796	586	107
C_p−C−C_N	596	527	120	-	-	-	-
N−C_N−H	392	293	116	O−C_O−H	324	293	116
C*−C *−H	314	293	120	C*−C*−H	240	293	127

oxazole				isoxazole			
	k^b (kJ/mol rad^{-2})		θ_0		k^b (kJ/mol rad^{-2})		θ_0
IC	QMD	AMBER	(deg)	IC	QMD	AMBER	(deg)
O−C_O−C_N	767	586	108	O−C_O−C	582	586	110
C_O−C_N−N	838	586	109	C_O−C−C_N	731	586	103
C_N−N−C	729	586	104	C−C_N−N	762	586	113
O−C−N	889	586	115	O−N−C_N	873	586	105
C_O−O−C	846	586	104	C_O−O−N	970	586	109
O−C_O−H	386	293	116	N−C_N−H	308	293	118
N−C−H	234	293	129	O−C_O−H	321	293	116
O−C−H	345	293	117	C−C_O−H	254	293	134
N−C_N−H	395	293	122	C_O−C−H	254	293	129
C_O−C_N−H	151	293	129	C_N−C−H	254	293	128
C_N−C_O−H	151	293	135	C−C_N−H	254	293	119

[a] Symbols refer to Figure 1.

azoles, the bending constants involving hydrogen-containing triplets range from 151 kJ/mol rad^{-2} (C_O−C_N−H and C_N−C_O−H in oxazole) to 395 kJ/mol rad^{-2} (N−C_N−H in oxazole), while the other triplets show force constants from 582 kJ/mol rad^{-2} (O−C_O−C in isoxazole) to 970 kJ/mol rad^{-2} (C_O−O−N in isoxazole). Conversely literature FF only accounts for two types of constants, 293 and 586 kJ/mol rad^{-2}, depending on whether the triplet contains a hydrogen atom or not.

A similar comparison cannot be easily made with the torsion constants reported in Table 3, because different functional forms are used to describe “hard” dihedral motions. In particular the AMBER FF[14] does not distinguish between “hard” and “soft” dihedrals, employing sums of cosine functions even for the former type. However the QMD model should result in a lower tendency of the aromatic rings to lose planarity, since literature torsional constants never exceed 15 kJ/mol, and the employed sinusoidal functions[14] are much smoother than the harmonic model obtained in the present work and reported in Table 3.

According to eq 31, the intramolecular potentials described by the QMD and AMBER parameters were complemented with an intermolecular part, whose parameters were taken from an OPLS-FA description, explicitly designed to reproduce, through MC simulations in the liquid phase, some thermodynamic properties (as density or heat of vaporization) of the target molecules.[53,54] The resulting FFs, which will again be labeled QMD and AMBER for simplicity, were employed in MD simulations, according to the details given in the previous section. Two sets of MD simulations for each molecule were performed at 298 K and 1 atm; all systems were equilibrated for 1 ns, and thermodynamic averages were taken on production runs of a further ns.

From the average thermodynamic quantities reported in Table 4, it appears that the proposed QMD-FF well couples with the OPLS intermolecular parameters, since it does not alter the liquid density nor the energy distribution by more than 3%.

Moreover Figure 2, where the radial distribution functions $g(r)$ are reported for the oxygen−oxygen pair, shows that the liquid structure resulting from the QMD model is in excellent agreement with the AMBER results, retaining some

Table 3. Fitted Parameters for Dihedral Angles as in Eq 38[a]

pyridine			furan		
IC	k^t (kJ/mol rad^{-2})	ϕ_0	IC	k^t (kJ/mol rad^{-2})	ϕ_0
$N-C_N-C-C_p$	94	0	$C_O-C-C-C_O$	125	0
C_N-N-C_N-C	99	0	$C-C_O-O-C_O$	151	0
$C-C_p-C-C_N$	76	0	$O-C_O-C-C$	220	0
$H-C-C_N-N$	63	180	$O-C_O-C-H$	82	180
$H-C_N-C-C_p$	77	180	$H-C_O-O-C_O$	55	180
C_N-N-C_N-H	124	180	$H-C-C-C_O$	54	180
$H-C-C_N-H$	44	0	$H-C-C-H$	11	0
$H-C_p-C-H$	40	0	-	-	-

oxazole			isoxazole		
IC	k^t (kJ/mol rad^{-2})	ϕ_0	IC	k^t (kJ/mol rad^{-2})	ϕ_0
$C-O-C_O-C_N$	174	0	$N-O-C_O-C$	172	0
$O-C_O-C_N-N$	197	0	$O-C_O-C-C_N$	166	0
$H-C_O-C_N-N$	57	180	C_O-C-C_N-N	194	0
C_O-C_N-N-C	187	0	$C-C_N-N-O$	188	0
$C_N-N-C-O$	198	0	$C_O-O-N-C_N$	164	0
$O-C_O-C_N-H$	97	180	$O-C_O-C-H$	73	180
$C_O-O-C-N$	199	0	$H-C-C_N-N$	52	180
$H-C_N-N-C$	57	180	$H-C_N-N-O$	124	180
$C-O-C_O-H$	51	180	$N-O-C_O-H$	80	180
$C_N-N-C-H$	102	180	C_O-C-C_N-H	63	180
$H-C_O-C_N-H$	9	0	$H-C^*-C^*-H$	10	0
$C_O-O-C-H$	56	180	-	-	-

[a] Symbols refer to Figure 1; C* refers to any aromatic carbon atom.

Table 4. Thermodynamic Results Obtained for Rigid Molecules at 1 atm and 298 K Employing Both QMD and AMBER FFs

	QMD			AMBER		
molecule	density (g/cm ^{3})	E_{inter} (kJ/mol)	E_{intra} (kJ/mol)	density (g/cm ^{3})	E_{inter} (kJ/mol)	E_{intra} (kJ/mol)
pyridine	0.974 ± 0.006	−39.5	33.7	0.970 ± 0.006	−39.3	35.4
furan	0.970 ± 0.010	−27.7	26.7	0.944 ± 0.009	−26.6	26.8
oxazole	1.165 ± 0.010	−45.0	26.0	1.130 ± 0.008	−44.4	25.9
isoxazole	1.130 ± 0.006	−42.2	23.1	1.098 ± 0.008	−40.8	23.6

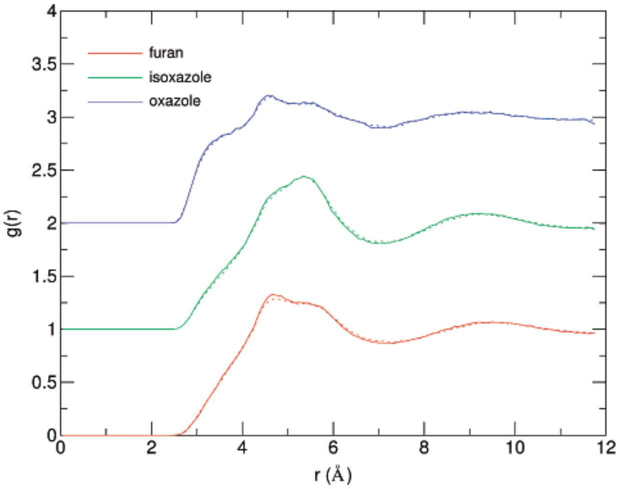


Figure 2. Oxygen−oxygen pair atomic correlation functions $g(r)$, computed for QMD-FF (solid lines) and for AMBER-FF (dashed lines). The isoxazole and oxazole curves are vertically shifted by 1 and 2, respectively.

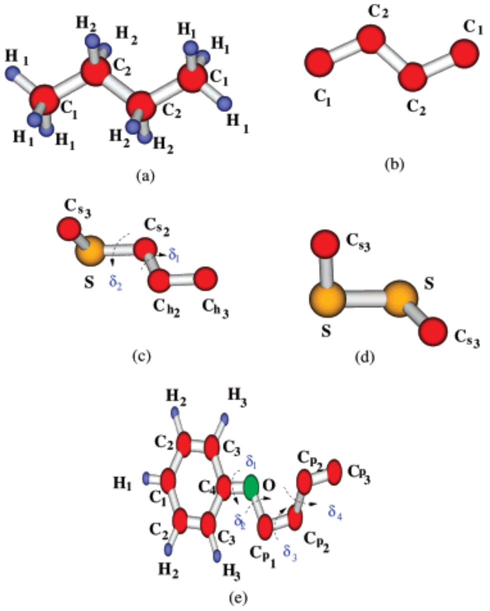


Figure 3. Test flexible molecules: (a) FA butane, (b) UA butane, (c) UA methyl propyl sulfide, (d) UA dimethyl disulfide, and (e) butoxybenzene.

small differences between different azoles. It is also worth noticing that the computed functions are also in agreement with those reported in ref 54, where a OPLS/AMBER FF was used in MC simulations without including deformations of planarity and therefore involving no torsional energy term. Similar results are found with the several $g(r)$ functions of pyridine. Concerning with the internal structure, the QMD and AMBER distance and angle distributions are very similar both in position and shape. Small differences are instead found in the shape of the dihedral distributions, being that the QMD bands are more sharp and localized. This indicates an increased stiffness of the aromatic rings, in agreement with the higher value of the QMD torsional parameters reported in Table 3.

The good agreement in the considered structural and thermodynamic properties suggests that the representation of the molecular structure given by the atomistic QMD model is quite correct. This encouraged us to extend the present approach to UA parametrizations of larger, flexible molecules, which is the main scope of the present paper.

4.2. Flexible Molecules and UA Description. Once the reliability of the QMD description of the molecular structure has been validated on rigid target molecules, the capability of the proposed method to describe the motion of “soft” ICs was tested on several molecules, namely *n*-butane, methylpropyl sulfide, dimethyl disulfide, and *n*-butoxybenzene. The adopted models for these target molecules are reported in Figure 3. Panels (a) and (b) show the two different models (FA and UA, respectively) employed for *n*-butane, panels (c) and (d) report the UA models for methyl-propyl sulfide and dimethyl disulfide, while in panel (e) the “hybrid” model (FA for aromatic hydrogens and UA for the aliphatic lateral chain) was employed for *n*-butoxybenzene.

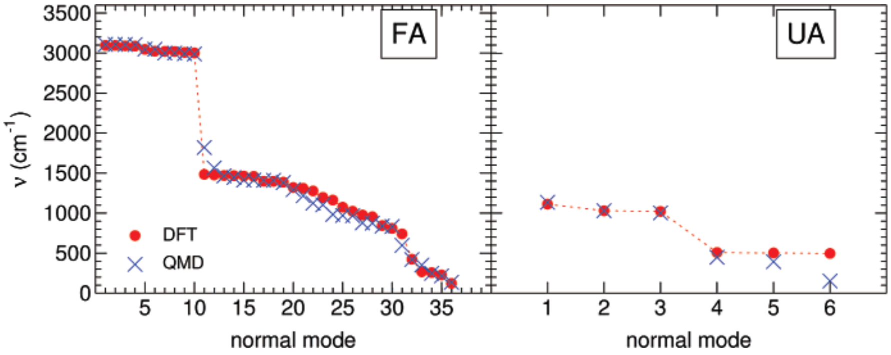


Figure 4. Normal modes frequencies computed by (red full circles) and predicted by the QMD-FF (blue crosses) for the butane molecule in the FA (left panel) and UA (right panel) model.

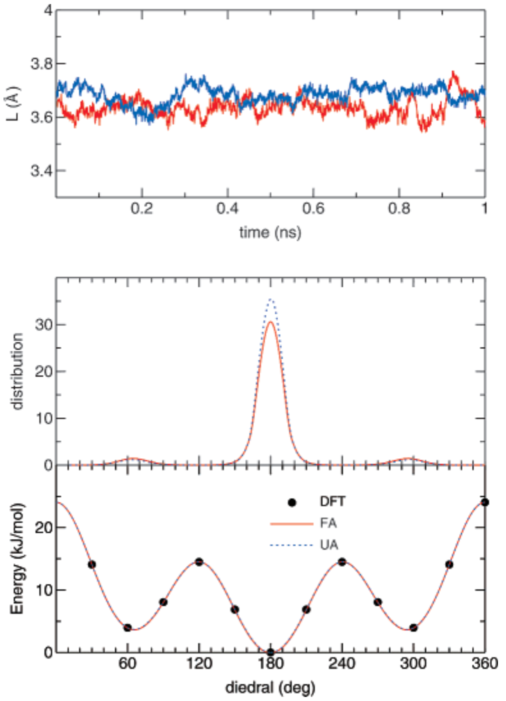


Figure 5. $C_1-C_2-C_2-C_1$ butane's dihedral in FA and UA models. Bottom panel: dihedral energy profile computed with DFT (black full circles) and QMD FA (red solid line) and UA (blue dashed line) potentials. Middle panel: dihedral distributions at 298 K and 1 atm computed in 1 ns FA (red) and UA (blue) models. Upper panel: butane elongation ($L = C_1-C_1$ distance) vs simulation time. Red and blue lines again refer to FA and UA models, respectively.

The energy profile exhibited by the "soft" degrees of freedom is rather flat and exhibits several local minima (see bottom panel of Figure 5) so that strongly distorted conformations may be populated, even at room temperature. This causes the equilibrium geometry and the related vibrational frequencies not to furnish sufficient data to parametrize the torsional potential, and several energy calculations at different geometries are needed to take into account large amplitude motions. Nevertheless, by imposing W' and W'' in eq 13 to be zero for each geometry different from the equilibrium one, the computational cost of the QM calculations is still reduced, since no vibrational frequency is needed except those in the minimum energy configuration. The inclusion of distorted geometries obtained by sampling the intramolecular energy surface imposing different values of the investigated "soft" dihedral and relaxing the other ones presents however some problems which may be clarified through a simple example.

The *n*-butane molecule in the UA approach (Figure 4, panel (b)) has six ICs: three bond distances, two bond angles, and one dihedral. The latter has to be considered as a "soft" IC. Let us suppose that the fitting includes two conformations, namely the staggered equilibrium ($C_1-C_2-C_2-C_1 = 180°$) and eclipsed ($C_1-C_2-C_2-C_1 = 0°$) whose relative energy E(eclip) $-$ E(stagg) is 24 kJ/mol. The central C_2-C_2 distance is different for the two conformations 2.90 Å and 2.95 Å, respectively, for the staggered and eclipsed, whereas the two bending angles change by about 3°. Despite the dihedral angle being by far the most evident geometrical change on going from staggered to eclipsed conformation, relevant energy contributions occur even for the small changes of the other ICs: bond lengths and angles account for 3.4 and 2.9 kJ/mol, respectively. Consequently the torsional energy term of eq 39 accounts for about 75% of the relative energy E(eclip) $-$ E(stagg). Therefore the resulting pure torsional potential (eq 39) describes a lower barrier (18 rather than 24 kJ/mol), being the remaining gap accounted for the energy terms of the bond distances and angles.

This (rather obvious) finding has the unpleasant consequence that a good description of the large amplitude torsional geometrical movements cannot be achieved with high accuracy by simple FFs. Indeed, by using a class I FF (i.e., no coupling term), the fraction of the torsional energy connected with the changes of the other IC is completely lost, because there is no reason for the bond lengths and angles to change during the internal rotation (frozen rotation). In fact the information linking the dihedral to the other ICs in the QM calculation is completely lost, since in central FFs the motion of one IC is independent from the value of the other ICs. The straightforward remedy for this problem would require the inclusion of a relevant number of coupling functions in eq 40, as done for example in the QMFF procedure,[29] with the consequence of increasing the number of functions in the FF. A more simple and direct solution is to ignore the changes of most of the ICs not directly involved in the internal rotation and, in case, retaining the changes of few pertinent ICs whose coupling terms with the dihedral are included in the FF. This route has the effect of ascribing

Table 5. Fitted Parameters for the Bond Stretching Potential of Eq 36 Computed for FA and UA Butane, Employing the FIRA[a]

model	IC	k^s (kJ/mol Å^{-2})	r_0 (Å)
FA	C_1-C_2	2249	1.53
UA	C_1-C_2	2594	1.53
FA	C_2-C_2	2146	1.53
UA	C_2-C_2	2353	1.53
FA	C_1-H_1	3101	1.10
FA	C_2-H_2	2996	1.11

[a] Symbols refer to panels (a) and (b) of Figure 4.

Table 6. Fitted Parameters for the Angle Bending Potential of Eq 37, Computed for FA and UA Models of Butane with the FIRA[a]

model	IC	k^b (kJ/mol rad^{-2})	θ_0 (deg)
FA	$C_1-C_2-C_2$	743	113
UA	$C_1-C_2-C_2$	927	113
FA	$C_2-C_1-H_1$	342	112
FA	$C_1-C_2-H_2$	401	110
FA	$C_2-C_2-H_2$	396	109
FA	$H^*-C^*-H^*$	334	106

[a] Symbols refer to panels (a) and (b) of Figure 4; C* and H* indicate every carbon and hydrogen atom in the ring, respectively.

Table 7. Fitted Parameters for Dihedral Angles of FA and UA Models of Butane as Defined in Eq 39 in the FIRA[a]

FA									UA		
$C_1-C_2-C_2-C_1$			$^*-C_2-C_2-H_2$			$H_1-C_1-C_2-^*$			$C_1-C_2-C_2-C_1$		
k^d (kJ/mol)	n	γ	k^d (kJ/mol)	n	γ	k^d (kJ/mol)	n	γ	k^d (kJ/mol)	n	γ
−14.132	0	0	-	0	0	-	0	0	−2.106	0	0
4.349	1	0	-	1	0	-	1	0	4.330	1	0
1.737	2	0	-	2	0	-	2	0	1.738	2	0
0.847	3	0	0.847	3	0	0.712	3	0	7.520	3	0
1.629	4	0	1.629	4	0	-	4	0	0.126	4	0
1.052	5	0	1.052	5	0	-	5	0	0.172	5	0
0.029	6	0	0.029	6	0	-	6	0	0.241	6	0

[a] Symbols refer to panels (a) and (b) of Figure 4; C*, H*, and * indicate any carbon, hydrogen, and carbon or hydrogen atoms, respectively.

the torsional energy to the torsional term (39) only, whereas in the QM calculation it is distributed on several ICs since all the ICs are in principle coupled to each other. This method, which has been implicitly adopted in partial parametrization of flexible molecules[24,35,38] will be called FIRA: frozen internal rotation approximation.

4.3. The UA Prototype: *n*-Butane. The parameters for both FA and UA descriptions of the *n*-butane molecule (reported in Tables 5−7) were obtained through eq 13 by choosing $W = 1$ for each energy, $W' = 0$ for each gradient, and $W''_{KL} = 1 \times 10^4$ and 0.5×10^4 for the diagonal ($K = L$) and off-diagonal ($K \neq L$) terms of the Hessian matrix in the equilibrium geometry. For all distorted geometries W'' was set to be null.

The torsional potentials around the C_1-C_2 and C_2-C_2 bonds were sampled with 7 points in the 0−180° range, but obviously only the latter was used to obtain the parameters in the UA description. By comparing the stretching and bending constants of the two models, reported in Tables 5 and 6, it appears that the increased size of the UA interaction sites shifts the k's to higher values. Conversely, the torsion around the C_2-C_2 bond, reported in Table 7, is described by almost the same FA and UA parameters, with the only exception of the $n = 0$ constant term, which does not alter the shape of the potential curve.

The comparison between DFT and QMD vibrational frequencies, reported in Figure 4, clearly shows that the proposed procedure is capable of removing the high frequencies from the FA description (i.e., those due to C−H stretching and bendings) and reproducing with good approximation those remaining in the UA molecule. An exception is made by the last UA normal mode, which essentially involves the low frequency torsion of the $C_1-C_2-C_2-C_1$ dihedral. However the energy profile of such torsion, reported in the bottom panel of Figure 5, is well represented by the two models, making us confident that the QMD potentials will be able to reproduce the correct population distribution for *n*-butane.

Simulations were carried out for both FA and UA models at 298 K and 1 atm, equilibrating the systems for almost 2 ns. Intermolecular interactions were modeled with the OPLS-AA[15,55] and UA[56] FF parameters for *n*-butane. The resulting average densities of 0.555 g/cm^3 and 0.584 g/cm^3, obtained for FA and UA models respectively, well agree with both recent MC simulation results (0.558 g/cm^3) and the experimental value of 0.5729 g/cm^3 (see ref 55 and references therein); the radial distribution functions, computed for both models, did not show any marked difference. The heats of vaporization (ΔH_{vap}), computed as suggested in refs 31, 55, and 56, are in good agreement with both OPLS-AA (5.00 kcal/mol) and experimental values (5.04 kcal/mol, see ref 55 and references therein), being 5.02 kcal/mol and 5.09 kcal/mol for QMD-FA and QMD-UA, respectively. The distribution of the $C_1-C_2-C_2-C_1$ dihedral and the time evolution of the end-to-end chain elongation L, reported in Figure 5, confirm the capability of the QMD-UA FF to account for the correct molecular structure for an aliphatic flexible chain. Indeed, the average L values of 3.64 Å and 3.68 Å, obtained for the FA and UA models, are very similar to recent MC results (3.67 Å), reported for liquid *n*-butane at the same temperature.

4.4. Two Sulfur Containing Molecules. The next two molecules considered to test the UA approximation are methyl-propyl sulfide (MPS) and dimethyl disulfide (DMDS), respectively, reported on panels (c) and (d) of Figure 3. The UA approximation implies that all methylene and methyl groups are treated as a single site coincident with the involved carbon atom. The parameters for both these molecules were obtained through the fitting by applying the same procedure and weights of the *n*-butane molecule: the resulting values are reported in Tables 8 and 9.

In the bottom panels of Figure 6 (a),(b), the resulting torsional potentials are compared to the computed energy data, for MPS and DMDS, respectively. It appears that the adoption of the UA model, in the FIRA approximation, does not alter the main features of the QM torsional curves. In particular the 90° minimum for the DMDS dihedral is well reproduced, and the small differences between the potential

Table 8. Fitted Parameters for Bond Stretching (Eq 36) and for Angle Bending (Eq 37) for UA Methyl-Propyl Sulfide (MPS) and Dimethyl Disulfide (DMDS), Employing the FIRA[a]

MPS			DMDS		
IC	k^s (kJ/mol Å^{-2})	r_0 (Å)	IC	k^s (kJ/mol Å^{-2})	r_0 (Å)
$C_{h3}-C_{h2}$	2476	1.53	$C_{s3}-S$	1697	1.84
$C_{h2}-C_{s2}$	2404	1.53	$S-S$	1217	2.09
$C_{s2}-S$	1424	1.84	-	-	-
$S-C_{s3}$	1728	1.83	-	-	-

MPS			DMDS		
IC	k^b (kJ/mol rad^{-2})	θ_0 (deg)	IC	k^b (kJ/mol rad^{-2})	θ_0 (deg)
$C_{h3}-C_{h2}-C_{s2}$	1013	112	$C_{s3}-S-S$	1086	103
$C_{h2}-C_{s2}-S$	724	115	-	-	-
$C_{s2}-S-C_{s3}$	2290	101	-	-	-

[a] Symbols refer to panels (c) and (d) of Figure 3.

energy curves with respect to dihedrals δ_1 and δ_2 of MPS are also retained.

MD simulations were performed by adding the OPLS intermolecular parameters designed for liquid sulfur compounds.[57] In this way the QMD and OPLS FFs differ for the intramolecular part, and a comparison of the results may give an indication as to the quality of our QMD FF. All simulations were carried out at 298 K and 1 atm, equilibrating both MPS and DMS systems for more than 2 ns.

In Figure 6 the resulting dihedral distributions are reported in the top of panel (a) and (b) for MPS and DMDS, respectively. It is worth noticing that both curves agree well with those reported in ref 57, where OPLS parameters were designed, confirming that the QMD UA approach can provide a correct sampling of the molecular configurations assumed in the condensed phase. Turning to the thermodynamic properties, it is apparent from Table 10 that the QMD results are in very good accord with both the OPLS and the experimental ones. It is fair to stress that the accuracy of the heats of vaporizations must be mainly ascribed to the

Table 9. Fitted Parameters for Dihedral Torsion Potential of Eq 39 Computed for UA Methyl-Propyl Sulfide (MPS) and Dimethyl Disulfide (DMDS), Employing the FIRA[a]

	MPS		DMDS
n	δ_1 k^d (kJ/mol)	δ_2 k^d (kJ/mol)	$C_{s3}-S-S-C_{s3}$ k^d (kJ/mol)
0	−2.899	−2.899	−5.069
1	4.070	3.604	6.010
2	2.404	3.263	16.805
3	7.462	4.194	3.219
4	0.017	0.012	0.342
5	0.036	0.096	0.263
6	0.181	0.130	0.045

[a] Symbols refer to panels (c) and (d) of Figure 3.

Table 10. Thermodynamic Properties of MPS and DMDS[a]

	MPS			DMDS		
	QMD-UA	OPLS-UA	exp	QMD-UA	OPLS-UA	exp
ρ (g/cm^3)	0.822	0.796	0.837	1.009	1.031	1.057
ΔH_{vap} (kJ/mol)	29.3	26.7	31.8	38.0	38.8	38.4

[a] OPLS and experimental values are reported in ref 57.

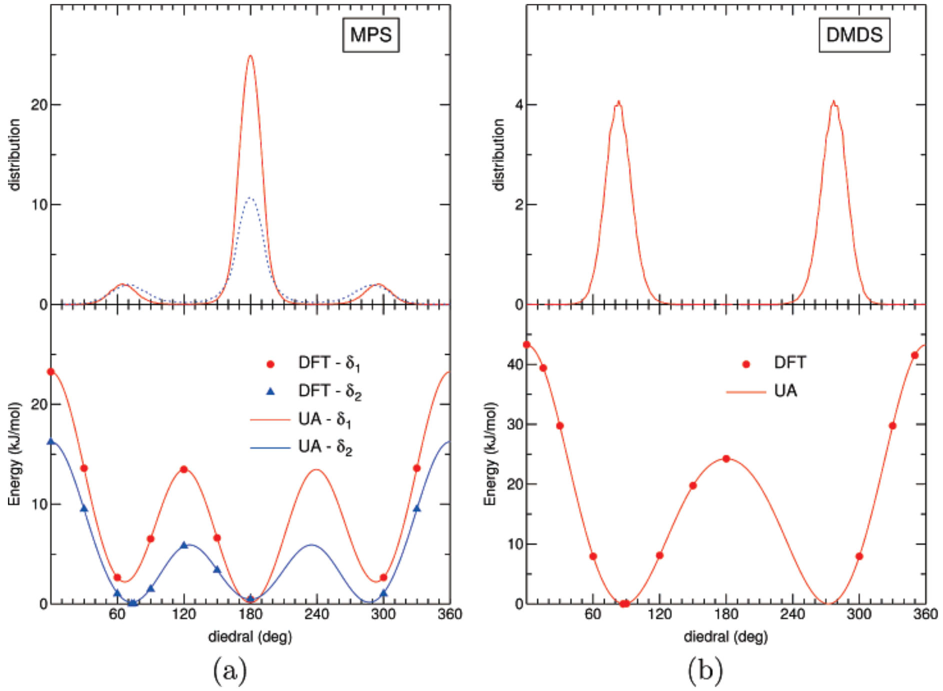


Figure 6. MPS and DMDS dihedrals. Bottom panels: (a) MPS dihedral energy profile computed with DFT (circles) and QMD UA potentials (solid line) for dihedrals δ_1 (red) and δ_2 (blue) of Figure 3 and (b) DMDS dihedral energy profile computed with DFT (circles) and QMD UA potentials (solid line). Top panels: (a) dihedral distributions at 298 K and 1 atm computed in 1 ns for dihedrals δ_1 (red line) and δ_2 (blue line) of MPS and (b) dihedral distributions at 298 K and 1 atm computed in 1 ns for DMDS dihedral.

Table 11. Fitted Parameters for Angle Bending Potential of Eq 37 Obtained for Both Coupled and Uncoupled Models of Butoxybenzene[a]

IC	k^b (kJ/mol rad^{-2})	θ_0 (deg)
C*−C*−C*	641	120
C*−C*−H*	320	120
C_3−C_4−O	786	120
C_4−O−C_{p1}	315	119
O−C_{p1}−C_{p2}	1548	108
C_{p1}−C_{p1}−C_{p2}	1071	114
C_{p2}−C_{p2}−C_{p3}	969	113

[a] Symbols refer to panel (e) of Figure 3; C* and H* stand for any aromatic carbon or hydrogen, respectively.

Table 12. Fitted Parameters for Dihedral Harmonic Torsional Potential of Eq 38 Obtained for Both for Coupled and Uncoupled Models of Butoxybenzene[a]

IC	k^t (kJ/mol rad $^{-2}$)	ϕ_0 (deg)
C*−C*−C*−C*	82	0
C*−C*−C*−H*	66	180
H*−C*−C*−H*	40	0
H*−C*−C*−O	43	0
C*−C*−C*−O	159	180

[a] Symbols refer to panel (e) of Figure 3; symbols C* and H* stand for any aromatic carbon or hydrogen, respectively.

Table 13. Fitted Parameters for Dihedral Angles of Butoxybenzene of Butane as Defined in Eq 39

δ_1 (C_3−C_4−O−C_{p1})				δ_2 (C_4−O−C_{p1}−C_{p2})			
n	γ	k^d (kJ/mol) uncoupled model	k^d (kJ/mol) coupled model	n	γ	k^d (kJ/mol) uncoupled model	k^d (kJ/mol) coupled model
0	0	0.557	0.794	0	0	0.279	0.397
2	0	−3.165	−2.924	1	0	12.387	8.679
4	0	−0.344	−0.543	2	180	−5.271	−2.587
6	0	0.275	0.120	3	0	6.735	4.356
-	-	-	-	4	180	0.299	0.356
-	-	-	-	5	0	0.045	0.008
-	-	-	-	6	180	−0.674	−0.388

δ_3 (O−C_{p1}−C_{p2}−C_{p2})				δ_4 (C_{p1}−C_{p2}−C_{p2}−C_{p3})			
n	γ	k^d (kJ/mol) uncoupled model	k^d (kJ/mol) coupled model	n	γ	k^d (kJ/mol) uncoupled model	k^d (kJ/mol) coupled model
0	0	0.279	0.397	0	0	0.279	0.397
1	0	1.889	1.970	1	0	3.676	3.641
2	0	2.374	2.359	2	0	1.855	1.609
3	0	7.882	7.785	3	0	7.035	6.775
4	0	0.346	0.470	4	0	0.198	0.119
5	0	−0.139	0.066	5	0	0.605	0.396
6	0	0.113	−0.378	6	0	0.456	0.386

Table 14. Fitted Parameters for Dihedral Couplings in Butoxybenzene Coupled Model as Defined in Eq 41

IC 1	IC 2	n	γ	m	α	k^{tc} (kJ/mol)
δ_1	δ_2	2	0	1	0	−1.1613
δ_1	δ_3	2	0	1	0	−0.1205
δ_2	δ_3	1	0	1	0	−0.9943

excellent capability of the intermolecular OPLS parameters to reproduce this quantity.

4.5. *n*-Butoxybenzene. The last target molecule *n*-butoxybenzene contains a rigid aromatic ring linked to a flexible aliphatic chain, a chemical situation frequently found in many liquid crystals and molecules of biological interest. In the adopted model, all aromatic hydrogen atoms were taken explicitly into account, while the methyl and methylene groups of the chain were modeled in a UA description, as shown in panel (e) of Figure 3. As done for butane, torsional energy profiles of each of the four dihedrals (δ_1−δ_4) were sampled with 8 points in the 0−180° range. All these calculations, together with the optimized geometry and frequencies, were used in the fitting procedure, with the same weighting factors reported for butane. The resulting parameters are reported in Tables 11−14. Since in preliminary fittings it was noted that the force constants of the "hard" ICs did not change significantly in going from the uncoupled to the coupled model, for the sake of simplicity all k^b and k^t of the coupled model were constrained to their uncoupled value during the fitting.

By looking at Tables 11 and 12, one can see as the bending and the "hard" torsion constants assume a broad range of values: 315−1548 and 40−159 J/mol rad^{-2}, respectively. This is not obvious since the atom triplets or quartets may be very similar at first sight. This may be regarded as further proof of QMD's capability to capture subtle different behavior of the ICs involving similar atom types. It is also important to point out that similar ICs belonging to different molecules result in similar parameters without

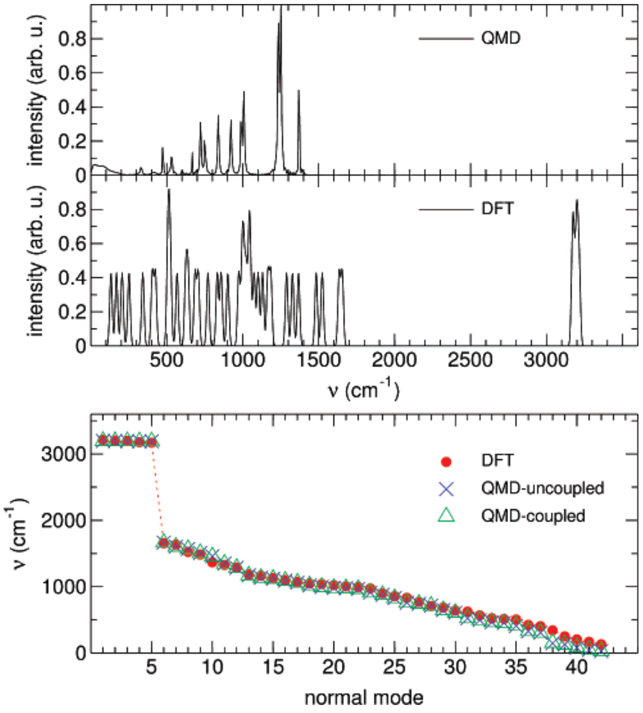
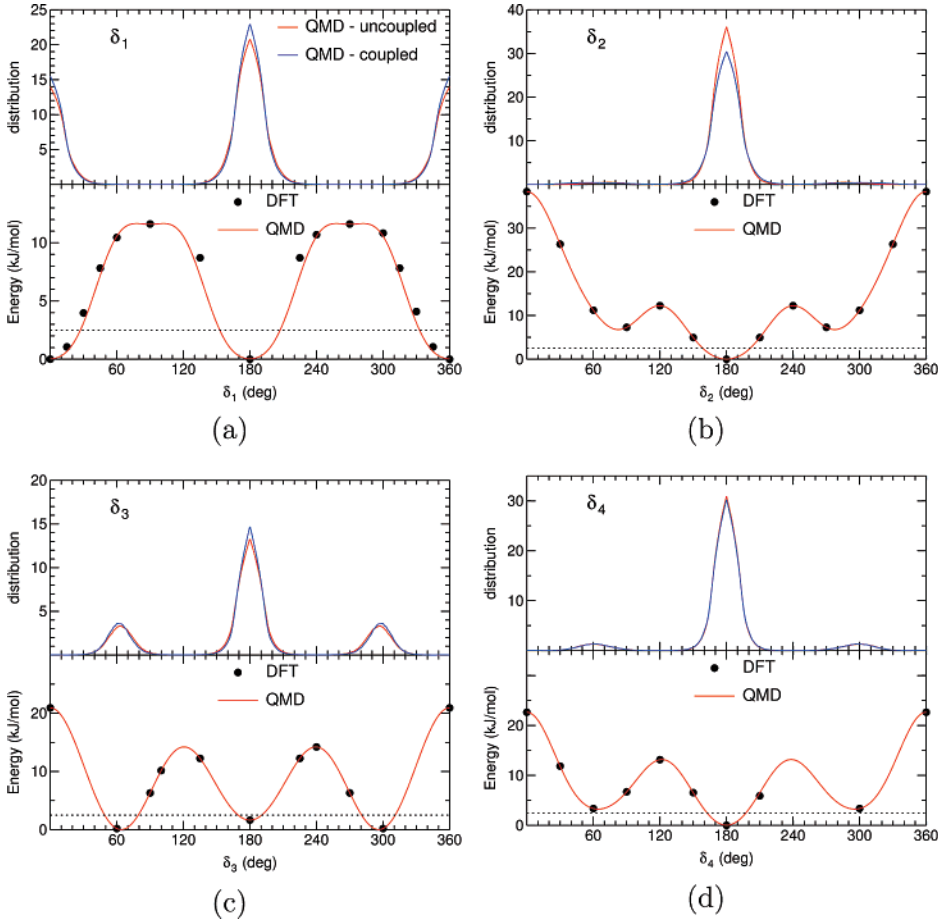


Figure 7. Vibrational modes of butoxybenzene. Bottom panel: comparison between DFT frequencies (red full circles) computed in the UA approximation according to eqs 28 and 29 and QMD predicted frequencies in the uncoupled (blue crosses) and coupled (green triangles) models. Upper panel: DFT computed power spectrum (bottom) and MD computed power spectrum at 298 K and 1 atm.



(a) (b) (c) (d)

Figure 8. Dihedral potentials and distributions for butoxybenzene. DFT computed energies and uncoupled QMD FF are reported in bottom subpanels of panels (a)−(d) for dihedrals $\delta_1-\delta_4$. Black dotted lines indicate the thermal energy at 298 K. In the upper subpanels are reported dihedral distributions computed from the MD trajectories produced with the uncoupled (red lines) and coupled QMD-FFs.

having imposed any constraint, as the bending constants of butane ($C_1-C_2-C_2$, 927 kJ/mol rad^{-2}) and butoxybenzene ($C_{p2}-C_{p2}-C_{p3}$, 969 kJ/mol rad^{-2}) or the "hard" dihedral constants of pyridine ($H-C-C_N-H$, 42 kJ/mol rad^{-2}) and butoxybenzene ($H^*-C^*-C^*-H^*$, 40 kJ/mol rad^{-2}). This seems to indicate that a certain level of transferability does exist, though this is to be verified depending on the molecule under study.

With regards to the torsional potentials, two different models have been adopted: an uncoupled and a coupled one (see eq 41), whose parameters are reported in Tables 13 and 14. The δ_4 dihedral was not coupled to the other dihedrals, and the parameters driving its torsion are slightly affected by the inclusion of the coupling terms between the other dihedrals. Conversely the inclusion of couplings between δ_1, δ_2, and δ_3 causes some changes in the parameters on the pure torsional terms (39), despite the low value of the coupling parameters $\simeq$ 1 kJ/mol.

The effect of the coupling on the resulting frequencies is almost negligible, as shown in the bottom panel of Figure 7, and the agreement with the DFT computed vibrational frequencies is good for both models. It is worth noticing the absence, in both coupled and uncoupled models, of nonbonded interactions between sites which are separated by more than four bonds, as for example the methyl group C_{p3} and the carbon atoms of the ring. In fact the presence of these terms couples all the ICs involving the two sites affected by the nonbonded interactions, introducing off diagonal elements which are not easily controlled during the fitting. Therefore we have chosen to include such terms only when strictly necessary, to prevent, for example, unphysical "curling" of the aliphatic chain over the ring. Due to the relative short length of the butoxybenzene chain, this was found to never be the case. However for longer chain lengths, as for example in the alkyl cyanobiphenyl series (with $n > 4$), where nonbonded interactions between chain sites and aromatic rings have been found to be necessary, a simple method of introducing them has been devised in our laboratory.[23,24,26]

The coupled and uncoupled sets of intramolecular parameters obtained for butoxybenzene were complemented with the OPLS[15] intermolecular parameters, and MD simulations were performed at 298 K and 1 atm. Owing to the increase of molecular dimensions and, consequently, to the time range needed for phase equilibration, it would be preferable to augment the equilibration time to at least 5 ns. To maintain the computational expense acceptable, the time step was increased to 1 fs, and energy conservation was ensured by constraining the stretching motions to equilibrium value during the simulations. The consequence of this choice are evident in the upper panel of Figure 7, where the power spectrum resulting from MD runs is compared to that predicted by the DFT calculations: the agreement is satisfactory for all frequencies except those corresponding to aromatic C−H stretching or C−C skeletal bands, which obviously disappear in the constrained simulation run.

Finally the effect of the coupling terms was checked on the dihedral's average distributions, along the MD trajectories. In Figure 8 the energy profiles and the dihedral distributions are reported for both uncoupled and coupled models. The effect of the coupling terms is negligible on the vibrational frequencies but alters the dihedral distributions of δ_1 and δ_2, which are the most coupled dihedrals. However, the differences in distribution are rather small, since the couplings between butoxybenzene's dihedrals are best appreciated in high-energy unfavorable conformations, which are not populated at room temperature.

5. Conclusions

An intramolecular parametrization of FFs, suitable for molecular simulations of condensed phase and based only upon QM calculations, is proposed here and validated. The main scope of the present work is not to provide general force fields to be put in standard MD packages but rather to implement a method which everyone can use in order to obtain a specific FF for the molecule under study. With this aim the present method has been implemented through the JOYCE program, a user-friendly Fortran code written by the authors and available upon request.

As a first benchmark, a group a rigid heteroaromatic molecules was chosen, whose description through literature FFs has shown to be accurate. The comparison of MD simulation results obtained with both the standard FF and the quantum mechanically derived (QMD) is favorable.

However the main scope of the proposed approach, rather than yielding very accurate FFs for small rigid molecules, is to provide intramolecular FFs for large (and often flexible) molecules, whose bonded parameters are less transferable or even not reported in the literature. For these reasons we have tested the QMD parametrizations on medium-size flexible molecules, modeled through representations of different complexity. Particular attention has been paid to the possibility of parametrizing bonded interaction between coarse grained sites grouping more than one atom, in view of applications to MD simulations of advanced materials condensed phases. In this sense *n*-butane can be seen as the smallest prototype of longer alkyl chains, which can be found in many liquid crystals or polymers. FA and UA reported parametrizations, employed in MD simulations, have shown that the QMD procedure is capable of reproducing many results achieved with widely employed literature FFs. Similar good results are then obtained for two sulfur containing molecules, again using the UA approach.

Finally the method was tested on *n*-butoxybenzene, a nonstandard molecule, whose intramolecular parameters (in particular those regarding the alkoxyl-aryl dihedral) are not readily available in literature databases. Also in this case the QMD-FF yields satisfactory results for dihedral distributions and vibrational frequencies which are in good agreement with the DFT values.

The present results have encouraged us to apply the reported procedure to some large liquid crystals forming molecules. The obtained QMD intramolecular potentials will be joined with intermolecular FFs produced through the FRM approach, recently devised in our group to compute the interaction energy between two large molecules. In fact it is worth pointing out that the proposed intra- and interparametrization procedure can be applied to any target molecule, regardless of its dimensions. In such a way the whole FF is obtained by a first principles approach, without the aid of any experimental data. This will allow us to perform MD or MC simulations with FFs specifically suited on the target molecules, thus accomplishing an important step toward predictivity. Such calculations are currently in progress in our laboratory.

References

(1) Allen, M.; Tildesley, D. *Computer Simulation of Liquids;* Clarendon: Oxford, U.K., 1987.

(2) Frenkel, D.; Smith, B. *Understanding Molecular Simulations;* Academic Press: San Diego, CA, 1996.

(3) Chávez-Páez, M.; dePablo, L.; dePablo, J. *J. Chem. Phys.* **2001**, *114*, 10948.

(4) Hackett, E.; Manias, E.; Giannelis, E. *Chem. Mater.* **2000**, 12, 2161.

(5) Colmenero, J.; Alvarez, F.; Arbe, A. *Phys. Rev. E* **2002**, *65*, 041804.

(6) Harmandaris, V.; Mavrantzas, V.; Theodorou, D.; Kroeger, M.; Ramirez, J.; Oettinger, H.; Vlassopoulos, D. *Macromolecules* **2003**, 36, 1376.

(7) *Computer Simulations of Liquid Crystals and Polymers NATO ASI series;* Pasini, P., Zannoni, C., Zumer, S., Eds.; Kluwer: Dordrecht, 2005.

(8) Care, C.; Cleaver, D. *Rep. Prog. Phys.* **2005**, *68*, 2665.

(9) Wohlert, J.; Olle, E. *J. Chem. Phys.* **2006**, *125*, 204703.

(10) Müller, M.; Katsov, K.; Schick, M. *Phys. Rep.* **2006**, *434*, 113.

(11) Brooks, B. R.; Bruccoeri, R. E.; Olafson, B. D.; States, D. J.; Swaminanthan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(12) Halgren, T. *J. Comput. Chem.* **1996**, *17*, 490.

(13) Hermans, J.; Berendsen, H.; van Gusteren, W.; Postma, J. *Biopolymers* **1984**, *23*, 1.

(14) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollmann, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

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

(16) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. *J. Comput. Chem.* **1997**, *18*, 1955.

(17) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* **2004**, *25*, 1157.

(18) Zannoni, C. *J. Mater. Chem.* **2001**, *11*, 2637.

(19) Voth, G. A. *J. Chem. Theory Comput.* **2005**, *2*, 463.

(20) Yang, L.; Tan, C.; Hsieh, M.; Wang, J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollmann, P.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 13166.

(21) Amovilli, C.; Cacelli, I.; Campanile, S.; Prampolini, G. *J. Chem. Phys.* **2002**, *117*, 3003.

(22) Amovilli, A.; Cacelli, I.; Cinacchi, G.; De Gaetani, L.; Prampolini, G.; Tani, A. *Theor. Chim. Acc.* **2007**, *117*, 885.

(23) Bizzarri, M.; Cacelli, I.; Prampolini, G.; Tani, A. *J. Phys. Chem. A* **2004**, *108*, 10336.

(24) Cacelli, I.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2005**, *109*, 3531.

(25) De Gaetani, L.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2006**, *110*, 2847.

(26) Cacelli, I.; De Gaetani, L.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2007**, *111*, 2130.

(27) Prampolini, G. *J. Chem. Theory Comput.* **2006**, *2*, 556.

(28) Maple, J.; Dinur, U.; Hagler, A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350.

(29) Maple, J.; Hwang, M.-J.; Stockfish, T.; Dinur, U.; Waldman, M.; Ewig, C.; Hagler, A. *J. Comput. Chem.* **1994**, *15*, 162.

(30) Palmo, K.; Mannfors, B.; Mirkin, N.; Krimm, S. *Biopolymers* **2003**, *68*, 383.

(31) Kaminski, G.; Jorgensen, W. *J. Phys. Chem.* **1996**, *100*, 18010.

(32) Dasgupta, S.; Brameld, K.; Fan, C.-F.; Goddard, W., III *Spectrochim. Acta, Part A* **1997**, *53*, 1347.

(33) Chelli, R.; Cardini, G.; Procacci, P.; Righini, R.; Califano, S. *J. Chem. Phys.* **2000**, *113*, 6851.

(34) Gontrani, L.; Ramondo, F.; Caminiti, R. *Chem. Phys. Lett.* **2006**, *422*, 256.

(35) Adam, C.; Clark, S.; Wilson, M.; Ackland, G.; Crain, J. *Mol. Phys.* **1998**, *93*, 947.

(36) Namba, A.; Léon, S.; da Silva, G.; Alemán, C. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 235.

(37) Ceccarelli, M.; Procacci, P.; Marchi, M. *Comput. Mater. Sci.* **2001**, *20*, 318.

(38) Berardi, R.; Muccioli, L.; Zannoni, C. *ChemPhysChem.* **2004**, *5*, 104.

(39) LaPenna, G.; Catalano, D.; Veracini, C. A. *J. Chem. Phys.* **1996**, *105*, 7097.

(40) Dasgupta, S.; Yamasaki, T.; Goddard, W., III *J. Chem. Phys.* **1996**, *104*, 2898.

(41) Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.

(42) Peng, C.; Ayala, P.; Shlegel, H.; Frisch, M. *J. Comput. Chem.* **1996**, *17*, 49.

(43) Dasgupta, S.; Goddard, W., III *J. Chem. Phys.* **1989**, *90*, 7207.

(44) Bakken, V.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 9160.

(45) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipies in Fortran 77;* Cambridge University Press: Cambridge, U.K., 1992.

(46) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648.

(47) 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 A.1)*; Gaussian Inc.: Pittsburgh, PA, 2003.

(48) Paschen, D.; Geiger, A. *MOSCITO 3.9;* Department of Physical Chemistry: University of Dortmund, 2000.

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

(50) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(51) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, A.; Lee, H.; Pedersen, L. J. *Chem. Phys.* **1995**, *103*, 8577.

(52) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *55*, 3336.

(53) Jorgensen, W.; McDonald, N. *J. Mol. Struct.* **1998**, *424*, 145.

(54) McDonald, N.; Jorgensen, W. *J. Phys. Chem. B* **1998**, *102*, 8049.

(55) Thomas, L.; Christakis, T.; Jorgensen, W. *J. Phys. Chem. B* **2006**, *110*, 21198.

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

(57) Jorgensen, W. *J. Phys. Chem.* **1996**, *90*, 6397.

CT700113H

JCTC Journal of Chemical Theory and Computation

Equilibrium Geometries and Structural Stability of the Al_mNa_n ($m = 2-4$; $n = 1-8$) Clusters

Hidenori Matsuzawa,*[,†] Kazuhiro Sato,[‡] Takuji Hirata,[‡] Koichi Ui,[‡] and
Nobuyuki Koura[‡]

Department of Life and Environmental Sciences, Faculty of Engineering, Chiba Institute of Technology, 2-1-1 Shibazono, Narashino, Chiba 275-0023, Japan, and Department of Pure and Applied Chemistry, Faculty of Science and Technology, Tokyo University of Science, Yamazaki, Noda, Chiba 278-8510, Japan

Received November 19, 2006

Abstract: The stable geometries and formation processes of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters were investigated using the density functional theory (DFT). The Al_m ($m = 2-4$) structures are maintained in the clusters. The Na atoms are attached to the Al−Al bond or Al plane for less than $n = 4$ in the Al_mNa_n ($m = 2-4$) clusters. The odd electron of the attached Na atom is transferred to the Al_m part for $n \leq 4$ or 5 in the Al_mNa_n ($m = 2-4$) clusters since the Al_m part becomes stable. The Na−Na bonding is formed by the attached Na atom after saturation of the Al−Al bonds or Al atoms. The Al_2Na_5, Al_3Na_5, Al_3Na_6, Al_3Na_7, and Al_3Na_8 clusters have a characteristic structure. The Na wing is formed in the Al_3Na_n ($n = 5-8$) clusters. The 2S shell containing the *3s* orbital of the Na atom and the *3p* orbital of the Al atom becomes stable before the occupation of the 1D shell because the electrons are delocalized on the Na plane for $n \geq 5$ in the Al_mNa_n ($m = 2-4$) clusters. The stability of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters was evaluated by comparison of the vertical ionization potential (IP), HOMO−LUMO gap, adsorption energy of the Na atom, and binding energy per atom.

1. Introduction

There have been many studies on the geometrical structures and physical properties of bimetallic clusters. Aluminum−sodium (Al−Na) bimetallic clusters have been produced by a molecular beam technique.[1] Recently, the geometries and physical properties of some Al−Na clusters have been examined by physical chemistry researchers. Kanhere and his co-workers reported the stable geometries of the Al−Na clusters, $AlNa_n$ ($n = 1-10$),[2,3] Al_nNa ($n = 1-12$),[4] Al_nNa_2 ($n = 1-12$),[4] Al_2Na,[5] and Al_4Na_4[6], using Car−Parrinello molecular dynamics (CPMD), the quadratic configuration interaction singles and doubles electron correlation (QCISD) method, the density-based molecular dynamics (DBMD) method, and local density approximation (LDA). They have also reported the physical properties of these Al−Na clusters, for example, the highest occupied molecular orbital (HOMO)−lowest unoccupied molecular orbital (LUMO) gap, ionization potential, electron affinity, hardness, and polarizability using the B3LYP and the Vosko-Wilk-Nusair (SVWN) calculations. The geometry and stability of $Al_{13}Na$[7,8] have been examined using the DFT and ab initio molecular dynamics. The stability of the Al−Na clusters has been discussed on the basis of the spherical jellium model for metallic clusters in these reports. An electronic shell closure effect known for simple metal clusters with 40 valence electrons is found in the $Al_{13}Na$ cluster. We have also systematically studied the geometrical and electronic structures of the Al−Na cluster. The geometrical and electronic structures of the Al_n-Na ($n = 1-4$) clusters with the restricted open-shell Hartree−Fock (ROHF) calculations using the 6-31G* basis set have been reported.[9] Some stable structures of the ground and excited states of both Al_nNa and Al_nNa^+ ($n = 1-4$) were described in our previous report. The stability, ionization

* Corresponding author phone: +81-47-454-9600; fax: +81-47-454-9689; e-mail: matuzawa_h@excite.co.jp.

[†] Chiba Institute of Technology.

[‡] Tokyo University of Science.

10.1021/ct600339u CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/15/2007

potential, and formation processes of these clusters have also been discussed. It was found that the Al_n cluster parts remain in the stable Al_nNa cluster, and the charge transfer from the Na atom to the Al_n part occurs.

In this study, the small size clusters, Al_mNa_n ($m = 2-4$; $n = 1-8$), are examined, because we are interested in the early formation process, the geometrical feature, and the stability of the Al−Na cluster. The formation process of the Na cluster is expected in the large size Al−Na clusters, which have a high number of Na atoms. We considered the characteristics of the mixed cluster in the small size clusters. There are few experimental studies of the small size of the Al−Na cluster. It might be difficult to technically select the small size cluster. There are few studies of the formation process of the Al−Na cluster based on the geometric and electronic structures. The charge transfer from the Na atom to the Al atoms is expected due to the difference in the electron negativities between the Al and Na atoms. We are interested in the effect of the charge transfer on the geometry and stability of the Al−Na bimetallic cluster. It is suggested that the adsorption site of the added Na atom is determined on the basis of the electronic structure of the cluster before the addition of a Na atom. In this paper, the geometries and the electronic states of the stable Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters using the DFT calculations are reported. We assumed that the Na atom is adsorbed into the Al_mNa_{n-1} clusters as the number of Na atoms increases in the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters. The calculated vertical ionization potentials, the HOMO−LUMO gap, the adsorption energy of the Na atom, and the binding energy per atom of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters are used to discuss the stability of the clusters.

2. Calculation

The possible geometric and electronic structures of Al_mNa_n ($m = 2-4$; $n = 1-8$) were examined using the density functional theory (DFT) calculations. The initial geometries of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters were assumed as follows. In the small Al_2Na_n clusters ($n = 1-5$), the Na atom was adsorbed by the Al−Al bond and Al atom. In addition, the Na atom was attached to the Al−Na−Al, Na−Al−Na, and Na−Na−Na plane in the large-sized clusters. Three types, i.e., on bond, on atom, and on plane, of the Na atom adsorption on the Al_mNa_{n-1} clusters were performed for the Al_3Na_n and Al_4Na_n clusters when the initial geometries were assumed as in the Al_2Na_n cluster. The addition of the Na atom was performed for all the Al_mNa_{n-1} isomers. The calculations were performed by Becke's three-parameter hybrid function using the Lee−Yang−Parr correlation functional (B3LYP)[10−12] method with the 6-311G* basis set. The B3LYP method with the cc-pVTZ[13] basis set was used for the confirmation of the true minimum for some clusters. The geometries of Al_mNa_n ($m = 2-4$; $n = 1-8$) were optimized using the energy gradient method. The stability of the optimized geometry was confirmed by a frequency analysis. The programs used were the GAUSSIAN 98[14] and GAUSSIAN 03[15] program packages on a COMPAQ Alpha 4100 at the Chiba Institute of Technology (CIT), an HP Exemplar V2500 at Hokkaido University, and a Fujitsu VPP5000/3 at the Tokyo University of Science (TUS). The symmetry was first assumed to be C_1, and, then under the possible high-symmetry found in the preliminary calculations, the geometry of each cluster was reoptimized. Spin multiplicities of the singlet and triplet for an even electron system, and doublet and quartet for an odd electron system, were considered in the geometry optimization. The relative energy of each isomer for the ground state was obtained from the ΔSCF method with the total energy after the zero-point correlation. The initial geometries of the examined clusters are determined on the basis of some precalculations. The vertical ionization potential was estimated by the difference in the total energy between the neutral and cationic clusters for the optimized geometry of a neutral cluster. We define the adsorption energy of the Na atom as $E_{Na} = -[E(Al_mNa_n) - E(Al_mNa_{n-1}) - E(Na)]$ and the binding energy per atom as $E_b = -[E(Al_mNa_n) - mE(Al) - nE(Na)]/(m+n)$.

3. Results and Discussion

3-1. Stable Structures of the Al_2Na_n ($n = 1-8$) Clusters. The ground states of the Al_2Na and Al_2Na_2 clusters have been reported.[4,5,9] There has been no study on the geometric and electronic structures of the Al_2Na_n ($n = 3-8$) clusters to the best of our knowledge. The stable geometries of the Al_2Na_n ($n = 1-8$) cluster are shown in Figure 1. The electron configurations and some term energies of the Al_2Na_n ($n = 1-8$) clusters are listed in Table 1. We assume that the Al_2Na_n cluster forms due to the adsorption of a Na atom on the Al_2Na_{n-1} cluster. In the small size clusters, a Na atom would directly adsorb on the Al_2 molecule, as the Al−Al bond is maintained in all of the obtained stable Al_2Na_n ($n = 1-8$) clusters. The relationship between the formation process and geometric and electronic features is discussed on the basis of this assumption. Two stable structures with an isosceles triangular shape for the Al_2Na cluster are found in our calculations. The $^2A_1$ (C_{2v}) **(2-1)** state is the ground state of the Al_2Na cluster. The Al−Al bond length of 2.687 Å in **2-1** is slightly shorter than that of 2.765 Å for the $^3\Pi_u$ ($D_{\infty h}$) state or the experimental result (2.701 Å)[18] of the Al_2 molecule. The Al−Al bond length (2.44 Å: BPW91/6-311G**)[17] in the Al_2Li cluster is also shorter than that of the Al_2 molecule. It is found that the Al−Al bond becomes shorter than that of the Al_2 molecule due to the adsorption of the alkali metal. The $2a_1$ orbital of **2-1** contains the components of the Al−Al π-bonding and *3s* orbital of the Na atom, and the paired electrons in this orbital are distributed in the Al−Na−Al plane. Therefore, the Al−Al bond becomes shorter because the odd electron of the Na atom is taken into the $1\pi_u$ orbital of the Al_2 molecule. The spin density of 0.51 for the Al atom supports the electron transfer from the Na atom to the Al atoms. The linear Al−Al−Na cluster is the low-lying state based on our calculations. The stable structure of the linear Al−Na−Al cluster was not obtained.

The $^1A_1$ (C_{2v}) **(2-2)**, $^2A_1'$ (D_{3h}) **(2-3)**, and $^1A_{1g}$ (D_{4h}) **(2-4)** states of the Al_2Na_2, Al_2Na_3, and Al_2Na_4, respectively, are the most stable state of each cluster. It was found that the Na atom attaches to the Al−Al bond in the Al_2Na_n ($n = 1-4$) clusters. In the Al_2Na_2 **(2-2)** cluster, the dihedral angle

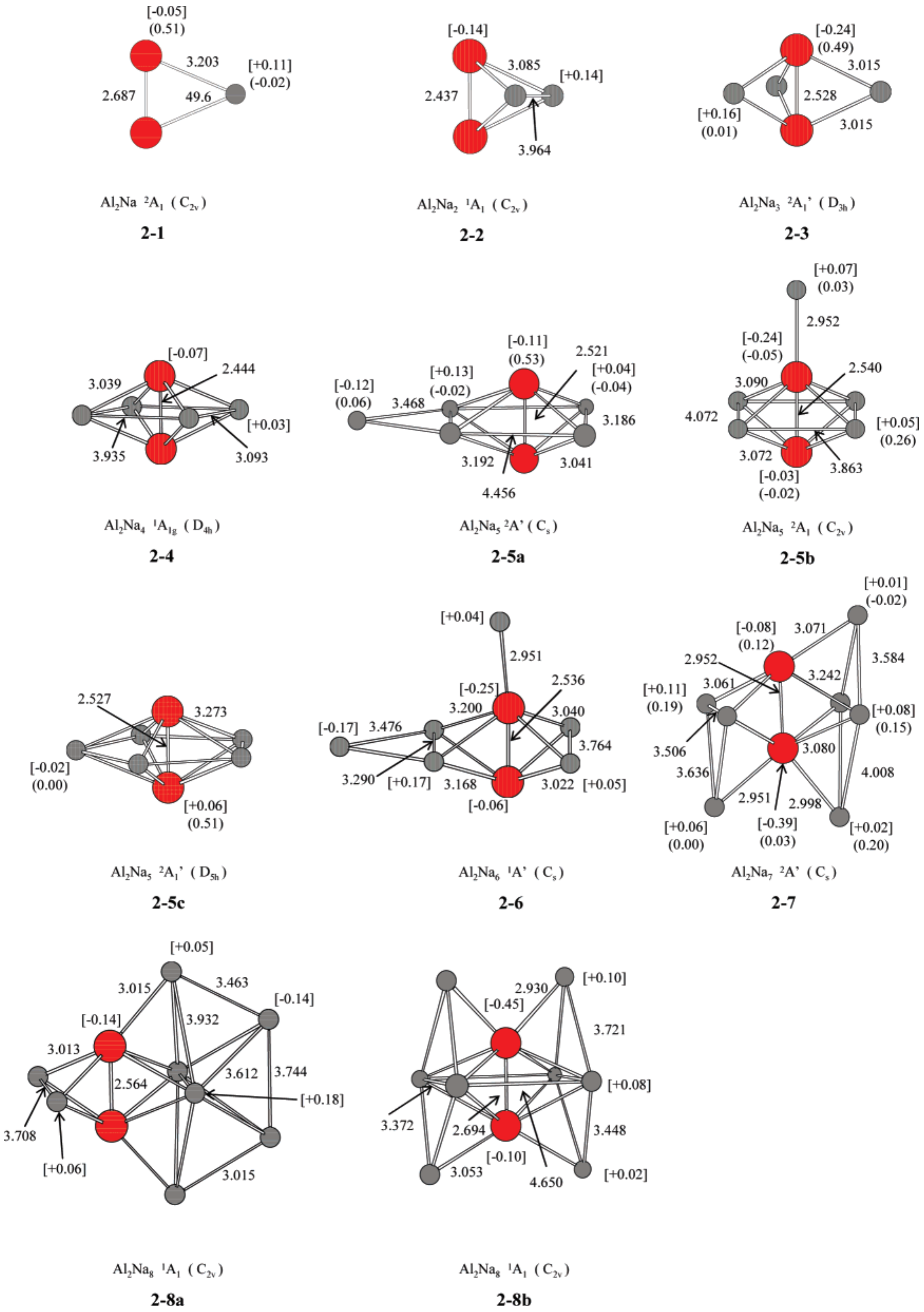


Figure 1. Optimized geometries of the Al_2Na_n ($n = 1-8$) clusters with some bond lengths (Å) and angles (deg). Red and gray circles are Al and Na atoms, respectively. Mulliken charges in square brackets and the spin densities in parentheses are also given.

for Na−Al−Al−Na becomes 88.7° because the odd electron of the Na atom is taken into the $1b_1$ orbital of the Al_2Na cluster, which has the out-of-plane Al−Al π-bonding character. It is required that the odd electron of the Na atom is transferred to the Al_2 part in order to form the Al_2Na_3 and Al_2Na_4 clusters because both the SOMO of **2-3** and the HOMO ($2a_{1g}$) of **2-4** have a large Al−Al σ-bonding character. The spin density of 0.49 for the Al atom of **2-3** supports this electron transfer. The odd electron of a Na atom is transferred to the Al_2 molecule because the electron negativity of an Al atom is greater than that of a Na atom.

The Al_2Na_4 structure is maintained in the ground states

Table 1. Electronic States, Symmetries, Electron Configuration, and Term Energies of Al_2Na_n ($n = 1-8$) Clusters

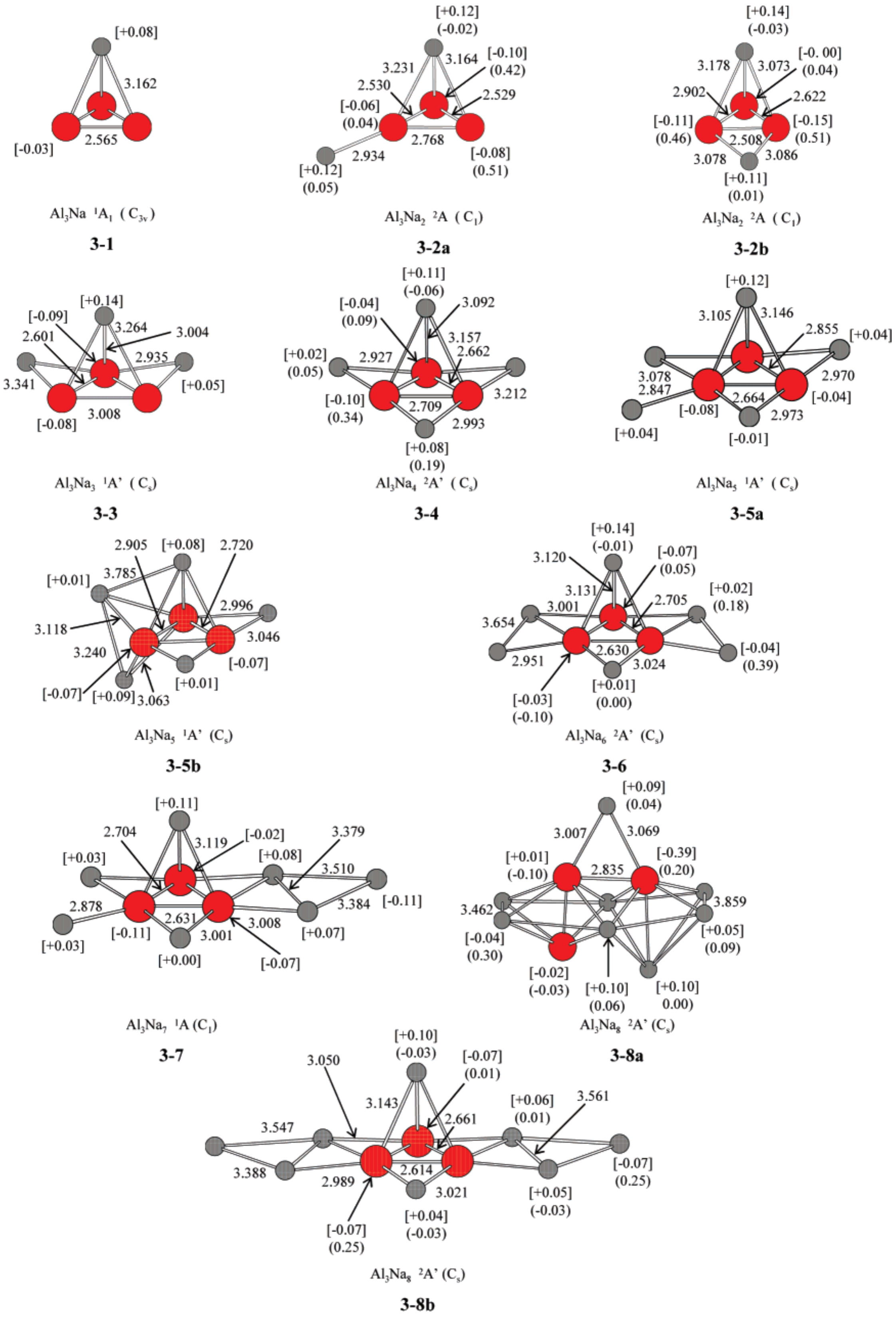
		electronic state	symmetry	electron configuration[a]	T_e/kJ·mol^{-1}
Al_2Na	**2-1**	$^2A_1$	C_{2v}	(core)$(1a_1)^2(1b_2)^2(2a_1)^2(3a_1)^1$	0.0
Al_2Na_2	**2-2**	$^1A_1$	C_{2v}	(core)$(1a_1)^2(1b_2)^2(2a_1)^2(1b_1)^2$	0.0
Al_2Na_3	**2-3**	$^2A_1'$	D_{3h}	(core)$(1a_1')^2(1a_2'')^2(1e'_+)^2(1e'_-)^2(1a_1')^1$	0.0
Al_2Na_4	**2-4**	$^2A_{1g}$	D_{4h}	(core)$(1a_{1g})^2(1a_{2u})^2(1e_{u+})^2(1e_{u-})^2(2a_{1g})^2$	0.0
Al_2Na_5	**2-5a**	$^2A'$	C_s	(core)$(1a')^2(2a')^2(3a')^2(1a'')^2(4a')^2(5a')^1$	0.0
	2-5b	$^2A_1$	C_{2v}	(core)$(1a_1)^2(2a_1)^2(1b_2)^2(1b_1)^2(3a_1)^2(4a_1)^1$	0.6
	2-5c	$^2A_1'$	D_{5h}	(core)$(1a_1)^2(2a_1)^2(1b_2)^2(1b_1)^2(3a_1)^2(4a_1)^1$	10.4
Al_2Na_6	**2-6**	$^1A'$	C_s	(core)$(1a')^2(2a')^2(3a')^2(1a'')^2(4a')^2(5a')^2$	0.0
Al_2Na_7	**2-7**	$^2A'$	C_s	(core)$(1a')^2(2a')^2(3a')^2(1a'')^2(4a')^2(5a')^2(6a')^1$	0.0
Al_2Na_8	**2-8a**	$^1A_1$	C_{2v}	(core)$(1a_1)^2(1b_2)^2(2a_1)^2(1b_1)^2(3a_1)^2(4a_1)^2(2b_2)^2$	0.0
	2-8b	$^1A_1$	C_{2v}	(core)$(1a_1)^2(1b_2)^2(2a_1)^2(1b_1)^2(3a_1)^2(4a_1)^2(2b_2)^2$	23.8

[a] The "core" in parentheses means the core electrons.

of the Al_2Na_n ($n = 5-8$) clusters. It is suggested that the Al_2Na_4 cluster is geometrically stable. Three stable structures of the Al_2Na_5 cluster were obtained. The $^2A'$ (C_s) (**2-5a**), $^2A_1$ (C_{2v}) (**2-5b**), and $^2A_1'$ (D_{5h}) (**2-5c**) states are formed by the attachment of the Na atom to the Na−Na bond, the Al atom, and the Al−Al bond of **2-4**, respectively. The $^2A'$ (C_s) (**2-5a**) state is the most stable state. The $^2A_1$ state (**2-5b**) was obtained when the Na atom approached the Na−Al−Na plane of **2-4** during the geometry optimization. The two geometries of **2-5a** and **2-5b** are shown to have essentially similar stabilities. The $^2A_1'$ (D_{5h}) state (**2-5c**) locates 10.4 kJ/mol higher than **2-5a** on the potential energy surface. The stability of the Al_2Na_5 cluster will be discussed in a later section. The $^1A'$ (C_s) **(2-6)** state is the most stable in the Al_2Na_6 isomers. This geometry is formed by the attachment of a Na atom to the $^2A'$ (C_s) state (**2-5a**) or $^2A_1$ state (**2-5b**) of the Al_2Na_5 cluster. This stable geometry supports the fact that the attachment of the Na atom to both the Al atom and the Na−Na bond produces a stable structure. The paired electrons of the HOMO containing the *3s* character of the Na atoms are mainly distributed on the Na_3 plane. This means that the odd electron is transferred not to the Al−Al part but to the Na plane in the Al_2Na_6 cluster. The attachment of the Na atom to the Na−Al−Na plane of **2-5b** produces the stable structure (no figure) located 11.7 kJ/mol higher than the ground state. The transition state of the Al_2Na_6 cluster is obtained when two Na atoms add to each of the two Na−Al−Na planes of the Al_2Na_4 cluster. In the $^2A'$ (C_s) **(2-7)** state of the Al_2Na_7 cluster, three Na atoms are attached to the Na−Al−Na planes of the Al_2Na_4 structure. This state is formed due to the approach of the Na atom to one of the local minima of the Al_2Na_6 cluster. A large deformation is required when the $^2A'$ (C_s) **(2-7)** state is formed due to the attachment of the Na atom to the ground state of the Al_2Na_6 cluster. The equilibrium structure of the pentagonal bipyramidal shape, i.e., the $^4A_1$ (D_{5h}) state, is the transition state because six imaginary frequencies are found. Therefore, it is considered that the approach of the Na atom to only the Al−Al bond or Na plane is impossible in the Al_2Na_7 cluster. The most stable state of the Al_2Na_8 cluster is the $^1A_1$ (C_{2v}) **(2-8a)** state. A large deformation of the Al_2Na_7 attached to the Na atom is required for the formation of this state. Actually, the $^1A_1$ (C_{2v}) **(2-8a)** state is obtained for the Al_2Na_7 cluster with the addition of a Na atom to the Na−Na bond. It is suggested that the stable part with the Na atoms is formed in this cluster. The wave functions of the $2a_1$ orbital (fifth HOMO) distribute in the hexagonal bipyramid part of four Na atoms and two Al atoms. This orbital is formed due to the mixture of the *3s* orbital of six Na atoms and the π orbital of Al_2. The paired electrons of this orbital are delocalized in the hexagonal bipyramid shape. The attachment of the Na atom to the leftover Na−Al−Na plane of the Al_2Na_7 cluster produces the $^1A_1$ state **(2-8b)** of the Al_2Na_8 cluster, which is located 23.8 kJ/mol higher than the $^1A_1$ (C_{2v}) **(2-8a)** state.

The Al−Al bond is maintained in all of the stable Al_2Na_n ($n = 1-8$) clusters. The Na atom approaches the Al−Al bond of the Al_2Na_{n-1} cluster up to $n = 4$. It was found that the transfer of the odd electron from the attached Na atom plays an important role in the formation process of these clusters. The high spin densities of the Al atoms in the Al_2Na (**2-1**) and Al_2Na_3 (**2-3**) clusters mean that the odd electron is localized on the Al−Al bonding orbital. Up to $n = 4$, a small odd−even alternation of the Al−Al bond length is found. The Al−Al bond lengths of the even numbered system are slightly shorter than those of the odd numbered system because the Al−Al bonding orbital is doubly occupied in the even numbered system. The Al−Al bond becomes stable due to the electron transfer from the Na atom to the Al atoms in small-sized Al_2Na_n clusters. The Al_2Na_4 cluster is expected to be geometrically stable because the molecular orbitals including the *3p* characters of the Al atom are occupied. The valence electrons are delocalized on the Na plane for $n \geq 6$ in the Al_2Na_n clusters.

3-2. Structures of the Al_3Na_n ($n = 1-8$) Clusters. Figure 2 shows the stable structures of the Al_3Na_n ($n = 1-8$) clusters. The most stable state of the Al_3Na cluster is the $^1A_1(C_{3v})$ **(3-1)** state with the tetrahedral shape. The same structure was obtained as the ground state in the PW91 calculation using the LanL2DZ basis set[4] and in the ROHF calculation.[9] The Al−Al bond lengths have not changed from those of the Al_3 cluster ($^2A_1$ state, 2.535 Å) after the adsorption of a Na atom. The Mulliken charges of the Na atom and Al atoms are small. This means that a small amount of charge is transferred from the Na atom to the Al_3 plane through the HOMO ($2a_1$), in which the *3s* orbital of the Na atom and the out-of-plane π orbital of the Al_3 cluster are mixed. The bonding of the Na atom in the $^2A(C_1)$ **(3-2a)** state is the "on-atom" type. On the other hand, that in the $^2A(C_1)$ **(3-2b)** state is the "on-bond" type. The odd electron



Al_3Na $^1A_1$ (C_{3v})

3-1

Al_3Na_2 ^2A (C_1)

3-2a

Al_3Na_2 ^2A (C_1)

3-2b

Al_3Na_3 $^1A'$ (C_s)

3-3

Al_3Na_4 $^2A'$ (C_s)

3-4

Al_3Na_5 $^1A'$ (C_s)

3-5a

Al_3Na_5 $^1A'$ (C_s)

3-5b

Al_3Na_6 $^2A'$ (C_s)

3-6

Al_3Na_7 ^1A (C_1)

3-7

Al_3Na_8 $^2A'$ (C_s)

3-8a

Al_3Na_8 $^2A'$ (C_s)

3-8b

Figure 2. Optimized geometries of the Al_3Na_n ($n = 1-8$) clusters.

of **3-2a** is localized on two Al atoms which are without the bonding of the Na atom, and that of **3-2b** is localized on two Al atoms which are bonded to the Na atom. There is essentially no energy difference between **3-2a** and **3-2b** though it is expected that the "on-bond" type is more stable than the "on-atom" type based on the result of the Al_2Na_n cluster.

Two or three Na atoms are attached to the Al−Al bond of the Al_3 plane in the ground states of the Al_3Na_3 and Al_3Na_4 clusters. The $^1A'(C_s)$ **(3-3)** state is formed due to the

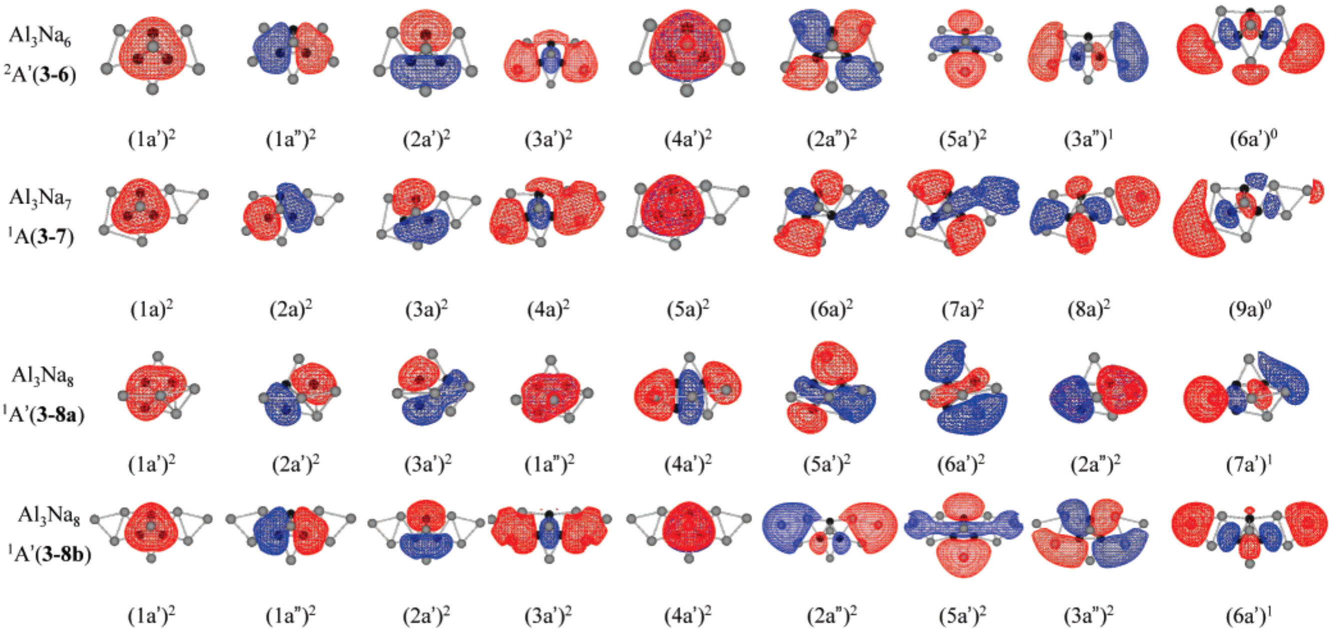


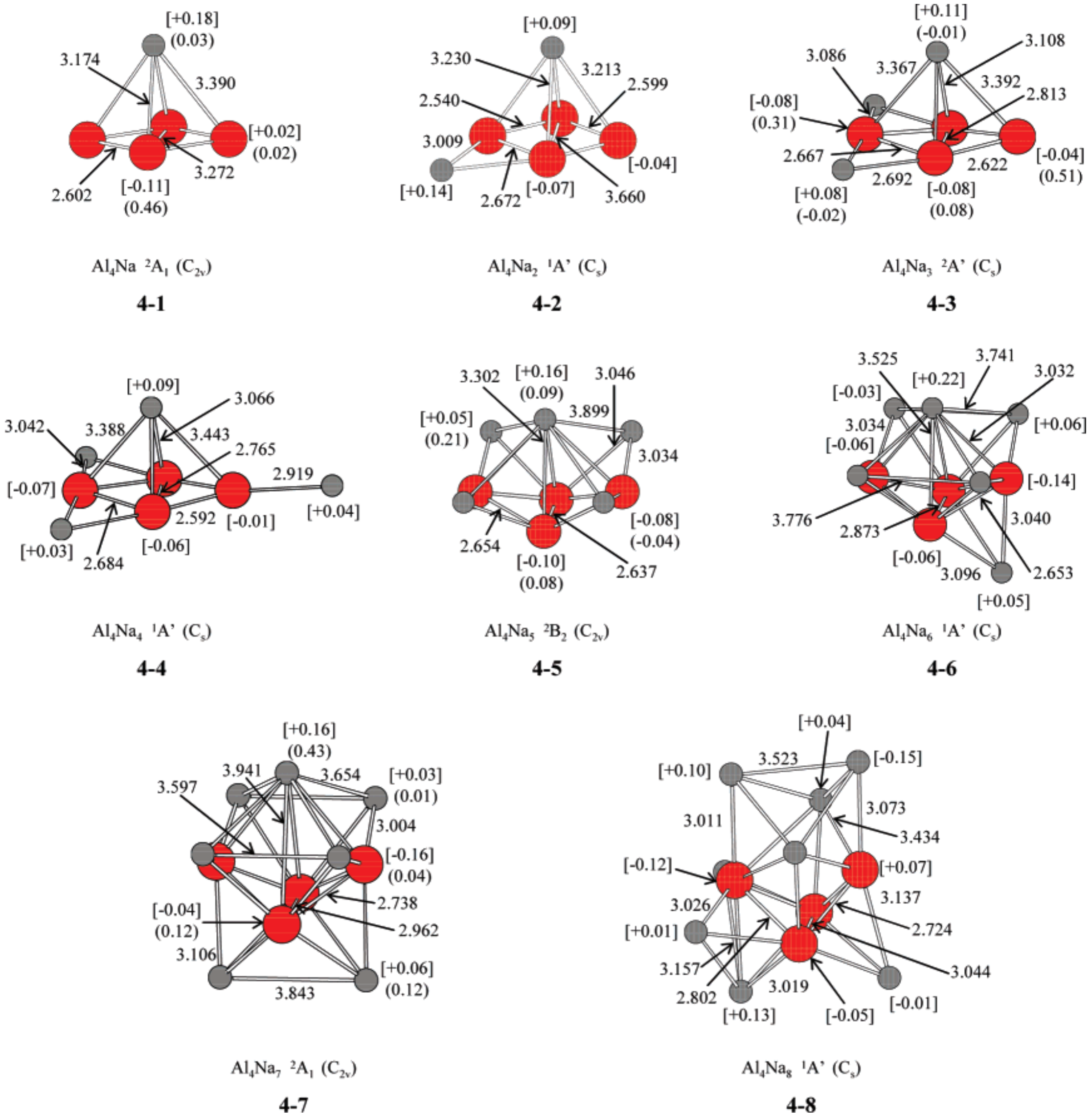
Figure 3. Valence molecular orbitals of the Al_3Na_6, Al_3Na_7, and Al_3Na_8 clusters.

addition of a Na atom to the $^2A(C_1)$ **(3-2b)** state of the Al_3Na_2 cluster. Furthermore, the $^2A'(C_s)$ **(3-4)** state is formed due to the addition of a Na atom to the $^1A'(C_s)$ **(3-3)** state of the Al_3Na_3 cluster. The spin densities of the corresponding Al and Na atoms of **3-4** are 0.34 and 0.19, respectively. Therefore, it is found that the odd electron is distributed to one of the Al−Na−Al planes. The Na atom is bonded to the Al atom of the Al_3Na_4 in the ground state, i.e., $^1A'(C_s)$ **(3-5a)**, of the Al_3Na_5 cluster. The two adsorption cases of the Na atom to the Al−Al bond [$^1A'(C_s)$ **(3-5b)**] and the Al_3 plane [$^1A'(D_{3h})$ **(3-5c)**] are located 1.44 and 4.73 kJ/mol higher than the ground state, respectively. That is to say, three adsorption types of the Na atom to the Al_3 cluster, i.e., on-atom, on-bond, and on-plane, produce a similar stability for the cluster. In the Al_2Na_n clusters, the charge transfer from the added Na atom to the Al_2 molecule is important for producing a stable cluster, and the charge transfer occurs with the on-bond approach of the Na atom. In the Al_3Na_n (n = 1−5) clusters, it is suggested that the adsorption type is less important as there is a slight charge transfer from the Na atom to the Al_3 cluster.

The most stable state of the Al_3Na_6 cluster is the $^2A'(C_s)$ **(3-6)**, in which the Na atom is bonded to the Al atom of the Al_3Na_5 **(3-5a)** cluster. The Na−Na bond begins to be formed in this state. The small negative charges (−0.04) of the Na atoms are also found. The spin densities of the four Na atoms (0.39 or 0.18) bonded to the Al atoms are greater than those of the other Na atoms. These high spin densities mean that the odd electron of the SOMO is distributed to the Na−Na bonds. In the ^1A (C_1) **(3-7)** state of the Al_3Na_7 cluster, the Na_3 plane is formed due to the addition of the Na atom to the Al_3Na_6 **(3-6)** cluster. The $^1A'(C_s)$ **(3-8a)** state locates 0.44 kJ/mol lower than the $^1A'(C_s)$ **(3-8b)** state. A similar stability between the $^1A'(C_s)$ **(3-8a)** and $^1A'(C_s)$ **(3-8b)** states is found. The Al_2Na_4 skeleton is included, and the Al_3 plane is broken in **3-8a**. It seems that the origin of this cluster is **3-5b** of the Al_3Na_5 cluster. The $^1A'(C_s)$ **(3-8a)** state is formed due to the adsorption of the Na atom on the Al_3Na_7 isomer (17.3 kJ/mol higher than the ground state) after the addition of a Na atom to one of the Al_3Na_6 isomers, which is located 36.4 kJ/mol higher than the ground state. The $^1A'(C_s)$ **(3-8b)** state is formed due to the addition of the Na atom to the Na−Na bond of the Al_3Na_7 cluster. The Na atoms located on top of the Na wing have negative charges as in the Al_3Na_6 **(3-6)** and Al_3Na_7 **(3-7)** clusters. The spin densities of these Na atoms increase because the SOMO includes the *3s* orbital characters of these Na atoms. In the Al_3Na_6 **(3-6)**, Al_3Na_7 **(3-7)**, and Al_3Na_8 **(3-8b)** clusters, the Na wing grows together with the increasing number of Na atoms. The Al−Al bond lengths and Mulliken charges of the Al atoms are not changed with the number of Na atoms.

The ground states from the Al_3Na_6 to Al_3Na_8 clusters maintain the Na planes. The molecular orbitals of these ground states are shown in Figure 3. The 3a′, 4a, and 3a′ orbitals of **3-6**, **3-7**, and **3-8b**, respectively, include the *3s* orbital character of the Na atoms. It is suggested that these orbitals corresponding to the 2S shell of the shell model become stable because the electrons are delocalized on the Na plane. The spin densities of the Al_3Na_6 and Al_3Na_8 clusters show no charge transfer from the Na atoms. It is suggested that these clusters become stable due to the delocalization of the valence electrons to the wing Na atoms.

3-3. Stable Structures of the Al_4Na_n (n = 1−8) Clusters. We assumed that the Al_4Na_n (n = 1−8) clusters are formed due to the adsorption of the Na atom on the Al_4Na_{n-1} (n = 1−8) clusters as in the other examined Al−Na clusters. Figure 4 shows the stable structures of the Al_4Na_n (n = 1−8) clusters with some bond lengths, angles, Mulliken charges, and spin densities. The electron configurations of the Al_4Na_n (n = 1−8) clusters are listed in Table 3. The most stable structure of the Al_4Na is formed due to the addition of the Na atom on the rhombus Al_4 cluster from a direction perpendicular to the Al_4 plane. It has been reported that the ground state of the Al_4 cluster is the $^3B_{1u}$ or $^3B_{1g}$ state of the planar rhombus structure.[9,19] The ground state of Al_4Na is the $^2A_1$ (C_{2v}) **(4-1)** state, though we have reported that the



Al_4Na $^2A_1$ (C_{2v})
4-1

Al_4Na_2 $^1A'$ (C_s)
4-2

Al_4Na_3 $^2A'$ (C_s)
4-3

Al_4Na_4 $^1A'$ (C_s)
4-4

Al_4Na_5 $^2B_2$ (C_{2v})
4-5

Al_4Na_6 $^1A'$ (C_s)
4-6

Al_4Na_7 $^2A_1$ (C_{2v})
4-7

Al_4Na_8 $^1A'$ (C_s)
4-8

Figure 4. Optimized geometries of the Al_4Na_n (n = 1−8) clusters.

Table 2. Electronic States, Symmetries, Electron Configuration, and Term Energies of Al_3Na_n (n = 1−8) Clusters

		electronic state	symmetry	electron configuration[a]	T_e/kJ·mol^{-1}
Al_3Na	**3-1**	$^1A_1$	C_{3v}	(core)$(1a_1)^2(1e_+)^2(1e_-)^2(2a_1)^2(3a_1)^2$	0.0
Al_3Na_2	**3-2a**	^2A	C_1	(core)$(1a)^2(2a)^2(3a)^2(4a)^2(5a)^2(6a)^1$	0.0
	3-2b	^2A	C_1	(core)$(1a)^2(2a)^2(3a)^2(4a)^2(5a)^2(6a)^1$	0.9
Al_3Na_3	**3-3**	$^1A'$	C_s	(core)$(1a')^2(1a'')^2(2a')^2(3a')^2(4a')^2(2a'')^2$	0.0
Al_3Na_4	**3-4**	$^2A'$	C_s	(core)$(1a')^2(1a'')^2(2a')^2(3a')^2(4a')^2(2a'')^2(5a')^1$	0.0
Al_3Na_5	**3-5a**	$^1A'$	C_s	(core)$(1a')^2(1a'')^2(2a')^2(3a')^2(4a')^2(2a'')^2(5a')^2$	0.0
	3-5b	$^1A_1'$	D_{3h}	(core)$(1a_1')^2(1e'_+)^2(1e'_-)^2(1a_2'')^2(2a_1')^2(2e'_+)^2(2e'_-)^2$	4.7
Al_3Na_6	**3-6**	$^2A'$	C_s	(core)$(1a')^2(1a'')^2(2a')^2(3a')^2(4a')^2(2a'')^2(5a')^2(3a'')^1$	0.0
Al_3Na_7	**3-7**	^1A	C_1	(core)$(a)^2(a)^2(a)^2(a)^2(a)^2(a)^2(a)^2(a)^2$	0.0
Al_3Na_8	**3-8a**	$^2A'$	C_s	(core)$(1a')^2(2a')^2(3a')^2(1a'')^2(4a')^2(5a')^2(6a')^2(2a'')^2(7a')^1$	0.0
	3-8b	$^2A'$	C_s	(core)$(1a')^2(1a'')^2(2a')^2(3a')^2(4a')^2(2a'')^2(5a')^2(3a'')^2(6a')^1$	0.4

[a] The "core" in parentheses means the core electrons.

$^4B_2$ (C_{2v}) state is the more stable structure using the ROHF level of calculations.[9] The $^4B_2$ state is located 22.5 kJ/mol higher than the $^2A_1$ (C_{2v}) state using the B3LYP calculations. After the adsorption of the Na atom, the Al_4 part is nearly planar, and the dihedral angle of the Al_4 plane is 179.4°. The Al−Al bond length of 2.602 Å is slightly shorter than that of the Al_4 (2.657 Å). The diagonal Al−Al distance of 3.272 Å is also shorter than that of the Al_4 (3.417 Å) because the SOMO ($4a_1$) of the Al_4Na cluster has the Al−Al σ-bonding character. The positive charge of the Na atom and high spin densities of the Al atoms suggest an electron transfer from the Na atom to the Al_4 cluster. The Na atom is adsorbed on the Al−Al bond in the stable structures of the Al_4Na_2 and Al_4Na_3. The $^2A'$ (C_s) (**4-2**) state of Al_4Na_2

Table 3. Electronic States, Symmetries, Electron Configuration, and Term Energies of Al_4Na_n ($n = 1-8$) Clusters

		electronic state	symmetry	electron configuration[a]
Al_4Na	**4-1**	$^2A_1$	C_{2v}	$(core)(1a_1)^2(1b_1)^2(1b_2)^2(2a_1)^2(3a_1)^2(1a_2)^2(4a_1)^1$
Al_4Na_2	**4-2**	$^2A'$	C_s	$(core)(1a')^2(2a')^2(1a'')^2(2a'')^2(3a')^2(4a')^2(5a')^2$
Al_4Na_3	**4-3**	$^2A'$	C_s	$(core)(1a')^2(2a')^2(1a'')^2(3a')^2(4a')^2(2a'')^2(5a')^2(6a')^1$
Al_4Na_4	**4-4**	$^1A'$	C_s	$(core)(1a')^2(2a')^2(1a'')^2(3a')^2(4a')^2(2a'')^2(5a')^2(6a')^2$
Al_4Na_5	**4-5**	$^2A_2$	C_{2v}	$(core)(1a_1)^2(1b_1)^2(1b_2)^2(2a_1)^2(3a_1)^2(1a_2)^2(4a_1)^2(2b_1)^2(2b_2)^1$
Al_5Na_6	**4-6**	$^1A'$	C_s	$(core)(1a')^2(2a')^2(la'')^2(3a')^2(4a')^2(5a')^2(2a'')^2(6a')^2(3a'')^2$
Al_4Na_7	**4-7**	$^2A_1$	D_{2v}	$(core)(1a_1)^2(1b_2)^2(1b_1)^2(2a_1)^2(3a_1)^2(4a_1)^2(1a_2)^2(2b_2)^2(2b_1)^2(5a_1)^1$
Al_4Na_8	**4-8**	$^1A'$	C_s	$(core)(1a')^2(2a')^2(1a'')^2(3a')^2(4a')^2(5a')^2(2a'')^2(6a')^2(7a')^2(3a'')^2$

[a] The "core" in parentheses means the core electrons.

has been reported in an earlier paper.[20] In the $^2A'$ (C_s) (**4-2**) state, the molecular orbital character of the Al_4 part might be like that of the stable Al_4^{2-} cluster.[19] This suggests that the odd electrons are transferred from two Na atoms to the Al_4 part though the small positive charges of two Na atoms remain. In the $^2A'$ (C_s) (**4-3**) state of the Al_4Na_3 cluster, two Na atoms are attached to the Al−Al bond of the Al_4 plane. The spin densities of 0.51 for the two Al atoms mean that an odd electron transfers from the Na atom to the Al atoms after the attachment of the Na atom to the Al−Al bond. The $^1A'(C_s)$ (**4-4**) state is the ground state of the Al_4Na_4 cluster though that is different from **4-4** with the CCD/6-31G(d, p) level of calculations.[21] In the $^1A'(C_s)$ (**4-4**) state, the added Na atom is bonded to the Al atom (on-atom). A small positive charge of the Na atom bonded to the Al atom is found. Furthermore, there is no change in the Mulliken charges of the Al_4 plane from those of the Al_4Na_3 cluster. These charge distributions suggest a small electron transfer in **4-4**. We obtained two isomers of the Al_4Na_4 cluster. One is the ^1A (C_1) state (no figure), in which the added Na atom is bonded to the Al−Al bond (on-bond). This isomer is 5.87 kJ/mol higher than **4-4**. It is suggested that the adsorption type of the Na atom to the Al_4Na_3 cluster is less important when forming the stable Al_4Na_4 cluster as in the Al_3Na_2 and Al_3Na_5 clusters. The others are located 28.0 kJ/mol higher than **4-4**. In this isomer, two Na atoms are attached to the Al−Na−Al plane in the octahedral shape of the Al_4Na_2 cluster.

The most stable state of the Al_4Na_5 cluster is the $^2B_2$ (C_{2v}) (**4-5**) state. The spin densities of 0.21 of each Na atom mean that the odd electron is localized on the Na atoms. The SOMO has the large *3s* characters of the four Na atoms. The dihedral angle of the Al_4 frame becomes 142.8° from 172.0° for the Al_4Na_4 (**4-4**) cluster. The distance of the central Na atom to each of the four Na atoms is 3.899 Å that means a weak Na−Na bond. Therefore, it is suggested that the formation of the Na−Na bond starts from this cluster. In the Al_4Na_6 and Al_4Na_7 clusters, the Na atoms are attached to the Al_3 plane in the Al_4Na_5 frame. The $^1A'$ (C_s) (**4-6**) state of the Al_4Na_6 cluster is the most stable state. The wave function of the HOMO distributes to the tetrahedral Na atoms on the Al_4 frame. The Na−Na distance between the central Na atom and four apical Na atoms is shorter than that of the Al_4Na_5 cluster. The Mulliken charges of the two Na atoms that formed the Al_4Na_5 skeleton are slightly negative. On the other hand, the positive charge of the central Na atom of the Al_4Na_5 skeleton is enhanced. By the addition of the Na atom to the Al_3 plane of **4-6**, the most stable state of the Al_4Na_7 cluster, $^2A_1$ (C_{2v}) (**4-7**), is formed. The Na−Na distance between the central Na atom and four apical Na atoms is shorter than that of the Al_4Na_6 cluster. The ground state of the Al_4Na_8 is the $^1A'$ (C_s) (**4-8**) state. The boat form of the Al_4 part is maintained in this state. It seems that the Na cluster part is distorted after the attachment of the Na atom to the Na−Na bond. The small negative charge of the Na atom is found as in the Al_2Na_8 and Al_3Na_8 clusters. The Al_4Na_8 cluster is a 20-valence electrons system, which produces a stable structure.

In the Al_4Na_n ($n = 1-8$) clusters, the Na atom is first attached to the planar Al_4 cluster from the perpendicular plane. From the Al_4Na_2 to the Al_4Na_5 cluster, the Na atom essentially adsorbs on the Al−Al bond. The Na atom is added to the Al_3 plane from the opposite side of the Na_5 part in the Al_4Na_6 and Al_4Na_7 clusters. In the Al_4Na_8 cluster, the Na atoms are attached to the Al−Al bond, Al atom, and Al_3 plane, and the Na_4 part is formed. From the view point of the cluster growth from Al_4Na_{n-1} to Al_4Na_n, after the formation of the Al_4Na_5 cluster, the Na atom adsorbs on the Al_4Na_{n-1} cluster to maintain the Al_4Na_5 structure. The Na−Na bond begins to be formed from the Al_4Na_5 cluster. The negative charge of the Na atom is found in the Al_4Na_8 cluster as in the Al_2Na_8 and Al_3Na_8 clusters.

3-4. Stability of the Al_mNa_n ($m = 1-4$; $n = 1-8$) Clusters. Figure 5 shows the vertical ionization potential (IP) of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters. The HOMO−LUMO gap, the adsorption energy of Na atom, and the binding energy per atom of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters are shown in Figures 6−8, respectively. The stability of the clusters is discussed on the basis of the molecular orbital theory and shell model in this section. Three IP peaks (Al_2Na_2, Al_2Na_6, and Al_2Na_8) are found in Figure 5(a). The HOMO−LUMO gaps, the adsorption energy of Na atom, and the binding energy per atom of these clusters are also high in Figures 6(a), 7(a), and 8(a), respectively. The Al_2Na_2 is an 8-valence electron system. The molecular orbitals that correspond to the 1S and 1P shell of the shell model are occupied in this cluster. In the Al_2Na_3 and Al_2Na_4 clusters, the electrons are taken into the 2S shell containing the Al−Al σ-bonding character before the occupation of the 1D shell.

We tried to investigate the stability of the Al_2Na_5 cluster before the discussion about the stability of the Al_2Na_6 cluster. In the $^2A'$ state (**2-5a**), the 4a′ and 5a′ orbitals corresponded to the 2S shell and one of the 1D shells, respectively, contain the Al−Al σ-bonding and 3s orbital components of the plane Na atoms. The odd electron of the attached Na atom is

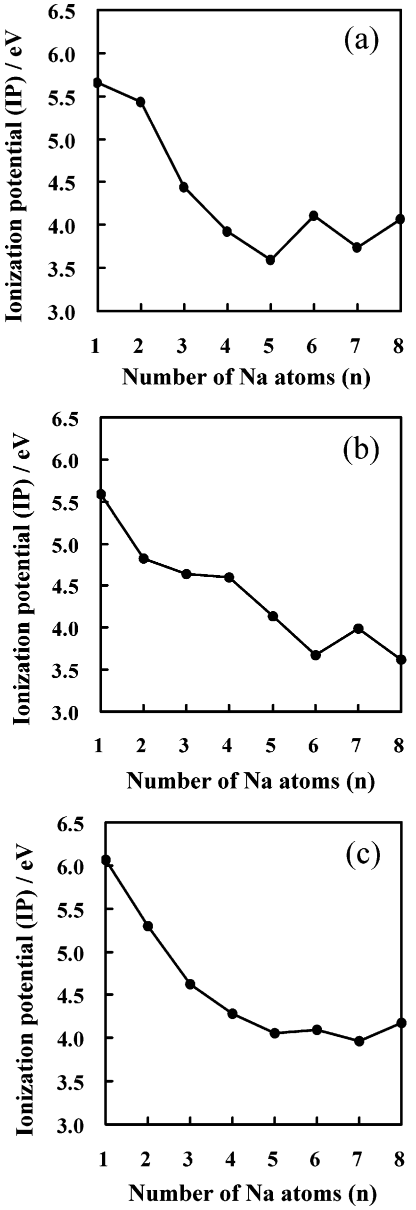


Figure 5. Vertical ionization potentials (IP) of the (a) Al_2Na_n, (b) Al_3Na_n, and (c) Al_4Na_n clusters with the number of Na atoms (*n*).

transferred to the Al−Al bond due to the addition of the Na atom perpendicular to the Al−Al bond (in plane of the Na atoms). The spin density (0.53) of the Al atom supports this electron transfer. The electron transfer might occur in order to form the stable Al−Al part. On the other hand, the paired electrons of the $3a_1$ orbital (2S shell) of **2-5b** contribute to the stability of the Al−Al part due to the addition of the Na atom to the Al atom (on atom). The $4a_1$ orbital (1D shell, SOMO) has the large *3s* character of the four Na atoms. The spin density of 0.26 for the four Na atoms shows that the odd electron is distributed on the plane. It is suggested that this structure becomes stable due to the odd electron delocalized to the four Na atoms. A stable structure is not obtained due to the addition of the Na atom to the Na−Al−Na plane of the Al_2Na_4 cluster. Therefore, two formation cases of the stable structure are found; one is the electron transfer to the Al−Al bond, and another is delocalization of the odd electron to the four Na atoms.

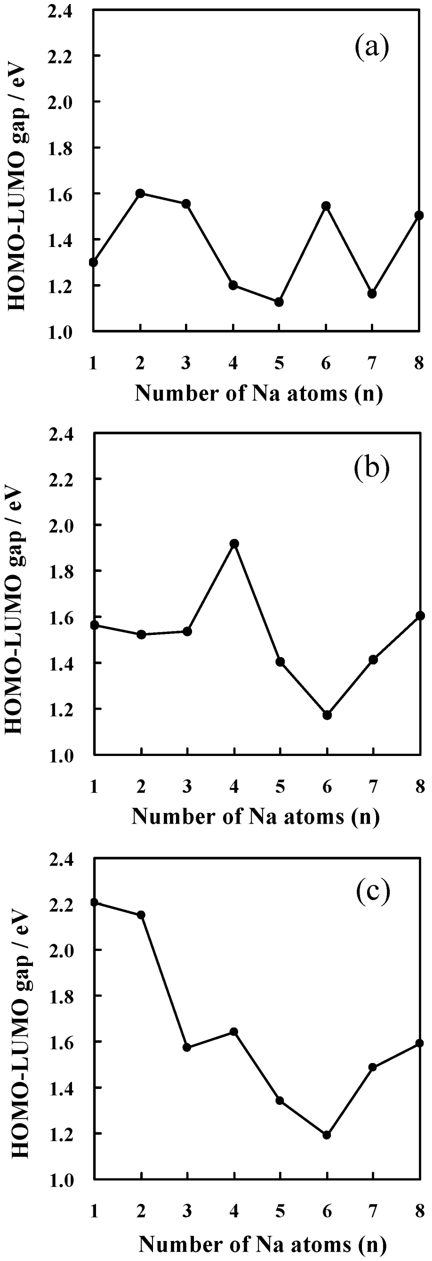


Figure 6. HOMO−LUMO gap of the (a) Al_2Na_n, (b) Al_3Na_n, and (c) Al_4Na_n clusters with the number of Na atoms (*n*).

The Al_2Na_6 cluster is more stable than the neighbor clusters. It is found that both the 4a′ (2S shell) and 5a′(1D shell) are occupied. The 4a′ orbital contains the Al−Al σ-bonding character, and the 5a′ orbital contains the 3s orbital characters of the Na atoms and the Al−Na σ-bonding character of the added Na atom and the Al_2 molecule. That is, both the Al−Al part and the Na plane become stable. In the Al_2Na_8 (**2-8a**) cluster, the $2a_1$ (1P), $3a_1$(2S), and $2b_2$-(1D) orbitals have the *3s* orbital character of the Na atoms in the hexagonal bipyramid shape. It suggested that this cluster becomes stable due to the valence electrons delocalized to the Na atoms. The addition of the Na atom to the Na−Al−Na plane, which forms the stable 1D shell, is found in the Al_2Na_7 and Al_2Na_8 (**2-8b**).

The IP of the Al_3Na_4 in Figure 5(b) is larger than those of the Al_3Na_n ($n = 5-8$) cluster though it is an open shell system. The HOMO−LUMO gap of the Al_3Na_4 cluster is

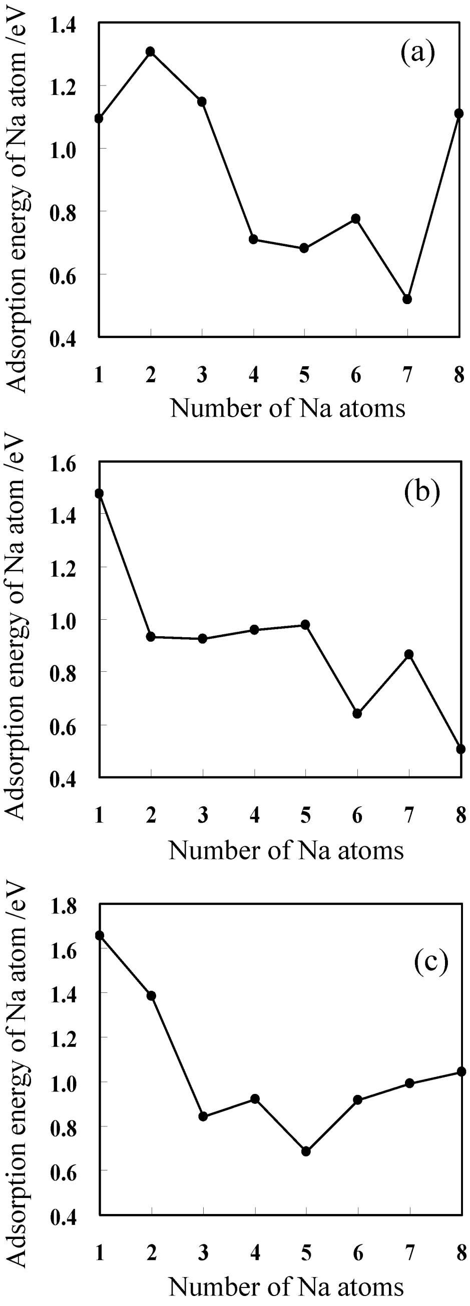


Figure 7. Adsorption energy of the Na atom (E_{Na}) of the (a) Al_2Na_n, (b) Al_3Na_n, and (c) Al_4Na_n clusters with the number of Na atoms (n).

also very large. These results suggest that the C_{3v} symmetry of the Al_3Na_4 anion cluster is stable because the 2a″ and 5a′ of the neutral cluster are degenerated when the anion cluster is formed. The 4a′ orbital corresponding to the 2S shell contains the *3s* orbital characters of the Na atoms and σ-bonding of the Al atoms. The molecular orbitals containing the *3s* and *3p* characters of the Al atoms, from 1a′ to 3a′, are already occupied. In the Al_3Na_n clusters of $n = 6-8$, the 2S shell is more stable than one of the 1P shells and 1D shells because the electrons are distributed to the Na wing. The explanation for the formation of the Na wing is difficult on the basis of the jellium. In the monometallic cluster, i.e., Na_n and Al_n, the electron configuration of 1S, 1P, 1D, and 2S is found based on this model. In the Al−Na cluster, the 1S and 1P shells containing the *3s* or *3p* orbital character of the Al atom become stable due to the electron transfer from the Na atom. The 2S shell containing the *3s* orbital of the Na atom and one of the *3p* orbitals of the Al atom is also stable. It is considered that the formation of the sphere-shaped 2S shell is difficult due to the contribution of the *3p* orbital of the Al atom. The same tendency is found in the 2S shell of the Al_2Na_n ($n = 1-8$) cluster. The shape of the 2S shell might be changed to spherical in the case of the large Al−Na cluster containing a large number of Na atoms. After the Na atoms attach to the Al−Al bond of the Al_3 part, the addition of the Na atom on the Al−Na−Al plane, which forms the 1D shell or sphere-shaped 2S shell, does not occur because the odd electron of the Na atom added to the Na−Na bond is transferred to the Al−Al bond, for example, the $^2A'$ state (**2-5a**) of the Al_2Na_5 cluster.

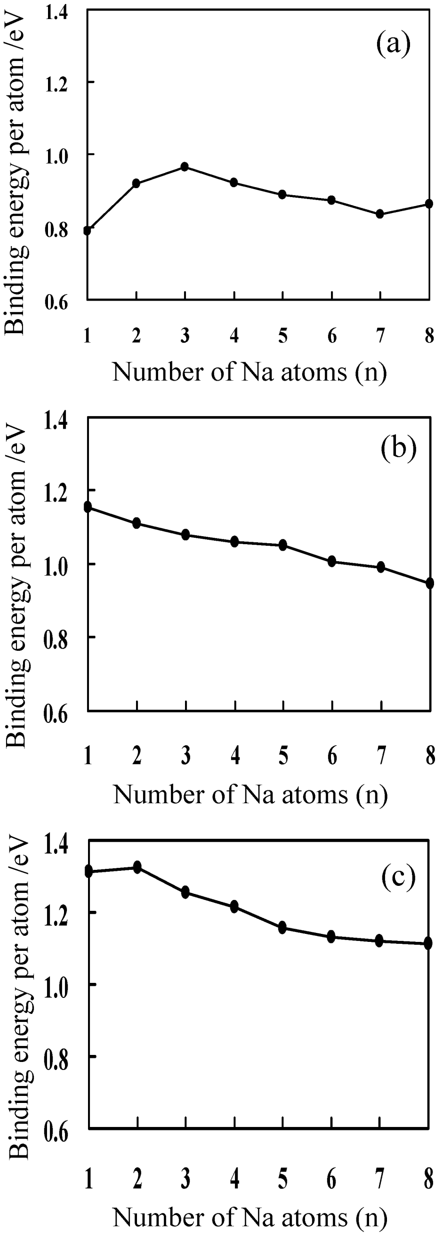


Figure 8. Binding energy per atom (E_b) of the (a) Al_2Na_n, (b) Al_3Na_n, and (c) Al_4Na_n clusters with the number of Na atoms (n).

There are two characteristics of the IP of the Al_4Na_n clusters in Figure 5(c). One is the monotonic decrease for *n*

< 5, and another is the small alternation for $n \geq 5$. The HOMO−LUMO gap of the Al_4Na_n clusters clearly shows the stability of the cluster. The Al_4Na_2, Al_4Na_4, and Al_4Na_8 clusters are relatively stable. In the Al_4Na_2 cluster, the HOMO is the 2S shell. Recently, the stability of the Al_4Na_4 cluster is described as a metalloaromatic compound. The character of the 2S shell is not clear in the Al_4Na_4 cluster. The 3a′ and 5a′ orbitals include the 2S shell character. The Al_4Na_8 cluster is stable because this cluster is a 20-valence electron system, which produces a stable structure based on the jellium model. In this cluster, the 2S shell (4a′) becomes more stable than the 1D shell as in the Al_mNa_n ($n = 2,3$; $n = 1-8$), and the five 1D shells are occupied. The contribution of the *3p* orbital of the Al atom in the Al_4Na_n cluster is different from those of the Al_2Na_n and Al_3Na_n clusters. It is suggested that the hybridized *3p* orbitals of the stereoscopic Al_4 structure produce the spherical shape of the 2S shell.

The stability of the Al_2Na_6, Al_3Na_4 ($Al_3Na_4^-$), and Al_4Na_4 clusters is explained on the basis of the monovalence electron system as in the $AlNa_7$ cluster.[2,3] On the other hand, an analysis of the stabilities for the Al_2Na_2 and Al_4Na_8 clusters are impossible because they have 4 and 12 electrons based on the monovalence system, respectively. They have 8- and 20-valence electrons based on the trivalence electron system per Al atom. Neither the monovalence system nor trivalence system explains the stability of the Al_2Na_8, Al_3Na_7, and Al_4Na_2 clusters. We will try to investigate the stability of the Al−Na clusters, especially the Al_m system, based on the stability and distribution of the 2S shell containing the *3s* orbital of the Na atom and the *3p* orbitals of the Al atom.

4. Conclusion

The formation process of the Al_mNa_n ($m = 2-4$; $n = 1-8$) was investigated using the B3LYP method with the 6-311G* basis sets. The stable structures of the Al_mNa_n ($m = 2-4$; $n = 1-8$) clusters were discussed on the basis of the assumption that the Na atoms adsorbed on the Al_mNa_{n-1} clusters because the Al_mNa_{n-1} part was remained in most of the stable structures of the Al_mNa_n cluster. The attachment of the Na atom first occurs for the Al−Al bond (or Al plane) in the formation of the Al_mNa_n cluster. The odd electron of the attached Na atom is transferred to the Al_m part for $n \leq 4$ or 5 in the Al_mNa_n ($m = 2-4$) clusters since the Al_m part becomes stable. The Al_2Na_4 structure, in which the molecular orbitals formed by the component of the Al atoms are doubly occupied, is maintained for $n \geq 5$. The stable Al_3Na_4 structure, in which the Al−Al bond is saturated by the Na atoms, is also maintained for $n \geq 5$. In the Al_4Na_n clusters, the Na atoms are attached to the Al−Al bond, and the Al_4 plane maintained the planar Al_4 structure for $n \leq 4$. The formation of the in-plane Na−Na bond started from $n \geq 5$ or 6 for the Al_mNa_n ($m = 2-4$) clusters. The Al_2Na_5, Al_3Na_5, Al_3Na_6, Al_3Na_7, and Al_3Na_8 clusters have a characteristic structure. The Na wing is formed in the Al_3Na_n ($n = 5-8$) clusters. The stable 2S shell containing the *3s* orbital of the Na atom and *3p* orbital of the Al atom is formed in $n \geq 5$ or 6 for the Al_mNa_n ($m = 2-4$) clusters. It is considered that the 2S shell becomes stable before the occupation of the 1D shell because the electrons are delocalized on the Na plane.

References

(1) Nakajima, A.; Hoshino, K.; Naganuma, T.; Sone, Y.; Kaya, K. *J. Chem. Phys.* **1991**, *95*, 7061−7066.

(2) Dhavale, A.; Shah, V.; Kanhere, D. G. *Phys. Rev. A* **1998**, *57*, 4522−4527.

(3) Zope, R. R.; Blundell, S. A.; Guet, C.; Baruah, T.; Kanhere, D. G. *Phys. Rev. A* **2001**, *63*, 043202−1−043202−8.

(4) Dhavale, A.; Kanhere, D. G.; Blundell S. A.; Zope, R. R. *Phys. Rev. B* **2001**, *65*, 085402−1−085402−9.

(5) Ishikawa, Y. *Chem. Phys. Lett.* **1993**, *213*, 527−530.

(6) Chacko, S.; Deshpande, M.; Kanhere, D. G. *Phys. Rev. B* **2001**, *64*, 155409−1−155409−6.

(7) Kumar, V. *Phys. Rev. B* **1998**, *57*, 8827- 8829.

(8) Khanna, S. N.; Rao, B. K.; Jena, P. *Phys. Rev. B* **2002**, *65*, 125105−1−125105−5.

(9) Matsuzawa, H.; Hanawa, T.; Suzuki, K.; Iwata, S. *Bull. Chem. Soc. Jpn.* **1992**, *65*, 2578−2587.

(10) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(11) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(12) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(13) Woon, D. E.; Duning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358−1371.

(14) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, H. B.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, D. J.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *GAUSSIAN 98, Revision A.7 ed.*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(15) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, H. B.; 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, B.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.;

Liashenko, A., Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *GAUSSIAN 03, Revision B.03 ed.*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(16) Cheng, H.-P.; Barnett, R. N.; Landman, U. *Phys. Rev. B* **1993**, *48*, 1820−1824.

(17) Rao B. K.; Jena, P. *J. Chem. Phys* **2000**, *113*, 1508−1513.

(18) Pettersson, L. G. M.; Bauschlicher, Jr., C. W.; Halicioglu, T. *J. Chem. Phys* **1987**, *87*, 2205−2213.

(19) Zhan, C-G.; Zheng, F.; Dixon, D. A. *J. Am. Chem. Soc.* **2002**, *124*, 14795−14803.

(20) Shetty, S.; Kanhere, D. G.; Pal, S. *J. Phys. Chem. A* **2004**, *108*, 628−631.

(21) Boldyrev, A. I. A.; Kuznetsov, E. *Inorg. Chem.* **2002**, *41*, 532−537.

CT600339U

JCTC Journal of Chemical Theory and Computation

Assessment of Approximate Density Functional Methods for the Study of the Interactions of Al(III) with Aromatic Amino Acids

E. Rezabal,*,† T. Marino,‡ J. M. Mercero,† N. Russo,‡ and J. M. Ugalde†

Kimika Fakultatea, Euskal Herriko Unibertsitatea and Donostia International Physics Center (DIPC), P. K. 1072, 20080 Donostia, Euskadi, Spain, and Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite - Centro d'Eccellenza MIUR, Università della Calabria, I-87030 Arcavacata di Rende (CS), Italy

Received February 2, 2007

Abstract: Four approximate Density Functional Theory methods, the standard hybrid B3LYP functional, the hybrid mPW1PW91 functional designed to account for van der Waals forces, the one-parameter meta hybrid TPSSh functional, the general-purpose meta hybrid MPWB1K functional and one Molecular Orbital Theory method, the standard Moller−Plesset perturbation theory up to second-order MP2, have been assessed for studying the complexation modes of the highly acidic Al(III) cation with the three aromatic amino acids, phenylalanine (Phe), tyrosine (Tyr), and tryptophan (Trp). Based on their performance toward the prediction of the geometrical structure of a number of lowest energy isomers and their relative binding energies, it is concluded that the B3LYP approximate functional renders the desired accuracy at the minimum computational cost.

1. Introduction

We concur with others[1] that Density Functional Theory (DFT) implementations are the most promising ab initio quantum mechanical methods for the computational study of large compounds, in general, and biologically relevant structures in particular.[2] However, it should be pointed out that DFT methods constitute a family of methods rather than a single method. Although Perdew's *Jacob's ladder* approach enables rationalizing the quality of the various DFT methods, a precise prescription to assess the quality of a given DFT method is still lacking. This is why given a particular system, selecting the most appropriate DFT method is so difficult and why validation and assessment of the various approximate DFT methods remain crucial for the reliability of the obtained data.

For biological systems, validation of the theoretical methods can be made by comparison of the calculated data against experimental data for a collection of *molecules containing atoms* commonly found in biomolecules.[2−4] Alternatively, one could also carry out the validation by selecting a number of relevant biomolecules themselves and then comparing the calculated DFT data against other accurate molecular structure methods.[5,6]

Our primary interest here is to substantiate the selection of a computationally cost-effective and accurate enough DFT method for the investigation of the aluminum metalloproteins. The relevance of aluminum in protein environments relies on the fact that aluminum, the most abundant metal on Earth−about 8% of the Earth's crust, is known to be toxic in biological environments. Thus, it is highly toxic to plant roots,[7] especially under acidic soil conditions, and it has also been related to several neurological disorders.[8] Since the bioavailability of aluminum has increased considerably, due to human intervention,[9,10] interest toward its biochemistry has increased recently.[11] However, despite this research effort, the molecular basis of the toxicity of aluminum are

* Corresponding author e-mail: pobrease@sc.ehu.es.

† Euskal Herriko Unibertsitatea and Donostia International Physics Center.

‡ Università della Calabria.

10.1021/ct700027n CCC: $37.00
Published on Web 07/03/2007

Table 1. Number of Isomers Characterized for the Aluminum Aromatic Amino Acids Complexes

	non cation-π				cation-π			
	bidentate					bidentate		
				monodentate	tridentate			
	N/O	N/OH	O/O	O	N/O/ring	N/ring	O/ring	total
Phe	2	1	1	1	1		2[a]	8
Tyr	2	1	1	1	1	1	2[a]	9
Trp	2				3[b]	2[c]	4[d]	11

[a] One isomer is charge solvated (cs) and the other is zwitterionic (zw). [b] Two isomers have covalent Al−ring bonding. [c] One isomer has covalent Al−ring bonding. [d] Two isomers are charge solvated (cs) and two isomers are zwitterionic (zw).

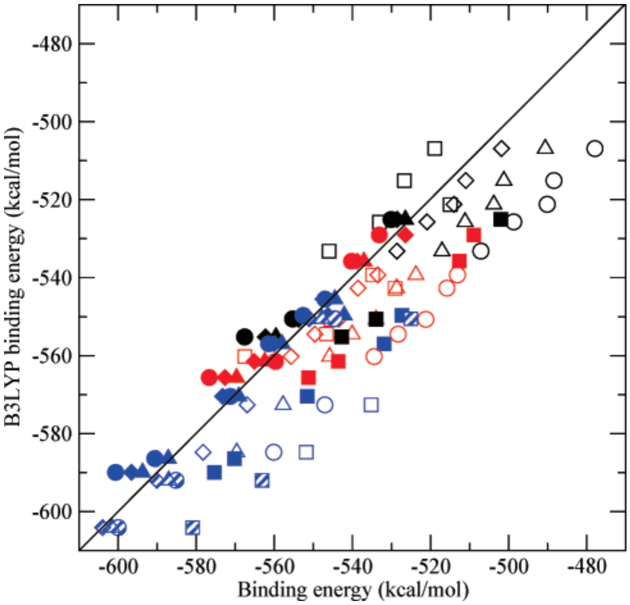


Figure 1. B3LYP binding enthalpies are presented on the *y*-axis, and binding enthalpies obtained by other methods on the *x*-axis, in kcal/mol. Black symbols regard Phe complexes, while red and blue symbols stand for Tyr and Trp ones, respectively. Squares represent MP2 results, triangles the mPW1PW91, circles the MPWB1K, and diamonds the TPSSh. Finally, empty figures symbolize the complexes without cation-π interaction and the filled ones the complexes with cation-π interaction. The striped blue symbols regard the Trp structures where the metal interacts covalently with the ring

still largely unknown. Consequently, a detailed investigation of the interactions of aluminum with amino acid residues constitutes an important piece of information that could help unveiling the behavior of the metal in protein environments.

The intricate three-dimensional structures that metalloproteins assume are largely determined by a delicate balance of a myriad of interactions between the residues and the protein's backbone with the metal.[12] Proteins containing aromatic amino acids rank high on the complexity scale of these interactions for they foster, in addition to ordinary covalent, ionic, and charge-transfer interactions, cation-π interactions between the metal and the aromatic part of their residues. Therefore, the aromatic amino acids (AAA) are well suited for assessing the reliability of a method for metalloprotein studies. Consequently, selection of a particular level of theory for reliable studies on these complex systems is more subtle than for the metal−amino acid complexes bearing less variety of interactions. Even more, experimental data concerning the binding energies of AAAs and cations, like Na$^+$, Li$^+$, and K$^+$,[13−16] and late transition metals,[17−30] Ag$^+$, Cu$^+$, and Zn^{2+}, can be found in the literature, suggesting that sensible data for aluminum could also be obtained similarly for the benefit of a reliable assessment. These amino acids, phenylalanine (Phe), tyrosine (Tyr), and tryptophan (Trp), account for 8.4% of the amino acids in proteins,[31] and the 26% of all Trp residues are known to be involved in energetically significant cation-π interactions.[32] Protein Data Bank (PDB) research has revealed the cation-π interaction to be widespread, also within the proteins.[32,33]

On the other hand, it is worth noticing that cation-π interactions in protein environments have been extensively studied for metals other than aluminum.[34] Thus, the open chemical literature contains numerous studies that underline their importance for the stabilization of the protein's geometry.[35−37] Several studies have also been devoted to investigate the nature of the cation-π interactions. Current consensus suggests that they are dominated by electrostatic forces[32,38,39] and cation-induced polarization terms,[40] which correlate with the magnitude of the quadrupole moment of the aromatic ring and the molecular polarizability of the aromatic compounds, respectively.[41−43]

Aluminum is expected to interact predominantly with the most electronegative parts of the AAAs, namely, the aromatic ring of the side chain and the carboxylate oxygens and the N atom of the backbone, giving rise to several isomeric structures, some of which present cation-π interactions and some of which do not. Finding a method that predicts accurately the relative energies between all these isomers is crucial for the subsequent biochemical interpretation. Consequently, our goal for the present study is to assess the reliability of a number of computational implementations of DFT for predicting reliably the relative stabilities of aluminum metalloproteins.

2. Methods

Among the long list of approximate density functionals, we have selected four. In first place comes the venerable B3LYP hybrid DFT approximate functional[44] which consists of the B3 exchange functional,[45] the LYP correlation functional,[46] and a 20% of exact exhange. It has already been well-established that this density functional implementation gives excellent results for most chemical systems[47] including cation-π interactions.[30,37,48−50] Second, we will consider the mPW1PW91 functional of Adamo and Barone.[51] This

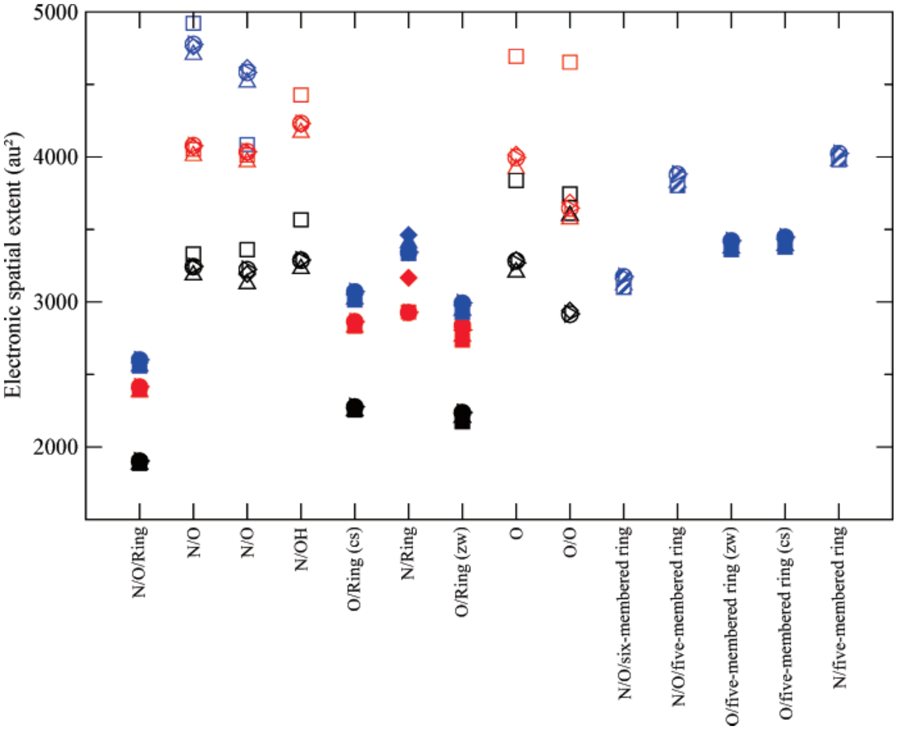


Figure 2. The electronic spatial extent as a function of the aluminum coordination mode for the three aromatic amino acids, black (Phe), red (Tyr), and blue (Trp). Empty and filled symbols represent complexes without and with intramolecular cation-π interaction, respectively. Squares stand for MP2, up triangles for mPW1PW91, circles for MPWB1K, diamonds for TPSSh, and right triangles for B3LYP levels of theory. The striped symbols stand for the Trp−aluminum complexes having a covalent bond between the metal on one aromatic carbon atom.

functional, which was specifically designed to account for van der Waals interactions, has been found to give excellent results, as confronted with experimental data, for molecules with intramolecular cation-π interactions.[29,52,53] Third, we have selected the promising hybrid meta approximate functional of Tao, Perdew, Staroverov, and Scuseria[54] which has been reported to provide highly accurate descriptions of a number of diverse systems and their properties.[55] Fourth, we have chosen to consider the MPWB1K hybrid *meta* approximate functional of Zhao and Thruhlar,[56−58] as a representative of the new-generation general-purpose functionals for applications in thermochemistry, kinetics, and noncovalent interactions.[3]

Additionally, we have also carried out full quantum mechanical analysis of the interactions between aluminum and the aromatic amino acids at the MP2 level of theory for it has been reported that it yields very satisfactory agreement with available spectroscopic experimental data of many biological molecules.[59]

All the calculations carried out in this research were performed with the Gaussian 03 code.[60] The structures were fully optimized at the B3LYP, mPW1PW91, TPSSh, and MP2 levels of theory, using the standard all-electron 6-31+G(d,p) basis for the aluminum ion, and the compact effective core potentials and shared-exponent basis set of Stevens, Basch, Krauss, and Jasien (SBKJ)[61] for C, N, O, and H. Gresh et al.[62−64] found that this pseudopotentials/all-electron basis set combination for the ligand and the metal cation, respectively, represents a very well balanced compromise between accuracy and computational efficiency. This method has been widely used in our group and has shown to be adequate for this type of calculations.[65−70] This basis set will hereafter be referred to as SBKJ/*+. Subsequent vibrational frequency analysis confirmed that structures were stable minima on their corresponding potential energy surface.

Single point calculations, at the optimized geometries, were carried out, with the considerably larger 6-311++G(2df,2p) basis set, in order to improve the binding energies of the species considered, for the B3LYP, mPW1PW91, TPSSh, and MP2 levels of theory. For the MPWB1K approximate functional persistent geometry convergence problems were encountered. Consequently, at this level of theory single point calculation with the 6-311++G(2df,2p) basis set were carried out at the optimized B3LYP geometries. Nevertheless, it is worth mentioning that for all cases in which geometry convergence was achieved with the MPWB1K functional, the resulting structures were found to be remarkably similar to their corresponding optimized B3LYP ones. In the present work, the binding energy between an aromatic amino acid (AAA) and the aluminum cation is defined as the enthalpy change (ΔH) of the following process:

$$\mathrm{Al}^{3+} + \mathrm{AAA} \rightarrow (\mathrm{Al} - \mathrm{AAA})^{3+} \qquad (1)$$

The measurement of the performance of the various levels of theory for the structural characterization task will be carried out by the analysis of the electronic spatial extent of the optimized geometry of all the properly characterized structures. The electronic spatial extent is defined as the

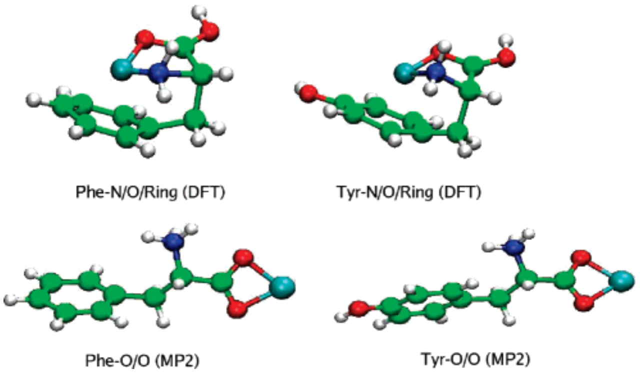


Figure 3. The most stable isomer of the Phe−Al $^{+3}$ and Tyr−Al^{+3} complexes at the three B3LYP, mPW1PW91, MPWB1K, and TPSSh DFT levels of theory, top panel, and the most stable isomer of the Phe−Al^{+3} and Tyr−Al^{+3} complexes at the MP2 level of theory, bottom panel.

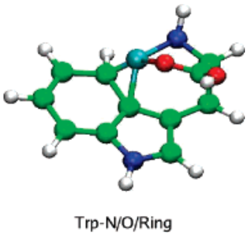


Figure 4. The most stable isomer of the Trp−Al^{+3} complex

expectation value, $\langle r^2 \rangle$, of the square modulus of the electronic vector **r** by the following equation

$$\langle r^2 \rangle = \int r^2 \rho(\mathbf{r}) d\mathbf{r} \qquad (2)$$

where $\rho(\mathbf{r})$ is the ground state electron density of the optimized structure.

3. Results and Discussion

We have fully characterized 28 stable minima structures for the aluminum aromatic amino acid complexes. Table 1 shows the rich structural variety of these isomers. In particular, 12 of them bear an intramolecular cation-π interaction. Remarkably, no stable structure was found neither for the N/OH or O/O bidentate nor for the O monodentate bonding modes of aluminum with tryptophan. Aluminum interacting with Trp is very prone to form bidentate complexes, where one of the interactions corresponds to a cation-π interaction, and the other one to the charge-transfer interaction with the carboxylic oxygen atom. Notice that of the four isomers found with this bonding mode, two are charged solvated, while the remaining two are charge separated zwitterionic like complexes. Additionally, Trp has another remarkable feature when interacting with aluminum. Namely, when the aminic nitrogen and the aromatic ring lie in the coordination sphere of aluminum, the metal can bind covalently one of the carbon atoms of the aromatic ring. Both the tridentate N/O/ring and the bidentate N/ring bonding modes of Trp show this feature.

Phenylalanine and tyrosine behave quite similarly in the metalation process by aluminum ion, as observed from Table 1. The only salient difference is that Phe does not form an N/ring bidentate cation-π complex with aluminum. Tryptophan behaves distinctively with respect to the remaining AAAs toward aluminum complexation.

The binding energies, as defined in eq 1, obtained with all five methods assessed for all the 28 structures have been arranged in Figure 1. We have chosen to represent the B3LYP binding energies on the *y*-axis and the mPW1PW91, MPWB1K, TPSSh, and MP2 binding energies on the *x*-axis.

Regarding the binding energies of the structures without cation-π intramolecular interaction (empty symbols in Figure 1) one can conclude that all four selected DFT approximate functionals behave similarly. It is worth noting that the empty triangles (mPW1PW91 binding energies), empty circles (MPWB1K binding energies), and emptly diamonds (TPSSh binding energies) lie parallel to the diagonal. Namely, mPW1PW91, MPWB1K, and TPSSh predict binding energies similar to B3LYP, and their relative binding energies are also predicted to be similar, for all 12 aluminum complexes with the three aromatic amino acids with no intramolecular cation-π interactions. Comparison with MP2 is also very satisfactory for all the structures except for the two N/O bidentate complexes of Trp. Observe that the blue empty squares of Figure 1 lie far off the diagonal.

The electronic spatial extents as measured by the expectation value, $\langle r^2 \rangle$, of the square modulus of the electronic vector **r** of the 28 structures characterized at the five levels of theory assessed are shown in Figure 2. Inspection of the figure reveals three salient points. First, the MP2 optimum geometry of the least stable N/O bidentate Trp−aluminum complex differs substantially from the three, B3LYP, mPW1PW91, and TPSSh, optimum DFT geometries. Observe Figure 2, where the right side blue empty square lies below the blue triangle, circle, and diamond. Second, the MP2 optimum geometries of both the O/O bidentate and the monodentate Phe−aluminum complexes are also different with respect to their corresponding B3LYP, mPW1PW91, and TPSSh optimum DFT geometries. Third, similarly, the MP2 optimum geometries of both the O/O bidentate and the monodentate Tyr−aluminum complexes differ with respect to their DFT counterparts.

Inspection of Figure 2 reveals one-fourth salient point. Namely, that the optimum geometries of the complexes bearing an intramolecular cation-π interaction are very similar for all five methods assessed. In particular we would like to point out the remarkable similarity among the optimum DFT structures. This is reflected also in the calculated binding energies. Observe how close to the diagonal (B3LYP binding energies) lie in Figure 1 the filled triangles (mPW1PW91 binding energies), the circles (MPWB1K binding energies), and the diamonds (TPPSh binding energies). Nevertheless, it is worth noticing that the N/ring bidentate Tyr−aluminum complex was not found within the mPW1PW91 level of theory.

The optimized MP2 geometries agree well with the B3LYP ones for the aluminum aromatic amino acid complexes characterized by the presence of cation-π interaction. Although the B3LYP overbinds all complexes by some 20

Table 2. Mean Absolute Deviations from the B3LYP Results of the Binding Energies $\epsilon^{\Delta H}$, in kcal/mol, and of the Electronic Spatial Extension $\epsilon^{\langle r^2\rangle}$, in au^2 [a]

	mPW1PW91		MPWB1K		MP2		TPSSh	
	$\epsilon^{\Delta H}$	$\epsilon^{\langle r^2\rangle}$	$\epsilon^{\Delta H}$	$\epsilon^{\langle r^2\rangle}$	$\epsilon^{\Delta H}$	$\epsilon^{\langle r^2\rangle}$	$\epsilon^{\Delta H}$	$\epsilon^{\langle r^2\rangle}$
cation-π	2.52	46.87	5.31	3.65	20.58	45.96	4.06	33.83
no cation-π	15.29	121.19	28.35	2.11	12.86	346.75	4.48	16.92
overall	7.88	78.72	15.18	2.99	17.27	174.87	4.33	25.19

[a] The aluminum−AAA complexes bearing an intramolecular cation−π interaction and for those not bearing it. The overall mean absolute deviations are also given.

kcal/mol, the predicted relative stability of the various isomers is the same for both levels of theory.

Considering Trp complexes where the metal interacts covalently with a C atom of the indole side chain (blue stripped symbols), both mPW1PW91 and MPWB1K functionals predict a slightly smaller (~3 kcal/mol and ~7 kcal/mol, respectively) binding energy with respect to the B3LYP mark. The TPSSh binding energies, on the other hand, lie within ~0.5 kcal/mol with respect to the corresponding B3LYP marks. The MP2 optimized geometries for these complexes are similar to the DFT ones. However, the MP2 binding energies are ~30 kcal/mol smaller than the corresponding B3LYP values, although the predicted relative stability remains unaltered.

One final general observation is that isomers bearing a cation-π interaction are more stable than the ones that do not, irrespective of the DFT level of theory. The relative stability of the formers is slightly greater at mPW1PW91, MPWB1K, and TPSSh levels of theory, as compared with B3LYP. This difference is minimum in the later case. This feature has been reported by Dunbar[52] for monovalent Na^+, Mg^+, Al^+, and the first row transition-metal cations.

Regarding MP2 results, the preference for the cation-π bearing structures depends on the actual amino acid considered. For example, for both Phe− and Tyr−aluminum complexes, MP2 favors the bidentate O/O coordination mode for the aluminum shown in Figure 3. Conversely, all the three DFT approximate methods considered in the present study predict a tridentate N/O/ring structure for the lowest energy isomer. It is worth emphasizing that it is the latter coordination mode, the tridentate one, the one that has been observed in previous structural experimental characterizations[29,30,50] of complexes between aromatic amino acids and Li^+, Na^+, K^+, Cu^+, and Ag^+ cations. The better agreement with experiments carried out with related complexes sheds some confidence on the improved performance of the DFT methods over MP2 for these aluminum complexes. Additionally, we have carried out CCSD(T) single point calculations for these two Al(III)−Phe structures which have conformed the higher stability of the N/O/ring coordination mode with respect to the O/O coordination mode.

Instead, for aluminum−Trp complexes all four methods assessed concur that the lowest energy structure bears a cation-π interaction. The most stable conformation, for all the theory levels studied, is predicted to be the tridentate N/O/ring complex where the aluminum interacts covalently with one of the carbon atoms of the indole six-membered ring depicted in Figure 4

4. Conclusions

Table 2 shows the mean absolute deviations calculated for the two properties scrutinized in the present study for all the 28 isomers characterized as stable minima for the complexes involving aluminum and the aromatic amino acids. The reference data set has been chosen to consist of the B3LYP results. It is observed that for the relative binding energies the DFT methods scrutinized give similar and satisfactory deviations, in particular for the complexes bearing an intramolecular cation-π interaction. Additionally, TPSSh does very much like B3LYP also for the complexes not bearing cation-π interactions. The MP2 results deviate four times as much than TPSSh, five times as much than MPWB1K, and ten times as much than mPW1PW91 for the binding energy of the complexes bearing an intramolecular cation-π interaction.

The deviation for the electronic spatial extent is smaller for complexes bearing an intramolecular cation-π interaction than for those that do not. Although this deviation might appear large, it is worth noticing that, in the worst case, the electronic spatial extent for the structures without a cation-π interaction, $\epsilon^{\langle r^2\rangle} = 346$ au^2, deviates less than 10% from the mean electronic spatial extent, ~3500 au^2.

Therefore, we can conclude that for aluminum aromatic amino acid complexes no significant improvement is gained by using neither mPW1PW91, MPWB1K, or TPSSh nor MP2, as seen in the literature for other related systems.[30] Besides, geometry optimization and frequency calculations at the MP2 level of theory requires a considerable computational effort, and MPWB1K suffers from poor geometry convergence. Consequently, both B3LYP and TPSSh methods were concluded to be a good compromise between cost and accuracy for the study of aluminum−AAA complexes, involving isomers with and without cation-π interactions.

Acknowledgment. This research was funded by Euskal Herriko Unibertsitatea (the University of the Basque Country), Gipuzkoako Foru Aldundia (the Provincial Government of Guipuzkoa), and Eusko Jaurlaritza (the Basque Government). The SGI/IZO−SGIker UPV/EHU (supported by the National Program for the Promotion of Human Resources within the National Plan of Scientific Research, Development and Innovation − Fondo Social Europeo and MCyT) is greatfully acknowledged for assistance and generous allocation of computational resources. Financial support from the Universitá degli Studi della Calabria and Regione Calabria (POR Calabria 2000/2006, misura 3.16, progetto PROSICA) is also acknowledged. Authors wish to thank Dr. A. Rimola

(Universitat Autonoma de Barcelona) for his critical reading of the manuscript.

Supporting Information Available: Optimized geometries of all characterized stable isomers as well as their relative energies at the five levels of theory. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062201.

(2) Riley, K. E.; Holt, B. T. O.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 407.

(3) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

(4) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.

(5) Dudev, T.; Lim, C. *J. Phys. Chem. B* **2000**, *104*, 3692.

(6) Johnson, E. R.; Becke, A. D. *Chem. Phys. Lett.* **2006**, *432*, 600.

(7) Williams, R. *Coord. Chem. Rev.* **2002**, *228*, 93.

(8) Yokel, R. *Neurotoxicology* **2000**, *21*, 813.

(9) Exley, C.; Esiri, M. M. *J. Neurol. Neurosurg. Psychiatry* **2006**, *77*, 877.

(10) Exley, C.; Begum, A.; Woolley, M. P.; Bloor, R. N. *Am. J. Med.* **2006**, *119*, 276.

(11) Exley, C. *J. Inorg. Biochem.* **2005**, *99*, 1747.

(12) Ruan, C.; Yang, Z.; Hallowita, N.; Rodgers, M. T. *J. Phys. Chem. A* **2005**, *109*, 11539.

(13) Gapeev, A.; Dunbar, R. *J. Am. Chem. Soc*. **2001**, *123*, 8360.

(14) Gapeev, A.; Dunbar, R. *Int. J. Mass Spectrom.* **2003**, *227*, 825.

(15) Kish, M. M.; Ohanessian, G.; Wesdemiotis, C. *Int. J. Mass Spectrom.* **2003**, *227*, 509.

(16) Ruan, C.; Rodgers, M. *J. Am. Chem. Soc.* **2004**, *126*, 14600.

(17) Cerda, B.; Wesdemiotis, C. *J. Am. Chem. Soc.* **1995**, *117*, 9734.

(18) Hoyau, S.; Ohanessian, G. *J. Am. Chem. Soc.* **1997**, *119*, 2016.

(19) Gatlin, C. L.; Turecek, F.; Vaisar, T. *J. Mass Spectrom.* **1995**, *30*, 1605.

(20) Gatlin, C. L.; Turecek, F.; Vaisar, T. *J. Am. Chem. Soc.* **1995**, *117*, 3637.

(21) Wen, D.; Yalcin, R.; Harrison, A. G. *Rapid Commun. Mass Spectrom.* **1997**, *8*, 749.

(22) Lei, Q.; Amster, I. J. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 722.

(23) Yalcin, T.; Wang, J.; Wen, D.; Harrison, A. G. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 749.

(24) Lavanant, H.; Hoppilliard, Y. *J. Mass Spectrom*. **1997**, *32*, 1037.

(25) Lavanant, H.; Hecquet, E.; Hoppilliard, Y. *Int. J. Mass Spectrom.* **1999**, *185/186/187*, 11.

(26) Talaty, E. R.; Perera, B. A.; Gallardo, A. L.; Barr, J. M.; Stipdonk, M. J. V. *J. Phys. Chem. A* **2001**, *105*, 8059.

(27) Shoeib, T.; Cunje, A.; Hopkinson, A.; Siu, K. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 408.

(28) Shoeib, T.; Siu, K. M.; Hopkinson, A. C. *J. Phys. Chem. A* **2002**, *106*, 6121.

(29) Polfer, N. C.; Oomens, J.; Moore, D. T.; Helden, G.; Meijer, G.; Dunbar, R. *J. Am. Chem. Soc.* **2006**, *128*, 517.

(30) Polfer, N.; Oomens, J.; Dunbar, R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2744.

(31) Meadows, E.; De Wall, S.; Barbour, L. J.; Gokel, G. W. *J. Am. Chem. Soc.* **2001**, *123*, 3092.

(32) Gallivan, J.; Dougherty, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459.

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

(34) De Wall, S. L.; Meadows, E. S.; Barbour, L. J.; Gokel, G. W. *Proc. Natl. Acad. Sci.* **2000**, *97*, 6271.

(35) Ma, J.; Dougherty, D. *Chem. Rev.* **1997**, *97*, 1303.

(36) Gokel, G. W.; DeWall, S. L.; Meadows, E. S *Eur. J. Org. Chem*. **2000**, 2967.

(37) Rimola, A.; Rodríguez-Santiago, L.; Sodupe, M. *J. Phys. Chem. B* **2006**, *110*, 24189.

(38) Mecozzi, S.; West, A.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10566.

(39) Dougherty, D. A. *Science* **1996**, *271*, 163.

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

(41) Garau, C.; Frontera, A.; Quinonero, D.; Ballester, P.; Costa, A.; Deya, P. M. *J. Phys. Chem. A* **2004**, *108*, 9423.

(42) Garau, C.; Frontera, A.; Quinonero, D.; Ballester, P.; Costa, A.; Deya, P. M. *Chem. Phys. Lett.* **2004**, *392*, 85.

(43) Garau, C.; Frontera, A.; Quinonero, D.; Ballester, P.; Costa, A.; Deya, P. M. *Chem. Phys. Lett.* **2004**, *399*, 220.

(44) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(45) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(46) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(47) Mercero, J. M.; Matxain, J. M.; Lopez, X.; York, D. M.; Largo, A.; Eriksson, L. A.; Ugalde, J. M. *Int. J. Mass Spectrom.* **2005**, *240*, 37.

(48) Reddy, A. S.; Sastry, G. N. *J. Phys. Chem. A* **2005**, *109*, 8893.

(49) Zhang, S.; Liu, L.; Fu, Y.; Guo, Q. *J. Mol. Struct.* **2005**, *757*, 37.

(50) Ryzhov, V.; Dunbar, R. C.; Cerda, B.; Wesdemiotis, C. *Am. Soc. Mass Spectrom.* **2000**, *11*, 1037.

(51) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.

(52) Dunbar, R. C. *J. Phys. Chem. A* **2002**, *106*, 7328.

(53) Oliveira, G.; Martin, J.; Proft, F.; Geerlings, P. *Phys. Rev. A* **1999**, *60*, 1034.

(54) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(55) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

(56) Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 19046.

(57) Zhao, Y.; Thrular, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(58) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701.

(59) Gerber, R. B.; Chaban, G. M.; Gregurick, S. K.; Brauer, B. *Biopolymers* **2003**, *68*, 370.

(60) 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. G*aussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(61) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612.

(62) Garmer, D. R.; Gresh, N. *J. Am. Chem. Soc.* **1994**, *116*, 3556.

(63) Gresh, N.; Stevens, W. J.; Krauss, M. *J. Comput. Chem.* **1995**, *16*, 843.

(64) Gresh, N.; Garmer, D. R. *J. Comput. Chem.* **1996**, *17*, 1481.

(65) San Sebastian, E.; Mercero, J. M.; Roland H. Stote, A. D.; Cossío, F. P.; Lopez, X. *J. Am. Chem. Soc.* **2005**.

(66) Mercero, J. M.; Fowler, J. E.; Ugalde, J. M. *J. Phys. Chem. A* **1998**, *102* (35), 7006.

(67) Mercero, J. M.; Matxain, J. M.; Rezabal, E.; Lopez, X.; Ugalde, J. M. *Int. J. Quantum Chem.* **2004**, *98*, 409.

(68) Mercero, J. M.; Fowler, J. E.; Ugalde, J. M. *J. Phys. Chem. A* **2000**, *104*, 7053.

(69) Rulised, L.; Vondrasek, J. *J. Inorg. Biochem.* **1998**, *71*, 115.

(70) Mercero, J. M.; Mujika, J. I.; Matxain, J. M.; Lopez, X.; Ugalde, J. M. *Chem. Phys.* **2003**, *295*, 175.

CT700027N

JCTC Journal of Chemical Theory and Computation

Peptide Hydrolysis in Thermolysin: Ab Initio QM/MM Investigation of the Glu143-Assisted Water Addition Mechanism

Jochen Blumberger,*,† Guillaume Lamoureux,*,‡ and Michael L. Klein

Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, 231 S. 34th Street, Philadelphia, Pennsylvania 19104-6323

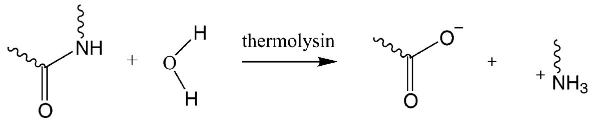
Received March 31, 2007

Abstract: Thermolysin (TLN) is one of the best-studied zinc metalloproteases. Yet the mechanism of action is still under debate. In order to investigate the energetic feasibility of the currently most favored mechanism, we have docked a tripeptide to the active site of TLN and computed the free energy profile at the quantum mechanics/molecular mechanics level of theory. The mechanism consists of three distinct steps: (i) a Zn-bound water molecule is deprotonated by Glu143 and attacks the carbonyl bond of the substrate; (ii) Glu143 transfers the proton to the amide nitrogen atom; (iii) the nitrogen atom is protonated and the peptide bond is irreversibly broken. The free energy barriers for steps i and iii have almost equal heights, 14.8 and 14.7 kcal/mol, respectively, and are in good agreement with the effective experimental activation barrier obtained for similar substrates, 12.1−13.6 kcal/mol. Transition state stabilization for nucleophilic attack is achieved by formation of a weak coordination bond between the substrate carbonyl oxygen atom and the Zn ion and of three strong hydrogen bonds between the substrate and protonated His231 and two solvent molecules. The transition state for the nucleophilic attack (step i) is more tightly bonded than the enzyme−substrate complex, implying that TLN complies with Pauling's hypothesis regarding transition-state stabilization. Glu143, at first unfavorably oriented for protonation of the amide nitrogen atom, displayed large structural fluctuations that facilitated reorganization of the local hydrogen-bond network and transport of the proton to the leaving group on the nanosecond time scale. The present simulations give further evidence that Glu143 is a highly effective proton shuttle which should be assigned a key role in any reaction mechanism proposed for TLN.

1. Introduction

Thermolysin (TLN) is an extracellular zinc endoprotease of bacterial origin that catalyzes peptide-bond hydrolysis specifically on the N-terminus side of large hydrophobic residues (Scheme 1).[1,2]

Scheme 1



The mechanism of action of TLN is assumed to be similar for all families of the thermolysin clan including the important peptidases of higher organisms such as carboxypeptidase A, angiotensin converting enzyme, enkephalinase, collagenase, and neprilysin. The latter, expressed in the

* Corresponding authors. Phone: ++44-(0)1223-763872 (J.B.); 514-848-2424 ext. 5314 (G.L.). fax: ++44-(0)1223-336362 (J.B.); 514-848-2868 (G.L.). e-mail: jb376@cam.ac.uk (J.B.); glamoure@alcor.concordia.ca (G.L.).

† Present address: Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

‡ Present address: Department of Chemistry and Biochemistry, Concordia University, 7141 Sherbrooke Street West, Montréal H4B 1R6, Canada.

10.1021/ct7000792 CCC: $37.00
Published on Web 08/18/2007

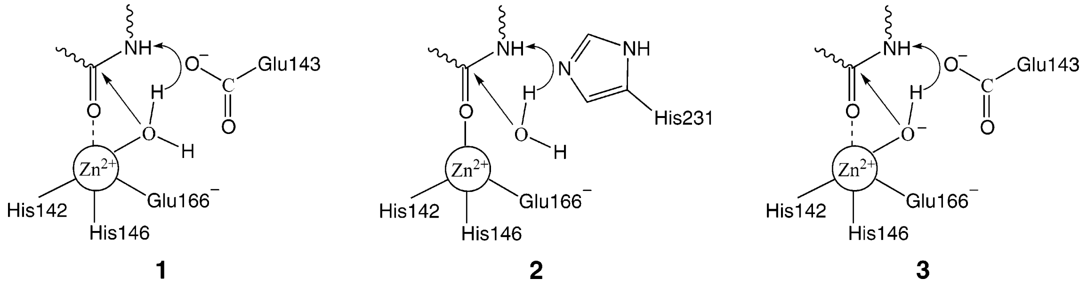


Figure 1. Three mechanisms for peptide bond hydrolysis in thermolysin: (1) Glu143-assisted addition of a water molecule, (2) His231-assisted addition of a water molecule, and (3) Glu143-assisted addition of a hydroxide ion.

kidneys and the central nervous system, is involved in the catabolism of active peptides. Neprilysin inhibitors have attracted great interest as they show antinociceptive and antihypertensive properties.[3] Due to its high thermostability, tolerance to organic solvents, and specifity against hydrophobic residues, TLN is also used in industrial processes, for example, as a catalyst for the hydrolysis and synthesis of a precursor of aspartame and in the continuous synthesis of peptide derivates (see refs in ref 4).

The first crystal structure of TLN was solved in 1972 by Matthews and co-workers,[5] and a number of TLN[6] and TLN-inhibitor structures[7−9] have been reported since then. The Zn^{2+} ion is coordinated with two histidine ligands, His142 and His146; one glutamate residue, Glu166; and one water molecule (see Figure 1). Glu166 can bind monodendate[5] or bidendate,[6] and Zn is accordingly either 4- or 5-fold coordinated. Catalytically active second shell residues include Glu143,[10−13] His231,[10,14] Asp226, and Tyr157.[15] The histidine and glutamate residues form the “HEXXH + E” zinc-binding motif that is conserved within the thermolysin clan. Analyses of TLN-inhibitor complexes have also given information about the residues involved in substrate recognition and stabilization of the enzyme substrate complex (ES).[16]

Despite the wealth of crystallographic[5−9] and kinetic[10−15,17−19] data, the reaction mechanism for thermolysin-catalyzed peptide bond cleavage is still under debate.[1,2] In the “hydroxide” mechanism (mechanism 1 in Figure 1), the reactive nucleophile is generated by deprotonation of the Zn-bound water molecule. Glu143 accepts the proton and transfers it to the amide nitrogen. This mechanism is supported by many crystal structures of enzyme−inhibitor complexes,[9] and most convincingly by the fact that Glu143 mutation causes an almost total loss of enzymatic activity in neutral proteases[10,11] and related Zn endopeptidases.[12,13] The “hydroxide” mechanism was challenged by Mock et al.[17,18] who found that the logarithm of the catalytic rate constant, log k_{cat}, increases linearly with increasing pH in the range pH = 5−8 and saturates after a pH of 8.26 is reached. This behavior is indicative of an acidic species with $pK_a = 8.26$ that actively participates in the reaction in its *deprotonated* form for the entire range of pH values. Since the “hydroxide” mechanism cannot convincingly account for this observation, Mock suggested that neutral His231 is this particular residue, which in place of Glu143 acts as general base and deprotonates an unbound water molecule (mechanism 2 in Figure 1). This proposal is rather controversial because it cannot explain the dramatic loss of activity of Glu143 mutants. Thereafter, Lipscomb and Sträter[1] suggested that Mock’s kinetic data can be accounted for by a slight modification of the original “hydroxide” mechanism: not His231 but the Zn-bound water molecule was assigned a $pK_a = 8.26$, implying that the reactive oxygen moiety in the ES complex is hydroxide rather than water (mechanism 3 in Figure 1).

As implied by its name, thermolysin is resistant to high temperatures. The thermostability is due in part to four Ca^{2+} ions in the interior of the protein that prevent large conformational fluctuations.[20] Dynamical effects due to protein motion being intensively discussed in the current literature[21−23] are expected to be small for thermolysin. This suggests that computer simulations, which are capable of probing the enzyme dynamics on the pico- and nanosecond time scale, can be helpful in supporting or excluding one of the mechanistic interpretations drawn from experiments.

The computation of accurate free energy profiles for enzymatic reactions still represents a major challenge. The less time-consuming quantum mechanics/molecular mechanics (QM/MM) approaches that use standard semiempirical methods for the QM part are not expected to be accurate enough to distinguish between the three mechanistic cases. Indeed, the barrier for the Glu143-assisted peptide bond cleavage (mechanism 1 in Figure 1) was estimated to be more than 40 kcal/mol at the AM1/AMBER level of theory,[16] which is more than twice as large as experimental estimates. Gas-phase modeling of enzymatic reactions with fairly accurate density functionals but rather poor representation of the enzymatic environment is at the other extreme. In the gas-phase model of Pelmenschikov et al.,[24] the substrate, all first-shell residues, and certain second-shell residues were treated at the B3LYP level of theory and the enzymatic environment was replaced by a continuum model. The free energy barrier obtained for mechanism 1 was in remarkably good agreement with experimental results, and the measured mutation effects could be reasonably well reproduced. However, the gas-phase modeling did not yield a stable tetrahedral intermediate. Instead, a one-step reaction with a single barrier was reported, which is rather unusual for peptide-bond hydrolysis reactions.

The approach we choose in this work combines the advantages of explicit density functional calculations and full atomistic representation of the enzymatic environment at a finite temperature. Starting from the crystal structure of the apoenzyme, a tripeptide for which kinetic data have been measured is docked to the active site. After selection of a low-energy docking structure, the energetic feasibility of mechanism 1 is investigated by computing the full free energy profile for the Glu143-assisted peptide-bond cleavage. We find that in the ES complex the substrate is bound to the enzyme merely through hydrogen bonds, but it is not coordinated to Zn. As the reaction proceeds, a tight bond between the substrate and Zn is formed, indicating that the metal ion plays an essential role in transition state stabilization. The free energy profile is computed for three distinct steps: (i) nucleophilic attack, (ii) transfer of the proton to the leaving amide group, and (iii) protonation of the amide nitrogen and break of the peptide bond. The barrier height for step i is in good agreement with experimental results and overestimated by 1.2−2.7 kcal/mol depending on the substrate used in experiments. Our computations give evidence that mechanism 1 is indeed energetically feasible and one possbile mechanism of action of thermolysin.

This paper is organized as follows. In section 2, we describe the computational methods used for substrate docking, gas-phase calculations, and classical and QM/MM simulations. In section 3, the energy profile for a gas-phase model of the enzymatic reaction is presented. The BLYP density functional and pseudopotentials used for the full enzymatic reaction are validated by comparison to B3LYP and all-electron calculations. The results from the peptide docking procedure are analyzed, and the choice of a specific structure for the ES complex is explained. Using constrained QM/MM simulation, the free energy profiles for formation of the tetrahedral intermediate (reaction i) and for protonation of the leaving group followed by peptide bond cleavage (reaction iii) are computed. The thermodynamics of the rearrangement of the tetrahedral intermediate (reaction ii) are characterized using biased classical simulation. The paper is concluded in section 4.

2. Computational Methods

Gas-Phase Calculations. Gas-phase calculations were carried out for a minimal model of the active site. The zinc ion is ligated by one water molecule, two imidazoles, and one formate ion, the last three ligands replacing residues His142, His146, and Glu166 of the enzyme. The conserved residue Glu143 is modeled by formate and the substrate by *N*-methylacetamide (NMA). For calculation of the energy profile shown in Figure 2, the distance between Zn and the carbonyl oxygen atom of NMA (ZnO_s) was fixed while all remaining atomic positions were optimized until the default convergence criteria were reached. Geometries for the ES complex, tetrahedral intermediate (TI), and product (P) were obtained from unconstrained optimizations. The structure for transition state 1 (TS1) was approximated from geometry optimization with the distance between the oxygen atom of water and the carbonyl carbon atom of NMA (C_sO_w) fixed to a value that closely corresponds to the maximum of the potential of mean force (PMF) in the enzymatic system, 1.85 Å. Similarly, the structure of transition state 2 (TS2) was obtained by optimization with the distance between the transferred hydrogen atom and the nitrogen atom (H_1N_s) fixed to 1.30 Å. Calculations with the CPMD code[25] were carried out using the BLYP density functional,[26,27] Troullier−Martins

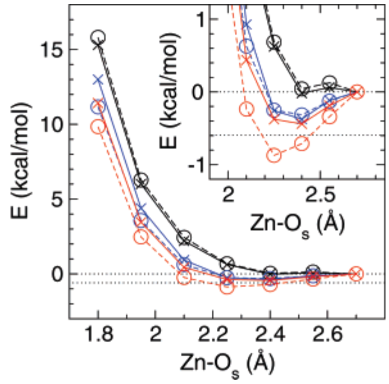


Figure 2. Potential energy profile for *N*-methylacetamide (NMA) binding to a model cofactor of thermolysin in the gas phase. ZnO_S denotes the distance between Zn and the oxygen atom of the carbonyl group of NMA. CPMD/TM/cut70//CPMD/TM/cut70 (black ○), CPMD/TM/cut100//CPMD/TM/cut70 (black ×), BLYP/6-31G(d,p)//BLYP/6-31G(d,p) (blue ○), BLYP/TZVP//BLYP/6-31G(d,p) (blue ×), B3LYP/6-31G(d,p)//BLYP/6-31G(d,p) (red ○), and B3LYP/6-311++G(d,p)//B3LYP/6-31G(d,p) (red ×). All energies are relative to the potential energy at a distance of 2.7 Å. The dotted lines represent the thermal energy k_BT at 300 K. See sections 2 and 3.1 for details.

pseudopotentials,[28] and a reciprocal kinetic energy cutoff of 70 Ry. The final energies were refined using a larger cutoff of 100 Ry. For Zn, a pseudopotential similar to that of ref 29 was used. The pseudostates were generated for the [Ar] $3d^{10}4s^{1.75}4p^{0.25}$ reference configuration. The 3d and 4s electrons were treated as valence and the angular momentum channels of Zn were 3d and 4s with pseudization radii of 1.8 au. See ref 30 for specification of pseudopotentials of second row elements. To test the CPMD pseudopotential calculations, all-electron geometry optimizations were carried out at the BLYP/6-31G(d,p) and B3LYP/6-31G(d,p) levels of theory using the Gaussian program package.[31] The final potential energies were refined using the 6-311++G(d,p) basis.

Substrate Docking. The tripeptide Gly−Leu−Ala was capped with acetyl (Ace) at the N-terminal residue Gly and methylamide (Mam) at the C-terminal residue Ala. Neutral termini were chosen because charged termini would have created interactions not present in a longer peptide. The structure of the apoenzyme was taken from the Protein Data Bank (PDB), code 1LNF.[6] Hydrogen atoms were generated using the HBUILD facility of CHARMM,[32] and their positions were adjusted to minimize the energy. According to mechanism 1 of Figure 1, His231 was assumed to be protonated and Glu143 to be deprotonated. All other residues were protonated according to the respective protonation state in aqueous solution at pH = 7. The oxygen atom in the first coordination shell of Zn that is closest to Glu143 was modeled as a water molecule. The second first-shell oxygen atom closest to His231 was deleted because it was at the location of the substrate oxygen. Six further crystallographic water molecules interfering with the docking were selectively deleted. The docking of the substrate Ace−Gly−Leu−Ala−Mam into the enzyme pocket was carried out using successive energy minimizations for a four-dimensional representation of the substrate[33] as implemented in the CHARMM simulation package.[34] First, 20 independent conformations of the peptide were extracted from an independent 2 ns classical simulation of the peptide in explicit bulk water (one conformation every 100 ps). The orientation of each was systematically explored by rotating the peptide by 0, 90, 180, and 270° about 12 orientations covering the sphere. Each of these $48 \times 20 = 960$ rotated conformations was used as a starting peptide structure. Second, each peptide structure was superimposed to the 1LNF structure so that the substrate oxygen, O_s, was on top of the assigned zinc-binding site. The docking procedure avoids atomic clashes by embedding the substrate in a four-dimensional space. Each substrate atom s is given four spatial coordinates (x_s, y_s, z_s, and w_s) that are used to compute the distance with respect to each atom e of the enzyme: $r'_{es} = \sqrt{{r_{es}}^2 + {w_s}^2}$, where r_{es} is the distance in real, three-dimensional space and w_s the fourth component. The distance between two atoms s and t of the substrate is defined by $r'_{ts} = \sqrt{{r_{ts}}^2 + (w_s - w_t)^2}$. The potential energy was computed from the CHARMM22 empirical force field,[35] using the "embedded" r' distances instead of the real r distances. The force field was slightly modified to account for charge-transfer effects between zinc and its ligands. See ref 36 for the effective charges used. The $\{w\}$ coordinates were restrained to a reference value w^* by adding the penalty term $u(\{w\}) = k\sum_s(w_s - w^*)^2$ to the potential energy. This ensured that the substrate maintained a relative chemical integrity in real space, and that each substrate atom was driven into the real, three-dimensional enzyme pocket in a controlled manner. A series of successive energy minimizations of the $4N$ coordinates were performed (N the number substrate atoms) for decreasing values of w^* (from 5 to ~0.1 Å, in 11 geometric decrements of ×0.7) while the enzyme, Zn atom, and crystallographic water molecules were maintained rigidly in the 1LNF conformation. A final energy minimization was performed with all w's strictly enforced at 0 Å. The minimization procedure was repeated for a penalty constant k of 50, 30, and 10 kcal/mol/Å^2. In total, 3840 docking trials were performed.

Classical MD Simulations. Classical MD simulations were performed for the ES complex and the tetrahedral intermediate TI. The ES system was constructed by placing the best structure from the docking procedure in a 0.1 M KCl solution contained in a periodic box of dimensions 54 Å × 62 Å × 80 Å. The protein, substrate, and solvent were modeled with the Amber99 force field,[37] modifying the atomic charges of the zinc cofactor and its ligands by the same charge increments as in the docking procedure. The ES system was simulated at room temperature and pressure, using a 2 fs integration time step and constraining all bonds involving a hydrogen atom with the SHAKE algorithm.

The initial coordinates for the classical simulation of TI were taken from a free QM/MM simulation of length 2 ps. The simulation protocol is similar to the one of the ES complex. New RESP atomic charges were derived for a model of the TI form of the substrate computed at the B3LYP/6-31G* level (see Table 1). The model used was comprised only of TI in the gas phase: metal−ligand charge-transfer effects were not recalculated. To prevent the ligands from detaching from the zinc ion during the MD simulation, energy penalties were applied if any zinc−ligand distance was out of the range observed from QM/MM simulation of the tetrahedral intermediate. While affecting the short time dynamics of the first-shell ligands, the energy penalties are assumed to have a negligible effect on the nanosecond dynamics of second-shell ligands and backbone (which is the main focus of these classical MD simulations of the TI). In addition to free MD simulations, the TI form of the system was simulated using the adaptive biasing force (ABF) method[38,39] (see section 3.3 for the details on the reaction coordinate).

Table 1. Atomic Charges for the Oxy-Anion Gly(OH$^-$) and Leu Backbone Atoms Used in the Classical Simulations of TI and TI′[a]

residue	atom	Amber99	TI[b]
Gly(OH$^-$)	N	−0.4157	−0.38
	H	0.2719	0.23
	CA	−0.0252	−0.50
	HA1	0.0698	0.15
	HA2	0.0698	0.15
	C_s	0.5973	0.94
	O_s	−0.5679	−0.77
	O_w		−0.71
	H_2		0.32
Leu	N_s	−0.4157	−0.77
	H	0.2719	0.29
	CA	−0.0518	−0.024
	HA	0.0922	0.07
	C	0.5973	0.36
	O	−0.5679	−0.43

[a] For comparison, the charges from the original Amber99 force field for Gly and Leu residues are reproduced. All charges are in electrons. The RESP analysis was performed on the B3LYP/6-31G* electronic density of molecule Ace−Gly(OH$^-$)−Leu−Mam optimized at the B3LYP/6-31G* level. [b] The raw RESP charges are rearranged so that atomic charges of the Leu side chain and of groups Ace and Mam can retain their original values. The charge on the αC of Leu, −0.024e, corresponds to the combined RESP charges of αC and the hydrogen atom present in Gly but not in Leu, minus 0.074e, the total charge of the Amber99 Leu side chain. The total RESP charge on Ace, −0.10e, is transferred to the nitrogen of Gly, and the total RESP charge of Mam, −0.05e, is transferred to the carbonyl carbon of Leu.

QM/MM Simulations. The quantum region for QM/MM simulations included the metal ion; first-shell ligands His142, His146, and Glu166; the Zn-bound water molecule; second-shell residue Glu143; and the chemically active part of the substrate. The QM description was terminated at the αC position for protein residues and, for the substrate, at the αC position of Leu and βC position of Ala. The QM/MM boundary atoms were described by monovalent pseudopotentials.[40] There were 74 QM atoms, and the QM box dimensions were 33.07 × 33.07 × 34.07 au for nucleophilic attack and 35.07 × 35.07 × 35.07 au for protonation of the amide nitrogen atom. All remaining atoms of protein, substrate, and solvent were modeled with the Amber99 force field,[37] using the same system composition and atom topology as for the classical simulations described above. The interaction between the QM system and the MM system was computed using a Hamiltonian electrostatic coupling

Table 2. Energy Profile for Amide Bond Break in *N*-methylacetamide (NMA) Binding to a Gas-Phase Model of the Active Site of Thermolysin[a]

	TS1[b]	TI	TS2[c]	P
BLYP/TM/cut70//BLYP/6-31G(d,p)	22.1	23.9	14.9	−3.1
BLYP/TM/cut100//BLYP/6-31G(d,p)	21.8	23.5	14.3	−3.2
BLYP/6-31G(d,p)//BLYP/6-31G(d,p)	22.3	23.2[d]	11.6	−1.6
BLYP/6-311++G(d,p)//BLYP/6-31G(d,p)	22.2	24.1	14.3	−1.7
B3LYP/6-31G(d,p)//B3LYP/6-31G(d,p)	23.4	21.8	10.3	−0.6
B3LYP/6-311++G(d,p)//B3LYP/6-31G(d,p)	23.3	22.5	13.0	−0.5

[a] TS1, TI, TS2, and P denote optimized model structures for the transition state for nucleophilic attack, tetrahedral intermediate, transition state for protonation of the amide group, and product, respectively. All energies are relative to the energy of the optimized enzyme−substrate complex (ES) and are given in kilocalories per mole. See sections 2 and 3.1 for details. [b] Geometry optimized with distance C_sO_w fixed at 1.85 Å. [c] Geometry optimized with distance H_1N_s fixed at 1.30 Å. [d] Geometry optimized with distance C_sO_w fixed at 1.56 Å.

scheme.[41] QM/MM simulations were carried out with the CPMD code[25] using the BLYP density functional, Troullier−Martins pseudopotentials (see above), a reciprocal kinetic energy cutoff of 70 Ry, a fictitious mass of 700 au, and a time step of 5 au (0.1209 fs). After classical equilibration of the ES complex for 1 ns, a configuration was selected and equilibrated on the QM/MM potential energy surface for 5 ps at 300 K. During the first 1.5 ps of equilibration, separate Nosé−Hoover thermostats were used for QM atoms, protein atoms, and solvent atoms (water and ions). Final equilibration was carried out using a chain of Nosé−Hoover thermostats[42] for the entire system with a target temperature of 300 K. During MD, the bonds between terminating monovalent carbon atoms and the respective QM atoms were fixed at the equilibrium distances taken from classical MD. The PMF for nucleophilic attack was obtained using constrained MD simulations, by taking the average of the Lagrange multiplier constraining distance C_sO_w (denoted as r_1) at a given value. The mean force was computed for 13 windows bridging the equilibrium distances in the ES and TI* complex (* denoting a particular conformation of TI): 3.2, 3.0, 2.8, 2.6, 2.4, 2.2, 2.0, 1.9, 1.8, 1.7, 1.6, 1.5, and 1.4 Å. The initial configuration for each window was selected from an equilibrated configuration of the previous window and the initial snapshot of the first window (3.2 Å) from the free ES simulation. Each constrained MD simulation was equilibrated for about 1−2 ps using again separate thermostats for each subsystem. The next 5 ps were used for calculation of the configurational averages. Except for two windows close to the transition state, the forces averaged over 2.5 and 5 ps were virtually identical, indicating that equilibration and production times were sufficient. At $r = 1.8$ Å, data were averaged over 20 ps, and at $r = 1.7$ Å, the system was equilibrated for 4 ps and averages were taken over 6 ps. Computation of the PMF for protonation of the amide nitrogen atom was carried out similarly using the distance H_1N_s (denoted as r_2) as a constraint. The mean force was calculated for 10 windows with distances constrained to 2.0, 1.8, 1.6, 1.4, 1.35, 1.3, 1.25, 1.2, 1.1, and 1.0 Å.

3. Results and Discussion

3.1. Gas-Phase Hydrolysis. The enzymatic reaction has been modeled in the gas phase as described in section 2. The energies for formation of TS1, TI, TS2, and P are shown in Table 2 for different levels of theory. The main purpose of these calculations is to assess the performance of the BLYP functional used in QM/MM simulations. Validation relative to correlated wavefunction methods is not feasible because of the large size of the cofactor model. Instead, the BLYP calculations are compared to B3LYP calculations. The latter usually describes second-row chemistry better than BLYP, but unfortunately, it is computationally still too expensive for use in our plane wave code. The absolute values of the gas-phase energies should be interpreted with caution due to the simplicity of the model system chosen.

The energies computed with the plane wave basis set are fairly well converged at a moderate reciprocal space kinetic energy cutoff of 70 Ry (BLYP/TM/cut70, BLYP/TM/cut100). The change in energy is 0.1−0.6 kcal/mol when the cutoff is increased to 100 Ry. Similarly, small is the error due to the pseudopotentials used. The deviation of BLYP/TM/cut100 relative to all electron calculations at the BLYP/6-311++G(d,p) level of theory is not more than 0.6 kcal/mol for any reaction step except for formation of the product (1.5 kcal/mol). Remarkably, the optimizations with the BLYP functional did not give a stable minimum for TI. The optimized energy obtained by constraining the C_sO_w distance to a typical value in the transition state region of TS1, $r_1 =$ 1.85 Å, is *lower* than for a typical equilibrium distance of TI, $r_1 = 1.56$ Å, by $22.1 - 23.9 = -1.8$ kcal/mol (see Table 2 and Figure 3 for notation of atoms). The drastic underestimation of energies for stretching of the C_sO_w bond—leading to the instability of TI—is a striking illustration of the self-interaction error of the BLYP density functional. The inclusion of exact exchange at the B3LYP level of theory cures this problem and gives a barrier of $23.3 - 22.5 = 0.8$ kcal/mol for dissociation of the C_sO_w bond.

In the gas-phase model, the second reaction step, protonation of the amide nitrogen atom and dissociation of the amide bond, is predicted to be continuously downhill in energy in contrast with present QM/MM simulations (see section 3.3). Barrierless protonation of the leaving group was also reported for an extended gas-phase model of this reaction.[24] The missing barrier for the second step which involves large molecular rearrangements could be due to an oversimplified reaction coordinate which was identified here with the distance H_1N_s. However, also, the more advanced transition state search carried out in ref 24 did not yield a reaction barrier, indicating that barrierless protonation could be an artifact of the simple gas-phase models used. Note that this deficiency disappears when the full enzymatic environment is included in the calculations.

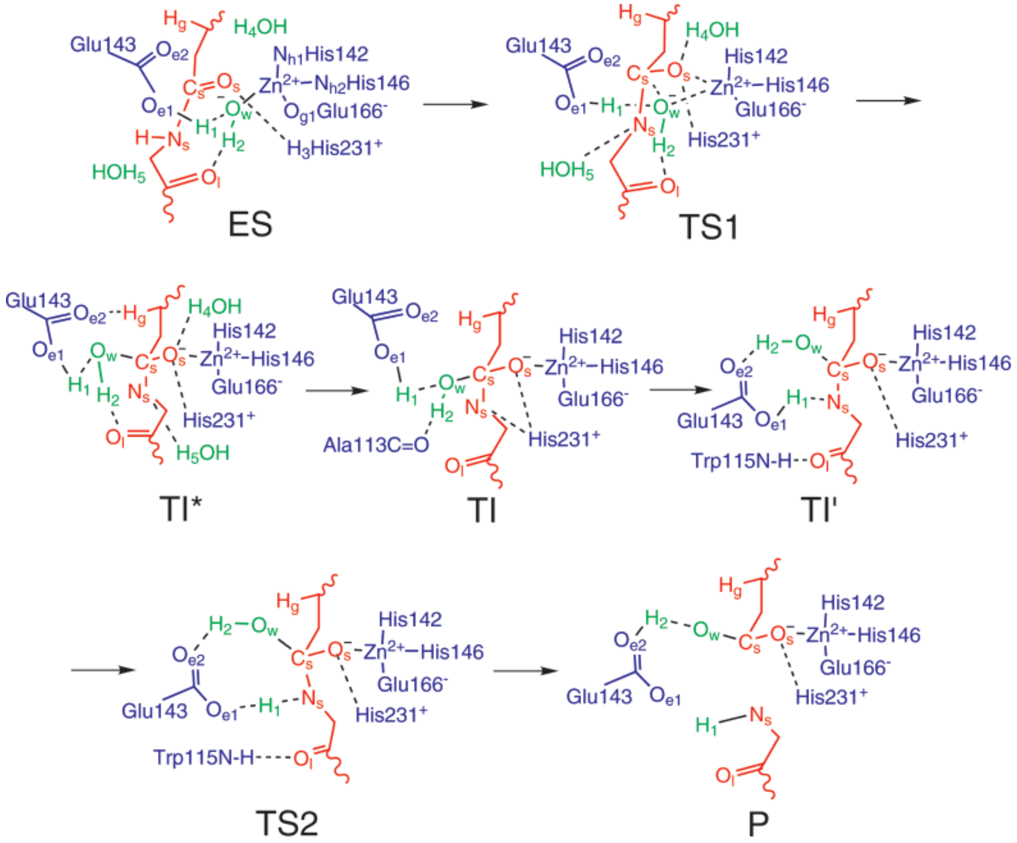


Figure 3. Reaction mechanism for Glu143-assisted peptide bond hydrolysis in thermolysin as obtained in this work. The substrate Ace−Gly−Leu−Ala−Mam is shown in red, catalytic residues of thermolysin in blue, and water molecules in green. Covalent and hydrogen bonds that become broken or formed during hydrolysis are depicted as solid and dashed black lines, respectively. ES denotes the enzyme−substrate complex; TS1 denotes the transition state for nucleophilic attack; TI*, TI, and TI′ denote three distinct conformations of the tetrahedral intermediate; TS2 denotes the transition state for protonation of the leaving group; and P denotes the product. See sections 2 and 3.3 for details.

To estimate the error of the electronic structure calculation in QM/MM simulations (carried out at the BLYP/TM/cut70 level), we have compared the energy profile of the gas-phase model reaction obtained at the BLYP/TM/cut70 level to the one obtained at the B3LYP/6-311++G(d,p) level of theory. The barrier for formation of TI appears to be underestimated by $23.3 - 22.1 = 1.2$ kcal/mol relative to B3LYP/6-311++G(d,p), while the reaction energy for TI formation is overestimated by $23.9 - 22.5 = 1.4$ kcal/mol. Accordingly, the energy of TS2 is overestimated by $14.9 - 13.0 = 1.9$ kcal/mol, and the energy of the product P underestimated by $-0.5 + 3.1 = 2.6$ kcal/mol. The error estimates for TS2 and P are probably not as reliable as for TS1 and TI because the gas-phase energy profile for the second reaction step is barrierless and qualitatively different from the one in the enzyme.

3.2. Enzyme−Substrate Complex. *Docking Structures.* Figure 4A shows a representative set of the energetically most favorable conformations of the substrate Ace−Gly−Leu−Ala−Mam docked to thermolysin. It appears that low-energy docking structures are consistently oriented with the N-terminus of the peptide pointing toward Arg203 (“up”). Although this conforts the choice of an ES conformation for the simulation of reaction 1 of Figure 1, it does not imply that the enzyme is inactive for peptides presented with an N-terminus “down” orientation. By forcing the positions of both water and substrate oxygen atoms, the system retains a relatively large conformational frustration. The extremities of the tripeptide can adopt many different conformations. The N-terminus of the peptide (at the top of Figure 4A) is either protruding away from the enzyme or buried into a small pocket, depending on which amino group of Arg203 binds the acetyl oxygen atom of Ace. The C-terminus is not forming any specific contacts with the protein and displays a considerable range of conformations.

The lowest-energy structure is shown in Figure 4B. It forms strong electrostatic interactions with residues Arg203, Asn112, and His231, and with the backbone of Trp115 and the hydroxyl group of Tyr157 (which are essential features observed for the ensemble of structures depicted in Figure 4A). The “buried” conformation of the N-terminus allows for an optimal interaction with the δ oxygen of Asn112. This interaction probably does not exist for a longer peptide chain as the latter would not fit into the small pocket. Under the electrostatic influence of positively charged His231 and negatively charged Glu143, the Gly−Leu peptide bond and the substrate water molecule are in an orientation favorable for nucleophilic attack.

Classical Molecular Dynamics Simulation. The conformation of Figure 4B was solvated and simulated with classical MD for 1 ns. In this conformation, the N-terminus of the substrate is stabilized by strong interaction of the acetyl

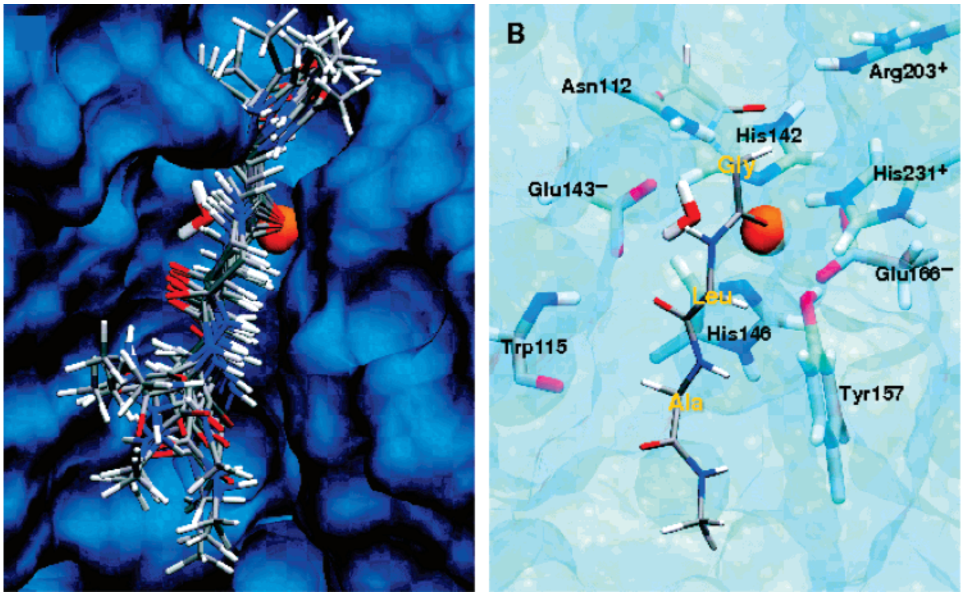


Figure 4. Low-energy conformations of the Ace−Gly−Leu−Ala−Mam tripeptide in the active site of thermolysin (PDB code 1LNF,[6] Ace = acetyl, Mam = methylamide). The two putative oxygen-binding sites of the 1LNF structure are forcibly occupied by a water molecule and the substrate oxygen. Panel A shows representative docking conformations with a total energy within 24 kcal/mol of the energy minimum. Panel B shows the lowest-energy structure only, along with important residues of the active site. The enzyme is represented by its solvent-accessible structure, and the zinc atom is depicted as an orange sphere. Color code: O, red; N, blue; C, gray; H, white. For clarity, the side chains of the substrate are not shown.

oxygen atom of Ace with charged residue Arg203. Preliminary multi-nanosecond simulations have shown that this interaction was generally stable, except in rare instances where one water molecule from the solvent inserts between Ace and Arg203. To prevent water from occasionally disrupting the position of the substrate, an energy penalty was applied if the Ace oxygen atom separated from the Arg203 nitrogen atoms by more than 3.5 Å.

During the 1 ns MD simulation, the carbonyl oxygen atom of the substrate, O_s, remained coordinated to the Zn ion at an average distance of 2.29 Å, and in strong interaction with protonated His231. Note that simulations at the QM/MM level predict that the binding of O_s to Zn is very weak (see below), indicating that the classical model is slightly overbinding. On the opposite side of the peptide bond to be cleaved, the amide hydrogen atom of Leu was interacting mostly with the carbonyl oxygen atom of Ala113. The carbonyl group of Leu was weakly interacting with the backbone of Trp115 and the amide hydrogen atom of Ala, with the hydroxyl group of Tyr157. The C-terminus of the substrate, exposed to the solvent, had no specific interactions with the protein.

Two minor rearrangements of the ES complex structure were observed. First, the binding mode of Glu166 changed from bidendate to monodendate, presumably because both the water oxygen atom O_w and the substrate carbonyl oxygen atom O_s are bound to Zn (as in the inhibitor complex). Second, the substrate water molecule reoriented to form a hydrogen bond with the carbonyl oxygen atom of Leu, O_l. This vacancy on Glu143 was rapidly filled by the amide hydrogen atom of Gly that left the vicinity of Asn112 to be replaced by a water molecule from the solution and by the polar hydrogen of the backbone of Ala113. Overall, the zinc−ligand distances stayed comparable to the crystal-structure values, in a coordination geometry combining characteristics from both the apo structure 1LNF and the inhibitor-bound structure 4TMN (see Table 3).

Table 3. Selected Bond Lengths during Peptide Bond Hydrolysis in Thermolysin[a]

	ES	TS1	TI*	TI′	TS2	P	E(cr)[b]	EI(cr)[c]
ZnO_s	3.63 (2.29)	2.15,[d] 2.85[e]	2.03	1.99	2.01	2.00	2.28	2.17
ZnO_w	1.97 (1.92)	2.58,[d] 2.08[e]	3.02	2.81	2.77	3.18	2.38	2.59
ZnO_{g1}	2.03 (1.91)	2.05	2.04	2.08	2.08	2.05	2.24	2.04
ZnO_{g2}	3.30 (2.82)	3.19	3.06	2.96	3.03	3.04	2.38	2.92
ZnN_{h1}	2.06 (2.11)	2.08	2.08	2.11	2.10	2.10	1.98	2.09
ZnN_{h2}	2.13 (2.25)	2.13	2.12	2.11	2.10	2.07	1.99	2.11
C_sO_w	3.13 (3.02)	1.80[f]	1.51	1.45	1.41	1.24		
C_sN_s	1.38	1.44	1.49	1.53	1.54	3.69		
H_1O_w	1.40	1.86	1.84	2.61	2.52	4.06		
H_1O_{e1}	1.15	1.02	1.02	1.02	1.31	4.64		
H_1N_s	4.00	3.85	3.98	2.00[f]	1.30[f]	1.00[f]		
H_2O_w	0.99	1.01	1.01	1.00	1.04	1.97		
H_2O_l	2.00	1.89	1.77	5.17	5.28	5.40		
H_2O_{e2}	3.94	4.24	4.25	1.88	1.59	1.01		
H_3O_s	2.17	1.94	1.95	1.88	1.89	2.05		
H_4O_s	2.23	1.84	1.84	nb[g]	nb[g]	nb[g]		
H_5N_s	2.50	2.16	2.12	nb[g]	nb[g]	nb[g]		
H_6O_{g2}	1.81	1.80	1.80	1.82	1.81	1.81		

[a] The data are obtained from unconstrained (ES, TI*) and constrained (TS1, TI′, TS2, P) QM/MM simulations. Distances in parentheses are obtained from free classical MD simulations. All bond lengths are given in Angstroms. See Figure 3 for notation of atoms and section 2 for simulation details. [b] Crystal structure of thermolysin, PDB code 1LNF.[6] ZnO_s and ZnO_w represent the distances of one single disordered water molecule or of two water molecules. [c] Crystal structure of thermolysin-inhibitor (ZFPLA) complex, PDB code 4TMN.[7] ZnO_s and ZnO_w denote the distances between Zn and the two oxygen atoms of the phosponamidate group. [d] Average over 4 ps after equilibration for 4 ps. [e] Average over the next 12 ps following the trajectory used in footnote *d*. [f] Distance constrained. [g] Not bonded.

Residues Arg203 and Tyr157 (see Figure 4B) have a stabilizing effect on the substrate. Indeed, preliminary studies using NMA as peptide model have revealed that a substrate that is too short does not provide enough secondary interactions to stabilize the orientation of the peptide bond during molecular dynamics. Although this structural instability may be aggravated by the deficiencies in the force field describing zinc ligation, it remains that the plane of the peptide bond to be broken matches the conformation of the reaction site better if the substrate has at least one additional peptide bond on each side.

QM/MM Simulations. The initial structure for QM/MM simulation of ES was taken from classical MD where the carbonyl oxygen atom of the substrate, O_s, was binding tightly to the Zn ion. Within the first picosecond of QM/MM simulation, the carbonyl oxygen atom was expelled from the first coordination shell. The ZnO_s distance fluctuated between 3 and 5 Å, with an average value of 3.6 Å (see Table 3). The remaining ligands (water, His142, His146, and Glu166) were found to bind tightly to the Zn ion, forming a 4-fold coordination sphere. The binding distances of these ligands are reasonably well reproduced with classical MD when compared to the QM/MM results.

Our observation that the substrate is not binding to Zn is not unusual for Zn peptidases. Related QM/MM studies of β-lactamase also showed that the substrate was not directly binding to the Zn atom.[43,44] In contrast to our result, Pelmenschikov et al. reported an equilibrium distance $r(ZnO_s) = 2.18$ Å for *N*-methylacetamide binding to a model cofactor for thermolysin (at 0 K) but noted that in the ES complex the interaction between Zn and the substrate via ZnO_s was the weakest along the entire reaction path.[24] To support the result of our QM/MM simulations, we have computed the potential energy for the gas-phase model as a function of the ZnO_s distance. As illustrated in Figure 2, the potential energy curve is indeed very flat, with a minimum at $r_m = 2.4$ Å. The weak interaction between Zn and the carbonyl oxygen is indicated by the very small energy required to stretch the ZnO_s bond from the minimum to a separation distance of 2.7 Å, $\Delta E_m = E(2.7\text{ Å}) - E(r_m) =$ 0.4 kcal/mol (BLYP/TZVP), 0.4 kcal/mol (B3LYP/6-311++G(d,p)), and <0.1 kcal/mol (BLYP/TM/cut70), respectively. At room temperature, this energy is smaller than k_BT, explaining the large fluctuations observed in QM/MM simulations.

Judging from the measured Michaelis−Menten constants, the binding of Cbz−Gly−Leu−P′$_2$ substrates to thermolysin is slightly exergonic: $K_M = 2.4$ to 20.6 mM.[45] The corresponding free energies are −2.3 to −3.6 kcal/mol. Our results indicate that the hydrogen bonds formed between the substrate and thermolysin contribute most to the binding free energy whereas the interaction between Zn and the carbonyl oxygen atom is very small.

In the ES complex, the substrate and the water molecule are almost perfectly aligned for nucleophilic attack (see structure ES of Figure 3). The distance between the reactive water molecule and the carbonyl carbon atom C_s is on average 3.13 Å. As suggested by our simulations, the active hydroxide nucleophile is already formed in the ES complex. Hydrogen atom H_1 is observed to switch back and forth frequently between the O_w and O_{e1} of Glu143 (see Figure 5A). The average bond lengths are 1.40 Å for H_1O_w and 1.15 Å for H_1O_{e1}. The carbonyl group of the substrate is also well preorganized through formation of a hydrogen bond with the protonated histidine His231. Tyr157, which was shown to have a modest catalytic effect,[15] does not form hydrogen bonds with the carbonyl oxygen atom O_s. The hydrogen atom of the phenol side chain is instead tightly bonded to the oxygen atom O_{g2} of Glu166 along the entire reaction path (see distance H_6O_{g2} in Table 3).

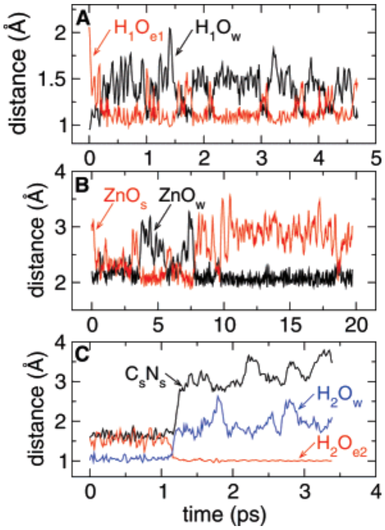


Figure 5. Fluctuations of selected bond lengths at 300 K (A) in the ES complex, (B) first transition state TS1, and (C) shortly after the second transition state TS2 is reached. In A, the fluctuations are obtained from equilibrium QM/MM simulation (no constraints), in B from QM/MM simulation with the distance C_sO_w constrained to 1.8 Å, and in C from QM/MM simulation with the distance H_1N_s constrained to 1.25 Å. See Figure 3 for notation of atoms and sections 2 and 3.3 for further details.

3.3. Enzymatic Hydrolysis. The mechanism for the Glu143-assisted hydrolysis of the substrate is summarized in Figure 3. Selected bond lengths for transition states and intermediates are given in Table 3 and shown in Figures 5 and 6. The free energy profiles for the individual reaction steps are illustrated in Figures 6 and 7 and combined into a single profile in Figure 8.

Transition State for Nucleophilic Attack (ES → TS1). Transformation of the ES complex into the tetrahedral intermediate was enforced by constrained QM/MM simulations using the distance C_sO_w as reaction coordinate r_1. The mean force and the corresponding free energy profile (PMF) are shown in Figure 7A and B. The mean force is close to zero at the average distance in the ES complex and increases almost linearly between 2.8 and 2.2 Å corresponding to a quadratic PMF. At 2.4 Å, hydrogen H_1 is fully transferred to Glu143 and strongly hydrogen-bonded to the reactive hydroxide ion. Remarkably, the average distances between Zn and first-shell ligands for TS1 differ by not more than 0.02 Å from the distances in the crystal structure of the enzyme−inhibitor complex[7] (see Table 3).

In the transition state region at around 1.8 Å, we observe two binding modes of the substrate (see Figure 5B). In the first 8 ps of QM/MM dynamics, the substrate is bonded to Zn (2.15 Å) while hydroxide is displaced (2.58 Å). This

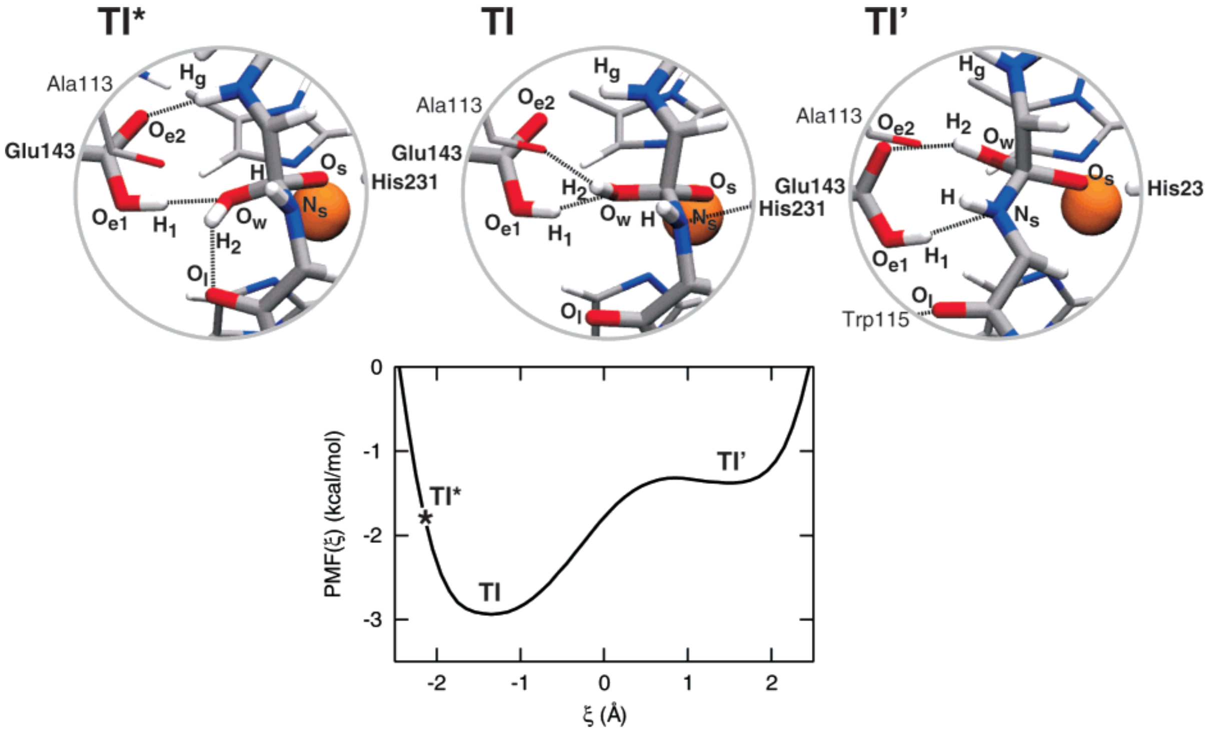


Figure 6. Reorganization of the tetrahedral intermediate described in terms of $\xi = d_{OO} - d_{ON}$, where d_{OO} is the distance beween O_l (the carbonyl oxygen of substrate residue Leu) and the center of mass of pair O_wH_2, and d_{ON} is the distance between the centers of mass of pairs $O_{e1}H_1$ and N_sH. Snapshots of representative conformations along the reorganization pathway are shown as insets. In free energy basin "TI", Glu143 is interacting with H_g and O_w, and His231 is interacting both with O_s and N_s. In basin "TI′", Glu143 is interacting with H_2 and N_s, and His231 is interacting with O_s only. Left-most structure (TI*) is a high-energy configuration in which the intrasubstrate hydrogen bond is partially formed. For clarity, the side chains of the substrate are not shown.

binding mode is favorable for reaction to the tetrahedral intermediate, and therefore the mean force is negative (Figure 7A, ○ at $r_1 = 1.8$ Å). During the following 12 ps of dynamics, the binding mode is reversed: the substrate is bonded weakly (2.85 Å), whereas hydroxide is bonded tightly (2.08 Å). In this case, the forward reaction is unfavorable, as indicated by a positive mean force (Figure 7A, × at $r_1 =$ 1.8 Å). Evidently, sufficient sampling of the transition between the two binding modes would require even longer simulations (100 ps or more) which is beyond current capabilities. For computation of the PMF, we took the average of the mean force in the two binding modes and considered the difference as statistical uncertainty. The structural rearrangement of the substrate in the transition state region is not affected by the binding mode. Besides formation of the fluctuating ZnO_s bond (3.63 Å in ES, 2−3 Å in TS1), the advancing formation of the oxy-anion is stabilized through the formation of two strong hydrogen bonds, one formed with protonated His231 (2.17 Å in ES, 1.94 Å in TS1) and one with a solvent molecule (2.23 Å in ES, 1.84 Å in TS1). Moreover, the increased negative charge on the amide nitrogen atom is stabilized by the formation of a weak hydrogen bond with a solvent molecule (2.50 Å in ES, 2.16 Å in TS1). In the transition state region, the carbonyl Π bond is partly broken, as indicated by the beginning pyramidalization of carbon and nitrogen atoms.

Tetrahedral Intermediate (TS1 → TI).* A decrease of the reaction coordinate from 1.8 to 1.5 Å led to a negative mean force or, equivalently, to a loss of free energy. At $r_1 = 1.5$ Å, the tetrahedral intermediate TI* is formed and separated from the reactant by a free energy barrier of 4.0 kcal/mol (* denotes a particular conformation of the tetrahedral intermediate; see below). The oxy-anion O_s is tightly bonded to Zn (2.03 Å), and the hydrogen bonds with His231 and the solvent molecule are fully formed. The oxygen O_w of the hydroxo group is only weakly bonded to Zn (3.02 Å), while the distances of the remaining first-shell ligands are almost unchanged when compared to those of the ES complex. The formation of TI* can therefore be viewed as a ligand exchange reaction where the Zn-bound nucleophile is replaced by the substrate carbonyl oxygen atom, leaving Zn 4-fold coordinated in the reactant and product states. The tetrahedral intermediate is a stable species in the enzyme as opposed to the model in the gas phase (see section 3.1).

Rearrangement of Tetrahedral Intermediate (TI → TI → TI′).* The hydrogen-bonding pattern of TI* as obtained from the relatively short QM/MM constrained dynamics is rather unfavorable for subsequent protonation of the leaving group (see structure TI* of Figure 3). The amide nitrogen atom N_s is initially hydrogen-bonded to a solvent molecule, while the nearest acidic protons, H_1 at Glu143 and H_3 at His231, are separated from N_s by as much as 4 Å. Atom H_1 of protonated Glu143 forms a hydrogen bond with O_w, while the carbonyl oxygen atom of Glu143 (O_{e2}) is bonded to the proton of the amide group of Gly, H_g. Glu143 cannot deliver H_1 to N_s because hydrogen atom H_2 is blocking the transfer, the latter forming a strong intramolecular hydrogen bond with O_l, the carbonyl oxygen atom of Leu. The rather large separation

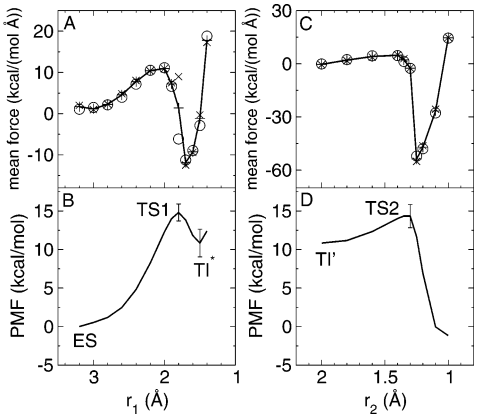


Figure 7. Computed mean force and free energy profile for Glu143-assisted hydrolysis of Ace−Gly−Leu−Ala−Mam catalyzed by thermolysin. (A) Mean force and (B) PMF for nucleophilic attack of the amide bond along the distance C_sO_w, r_1. (C) Mean force and (D) PMF for protonation of the amide nitrogen atom along the distance H_1N_s, r_2. The mean force averaged over the first and second halves of the QM/MM trajectory (5 ps/window in total) are denoted by circles (○) and crosses (×), respectively. (Two points in A were averaged over longer trajectories, see discussion in section 3.3). The average of the two values for the mean force is denoted + and used for the calculation of the PMF shown in B and D. The error bars in B and D denote the difference in free energy obtained when the first and the second halves of the trajectories are used for calculation of the mean force. The PMF in B is set equal to zero for the ES state at $r_1 = 3.2$ Å.

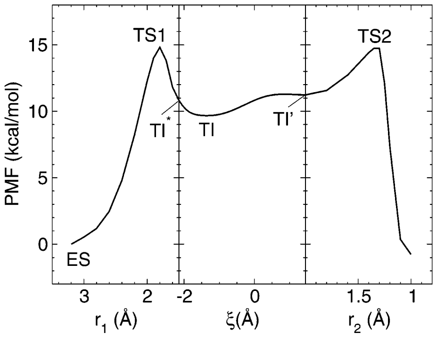


Figure 8. PMF for the full enzymatic reaction obtained by merging the PMFs shown in Figures 6 and 7. See section 3.3 for details.

distance between H_1 and N_s is a consequence of the "up" orientation of the substrate in the active site pocket (see section 3.2 and Figure 4). The backbone atoms between C_s and the N-terminus of the substrate are close to Glu143, while the backbone atoms between C_s and the C-terminus (including the C_sN_s peptide bond) are at the far end. In the inhibitor complex of ref 7, the substrate is aligned in the opposite direction, which facilitates direct protonation of N_s.

In order to probe the stability of the hydrogen-bonding pattern, we have carried out classical MD simulation for the tetrahedral intermediate. Within a few 100 picoseconds of dynamics, major reorganization of the hydrogen-bonding pattern was observed (see Figures 3 and 6). First, the intrasubstrate hydrogen bond H_2O_l broke, and a new hydrogen bond between H_2 and the carbonyl oxygen atom of Ala113 was formed, converting TI* into TI. Second, the hydrogen bond H_gO_{e2} broke, and a new hydrogen bond between H_2 and O_{e2} was formed. This rearrangement of TI into a structure we call TI′ brings H_1 into the vicinity of N_s. Due to the large amplitude motion of Glu143, H_1 only forms a transient hydrogen bond with N_s and frequently binds to O_l or reverts to form a hydrogen bond with O_w as in the TI structure. Figure 6 shows representative conformations of the hydrogen bond network as it rearranges into the reactive conformation TI′. The transition from TI to TI′ requires simultaneously a translation of Glu143 and a torsion of the substrate. The wide-range motion of Glu143 is facilitated by the presence of residues Ala113 and Trp115, whose backbone atoms are competing with atoms O_{e2} for hydrogen bonding with atoms H_g of Gly and O_l of Leu. Our analysis shows that Glu143 is indeed a highly effective proton shuttle, even in cases where the substrate is not favorably aligned for protonation.

The breaking of the intrasubstrate hydrogen bond H_2O_l ($TI^* \rightarrow TI$) is consistent with ab initio energy minimizations done on a mimic of the isolated substrate. At the B3LYP/6-31G* level, the minimum-energy structure has no intrasubstrate hydrogen bond and is 4 kcal/mol more stable than a structure optimized with atoms O_w and O_l kept 2.62 Å apart (i.e., forming a hydrogen bond as in the final snapshot of the QM/MM simulation). The breaking happens on the nanosecond time scale in the protein presumably because it requires a reorganization of the whole substrate.

The free energy corresponding to this reorganization is computed using the ABF method[38,39] for a reaction coordinate ξ equal to $d_{OO} - d_{ON}$. d_{OO} is the distance between atom O_l and the center of mass of pair O_wH_2. It describes the breaking of the intrasubstrate hydrogen bond and the gradual exposition of the amide nitrogen N_s to acidic residue Glu143. d_{ON} is the distance between the center of mass of pair $O_{e1}H_1$ and the center of mass of pair N_sH. ξ is minimum in the TI* structure when Glu143 is in its original, "high" position and when the intrasubstrate hydrogen bond is formed, and it is maximum in the TI′ structure when Glu143 is in position to protonate the amide nitrogen. The graph of Figure 6 shows the free energy profile obtained from 20 ns of ABF simulation. Two distinct groups of structures can be identified: TI structures, for which Glu143 binds substrate atoms H_g and O_w, and TI′ structures, for which Glu143 binds atoms H_2 and N_s. The TI′ structures are 1.5 kcal/mol less stable than structures TI.

The TI* structures obtained from the relatively short QM/MM simulations do not correspond to a minimum in the long classical simulations. The average value of ξ calculated from a 2 ps free QM/MM simulation of TI* is −2.15 Å. The ABF free energy at this value is 1.1 kcal/mol higher than the free energy of TI. Despite the coordination restraints preventing the Zn ligands from detaching, a water molecule from the solution inserted into the coordination shell of Zn during the third nanosecond of the ABF simulation. Although this insertion made the metal ion 6-fold coordinated, and slightly distorted its original coordination structure, it has a minor effect on the PMF of Figure 6. Indeed, we have extended the ABF simulation after manually expulsing the spurious

Table 4. Computed and Experimental Free Energies for Peptide Bond Hydrolysis Catalyzed by Thermolysin[a]

	TS1	TI*	TI	TI′	TS2	P
Antonczak et al.[b]	60		20		22	7
Antonczak et al.[c]	50		1		5	1
Antonczak et al.[d]	20.49		17.00		41.57	−3.91
Pelmenschikov et al.[e]	15.2		6.9		4.3	−8.8
this work[f]	14.8 ± 1.1	10.8 ± 1.8	9.7[g]	11.2[g]	14.7 ± 1.5	
this work[h]	16.0		8.3		12.8	
experiment[i]	13.6[j] 12.1[k]					

[a] TS1 denotes the transition state for nucleophilic attack; TI*, TI, and TI′ denote conformations of the tetrahedral intermediate; TS2 denotes the transition state for breakdown of the tetrahedral intermediate; and P denotes the product. All free energies are given in kilocalories per mole relative to the free energy of the enzyme−substrate complex. See section 2 for details. [b] Energy profile for the nonassisted hydrolysis of formamide estimated from Figure 5 of ref 53. QM/MM, AM1/AMBER. [c] Energy profile for the water-assisted hydrolysis of formamide estimated from Figure 5 of ref 53. QM/MM, AM1/AMBER. [d] Energy profile for the Glu143-assisted hydrolysis of the tripeptide Gly−Phe−Leu, taken from ref 16. QM/MM, AM1/AMBER. [e] Free energy profile for the Glu143-assisted hydrolysis of NMA, taken from ref 24. B3LYP gas-phase optimizations + self-consistent isodensity continuum model (SCI−PCM). [f] Potential of mean force (PMF) for the Glu143-assisted hydrolysis of Ace−Gly−Leu−Ala−Mam, Ace = acetyl, Mam = methylamide. CPMD/AMBER QM/MM, BLYP/TM/cut70 unless stated otherwise. The free energies are taken from Figure 8. [g] Estimated from classical free energy simulation. [h] Energy-corrected PMF. The correction is the energy difference between B3LYP/6-311++G(d,p) and BLYP/TM/cut70 levels of theory obtained for the model reaction in the gas phase (see Table 2). [i] Effective activation free enthalpy calculated from measured rate constants k_{cat} using the classical transition state formula. [j] Ref 45. Substrate: Cbz−Gly−Leu−Ala−OH, pH = 7, $k_{cat}/K_M = 78\,000$ s^{-1}M^{-1}, $K_M = 10.6$ mM [k] ref 49. Substrate: Fua−Gly−Leu−Ala−OH, $k_{cat}/K_M = 870\,000$ s^{-1}M^{-1}, K_M assumed to be the same as for Cbz−Gly−Leu−Ala−OH.

water molecule, and the PMF showed little change. While the free energy difference between structures TI and TI′ are probably reliable within less than k_BT, the free energy from structures TI* to TI (from $\xi = -2.15$ to -1.35 Å) could be less reliable because it involves the breaking of an intramolecular hydrogen bond (H_2O_l) for which the empirical force field was not explicitly parametrized.

Protonated residue His231 is forming a strong interaction with substrate atom O_s during most of the 20 ns of the biased classical MD simulation. For TI structures, His231 occasionally gets at close range to the amide nitrogen N_s, in a position where it could transfer a proton (see structure TI of Figure 6). However, if it were to protonate N_s, His231 would have to get reprotonated for the next catalytic cycle to proceed. Similarly, Glu143 would have to get deprotonated. There is no proton chain connecting His231 with Glu143 that could help restore protonation states. Deprotonation of Glu143 by a solvent molecule also does not seem likely as it is not exposed to the solvent. For this reason, we have not considered His231 as a potential proton donor facilitating the breakdown of TI′.

Transition State for Proton Transfer (TI′ → TS2). For the QM/MM simulation of the protonation of the amide group, we have taken a TI′ structure from the classical simulation, where H_1 of Glu143 forms a strong hydrogen bond with N_s. The proton transfer was enforced using again constrained QM/MM simulation and the distance r_2 between H_1 and N_s as geometrical constraint (see Figure 7C and D). The mean force is zero at 2.0 Å, increases slightly up to a small maximum value at 1.4 Å, and vanishes again in the transition state region 1.30−1.35 Å. The corresponding free energy barrier is 3.5 kcal/mol. At the second transition state TS2, the second proton transfer is about to begin, as indicated by the slightly elongated H_2O_w bond and the stronger interaction between H_2 and O_{e2}. Yet the amide bond is not significantly elongated.

Product Formation (TS2 → P). A decrease of the H_1N_s distance from 1.30 to 1.25 Å leads to an irreversible break of the amide bond. Virtually at the same time, H_2 is transferred from O_w to O_{e2} (see Figure 5C). The break of the peptide bond is indicated by a sudden jump of the mean force to a large negative value. After the peptide bond breaks, the carboxylate and amine groups separate quickly from one another (Figure 5C). The distance C_sN_s fluctuates between 3−4 Å for the remainder of the simulation. The mean force remains negative up until the approximate equilibrium distance of 1.05 Å is reached.

Evidently, the N-protonated tetrahedral intermediate is not a stable species; full protonation and breakage of the peptide bond occur virtually at the same time. A concerted breakdown of the tetrahedral intermediate was also reported for the hydrolysis of amide bonds in the gas phase[46,47] and in β-lactamase.[43] The reaction coordinate r_2 captures the structural reorganization for TI′ → TS2 well but cannot describe reversibly the spontaneous break of the peptide bond. Since we are primarily interested in reaction barriers, we have not carried out additional simulations to compute the free energy of the product. This could be done by taking the distance difference between H_1N_s and C_sN_s as a reaction coordinate instead of r_2.

Full Free Energy Profile. The free energy profile for the full hydrolysis reaction is shown in Figure 8 and summarized in Table 4. The PMFs of Figures 6 and 7B are merged by connecting the minimum TI* of the latter with the free energy at $\xi = -2.15$ Å of the former. This value of ξ is the average computed from a free QM/MM trajectory for TI* of length 2 ps. Similarly, the minimum of the PMF shown in Figure 7D is connected to the minimum denoted TI′ in Figure 6.

The free energy barrier for nucleophilic attack is 14.8 ± 1.1 kcal/mol, obtained as the difference in free energy at $r_1 = 1.8$ Å and $r_1 = 3.2$ Å. The uncertainty represents the difference in free energy obtained when averages were computed for the first and second half of the constrained trajectories, respectively (denoted by error bars in Figure 7). The major contribution to the statistical error is the uncertainty of the mean force at the transition state (see discussion above). The barrier increases to 16.0 kcal/mol if the underestimation of the gas-phase reaction barrier relative to the B3LYP density functional is taken into account (see section 3.1).

Formation of the tetrahedral intermediate TI reduces the free energy by 5.1 kcal/mol relative to TS1 or by 7.7 kcal/mol if corrected again relative to B3LYP gas-phase calculations. Reorganization of unreactive TI into the reactive TI′ conformation is reversible on the time scale of the enzymatic reaction. Protonation of the leaving group leads again to an increase in free energy by 3.5 kcal/mol to 14.7 ± 1.5 kcal/mol. Energy correction relative to B3LYP gives a value of 12.8 kcal/mol. However, as mentioned in section 3.1, the gas-phase reaction that was adopted to calculate the correction energy is not expected to model protonation as well as nucleophilic attack. The difference of the reaction barriers for nucleophilic attack and protonation, 0.1 kcal/mol at the BLYP level of theory, is significantly smaller than the statistical uncertainties, which prevents us from identifying the rate-limiting step of the overall reaction.

The free energy profile presented in Figure 8 does not include nuclear quantum effects. They are generally expected to slightly lower the barriers. For chorismate mutase, for instance, the zero-point energy correction to the reaction barrier was calculated to be −1.5 kcal/mol.[48] Nuclear tunneling effects are expected to play a minor role for nucleophilic attack since proton transfer from O_w to O_{e1} occurs spontaneously and well before transition state TS1 is reached. However, tunneling effects might be important for protonation of the nitrogen atom. In this case, the second barrier is expected to slightly decrease and the difference between the first and second reaction barrier would be more pronounced.

The largest barrier height of the reaction, 14.8 kcal/mol at the BLYP level of theory, is in good agreement with the experimental activation free energy for hydrolysis of the Gly−Leu peptide bond in Cbz−Gly−Leu−Ala−OH, 13.6 kcal/mol,[45] and Fua−Gly−Leu−Ala−OH, 12.1 kcal/mol[49] (pH = 7). When the barrier is corrected for the error of the BLYP functional (relative to B3LYP), +1.2 kcal/mol, and for the zero-point energies, assumed to be between −1 and −2 kcal/mol, our estimate for the barrier is 0.4−2.9 kcal/mol higher than the experimental activation free energies for the two substrates. The experimental values were obtained from the measured rate constants k_{cat} using the classical transition state formula and unity for the transmission coefficient. The latter is typically between 0.5 and 1 for enzymatic reactions.[50−52] Adopting a conservative value of 0.5, the experimental activation free energies decrease by 0.4 kcal/mol, leading to an increase of the overestimation of the experimental barrier by the same amount. The relatively large variation of experimental rate constants among substrates with different terminal groups implies that part of the discrepancy with experimental results could stem from the use of acetyl and *N*-methylamide as terminal groups of the substrate.

4. Conclusion

Starting from the crystal structure of the apoenzyme, a tripeptide was docked to the active site of thermolysin and the free energy profile for peptide bond cleavage computed using state-of-the-art QM/MM calculations. After the nucleophilic attack of deprotonated water, Glu143 and the oxy-anion substrate rearrange into a form where Glu143 is in position to protonate the amide group. Accounting for the free energy of this rearrangement, the barrier heights for nucleophilic attack and protonation of the leaving group are almost identical at the BLYP level of theory and overestimate the experimental activation free energies by 1.2−2.7 kcal/mol depending on the substrate used in the experiment. We expect that the barrier is overestimated by the same energy range if the error of the BLYP density functional (relative to B3LYP), zero-point energies, tunneling effects, and the deviation of the transmission coefficient from unity are included in the calculation.

The QM/MM molecular dynamics simulations carried out in this work indicate that the carbonyl oxygen atom is *not* coordinated to Zn in the ES complex. Stabilization of the transition state is achieved through formation of a weak bond between the carbonyl oxygen atom of the substrate and Zn and formation of a strong hydrogen bond with protonated His231. The formation of the oxy-anion is further stabilized by two hydrogen bonds with solvent molecules. Dynamical effects to catalysis are expected to be small for thermolysin due to the presence of Ca^{2+}, which makes the backbone motions very rigid.[20] Thermolysin therefore complies with Pauling's paradigm that enzymes accelerate rates because they bind the transition state better than the substrate and thereby lower the activation barrier.

In this study, we have shown that the Glu143-assisted water addition mechanism, which is best supported by many crystallographic and biochemical studies, is energetically feasible and has a free energy barrier that is slightly higher but still close to the effective barrier calculated from experimental measurements. We have further shown that the conserved residue Glu143 is a highly effective proton shuttle which is capable of transferring the proton to the leaving group even if the substrate is not ideally aligned. However, while intuitive, the mechanism investigated can explain the observed pH dependence of k_{cat} only if Glu143 is identified with the catalytic residue that has a $pK_a = 8.26$. Although Glu143 is not exposed to the solvent, this value seems rather high. Mock and Stanford[17] suggested that His231 could be this particular residue that in place of Glu143 deprotonates the reactive water molecule (mechanism 2 in Figure 1). Preliminary classical MD simulations show, however, that deprotonated His231 does not form a strong hydrogen bond with a water molecule in plane with the histidine ring as required for deprotonation. Instead, the ϵN atom of His231 is prone to form a hydrogen bond with the substrate. Although we cannot exclude Mock and Stanford's proposal on this basis, we agree with the authors of ref 1 that the His231-based mechanism is rather unlikely. The proposal that a Zn-bound hydroxide ion is the nucleophile rather than water (mechanism 3 in Figure 1) seems to be a more likely alternative that will be investigated in future work.

Acknowledgment. This research was supported by NIH under grant GM 67689. We want to thank Drs. Matteo Dal Peraro and Jérôme Hénin for helpful discussions and Dr. Axel Kohlmeyer for valuable advice on QM/MM molecular dynamics simulations. G.L. is a fellow of the Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT). J.B. acknowledges financial support from The Royal Society for a University Research Fellowship. High-performance computing resources were provided by the Pittsburgh Supercomputing Center.

References

(1) Lipscomb, W. N.; Sträter, N. *Chem. Rev.* **1996**, *96*, 2375.

(2) Parkin, G. *Chem. Rev.* **2004**, *104*, 699.

(3) Roques, B. P.; Noble, F.; Dauge, V.; Fournie-Zaluski, M. C.; Beaumont, A. *Pharmacol. Rev.* **1993**, *45*, 87.

(4) Marukami, Y.; Chiba, K.; Oda, T. A. H. *Biotechnol. Bioeng.* **2001**, *74*, 406.

(5) Matthews, B.; Jansonius, J. N.; Colman, P. M.; Schoenborn, B. P.; Dupourque, D. *Nature* **1972**, *238*, 37.

(6) Holland, D. R.; Hausrath, A. C.; Juers, D.; Matthews, B. *Protein Sci.* **1995**, *4*, 1955.

(7) Holden, H. M.; Tronrud, D.; Monzingo, A. F.; Weaver, L.; Matthews, B. *Biochemistry* **1987**, *26*, 8542.

(8) Bartlett, P. A.; Marlowe, C. K. *Biochemistry* **1987**, *26*, 8553.

(9) Matthews, B. *Acc. Chem. Res.* **1988**, *21*, 333.

(10) Toma, S.; Campagnoli, S.; De Gregoriis, E.; Gianna, R.; Margarit, I.; Zamai, M.; Grandi, G. *Protein Eng.* **1989**, *2*, 359.

(11) Kubo, M.; Mitsuda, Y.; Takagi, M.; Imanaka, T. *Appl. Environ. Microbiol.* **1992**, *58*, 3779.

(12) Devault, A.; Nault, C.; Zollinger, M.; Fournie-Zaluski, M.-C.; Roques, B. P.; Crine, P.; Boileau, G. *J. Biol. Chem.* **1988**, *263*, 4033.

(13) Corbeil, D.; Milhiet, P.-M.; Simon, V.; Ingram, J.; Kenny, A. J.; Boileau, G.; Crine, P. *FEBS Lett.* **1993**, *335*, 361.

(14) Beaumont, A.; O'Donohue, M. J.; Paredes, N.; Rousselet, N.; Assicot, M.; Bohuon, C.; Fournie-Zaluski, M.-C.; Roques, B. P. *J. Biol. Chem.* **1995**, *270*, 16803.

(15) Marie-Claire, C.; Ruffet, E.; Tiraboschi, G.; Fournie-Zaluski, M.-C. *FEBS Lett.* **1998**, *438*, 215.

(16) Antonczak, S.; Monard, G.; Ruiz-Lopez, M. F.; Rivail, J.-L. *J. Mol. Model.* **2000**, *6*, 527.

(17) Mock, W. L.; Stanford, D. J. *Biochemistry* **1996**, *35*, 7369.

(18) Mock, W. L.; Aksamavati, M. *Biochem. J.* **1994**, *302*, 57.

(19) Inouye, K.; Lee, S.-B.; Nambu, K.; Tonomura, B. *J. Biochem.* **1997**, *122*, 358.

(20) Feder, J.; Garrett, L. R.; Wildi, B. S. *Biochemistry* **1971**, *10*, 4552.

(21) Benkovic, S. J.; Hammes-Schiffer, S. *Science* **2003**, *301*, 1196.

(22) Schramm, V. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 604.

(23) Olsson, M. H. M.; Parson, W. W.; Warshel, A. *Chem. Rev.* **2006**, *106*, 1737.

(24) Pelmenschikov, V.; Blomberg, M. R. A.; Siegbahn, P. E. M. *J. Biol. Inorg. Chem.* **2002**, *7*, 284.

(25) *CPMD*, version 3.10; The CPMD consortium, MPI für Festkörperforschung and the IBM Zurich Research Laboratory: Zurich, Switzerland, 2005. http://www.cpmd.org (accessed Jul 2007).

(26) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(27) Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(28) Troullier, N.; Martins, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 1993.

(29) Rothlisberger, U. *ACS Sym. Ser.* **1998**, *712*, 264.

(30) Blumberger, J.; Klein, M. L. *Chem. Phys. Lett.* **2006**, *422*, 210.

(31) 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.01; Gaussian, Inc.: Wallingford, CT, 2004.

(32) Brünger, A. T.; Karplus, M. *Proteins* **1988**, *4* (2), 148,156.

(33) Crippen, G. M. *J. Comput. Chem.* **1982**, *3* (4), 471,476.

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

(35) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, M.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.

(36) Dal Peraro, M.; Spiegel, K.; Lamoureux, G.; De Vivo, M.; DeGrado, W. F.; Klein, M. L. *Struct. Biol.* **2007**, *157*, 444−453.

(37) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; S. Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. *AMBER 7*; University of California: San Francisco, 2002.

(38) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115* (20), 9169.

(39) Hénin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121* (7), 2904.

(40) Blumberger, J.; Klein, M. L. *J. Am. Chem. Soc.* **2006**, *128*, 13854.

(41) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(42) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.

(43) Dal Peraro, M.; Llarrull, L. I.; Rothlisberger, U.; Vila, A. J.; Carloni, P. *J. Am. Chem. Soc.* **2004**, *126*, 12661.

(44) Díaz, N.; Suárez, D.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2001**, *123*, 9867.

(45) Morihara, K.; Tsuzuki, H. *Eur. J. Biochem.* **1970**, *15*, 374.

(46) Bakowies, D.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 5712.

(47) Lopez, X.; Inãki Mujika, J.; Blackburn, G. M.; Karplus, M. *J. Phys. Chem. A* **2003**, *107*, 2304.

(48) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H.-J. *Angew. Chem., Int. Ed.* **2006**, *45*, 6856.

(49) Blumberg, S.; Vallee, B. L. *Biochemistry* **1975**, *14*, 2410.

(50) Nam, K.; Prat-Resina, X.; Garcia-Viloca, M.; Devi-Kesavan, L. S.; Gao, J. *J. Am. Chem. Soc.* **2004**, *126*, 1369.

(51) Roca, M.; Moliner, V.; Tunon, I.; Hynes, J. T. *J. Am. Chem. Soc.* **2006**, *128*, 6186.

(52) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186.

(53) Antonczak, S.; Monard, G.; Ruiz-Lopez, M. F.; Rivail, J.-L. *J. Am. Chem. Soc.* **1998**, *120*, 8825.

CT7000792

JCTC Journal of Chemical Theory and Computation

Spontaneous Formation of KCl Aggregates in Biomolecular Simulations: A Force Field Issue?

Pascal Auffinger,*,[†] Thomas E. Cheatham III,[‡] and Andrea C. Vaiana[†,§]

Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg, France, and Department of Medical Chemistry, Pharmaceutical Chemistry and Pharmaceutics and Bioengineering, University of Utah, Salt Lake City, Utah 84112

Received June 11, 2007

Abstract: Realistic all-atom simulation of biological systems requires accurate modeling of both the biomolecules and their ionic environment. Recently, ion nucleation phenomena leading to the rapid growth of KCl or NaCl clusters in the vicinity of biomolecular systems have been reported. To better understand this phenomenon, molecular dynamics simulations of KCl aqueous solutions at three (1.0, 0.25, and 0.10 M) concentrations were performed. Two popular water models (TIP3P and SPC/E) and two Lennard-Jones parameter sets (AMBER and Dang) were combined to produce a total of 80 ns of molecular dynamics trajectories. Results suggest that the use of the Dang cation Lennard-Jones parameters instead of those adopted by the AMBER force-field produces a more accurate description of the ionic solution. In the later case, formation of salt aggregates is probably indicative of an artifact resulting from misbalanced force-field parameters. Because similar results were obtained with two different water parameter sets, the simulations exclude a water model dependency in the formation of anomalous ionic clusters. Overall, the results strongly suggest that for accurate modeling of ions in biomolecular systems, great care should be taken in choosing balanced ionic parameters even when using the most popular force-fields. These results invite a reexamination of older data obtained using available force-fields and a thorough check of the quality of current parameters sets by performing simulations at finite (>0.25 M) instead of minimal salt conditions.

Introduction

Biomolecular systems are surrounded by solvent particles (including water molecules, cations, and anions), and this environment modulates to a significant degree the physicochemical properties of these systems.[1] Theoretical methods such as molecular dynamics (MD) simulations are often used to gain microscopic insight into the complex interplay of interactions between biomolecular species and solvent particles. These methods use empirical force fields specifically developed and validated by extensive use of experimental and high-level ab initio computational data. Given the importance of the various ionic species surrounding biomolecular systems, a significant effort has been put into fine-tuning various sets of Lennard-Jones (LJ) parameters for monovalent ions such as Na^+, K^+, and Cl (for example, see refs 2−6). These parameters have subsequently been included in major force-fields. Recently, some of these parameters, in conjunction with a choice of water models, have been evaluated by comparison of a large array of calculated structural and thermodynamic properties.[7] The authors note that the use of different parametrizations leads to a *large* dispersion of calculated properties resulting mainly from incomplete experimental knowledge of the structural features of ionic aqueous solutions at finite molarity. Therefore, they

* Corresponding author phone: (33) 388 41 70 49; fax: (33) 388 60 22 18; e-mail: p.auffinger@ibmc.u-strasbg.fr.

[†] Université Louis Pasteur de Strasbourg.

[‡] University of Utah.

[§] Present address: Los Alamos National Laboratory, MS K710, Los Alamos, NM 87545.

10.1021/ct700143s CCC: $37.00
Published on Web 08/18/2007

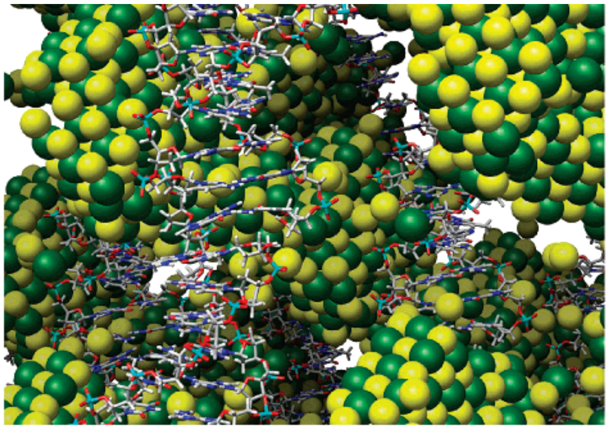


Figure 1. Spontaneous formation of NaCl aggregates in a simulation of a d(CGCGAATTCGCG)$_2$ duplex in 4 M salt solution using the AMBER adopted ion parameters and the TIP3P model. Shown are the unit cell (omitting the water) and images ±1 unit cell in each direction. The sodium and chloride ions are yellow and green, respectively.

did not reach a conclusive ranking of the investigated models. Indeed, finding criteria that allow one to select the most appropriate models and to unambiguously discard defective parameter sets are not straightforward unless some clear-cut artifacts can be identified (see for example, early studies that demonstrated the limits of truncation methods in the calculation of electrostatic interactions).[8] Similar artifacts are likely present in several recently published studies that in some cases reported the formation of salt aggregates in the vicinity of biomolecular systems.[9−13] For example, in a series of simulations exploring "A to B" DNA transitions in >1.0 M NaCl salt solution with AMBER force fields, clear formation of NaCl aggregates were observed (Figure 1; for computational details, see the Supporting Information). In fact, spontaneous and systematic formation of salt aggregates at concentrations around and below 1.0 M is not expected in NaCl and KCl electrolytes (the experimental solubility limits at 20 °C for KCl and NaCl are around 3.2[14,15] and 5.4 mol/L,[14] respectively). Interestingly, all these biomolecular simulations make use of the AMBER force-fields.[16]

To identify the parameters that may be involved in this atypical aggregation process, we performed MD simulations of model systems of aqueous KCl solutions at three different concentrations (1.0, 0.25, and 0.10 M) using two popular water models (TIP3P and SPC/E), as well as two Lennard-Jones (LJ) parameter sets for the K$^+$ cation. One of these parameter sets (Åqvist)[2] is widely used as a part of the parm99 force-field (and all earlier versions) delivered with the AMBER package.[16] The other is derived from the Dang and Kollman's work[17] and has been extensively tested in our MD investigations on nucleic acid fragments,[1,18−21] as well as in studies from other groups.

In this paper, we show that the monovalent cation parameters[2] that are part of the AMBER force-field are involved in the observed aggregation phenomena and that the chosen water model has no impact on the manifestation of this artifact. Hence, we suggest that current AMBER-adapted Åqvist parameters should no longer be used for simulation of ionic solutions because they may affect to an unknown degree the physicochemical properties of the investigated system. Instead, we observed that the K$^+$ parameters of Dang et al.[17] prevent the formation of salt aggregates. Hence, those or similar parameters should be more thoroughly tested and, if considered appropriate, used in replacement of the ones integrated in AMBER that are clearly imbalanced and not adapted for conducting long MD simulations.

Table 1. Characteristic Simulation Parameters

	Amber_ TIP3P	Amber_ SPC/E	Dang_ TIP3P	Dang_ SPC/E
~1.0 M 145 KCl pairs 7568 H_2O	5 ns	5 ns	5 ns	5 ns
~0.25 M 36 KCl pairs 7785 H_2O	5 ns	5 ns	5 ns	5 ns
~0.10 M 15 KCl pairs 7827 H_2O	10 ns	10 ns	10 ns	10 ns

Table 2. Lennard-Jones Parameters (r^* and ϵ) and Partial Charges *(q)* for the Water and Ion Models[a]

	model	q^b	r^* (Å)^c	ϵ (kcal/mol) ^{c}
water[d]	TIP3P	−0.8340	1.7683	0.1520
	SPC/E	−0.8476	1.7766	0.1553
K$^+$	Amber	+1	2.6580	0.000328
	Dang	+1	1.8687	0.100000
Cl$^-$	Amber	−1	2.4700	0.10
	Dang	−1	2.4700	0.10

[a] Note That the AMBER and Dang Parameters for the Cl$^-$ Ion Are Identical. [b] Partial charge for the oxygen atom of the water model and the monovalent ions. [c] r^* corresponds to the position of the Lennard-Jones minimum, and ϵ corresponds to the depth of this minimum. [d] For the TIP3P and SPC/E models, the OW−HW and HW−HW distances are constrained to 0.9572 and 1.5136 Å and to 1.0000 and 1.6330 Å, respectively.

Computational Methods

Twelve molecular dynamics (MD) simulations of aqueous KCl solutions at different ionic strength (1.0, 0.25, and 0.10 M) totaling 80 ns, each on a 5−10 ns scale, were carried out (Table 1). Two water models, TIP3P and SPC/E,[22] as well as two parameter sets for the K$^+$ cation, were used (Table 2). The first set, which contains K$^+$ parameters adapted from the work of Åqvist,[2] is extracted from the AMBER force-field.[23] The second set has been optimized for the SPC/E water model and is extracted from a work of Dang and Kollman.[24] The parameters for the Cl$^-$ anions, which have been used along with the SPC/E water model, are derived from the work of Smith and Dang.[3] Interestingly, these chloride parameters are implemented in the AMBER force field, although they have been adjusted to match the SPC/E (and not the TIP3P) water model.[3] The simulations performed here are named after the type of K$^+$ parameters (AMBER or Dang) and water models (TIP3P or SPC/E) that were used (see Table 1). Note that in the following, AMBER parameters refer to the Åqvist monovalent cation parameters adopted by the all AMBER force-field versions.

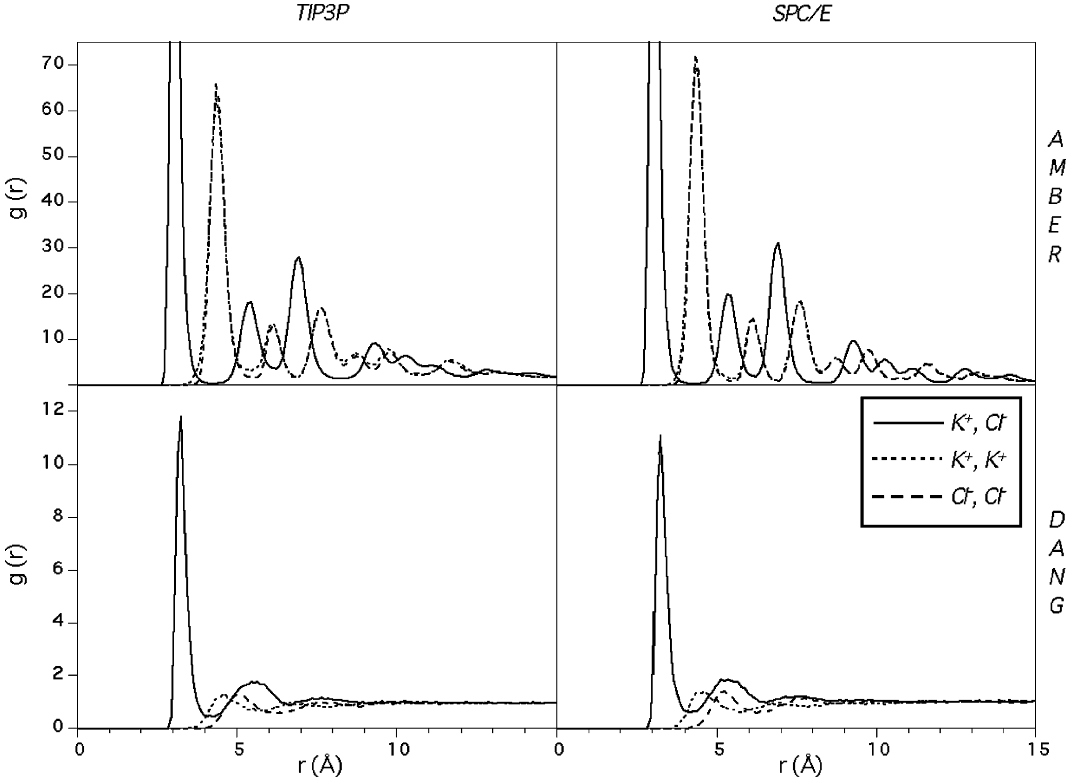


Figure 2. Ion−ion radial distribution functions calculated over the last 200 ps of the 5 ns long MD trajectories of 1.0 M KCl aqueous solutions generated by using the AMBER (top) and Dang (bottom) parameters for the K^+ cation and the TIP3P (left) and SPC/E (right) water parameters.

All simulations were run at constant temperature (298 K) and pressure (1 atm = 101.325 Pa) by using the PMEMD module of the AMBER 8.0 simulation package.[16,25] This module treats the long-range electrostatic interactions[26] with the particle mesh Ewald (PME) summation method. The chosen charge grid spacing is close to 1.0 Å, and a cubic interpolation scheme was used. A cutoff of 9 Å for the Lennard-Jones interaction and the Berendsen temperature-coupling scheme with a time constant of 0.4 ps were used. The trajectories were run with a 2 fs time step. The lengths of the simulations conducted with different parameter sets and at various ionic concentrations are given in Table 1. Note that the simulation times were doubled (from 5 to 10 ns) for the 0.10 M electrolyte solution to ensure a better sampling of the configurational space. The ions were initially placed such that no two ions were closer than 8, 12, or 16 Å in the 1.0, 0.25, and 0.10 M setups, respectively. In this manner, the ions are at the beginning of the simulations uniformly distributed in the simulation box (a cube with a ~62 Å edge). The number of water molecules per ion is ~26, ~108, and ~260 at the 1.0, 0.25, and 0.10 M concentrations, respectively.

Results

Ion−Ion Radial Distribution Functions at High Salt Concentration. The ion−ion radial distribution functions (RDF) derived from the *Amber_TIP3P_1.0M* and *Amber-_SPC/E_1.0M* simulations calculated over the last 200 ps of the 5 ns long trajectories display comparable regular patterns indicative of a highly ordered ionic structure (Figure 2). A visual examination of the MD trajectories allows the clear identification of a rapid aggregation process that leads to the formation of very large KCl clusters: after 5 ns of simulation, the largest of these clusters comprises more than 100 ions (or 50 KCl ion pairs), while a total of less than 10 ions remain unpaired (Figure 3 and Supporting Information). These clustered ions are arranged in a three-dimensional face-centered cubic lattice, typical of KCl, NaCl, or LiF[27] crystals. In these clusters, each ion is, on the average, surrounded by ~4 ions of opposite charge and by ~6 ions carrying the same charge (Table 3). Another characteristic of this crystal-like ionic arrangement resides in comparable cation−cation and anion−anion radial distribution functions (Figure 2). Indeed, one does not expected to find such ordered structures in a liquid phase.

A glimpse into the dynamics of formation of these “nanocrystals” is given by the K−Cl radial distribution functions calculated over four different 500 ps time intervals for the *Amber_TIP3P_1.0M* simulation (Figure 4). They indicate that, after 5 ns, the ionic aggregates are still growing, suggesting that no equilibrium was reached at this point. Hence, it can be inferred that for sufficiently long simulation times, all 290 ions present in the simulation box will aggregate and form a single nanocrystal. The profile of the RDF calculated over the first 200 ps is also remarkable in that it clearly indicates that the artifactual formation of ionic aggregates is difficult to observe in subnanosecond MD simulations.

On the contrary, no salt aggregation is observed in the *Dang_TIP3P_1.0M* and *Dang_SPC/E_1.0M* simulations. Here, the calculated ion−ion RDF’s are close to what is expected for a simulation of a dissociated salt solution

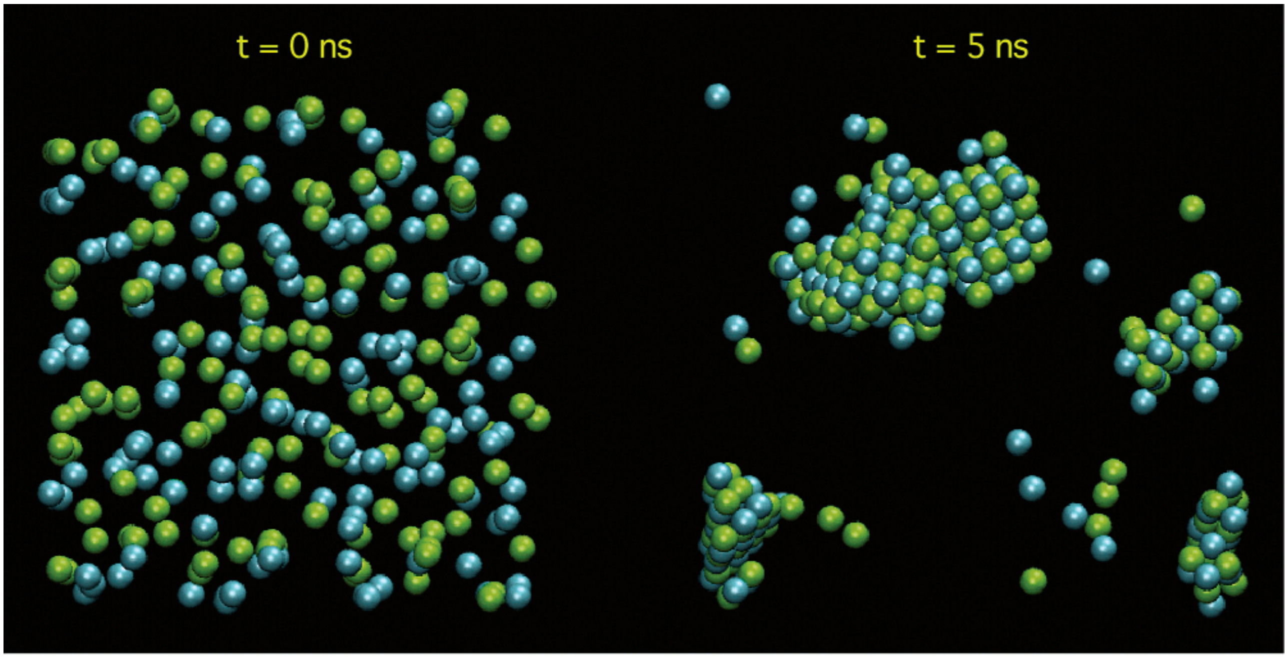


Figure 3. Initial (left) and final (right) configuration of the *Amber_TIP3P_1.0M* simulation. The K^+ and Cl^- ions are shown in green and cyan, respectively. For clarity, water molecules are not shown.

Table 3. First Maxima (r_{max}) of the Ion−Ion Radial Distribution Functions (See Figure 2) and Average Number of Ions (n) Present in Their First Coordination Shell

		K−K		K−Cl		Cl−Cl	
		r_{max} (Å)	n	r_{max} (Å)	n	r_{max} (Å)	n
1.0 M[a]	Amber_TIP3P	4.3	5.8	3.0	3.8	4.3	5.6
	Amber_SPC/E	4.3	5.9	3.0	3.9	4.3	5.7
	Dang_TIP3P	4.5	0.3	3.2	0.4	5.0	0.4
	Dang_SPC/E	4.6	0.3	3.2	0.4	5.2	0.4
0.25 M[a]	Amber_TIP3P	4.3	1.7	3.0	1.6	4.3	1.6
	Amber_SPC/E	4.2	1.1	3.0	1.3	4.3	1.1
	Dang_TIP3P	4.5	<0.1	3.2	0.1	5.3	<0.1
	Dang_SPC/E	4.5	<0.1	3.2	0.1	5.2	<0.1
0.10 M[b]	Amber_TIP3P	4.2	0.3	3.0	0.6	4.3	0.2
	Amber_SPC/E	4.3	0.1	3.0	0.5	4.8	0.1
	Dang_TIP3P	4.2	<0.1	3.2	0.1	5.1	<0.1
	Dang_SPC/E	4.5	<0.1	3.2	0.1	5.2	<0.1

[a] These values (1.0 and 0.25 M) have been calculated by using the last 500 ps of the 5 ns trajectories. [b] These values (0.10 M) have been calculated by using the last ns of the 10 ns trajectories.

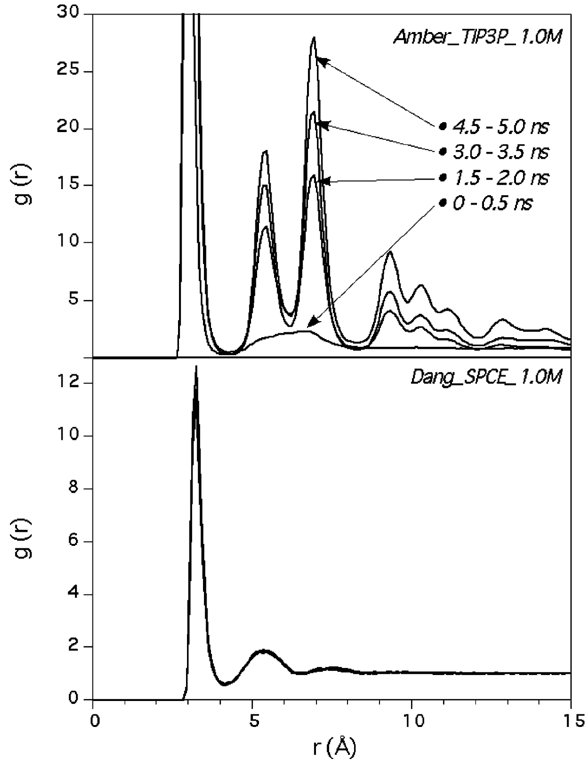


Figure 4. K−Cl radial distribution functions calculated over four 500 ps time intervals (see top panel) of the 5 ns long *Amber_TIP3P_1.0M* (top) and *Dang_SPC/E_1.0M* (bottom) trajectories.

(Figure 2). They display first and second peaks revealing the transient formation of contact and water-mediated ion pairs. The RDF's calculated at different time intervals are almost indistinguishable suggesting that the simulations have rapidly converged with respect to the distribution of ions in the simulation box (Figure 4). No formation of ion aggregates can be observed (Supporting Information). Interestingly, the average ion−ion contact distances are larger here than in the simulations where salt aggregates were observed. The average K^+ to K^+, K^+ to Cl^-, and Cl^- to Cl^- distances are increased by ~0.2, ~0.2, and ~0.8 Å, respectively, with only ~0.4 ions of the same and of opposite charge present in their first coordination shell (Table 3). These numbers suggest that, with the Dang parameters, almost no K−K and Cl−Cl contact ion pairs are formed even at the high 1.0 M salt concentration. The results appear to be strongly dependent on the type of parameters used for the K^+ ions but rather insensitive to the water model chosen. The dependence on the Cl^- parameters has not been explicitly evaluated here because we believe that these parameters were derived in a more consistent manner and that they work well in simulation of both salt solution and biomolecules and because there is

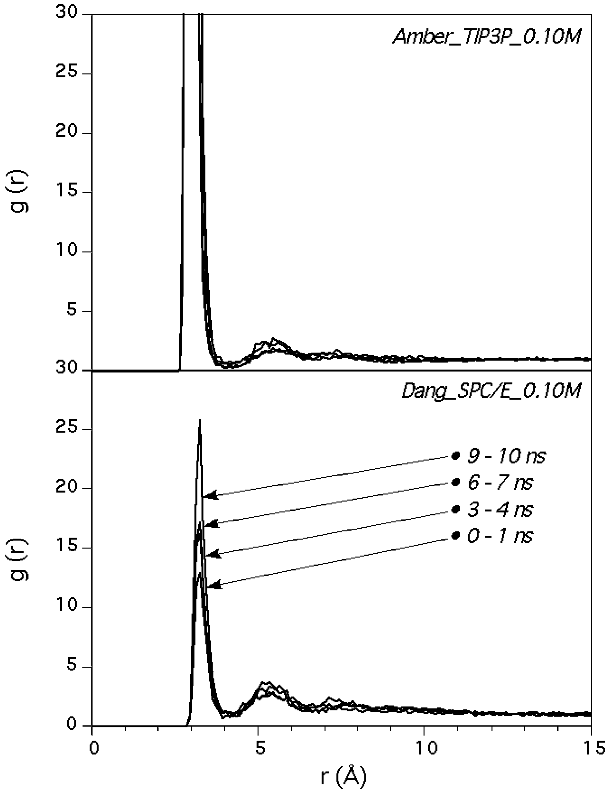


Figure 5. K−Cl radial distribution functions calculated over four 1 ns time intervals (marked in the bottom panel) of the *Amber_TIP3P_0.10M* (top) and *Dang_SPC/E_0.01M* (bottom) 10 ns long MD trajectories.

fairly good consensus on the use of the Smith and Dang parameters.[3]

Medium to Low Ionic Concentrations. At 1.0 M, the interpretation of the collected data is unambiguous. At 0.25 M, the formation of KCl aggregates, though less dramatic, is still clearly observable on the 5 ns time scale. However, at the low salt concentration (0.10 M) ion aggregation is not observed. Small clusters composed of up to five ions form and disaggregate relatively rapidly (Supporting Information). Indeed, at 0.10 M, it is more difficult to identify the formation of aggregates from a visual inspection of the trajectories because only 15 ion pairs are present in the simulation cell. Yet, the ion−ion RDF's calculated over four different 1 ns time intervals reveal a clear dependence on the type of K^+ parameters that were used (Figure 5). This is most clearly revealed by the height of the first peaks and associated coordination numbers. For example, the numbers of ions of opposite charge surrounding a given ion are five times larger when the AMBER rather than the *Dang* parameters are used (∼0.5 instead of ∼0.1), revealing an increased occurrence of KCl contact pairs (Table 3). The time evolution of the RDF's calculated over four different 1 ns time intervals for the *Dang_SPC/E_0.10M* trajectories (Figure 5) suggest that these simulations have not converged over the 10 ns time scale. Yet, for such diluted solutions, a rather slow convergence rate is expected. On the other hand, these results could be indicative of a phase transition associated with the AMBER parameters with a critical concentration between 0.10 (no aggregation) and 0.25 M (aggregation).

Table 4. First Maxima ($r_{\max}$) of the Ion−Water Radial Distribution Functions and First Coordination Shell Ion Hydration Number *(n)* Calculated for the Four 0.10 M Simulations[a]

	$K-O_w$		$Cl-O_w$	
	$r_{\max}$ (Å)	*n*	$r_{\max}$ (Å)	*n*
Amber_TIP3P_ (0.10M/1.0M)	2.73/2.74	6.2/2.2	3.23/3.24	7.0/2.4
Amber_SPC/E_ (0.10M/1.0M)	2.75/2.75	5.6/2.1	3.23/3.23	6.1/2.2
Dang_TIP3P_ (0.10M/1.0M)	2.82/282	7.2/6.9	3.24/3.24	7.6/7.1
Dang_SPC/E_ (0.10M/1.0M)	2.83/2.82	7.1/6.8	3.24/3.22	7.1/6.8
experimental	2.8	6.0−8.0	3.2	6.0−8.0

[a] Experimental values are derived from refs 28 and 29.

Ion−Water Coordination Numbers. The ion−water coordination numbers have been determined for the four simulations conducted at 0.10 M KCl (Table 4). The calculated $K-O_w$ coordination distances, although slightly different (AMBER ≈ 2.74 Å; *Dang* ≈ 2.83 Å), are close to the experimental consensus value of 2.8 Å.[28,29] Not surprisingly, the calculated $Cl-O_w$ coordination distances are identical in all simulations because the same parameters for Cl^- were used in all of them. The calculated 3.23 Å value is close to the experimental consensus value of 3.2 Å.[28,29] For both ions, the number of water molecules located in the first shell is larger when the *Dang* parameters are chosen, reflecting again the fact that the parameters adopted by AMBER favor the formation of ion pairs and aggregates. Interestingly, at higher concentrations the ion hydration number decreases as expected.

Discussion and Conclusion

AMBER Lennard-Jones Parameters for K^+ Favor the Rapid Formation of KCl Aggregates. The formation of salt aggregates in MD simulations of biomolecular systems has already been described in a few studies using KCl or NaCl salts.[9−13] However, these studies did not identify the origin of this phenomenon. Our investigation reveals that the Lennard-Jones parameters for the K^+ cation extracted from the AMBER force field[23] and derived from an early parametrization study[2] are likely at the origin of a rapid, irreversible, and unnatural formation of KCl aggregates at high (1.0 M), as well as near physiological (0.25 M), salt concentration. In addition, the simulations clearly show that the observed aggregation behavior is not dependent on the properties of two of the most widely used rigid water models (TIP3P or SPC/E).

To the best of our knowledge, no biomolecular simulations based on "non-AMBER" force-fields have reported such artifactual behavior. Feig and Pettitt,[30] who investigated the distribution of sodium and chlorine ions around DNA duplexes by comparing the AMBER and CHARMM force fields, used parameters developed by Roux[4] for the Na^+ and Cl^- ions, and did not report any strange behavior in the ionic atmosphere. Similarly, K^+ parameters extracted from a study by Dang and Kollman[24] did not lead, in this and earlier nanosecond long simulations from our group,[12,18,20,21,31,32] to detectable aggregation artifacts, while the use of the AMBER parameters resulted in rapid aggregation of K^+ and Cl^- particles.[12]

A large ensemble of MD studies of aqueous ionic solutions using various parameter sets and particle mesh Ewald (PME) summation methods for the treatment of long-range electrostatic interactions have been published, including simulations of LiF,[27] LiCl,[33−35] NaCl,[7,36−45] KCl,[15,40,45,46] RbCl,[35,45] CsCl,[45] NaBr, KBr, RbBr, CsBr,[45] and CsI.[35] Polarizable force-fields have also been used in other force-fields.[3,44,47−50] Among all these simulations, the use of the Smith and Dang[3] parameters is quite popular (at least for NaCl salts). As reported here, no ion aggregation has been reported in simulations using the Dang parameters even under high and supersaturated salt conditions.[38,42,43] AMBER parameters are rarely used in simulations of ionic aqueous solutions,[7,45] while they are recurrently used in MD simulations of biomolecular systems.[9−13] In one study, a comparison of calculated properties of a ∼1.0 M NaCl aqueous solution generated by using six different parameter sets revealed some level of aggregation for various force-fields including AMBER and GROMOS, while as expected, the Dang parameters did not lead to any detectable formation of ion clusters during 2 ns MD simulations.[7] In another study,[45] the GROMACS program[51] was used to simulate NaCl to CsCl and NaBr to CsBr aqueous salts at various concentrations ranging from 0.10 to 1.0 M. The Åqvist parameters were used for the Na^+, K^+, Rb^+, and Cs^+ cations and different parameters were used for the Cl^-[52] and Br^-[53] anions. The authors reported the formation of ion clusters for all salts at 1.0 M, but not at 0.10 M, in agreement with our own data. However, these clusters that comprise approximately one-third of all ions present in solution appear to be in rapid equilibrium with dissociated ions. Since the formation of ion aggregates was apparently not as dramatic as the one we observed in simulations conducted with the AMBER program and force-fields, these cluster formations were considered as representative of a nonideal behavior observed at the higher ionic concentrations.

Ionic aggregation was also observed in MD simulations of LiF,[27] LiCl,[33] and NaCl[36,37] at 1.0 M concentrations and above. The authors of these studies used self-developed[15,27,33] or GROMOS-adapted parameters[36,37] for the ions. For a 1.0 M solution of LiF, a phase separation was observed. The resulting data indicated that all ions had formed a large and unique cluster geometrically described as a face-centered cubic lattice, the same crystalline structure as that exhibited by LiF, NaCl, or KCl. Smaller clusters were observed in NaCl simulations, mainly, because simulation times below 0.5 ns limited the full formation of aggregates.[36,37] With self-developed parameters, ionic association in 1.0 M of KCl was still observed but was considered to be weak.[15]

More generally, from an experimental point of view, it can be stated that in dilute electrolyte solutions the tendency to aggregate is counterbalanced by thermal fluctuations. Above the saturation point, however, the number of water molecules per ion pair becomes too small to prevent initial ion nucleation followed in most cases by crystallization. From a theoretical point of view, instead, it appears that the interatomic potentials must be correctly balanced to reproduce these subtle equilibria. Any imbalance would lead to observable microscopic catastrophes such as physically improbable aggregation processes. Correct parametrization of three component systems (water, cation, anion) is certainly not straightforward because it involves the fine-tuning of ten different intermolecular potentials. The AMBER potentials by Åqvist[2] were obtained by fitting to experimental free energies of ion hydration, whereas those by Dang were constructed by fitting to gas-phase binding enthalpy data. A recent study devoted to the calculation of ion−ion potential of mean forces also addressed the respective qualities of the Åqvist (AMBER) and Dang models.[54] According to this author, “the Na^+ Lennard-Jones (LJ) parameters of Dang and Åqvist differ considerably with respect to each other. Thus, if only one experimental property is used to determine the LJ parameters, the determined LJ parameters become not necessarily unique. Hence, LJ parameters should ideally be optimized with respect to independent experimental properties to narrow down the ambiguity in the assessment of their values”.

Are Long Simulation Times Needed to Detect Artifacts? Short simulation times may lead to insufficient equilibration of the ionic atmosphere surrounding biomolecules. To achieve a “significant” level of equilibration, simulation times of tens of nanoseconds[55,56] and up to ∼500 ns[10] were suggested for the monovalent cation distribution within DNA grooves to converge. Indeed, short simulation times may significantly complicate the detection of ionic aggregation, as well as other potential artifacts that may only manifest themselves on the longer timescales because of accumulation of errors during the MD runs.[57] Yet, convergence times strongly depend on the type of properties and system investigated. For example, convergence of the ion−ion radial distribution functions is achieved in less than 1 ns for the *Dang_1.0M* simulations **(**Figure 4), while convergence of the same properties for the *Dang_0.10M* is not achieved after 10 ns (Figure 5). Similarly, ion aggregation is observable already after 0.5 ns for the *Amber_1.0M* simulations, while it is very difficult to detect this phenomenon in simulations conducted at low concentration **(**see Movies S1 and S2**)**. Indeed, the fastest equilibration times are probably obtained for the most homogeneous systems, (i.e., highly concentrated ionic solutions or pure water systems). On the other hand, equilibration is difficult to achieve for highly diluted electrolytes.[41,58] An extreme case of dilute solutions is related to “minimal salt conditions” and will be discussed in the following section.

Minimal Salt Strategies: Implications for Biomolecular Simulations. Salt effects should be taken into account with the greatest possible accuracy in MD simulations of biomolecular systems. This is especially true for highly charged nucleic acids. However, MD simulations of nucleic acid systems taking into account a complete representation of the salt environment are relatively infrequent (especially among AMBER users) because it is generally believed that the Lennard-Jones parameters for monovalent cations are more reliable than those for the highly polarizable chloride anion (see ref 59 and associated Supporting Information). Hence, a minimal salt strategy, in which only charge neutralizing cations are taken into account, was used by many groups to prevent the occurrence of anion related artifacts (see ref 1). Although such a strategy seems at first sight reasonable, its use was not based on a precise evaluation of the reliability of available parameter sets. The present investigation suggest that, at least for AMBER users and probably also for users of other force-fields (see ref 7), a misbalance in the ionic Lennard-Jones parameters is at the origin of the serious ionic aggregation problems described above that might affect to an unknown degree the quality of the generated MD trajectories. This misbalance might have its roots in the Lennard-Jones parameters for monovalent cations (Na^+, K^+, ...) as suggested by our data. Consequently, the community is strongly encouraged to reevaluate all the data collected using MD simulations of biomolecular systems in monovalent salt solution, especially data using the default parameters supplied with AMBER. Of particular concern are data related to the interaction of monovalent cations with nucleic acid bases (and other biomolecular fragments) because the $M^+\cdot\cdot\cdot O/N$ interactions are certainly affected to an unknown degree by the use of misbalanced ionic parameters. On the other hand, recent MD simulations successfully reproduced the nucleic acid anion binding sites observed in crystallographic data,[19] suggesting that the Dang Cl^- parameters, although certainly far from being perfect, can be used in MD simulations to reproduce salient features of the ionic atmosphere surrounding biomolecules.

Possible Application to Nucleation Studies. Interestingly, alteration of Lennard-Jones parameters has been used to initiate a nucleation process for NaCl that was subsequently investigated by using the path sampling method developed by Chandler and co-workers.[60] The authors modified the ion−water interactions to obtain an artificial system that crystallizes in a few tens of picoseconds. Hence, nucleation could be studied from simulations generated by using AMBER parameters. Although such trajectories do not correspond to realistic models of ionic solutions, this approach may still be used for getting insight into nucleation phenomena. This is especially true in view of the fact that the ion clusters seem to adopt the same ion ordering as in the crystalline state. Furthermore, phase transition points, such as those occurring at concentrations between 0.10 and 0.25 M in KCl aqueous solutions could be characterized from such MD simulations.[61] Indeed, dissolution of NaCl clusters or nucleation at an NaCl/water interface have been investigated by using the AMBER force-field parameters.[62,63] However, this is not the scope of the present investigation.

Which Monovalent Cation Parameter Sets Should be Used?. Unfortunately, no straightforward answer to this question can be provided. Current parameter sets are far from perfect and suffer from various drawbacks. In 1996, for instance, Lyubartsev and Laaksonen noted “reported RDF (Radial Distribution Function) or PMF (Potential of Mean Force) results appear to be quite often in contradiction to each other and show an apparent dependence on the used model. A general picture arises in several works: the anion-cation potential of mean force has usually two minima, first corresponding to the contact ion pairs (CIP configuration) and the second corresponding to the solvent separated ion pairs (SSIP configuration). However, the intensities, i.e., the relative importance of these minima, vary largely from work to work.”[38] Patra and Karttunen[7] reached the same conclusion by analyzing MD simulations of aqueous NaCl obtained by using six different ion parameter sets and four different water models and concluded that the observed uncertainty in calculated data reflects our current fragmentary experimental knowledge of the structural properties of ionic solutions at finite molarity.

It is not the scope of this study to develop new parameter sets. Nevertheless, on the basis of our data, it can be suggested that it would be worth abandoning ionic models that display any propensity toward anomalous aggregation (for instance AMBER; see ref 7) in favor of those leading to an “appropriate level” of dissociation (Dang), as spontaneous ion aggregation is not expected for molar aqueous solutions of KCl or NaCl salts. Control simulations with other alkali cation models (Li^+, Na^+, Rb^+, or Cs^+) included in the AMBER force-field were not performed. But we suspect that these parameters suffer from the same flaws because they have been parametrized in a similar manner.[45] A reevaluation of the performances of all available parameters in the context of three component electrolyte solutions should be undertaken with a special emphasis on this aggregation issue in the framework of a recent proposal devoted to create a set of descriptive parameters and measures allowing us to judge the “quality” and reliability of MD simulations.[64]

In conclusion, the combination of more precise experimental and theoretical studies will lead to a generation of force-fields free from such imbalanced interatomic potential terms. In this respect, polarizable force-fields will certainly be key players in allowing the generation of more accurate and informative biomolecular simulations.[44,47,49,50,65] Finally, the issues discussed above are not limited to nucleic acid systems but are also relevant for all other biomolecular systems including proteins, membranes, and ion channels,[66] for which the electrolytic environment plays a determining role.

Acknowledgment. The authors wish to thank Prof. Nina Grower for reading the manuscript, as well as for interesting discussions. A.C.V. was supported by the European contract HPRN-CT2002-00190 (CARBONA). Computer time was provided by the “Centre d’Etudes du Calcul Parallèle et de la Visualisation” (ULP). We would also like to thank Prof. Eric Westhof for his ongoing support.

Supporting Information Available: Computational details related to Figure 1 and two movies showing MD trajectories of 1.0 and 0.10 M KCl aqueous electrolytes generated with the Åqvist (AMBER) and Dang parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Auffinger, P.; Hashem, Y., Nucleic acid solvation: From outside to insight. *Curr. Opin. Struct. Biol.* **2007**, in press.

(2) Åqvist, J. Ion−water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021−8024.

(3) Smith, D. E.; Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **1994**, *100*, 3757−3765.

(4) Roux, B.; Prod'hom, B.; Karplus, M. Ion transport in the gramicidin channel: Molecular dynamics study of single and double occupancy. *Biophys. J.* **1995**, *68*, 876−892.

(5) Peng, Z.; Ewig, C. S.; Hwang, M. J.; Waldman, M.; Hagler, A. T. Derivation of class II force fields. 4. van der Waals parameters of alkali metal cations and halide anions. *J. Phys. Chem. A* **1997**, *101*, 7243−7252.

(6) Jensen, K. P.; Jorgensen, W. L. Halide, Ammonium, and alkali metal ion parameters for modeling aqueous solutions. *J. Chem. Theory Comput.* **2006**, *2*, 1499−1509.

(7) Patra, M.; Karttunen, M. Systematic comparison of force fields for microscopic simulations of NaCl in aqueous solutions: Diffusion, free energy of hydration, and structural properties. *J. Comput. Chem.* **2004**, *25*, 678−689.

(8) Auffinger, P.; Beveridge, D. L. A simple test for evaluating the truncation effects in simulation of systems involving charged groups. *Chem. Phys. Lett.* **1995**, *234*, 413−415.

(9) Mazur, A. K. Titration in silico of reversible B ↔ A transitions in DNA. *J. Am. Chem. Soc.* **2003**, *125*, 7849−7859.

(10) Cheatham, T. E. Simulations and modeling of nucleic acid stucture, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360−367.

(11) Savelyev, A.; Papoian, G. A. Electrostatic, steric, and hydration interactions favor Na(+) condensation around DNA compared with K(+). *J. Am. Chem. Soc.* **2006**, *128*, 14506−14518.

(12) Vaiana, A. C.; Westhof, E.; Auffinger, P. A molecular dynamics simulation study of an aminoglycoside/A-site RNA complex: Conformational and hydration patterns. *Biochimie* **2006**, *88*, 1061−1073.

(13) Savelyev, A.; Papoian, G. A. Inter-DNA electrostatics from explicit solvent molecular dynamics simulations. *J. Am. Chem. Soc.* **2007**, *129*, 6060−6061.

(14) Stephen, H.; Stephen, T. *Solubilities of Inorganic and Organic Compounds*; Macmillan: New York, 1963.

(15) Vieira, D. S.; Degrève, L. Molecular simulation of a concentrated aqueous KCl solution. *J. Mol. Struct.* **2002**, *580*, 127−135.

(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*, 1668−88.

(17) Dang, L. X.; Kollman, P. A. Free energy of association of the K+/18-crown-6 complex in water: a new molecular dynamics study. *J. Phys. Chem.* **1995**, *99*, 55−58.

(18) Auffinger, P.; Westhof, E. Water and ion binding around RNA and DNA (C,G)-oligomers. *J. Mol. Biol.* **2000**, *300*, 1113−1131.

(19) Auffinger, P.; Bielecki, L.; Westhof, E. Anion binding to nucleic acids. *Structure* **2004**, *12*, 379−388.

(20) Auffinger, P.; Westhof, E. Water and ion binding around $r(UpA)_{12}$ and $d(TpA)_{12}$ oligomers−Comparison with RNA and DNA $(CpG)_{12}$ duplexes. *J. Mol. Biol.* **2001**, *305*, 1057−1072.

(21) Auffinger, P.; Bielecki, L.; Westhof, E. Symmetric K+ and Mg2+ ion binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J. Mol. Biol.* **2004**, *335*, 555−571.

(22) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potential. *J. Phys. Chem.* **1987**, *97*, 6269−6271.

(23) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwel, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(24) 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.

(25) Case, D.; Darden, T. A.; Cheatham, T. E.; 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, CA, 2004.

(26) Darden, T.; Perera, L.; Li, P.; Pedersen, L. New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure* **1999**, *7*, R55−R60.

(27) Degrève, L.; Borin, A. C.; Mazzé, F. M.; Rodrigues, A. L. G. Molecular simulation of a phase separation in a non-primitive electrolyte solution. *Chem. Phys.* **2001**, *265*, 193−205.

(28) Marcus, Y. Ionic radius in aqueous solutions. *Chem. Rev.* **1988**, *88*, 1475−1498.

(29) Ohtaki, H. Ionic solvation in aqueous and nonaqueous solutions. *Monatsh. Chem.* **2001**, *132*, 1237−1268.

(30) Feig, M.; Pettitt, B. M. Sodium and chlorine ions as part of the DNA solvation shell. *Biophys. J.* **1999**, *77*, 1769−1781.

(31) Auffinger, P.; Bielecki, L.; Westhof, E. The Mg2+ binding sites of the 5S rRNA loop E motif as investigated by molecular dynamics simulations. *Chem. Biol.* **2003**, *10*, 551−561.

(32) Auffinger, P.; Westhof, E. RNA hydration: Three nanoseconds of multiple molecular dynamics simulations of the solvated $tRNA^{Asp}$ anticodon hairpin. *J. Mol. Biol.* **1997**, *269*, 326−341.

(33) Degrève, L.; Mazzé, F. M. Molecular simulation of LiCl aqueous solutions. *Mol. Phys.* **2003**, *101*, 1443−1453.

(34) Egorov, A. V.; Komolkin, A. V.; Chizhik, V. I.; Yushmanov, P. V.; Lyubartsev, A. P.; Laaksonen, A. Temperature and concentration effects on Li+-ion hydration. A molecular dynamics simulation study. *J. Phys. Chem. B* **2003**, *107*, 3234−3242.

(35) Du, H.; Rasaiah, J. C.; Miller, J. D. Structural and dynamic properties of concentrated alkali halide solutions: A molecular dynamics simulation study. *J. Phys. Chem. B* **2007**, *111*, 209−217.

(36) Degrève, L.; da Silva, F. L. B. Large ionic clusters in concentrated aqueous NaCl solution. *J. Chem. Phys.* **1999**, *111*, 5150−5156.

(37) Degrève, L.; Da, Silva, F. L. B. Detailed microscopic study of 1 M aqueous NaCl, solution by computer simulations. *J. Mol. Liquids* **2000**, *87*, 217−232.

(38) Lyubartsev, A. P.; Laaksonen, A. Concentration effects in aqueous NaCl solutions. A molecular dynamics simulation. *J. Phys. Chem.* **1996**, *100*, 16410−16418.

(39) Koneshan, S.; Rasaiah, J. C. Computer simulation studies of aqueous sodium chloride solutions at 298 K and 693 K. *J. Chem. Phys.* **2000**, *113*, 8125−8137.

(40) Chowdhuri, S.; Chandra, A. Molecular dynamics simulations of aqueous NaCl and KCl solutions: effects of ion concentration on the single-particle, pair, and collective dynamical properties of ions an water molecules. *J. Chem. Phys.* **2001**, *115*, 3732−3741.

(41) Lyubartsev, A. P.; Marcelja, S. Evaluation of effective ion−ion potentials in aqueous electrolytes. *Phys. Rev. E* **2002**, *65*, 041202.

(42) Uchida, H.; Matsuoka, M. Molecular dynamics simulation of solution structure and dynamics of aqueous sodium chloride solutions from dilute to supersaturated concentration. *Fluid Phase Equilib.* **2004**, *219*, 49−54.

(43) Bouazizi, S.; Nasr, S.; Jaidane, N.; Bellissent-Funel, M. C. Local order in aqueous NaCl solutions and pure water: X-ray scattering and molecular dynamics simulations study. *J. Phys. Chem. B* **2006**, *110*, 23515−23523.

(44) Wick, C. D.; Dang, L. X. Computational observation of enhanced solvation of the hydroxyl radical with increased NaCl concentration. *J. Phys. Chem. B* **2006**, *110*, 8917−8920.

(45) Chen, A. A.; Pappu, R. V. Quantitative characterization of ion pairing and cluster formation in strong 1:1 electrolytes. *J. Phys. Chem. B* **2007**, *111*, 6469−6478.

(46) Chang, T.-M.; Dang, L. X. Detailed study of potassium solvation using molecular dynamics techniques. *J. Phys. Chem. B* **1999**, *103*, 4714−4720.

(47) Grossfield, A.; Ren, P.; Ponder, J. W. Ion solvation thermodynamics from simulation with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125*, 15671−15682.

(48) Cavallari, M.; Cavazzoni, C.; Ferrario, M. Structure of NaCl and KCl concentrated aqueous solutions by ab initio molecular dynamics. *Mol. Phys.* **2004**, *102*, 959−966.

(49) Dang, L. X.; Schenter, G. K.; Glezakou, V. A.; Fulton, J. L. Molecular simulation analysis and X-ray absorption measurement of Ca^{2+}, K^{+}, and Cl^{-} ions in solution. *J. Phys. Chem. B* **2006**, *110*, 23644−23654.

(50) Lamoureux, G.; Roux, B. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B* **2006**, *110*, 3308−3322.

(51) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(52) Chandrasekhar, J.; Spellmeyer, D. C.; Jorgensen, W. L. Energy component analysis for dilute aqueous solutions of Li+, Na+, F-, and Cl-, ions. *J. Am. Chem. Soc.* **1984**, *106*, 903−910.

(53) Lybrand, T. P.; Ghosh, I.; McCammon, J. A. Hydration of chloride and bromide anions: Determination of relative free energy by computer simulation. *J. Am. Chem. Soc.* **1985**, *107*, 7793−7794.

(54) Gavryushov, S.; Linse, P. Effective interaction potentials for alkali and alkaline earth metal ions in SPC/E water and prediction of mean ion activity coefficients. *J. Phys. Chem. B* **2006**, *110*, 10878−10887.

(55) Ponomarev, S. Y.; Thayer, K. M.; Beveridge, D. L. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14771−14775.

(56) Varnai, P.; Zakrzewska, K. DNA and its counterions: A molecular dynamics study. *Nucleic Acids Res.* **2004**, *32*, 4269−4280.

(57) Auffinger, P.; Vaiana, A. C., Molecular dynamics simulations of RNA systems. In *Handbook of RNA Biochemistry*; Westhof, X. X., Bindereif, X. X., Schön, X. X., Hartmann, X. X., Eds.; Willey-VCH: Manheim, Germany, 2005; pp 560−576.

(58) Donnini, S.; Mark, A. E.; Juffer, A. H.; Villa, A. Incorporating the effect of ionic strength in free energy calculations using explicit ions. *J. Comput. Chem.* **2005**, *26*, 115−122.

(59) Krasovska, M. V.; Sefcikova, J.; Reblova, K.; Schneider, B.; Walter, N. G.; Sponer, J. Cations and hydration in catalytic RNA: Molecular dynamics of the hepatitis delta virus ribozyme. *Biophys. J.* **2006**, *91*, 626−638.

(60) Zahn, D. Atomistic mechanism of NaCl nucleation from an aqueous solution. *Phys. Rev. Lett.* **2004**, *92*, 040801.

(61) Debenedetti, P. G. Thermodynamics: When a phase is born. *Nature* **2006**, *441*, 168−169.

(62) Yang, Y.; Meng, S.; Xu, L. F.; Wang, E. G.; Gao, S. Dissolution dynamics of NaCl nanocrystal in liquid water. *Phys. Rev. E* **2005**, *72*, 012602.

(63) Yang, Y.; Meng, S. Atomistic nature of NaCl nucleation at the solid−liquid interface. *J. Chem. Phys.* **2007**, *126*, 044708.

(64) Murdock, S. E.; Tai, K.; M. H., N.; Johnston, S.; Wu, B.; Fangohr, H.; Laughton, C. A.; Essex, J. W.; Sansom, M. S. P. Quality assurance for biomolecular simulations. *J. Chem. Theory Comput.* **2006**, *2*, 1477−1481.

(65) Glezakou, V.; Chen, Y.; Fulton, J. L.; Schenter, G. K.; Dang, L. X. Electronic structure, statistical mechanical simulaitons, and EXAFS spectroscopy of aqueous potassium. *Theor. Chem. Acc.* **2006**, *115*, 86−99.

(66) Noskov, S. Y.; Berneche, S.; Roux, B. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature* **2004**, *431*, 830−4.

CT700143S

JCTC Journal of Chemical Theory and Computation

Demonstrated Convergence of the Equilibrium Ensemble for a Fast United-Residue Protein Model

F. Marty Ytreberg†

Department of Physics, University of Idaho, Moscow, Idaho 83844-0903

Svetlana Kh. Aroutiounian†

Department of Physics, Dillard University, 2601 Gentilly Blvd., New Orleans, Louisiana 70122

Daniel M. Zuckerman*

Department of Computational Biology, University of Pittsburgh, 3501 Fifth Avenue, Pittsburgh, Pennsylvania 15260

Received March 26, 2007

Abstract: Because of the time-scale limitations of all-atom simulation of proteins, there has been substantial interest in coarse-grained approaches. Some methods, like “resolution exchange’’ (Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Phys. Rev. Lett.* **2006**, *96*, 028105-1−4), can accelerate canonical all-atom sampling but require properly distributed coarse ensembles. We therefore demonstrate that full sampling can indeed be achieved in a sufficiently simplified protein model, as verified by a recently developed convergence analysis. The model accounts for protein backbone geometry, in that rigid peptide planes rotate according to atomistically defined dihedral angles, but there are only two degrees of freedom (ϕ and ψ dihedrals) per residue. Our convergence analysis indicates that small proteins (up to 89 residues in our tests) can be simulated for more than 50 “structural decorrelation times’’ in less than a week on a single processor. We show that the fluctuation behavior is reasonable, and we discuss applications, limitations, and extensions of the model.

1. Introduction

How simplified must a molecular model of a protein be for it to allow full canonical sampling? This question may be important to the solution of the protein sampling problem, the generation of protein structures properly distributed according to statistical mechanics, because of the well-known inadequacy of all-atom simulations, which are limited to submicrosecond timescales. Even small peptides have proven slow to reach convergence.[1] Sophisticated atomistic methods, moreover, which often employ elevated temperatures,[2−6] have yet to show they can overcome the remaining gap in timescales,[7] which is generally considered to be several orders of magnitude. On the other hand, because of the drastically reduced numbers of degrees of freedom and smoother landscapes, coarse-grained models (e.g., refs 8−26) may have the potential to aid the ultimate solution to the sampling problem, particularly in light of recently developed algorithms like “resolution exchange’’[27,28] and related methods.[29−31]

Although the resolution exchange approach, in principle, can produce properly distributed atomistic ensembles of protein configurations, it requires full sampling at the coarse-grained level.[27,28] While the potential for such full sampling has been suggested by some studies of folding and conformational change (e.g., refs 26 and 32), convergence has yet

[*] Corresponding author e-mail: dmz@ccbb.pitt.edu.

[†] The first two authors contributed equally.

10.1021/ct700076t CCC: $37.00
Published on Web 08/25/2007

to be carefully quantified in equilibrium sampling of folded proteins. How much coarse-graining really is necessary? What is the precise computational cost of different approaches? This paper represents a first step in answering these questions by studying a united-residue model with realistic backbone geometry.

We will require a quantitative method for assessing sampling. A number of approaches have been suggested,[1,33−36] but we rely on a recently proposed statistical approach which directly probes the fundamental configuration-space distribution.[1,37] The method does not require knowledge of important configurational states or any parameter fitting. In essence, the approach attempts to answer the most fundamental statistical question, "What is the minimum time interval between snapshots so that a set of structures will behave as if each member were drawn independently from the configuration-space distribution exhibited by the full trajectory?'' This interval is termed the "structural decorrelation time'', τ_{dec}, and the goal is to generate simulations of length $t_{\text{sim}} \gg \tau_{\text{dec}}$.

In this report, we demonstrate the convergence of the equilibrium ensemble for several proteins using a fast, united-residue model employing rigid peptide planes. The relative motion of the planes is determined by the *atomistic* geometry embodied in the ϕ and ψ dihedral angle rotations, as explained below. We believe such realistic backbone geometry will be necessary for success in resolution exchange studies. The use of geometric tables enables the rapid use of only two degrees of freedom per residue (ϕ and ψ), and one interaction site at the α-carbon. The simulations are therefore extremely fast. Gō interactions stabilize the native state, while permitting substantial fluctuations in the backbone.

After the model and the simulation approach are explained, the fluctuations are compared with experimental data from X-ray temperature factors and the diversity of NMR structure sets. The simulations are then analyzed for convergence and timing.

2. Coarse-Grained Model

The coarse-grained model used for this study was chosen to meet several criteria: (i) the fewest number of degrees of freedom per residue, (ii) the ability to use lookup tables for en-
hanced simulation speed, (iii) the stability of the native state, along with the potential for substantial non-native fluctuations, and (iv) the ability to allow the addition of chemical detail, as simply as possible. Thus, we chose a rigid peptide plane model with Gō interactions[9,38,39] and sterics based on α-carbon interaction sites as shown in Figure 1. The use of such a simple model, we emphasize, is consistent with our goal of understanding both the potential and the limitations of coarse models for statistically valid sampling. Once we have understood the costs associated with the present model, we can design more realistic models, as discussed below. In other words, we made no attempt to design the most chemically realistic coarse-grained model, although we believe the use of atomistic peptide geometry is an improvement over a coarse model we considered previously.[26]

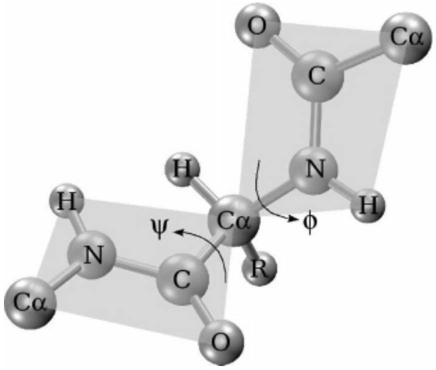


Figure 1. Rigid peptide plane model used this study. Note that, in the coarse-grained simulations, only α-carbons are represented, and the only degrees of freedom are ϕ and ψ. Other atoms are shown in the figure only to clarify the geometry and our assumption of rigid peptide planes.

Rigid peptide planes allow the use of only two degrees of freedom per residue, arguably the fewest that one would consider in such a model. Indeed, this is fewer than in a freely rotating chain, although admittedly our model requires somewhat more complex simulation moves, described below.

Gō interactions were used because they simultaneously stabilize the native state of the protein and also permit reasonable equilibrium fluctuations, as shown in an earlier study.[26] Given our interest in native-state fluctuations and the lack of a *universal* coarse-grained model capable of stabilizing the native state for *any* protein, Gō interactions are a natural choice for enforcing stability. Further, beyond the reasonable "local'' fluctuations shown below, the model also exhibits partial unfolding events which are expected both theoretically and experimentally.[40−42]

Because we see the present model as only a first step in the development of better models, it is important that it easily allows for the addition of chemical detail, such as Ramachandran propensities which require only the dihedral angles we use explicitly.[43] Furthermore, with a rigid peptide plane, the locations of all backbone atoms and the beta carbon are known implicitly. Thus hydrogen-bonding and hydrophobic interactions[15] can be included in the model with little effort. In other words, the "extendibility'' of the present simple model was a significant factor in its design.

2.1. Potential Energy of Model System. The total potential used in the model is given by

$$U = U^{\text{nat}} + U^{\text{non}} \tag{1}$$

where U^{nat} is the total energy for native contacts, and U^{non} is the total energy for non-native contacts.

For the Gō interactions, all residues that are separated by a distance *less* than a cutoff, R_{cut}, in the experimental structure are given native interaction energies defined by a square well

$$U^{\text{nat}} = \sum_{i<j}^{\text{native}} u^{\text{nat}}(r_{ij})$$

$$u^{\text{nat}}(r_{ij}) = \begin{cases} \infty & \text{if } r_{ij} < r_{ij}^{\text{nat}}(1-\delta) \\ -\epsilon & \text{if } r_{ij}^{\text{nat}}(1-\delta) \le r_{ij} < r_{ij}^{\text{nat}}(1+\delta) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where r_{ij} is the $C_\alpha - C_\alpha$ distance between residues i and j, r_{ij}^{nat} is the distance between the residues in the experimental structure, ϵ determines the energy scale of the native Gō attraction, and δ sets the width of the well. All residues that are separated by *more* than R_{cut} in the experimental structure are given non-native interaction energies defined by

$$U^{\mathrm{non}} = \sum_{i<j}^{\text{non-native}} u^{\mathrm{non}}(r_{ij})$$

$$u^{\mathrm{non}}(r_{ij}) = \begin{cases} \infty \text{ if } r_{ij} < (\rho_i + \rho_j)(1-\delta) \\ +h\epsilon \text{ if } (\rho_i + \rho_j)(1-\delta) \le r_{ij} < R_{\mathrm{cut}} \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

where ρ_i is the hard-core radius of residue i, defined as half the C_α distance to the nearest noncovalently bonded residue in the experimental structure, and h determines the strength of the repulsive interaction.

For this study, parameters were chosen to be similar to those in ref 26, that is, $\epsilon = 1.0$, $h = 0.3$, $\delta = 0.1$, and $R_{\mathrm{cut}} = 8.0$ Å.

2.2. Monte Carlo Simulation. The protein fluctuations were generated using Metropolis Monte Carlo,[44] although discontinuous molecular dynamics could also be used.[23] Trial configurations were generated by adding a random Gaussian deviate to the values of three sequential pairs of backbone torsions (three ϕ and three ψ angles). We found that changing six sequential backbone torsions maximizes the rate of convergence of the equilibrium ensemble (data not shown). The energy of the trial configuration was then determined using eq 1, and the conformation was accepted with probability min(1,e $^{-\Delta U/k_BT}$), where ΔU is the total change in potential energy of the system. The width of the Gaussian distribution for the generation of random deviates was chosen such that the acceptance ratio was about 40% for all simulations. The choice of temperature is discussed below.

2.3. Use of Lookup Tables. The speed of the coarse-grained simulation was enhanced by using lookup tables to avoid unnecessary computation. In general, lookup tables will increase memory usage, while decreasing the number of computations. Since memory is inexpensive and can be expanded easily, using as much memory as possible can be an effective technique for increasing the speed of simulations.

In our model, there are only two degrees of freedom per residue (ϕ and ψ), but the C_α distances, r_{ij}, must be computed to determine native and non-native interaction energies given by eqs 2 and 3. All peptide planes are considered to possess ideal, rigid geometry as determined by energy minimization of the all-atom OPLS forcefield[45] using the TINKER simulation package.[46]

Given a sequence of three residues (α carbons), we employed a lookup table to provide the Cartesian coordinates of the third residue, starting from the N-terminus, and its normal vector as a function of ϕ and ψ; see Figure 1. The table values assume that the first residue is at the origin and that the second residue is located on the z-axis. Once the coordinates for the third residue were determined via the lookup table, the fourth residue position was determined using the lookup table in conjunction with a coordinate rotation and shift. Continuing in this fashion, coordinates for the entire protein were determined.

The resolution of the lookup table is an important consideration, that is, the number of ϕ and ψ values for which Cartesian coordinates are stored. In our simulations, we tried resolutions as high as 0.1° and as low as 1.0° and found no difference between the results. Thus, all simulation results presented here use tables with a resolution of 1.0°.

2.4. Initial Protein Relaxation. One perhaps unexpected complication of the use of a rigid peptide plane model is that great care must be taken to relax the protein before the simulations can be performed. Although initial values of ϕ and ψ are obtained from the X-ray or NMR structure, there are slight deviations from planar/ideal geometry in a real protein. These deviations, while small, can accumulate rapidly to become very large differences in the Cartesian coordinate positions of the residues. Thus, the positions of residues near the beginning of the protein will be nearly correct, while the residues near the end of the protein will likely have large errors, compared to the experimental structure being modeled, which can create severe steric clashes or even incorrect protein topology. The severity of these “errors’’ necessitates the use of a relaxation procedure to generate a suitable starting structure, that is, a set of ϕ and ψ angles which, with our ideal-geometry peptide planes, leads to a topologically reasonable and relatively clash-free structure. Proteins larger than 100 residues are difficult to relax.

Before we detail our relaxation procedure, we note that the need for this additional calculation is an artifact of the simplicity of our model, which can be overcome. For instance, it is possible to include *flexible* peptide planes with the use lookup tables (libraries of plane configurations, in this case) without significantly increasing the computational cost of the model. Such an approach, which does not require initial relaxation, is currently under investigation with promising preliminary results (data not shown).

The relaxation procedure employed in the present study first uses the ϕ and ψ values directly obtained from the experimental structure. These dihedrals provide the initial (problematic) structure for a coarse-grained simulation. Because of the deviations from planarity described above, the root-mean-square deviation (RMSD) between the initial structure we create and the experimental structure tends to be large (~10 Å was not uncommon for the proteins in this study). To increase the number of native contacts and reduce the number of steric clashes, we next performed what we term “RMSD Monte Carlo’’ to relax the protein to a low RMSD structure. Trial moves for RMSD Monte Carlo were created as described above but accepted with probability min(1,e $^{-\Delta(\mathrm{RMSD})/k_BT_{\mathrm{RMSD}}}$), where $k_BT_{\mathrm{RMSD}} = 10^{-7}$ was chosen so that moves to a higher RMSD were rare. In other words, the energy function itself was not used in this initial phase.

Since residues near the beginning of the protein have less error in the starting structure than residues near the end, we used RMSD Monte Carlo in segments. The first twenty residues were relaxed until the RMSD was constant within a tolerance of 0.0001 Å, followed by the first forty, then the first sixty, and so on, until the RMSD of the entire protein was relaxed. The RMSD Monte Carlo simulation typically brought the RMSD of the simulated structure to less than 0.5 Å; however, there were generally still steric clashes, and some native contacts were still not present.

The final stage of relaxation was to do regular Metropolis Monte Carlo simulation (i.e., using energy), with a very low temperature. The low-temperature Monte Carlo eliminated all steric clashes and allowed the remaining native contacts to form.

Relaxation was performed until four criteria were met: (i) the number of native contacts in the relaxed structure was equal to that in the NMR or X-ray structure, (ii) no steric clashes were present, (iii) no non-native contacts were present, that is, $U^{\mathrm{non}} = 0$ in eq 3, and (iv) the RMSD was less than 1.0 Å. When these criteria were met, the structure was saved and used as the starting configuration in all future simulations of the protein.

3. Results and Discussion

Using the coarse-grained protein model described above, we generated and tested equilibrium ensembles for three proteins: barstar (PDB entry 1A19, residues 1−89), the N-terminal domain of calmodulin (PDB entry 1CLL, residues 4−75), and the binding domain of protein G (PDB entry 1PGB, residues 1−56).

For each protein, the initial simulation structure was generated, followed by RMSD and energy relaxation, as described above. Then, production runs of 2×10^9 Monte Carlo moves were performed with snapshots saved every 1000 moves, generating an equilibrium ensemble with 2×10^6 frames.

In an attempt to obtain consistent results for the three proteins, we chose the temperature of the simulation, k_BT, to be slightly below the unfolding temperature of the protein. The unfolding temperature was determined by running simulations over a broad range of temperatures and by studying the RMSD as a function of simulation time. The temperatures used in the simulations were $k_BT = 0.6$ for barstar, $k_BT = 0.4$ for calmodulin, and $k_BT = 0.5$ for protein G.

3.1. Speed of simulations. Because of the use of lookup tables for coordinate transformations, the small number of degrees of freedom, and the use of simple square potentials, equilibrium ensembles were generated very rapidly.

When run on one Xeon 2.4 GHz processor, 2×10^9 Monte Carlo moves with snapshots saved every 1000 steps took roughly 6 days for barstar, 4 days for calmodulin, and 3 days for protein G. Thus, less than a week was required to obtain well-converged simulations of these coarse-grained proteins.

The unfolding temperature for each protein was determined via short simulations of 10^7 Monte Carlo moves for around twenty different temperatures. This additional computational cost to determine the unfolding temperature was roughly 14 h for barstar, 10 h for calmodulin, and 7 h for protein G.

3.2. Protein Fluctuations. We first sought to determine whether fluctuations in the coarse-grained simulations are reasonable, bearing in mind that the model was designed for speed rather than chemical accuracy. Figure 2 shows the α-carbon relative root-mean-square fluctuation for three different proteins. Given the model's minimalist design, the figures show that there is reasonable qualitative agreement among the NMR, X-ray, and simulation data, with the notable exception of barstar, where the experimental peak at around residue 65 does not appear in the simulated results. Interestingly, for calmodulin, a standard Pearson correlation analysis indicates that the simulation data agree better with *both* experiments than the experiments do with one another.

We emphasize that the purpose of the current study is to understand the promise and limitations of coarse models for statistically valid sampling. We made no attempt to design the most chemically realistic coarse-grained model and, therefore, sought only rough qualitative agreement between experimental and simulated results. The protein systems reported here were not "cherry picked'' but intended to provide a representative picture of results obtainable in a minimalist model.

In addition, it should be noted that *none* of the three data sets in Figure 2a−c represents the true fluctuations in the protein, for different reasons. The X-ray temperature factor, in addition to thermal fluctuations, includes crystal lattice artifacts and other experimental errors.[48] NMR ensembles tend to be biased, perhaps severely, toward low-energy structures, and thus also do not represent equilibrium ensembles.[49] Finally, our simulation data is not accurate because of the lack of chemical detail in the forcefield.

The bottom panels of Figure 2 show the whole-molecule fluctuations exhibited throughout the trajectories. In addition to the ability to sample large conformational fluctuations, such as in the case of calmodulin and, to a lesser degree, for protein G, the trajectories are visibly more converged than is typically observed in atomistic simulations, where RMSD values rarely reach a plateau value, let alone sampling around that plateau value multiple times as would be desirable.

3.3. Convergence Analysis. The primary purpose of this report is to demonstrate the convergence of the equilibrium ensemble for a coarse-grained protein. The details of the convergence analysis are described in ref 37, so we will only briefly describe the method here.

Previously, Lyman and Zuckerman[1] developed an approach which classifies sampled conformations into the bins of a "structural histogram'' using the RMSD as a metric. While promising, the primary limitation of the method was the lack of a quantitative measure of the convergence.

In the method used here, convergence was analyzed by studying the variance of the structural histogram bin populations.[37] The new approach allows a rigorous *quantitative* estimation of convergence via the structural decorrelation time, τ_{dec}, given by the time between frames required for the variance to reach an analytically computable independent-sampling value. Intuitively and mathematically, τ_{dec} is the time interval between snapshots for which they behave as if each frame were sampled independently. If simulation times of $t_{\mathrm{sim}} \gg \tau_{\mathrm{dec}}$ are obtained, the equilibrium ensemble is considered converged.

Perhaps the most important feature of the convergence analysis for our study is that the method does not require

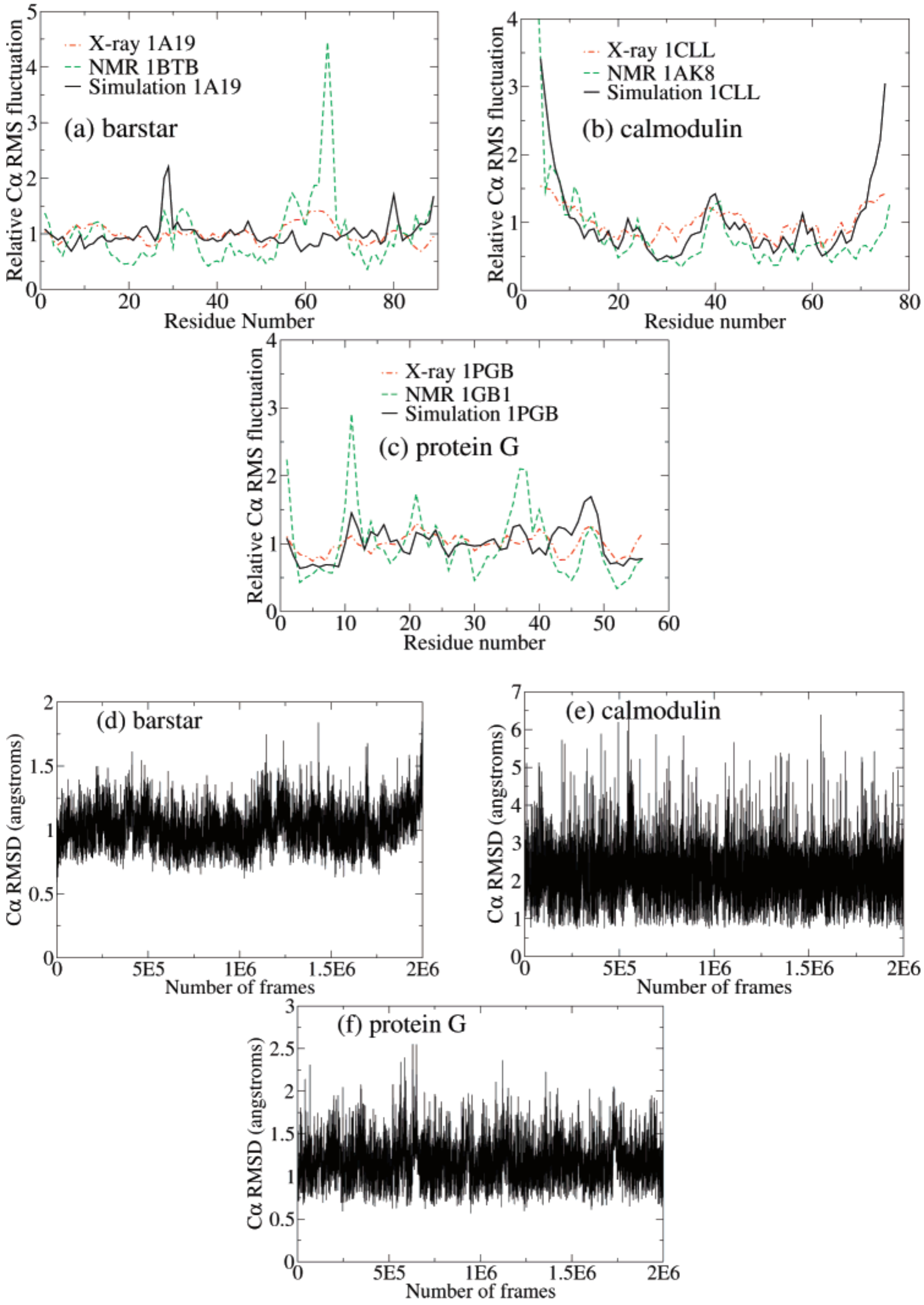


Figure 2. Relative α-carbon root-mean-square fluctuations for three different proteins: (a) barstar, (b) calmodulin, and (c) protein G. Each plot shows results from X-ray data (• - - •), the NMR ensemble (- - -), and the coarse-grained simulation (—). X-ray results were given by $\sqrt{3B/8\pi^2}$, where B is the temperature factor given in the PDB entry. NMR and simulation data were generated using the g_rmsf program in the GROMACS molecular simulation package;[47] each ensemble was aligned to the first structure in the corresponding trajectory. For each coarse-grained simulation, 2 × 10^9 Monte Carlo steps were performed with snapshots saved every 1000 steps, and the potential energy (1) was set up using the X-ray structure. Panels d−f show the corresponding whole-structure fluctuations as indicated by the RMSD from the experimental structures.

any prior knowledge of important states. Furthermore, there is no parameter-fitting or subjective analysis of any kind.

Figure 3 shows the convergence properties of the coarse-grained simulations using the same trajectories as in Figure 2. The ratio of the observed variance to the ideal variance for independent sampling is plotted as a function of the time between the configurations used to compute the observed variance. When this ratio decreases to 1, the structural decorrelation time τ_{dec} has been reached, as shown in the figure. The analysis indicates that each simulation is at least 50 times longer than the structural decorrelation time.

We conclude that, in less than a week of single-processor time, the equilibrium ensembles for these three proteins are well converged.

4. Conclusions

We have demonstrated the convergence of the equilibrium ensemble for a simple united-residue protein model. The

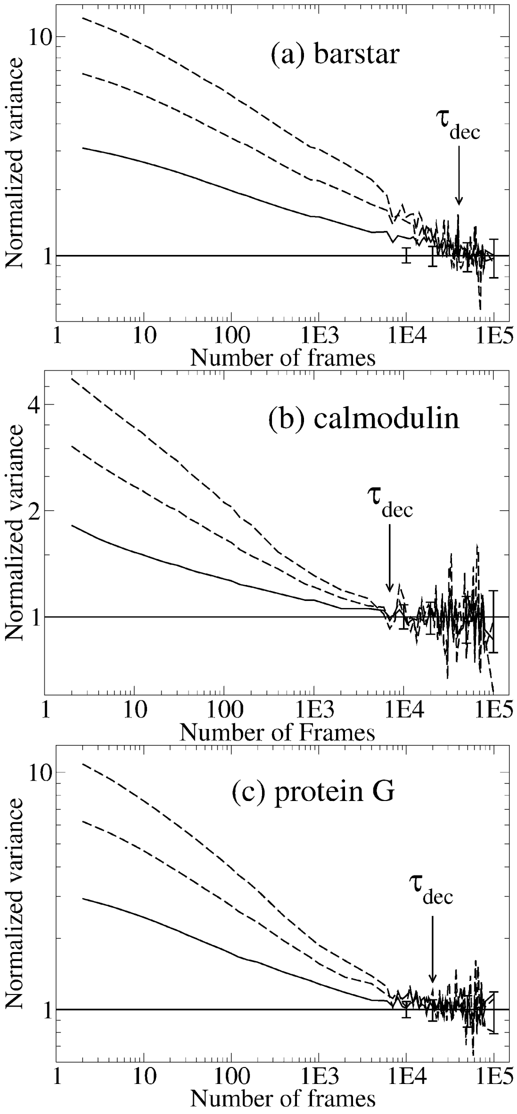


Figure 3. Convergence analysis for coarse-grained simulations of three different proteins: (a) barstar, (b) calmodulin, and (c) protein G. Each plot shows the convergence properties for the same trajectories as used for Figure 2, analyzed using the procedure in ref 37. The number of frames required to reach the value of one (horizontal line) is an approximation of the structural decorrelation time, τ_{dec}, and is shown on each plot. The three curves on each plot are results for different histogram subsample sizes[37] and demonstrate the robustness of the value of τ_{dec}. The plots predict that the decorrelation times are roughly 40 000 frames for barstar, 20 000 frames for calmodulin, and 30 000 frames for protein G. Note that the total number of frames generated for each protein during the simulation was 2×10^6. Thus, since each simulation was more than $50\tau_{dec}$ in length, we conclude that the equilibrium ensembles are well-converged. Error bars represent 80% confidence intervals in the expected fluctuations around the ideal value of one, based on the given trajectory length and the numerical procedure used to generate the solid curve.

model assumes rigid peptide planes, with atomistically correct geometry, and exhibits reasonable residue-level fluctuations based on the planes' geometry, Gō interactions, and sterics.

Most importantly, the results indicate *quantitatively* that carefully designed united-residue models have the potential to fully sample protein fluctuations. By using only 2 degrees of freedom per residue, lookup tables for coordinate transforms, and simple square well potentials, we were able to demonstrate that converged equilibrium ensembles can be obtained in less than a week of single-processor time. The quantitative convergence analysis indicates that more than 50 "decorrelation times'' were simulated in each case, indicating high-precision ensembles. In addition to application in resolution exchange sampling of all-atom models,[27,28] such speed opens up the long-term possibility of large-scale simulation of many proteins.

One important practical limitation of the ideal-peptide-plane geometry in the present model is the need to relax the initial structure. Proteins larger than 100 residues are difficult to relax. However, we have already begun investigating a flexible-plane model, incorporating precalculated libraries of plane configurations, which exhibits no such limitation and remains computationally affordable. We will report on the flexible model in the future.

Although the intrinsic atomistic geometry of the peptide plane was included in our model, it lacks chemical interactions. Yet because we obtained converged ensembles in such a short time, it is clear we can "afford'' extensions to the model which include realistic chemistry. For instance, additional potential energy terms such as Ramachandran propensities,[43] hydrophobic interactions,[15] and hydrogen-bonding can be included at small cost.

In addition to the potential for rigorous atomistic sampling,[27,28,50] it is important to note the general usefulness of coarse-grained models for generation of ad hoc atomistic ensembles. Specifically, upon generation of a well-sampled ensemble of coarse-grained structures, atomic detail can be added with existing software, such as those in refs 51 and 52. Once minimized and relaxed, these (now) atomically detailed structures form an ad hoc ensemble that may be of immediate use in docking[53,54] and homology modeling applications. Further, in principle, such structures can be reweighted into the Boltzmann distribution.[50]

In the long term, one can imagine a day when structural databases will be based not on single (static) structures but rather will collect ensembles, as envisioned in the authors' scheme for an "Ensemble Protein Database'' (http://www.epdb.pitt.edu/).

Acknowledgment. We thank Edward Lyman, Bin Zhang, Artem Mamonov, and David Richardson for helpful discussions. Funding was provided by the National Institutes of Health under Fellowship GM073517 (to F.M.Y.) and Grants GM070987 and ES007318 and by the National Science Foundation Grant MCB-0643456.

References

(1) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*, 164−172.

(2) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(3) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.

(4) Hansmann, U. H. E. *Chem. Phys. Lett*. **1997**, *281*, 140−150.

(5) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett*. **1999**, *314*, 141−151.

(6) Paschek, D.; Garcia, A. E. *Phys. Rev. Lett*. **2004**, *93*, 238105−1−4.

(7) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput*. **2006**, *2*, 1200−1202.

(8) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694−698.

(9) Ueda, Y.; Taketomi, H.; Gō, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.

(10) Tanaka, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3802−3806.

(11) Kuntz, I. D.; Crippen, G. M.; Kollman, P. A.; Kimmelman, D. *J. Mol. Biol*. **1976**, *106*, 983−994.

(12) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.

(13) Skolnick, J.; Kolinski, A.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5057−5061.

(14) Friedrichs, M. S.; Wolynes, P. G. *Science* **1989**, *246*, 371−373.

(15) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986−3997.

(16) Honeycutt, J. D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3526−3529.

(17) Monge, A.; Lathrop, E. J. P.; Gunn, J. R.; Shenkin, P. S.; Friesner, R. A. *J. Mol. Biol.* **1995**, *247*, 995−1012.

(18) Jernigan, R. L.; Bahar, I. *Curr. Op. Struct. Biol.* **1996**, *6*, 195−209.

(19) Zhou, Y.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14429−14432.

(20) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849−873.

(21) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874−887.

(22) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5871−5876.

(23) Smith, A. V.; Hall, C. K. *Proteins* **2001**, *44*, 344−360.

(24) Shimada, J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11175−11180.

(25) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896−10913.

(26) Zuckerman, D. M. *J. Phys. Chem. B.* **2004**, *108*, 5127−5137.

(27) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Phys. Rev. Lett*. **2006**, *96*, 028105−1−4.

(28) Lyman, E.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2006**, *2*, 656−666.

(29) Lwin, T. Z.; Luo, R. *J. Chem. Phys.* **2005**, *123*, 194904−1−10.

(30) Christen, M.; van Gunsteren, W. F. *J. Chem. Phys.* **2006**, *124*, 154106−1−7.

(31) Liu, P.; Voth, G. A. *J. Chem. Phys.* **2007**, *126*, 045106−1−6.

(32) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937−953.

(33) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. *Biochemistry* **1993**, *32*, 412−420.

(34) Straub, J. E.; Rashkin, A. B.; Thirumalai, D. *J. Am. Chem. Soc.* **1994**, *116*, 2049−2063.

(35) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins* **2002**, *48*, 487−496.

(36) Elmer, S. P.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 12760−12771.

(37) Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B*, in press.

(38) Ueda, Y.; Taketomi, H.; Gō, N. *Biopolymers* **1978**, *17*, 1531−1548.

(39) Gō, N.; Taketomi, H. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 559−563.

(40) Careaga, C. L.; Falke, J. J. *J. Mol. Biol.* **1992**, *226*, 1219−1235.

(41) Bai, Y.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192−197.

(42) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438*, 117−121.

(43) Lovell, S. C.; Davis, I. W.; Arendall, W. B., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437−450.

(44) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(45) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *117*, 11225−11236.

(46) Ponder, J. W.; Richard, F. M. *J. Comput. Chem.* **1987**, *8*, 1016−1024.

(47) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem*. **2005**, *26*, 1701−1718.

(48) Northrup, S. H.; Pear, M. R.; McCammon, J. A.; Karplus, M.; Takano, T. *Nature* **1980**, *287*, 659−660.

(49) Spronk, C. A. E. M.; Nabuurs, S. B.; Bonvin, A. M. J. J.; Krieger, E.; Vuister, G. W.; Vriend, G. *J. Biomol. NMR* **2003**, *25*, 225−234.

(50) Ytreberg, F. M.; Zuckerman, D. M. http://arxiv.org/abs/physics/0609194 (accessed Sept 22, 2006).

(51) Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V. *J. Comput. Chem.* **2004**, *25*, 712−724.

(52) de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins Struct. Funct. Genet.* **2002**, *51*, 21−40.

(53) Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. *J. Mol. Biol*. **1997**, *266*, 424−440.

(54) Shoichet, B. K. *Nature* **2004**, *432*, 862−865.

CT700076T

JCTC Journal of Chemical Theory and Computation

Coarse-Graining the Accessible Surface and the Electrostatics of Proteins for Protein−Protein Interactions

Francesco Pizzitutti,[†,‡] Massimo Marchi,*[,†] and Daniel Borgis*[,‡]

Commissariat à l'Énergie Atomique, DSV-DBJC-SBFM, CNRS URA 2096,
Centre d'Études Saclay, 91191 Gif-sur-Yvette Cedex, France, and Laboratoire
d'Analyse et Modélisation pour la Biologie et l'Environment, CNRS UMR 8587,
Université Evry-Val-d'Essonne, Boulevard François Mitterrand, 91405 Evry, France

Received May 21, 2007

Abstract: This study is concerned with the development and test of a coarse-grained representation specifically constructed for proteins and peptides, where each amino acid of the sequence is represented by a charged dipolar sphere. The model was parametrized from the physical properties of individual amino acids and applied to the study of the interaction between solvated proteins. Using an implicit solvent approach and our coarse-grained model, we computed the potential of mean force for the association of well-known proteins, such as the Cu−Zn superoxide dismutase, lysozyme, and basic pancreatic trypsin inhibitor, and a peptide, Aβ_7. The coarse-grained potentials of mean force were systematically compared with their all-atoms counterpart. For both the polar and nonpolar contributions to this potential, the results of our calculations show that the coarse-grained model provides a good approximation of the all-atoms potential when the distance between the molecule surfaces is greater than a solvent molecular diameter. For shorter distances and for specific interactions, like those found between the SOD monomers, the electrostatic desolvation effect appears to be underestimated by our coarse-grained representation. The possibility of a very short range all-atom refinement to better describe the interaction at close contact is explored. We find also that the most important contribution to the association free-energy comes from the hydrophobic solvent accessible surface area term, which is well reproduced by our coarse-grained approach.

I. Introduction

The development of molecular modeling has provided valuable insights to the understanding of interactions between biomolecules, proteins, and DNA. Because biological processes cover a broad range of time and length scales, progress needs to be made to adapt computational techniques to different levels of detail in the description of the systems. Indeed, in standard molecular dynamics (MD) simulations, solvent and biomolecules are represented at the atomic level. This implies that for an average size protein, the number of simulated atoms can easily reach a few tens of thousands atoms. Despite the ever increasing power of computers and the considerable effort undertaken to ameliorate molecular modeling algorithms, the time and length scales currently approachable in atomistic simulations are limited to a few hundred of angstroms and a few tens of nanoseconds, respectively. This means that even processes implying interactions between proteins and between proteins and DNA, crucial to life on earth, are currently out of reach of all-atoms (AA) computer simulations.

Although protein−protein interfaces have been widely studied[9,10] and calculations of free energy of association are now customary in the analysis of protein−ligand interac-

* Corresponding author e-mail: Massimo.Marchi@cea.fr (M.M.); Daniel.Borgis@univ-evry.fr (D.B.).

[†] Centre d'Études Saclay.

[‡] Université Evry-Val-d'Essonne.

10.1021/ct700121n CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/24/2007

tions,[11] the study of protein−protein associations involving all-protein and solvent atomic degrees of freedom are not easily manageable by an all-atom simulation. Indeed, the partial desolvation of the molecular surfaces occurring in the process of protein association requires very long simulation times.

A possible strategy to increase the length and time scales spanned by molecular simulation is to reduce the level of detail of the atomic systems. Formerly, the most common approach has been to remove non-relevant degrees of freedom from the simulated system: for example, see ref 1. As an example, in some of these coarse-grained (CG) models the protein amino acids are represented with one to six centers only, whereas the interaction energies among residues is knowledge-based, procured from the database of the many native structures of proteins available today. Such a class of models has been used in docking[2] and folding[3] studies. In particular, most of the docking methods are based on some energetic scoring function relying on a simplified picture of the interactions involved. This goes from simple pattern recognition and simplified electrostatics to all-atom force fields at some final stage of the minimization process. In this context, Zacharias[2] recently developed an empirical coarse-grained representation of proteins in terms of a few effective sites per residue. Following a different, but related approach, Klein and co-workers developed effective CG interaction potentials for phospholipids,[4] on the basis of results procured from all-atom models.

Because most of the biology occurs in a wet milieu, other CG methods have been developed where the explicit molecular representation of the solvent is replaced by a static mean-field representation. These approaches provide a simple way to compute the solvation free energy for any given solute. In the most common Poisson−Boltzmann surface-area (PB-SA) approximations, the polar solvent is represented as a structureless continuum medium of dielectric constant ϵ_s, whereas the solute is well described by a point-charge distribution embedded in a medium with dielectric constant ϵ_p. In general, the electrostatic boundary between solvent and solute, in this view, regions of high and low dielectric constant, respectively, is defined as the molecular surface accessible to the solvent. The nonpolar (hydrophobic) contributions to the solvation free energy are assumed to be proportional to the solvent accessible surface area (SASA) of the solute.

In this work, we present a CG model of proteins, similar in the spirit to that developed by Song.[7,8] Our primary objective here is the study of the protein−protein interactions. Indeed, the understanding of the forces between solvated proteins and, more generally, between biomolecules is crucial to phenomena such as cell signaling and protein crystallization, which are driven by noncovalent specific protein−protein interactions. Thus, our approach does not aim, at least at first, at dealing with protein−protein docking.

Ours is a so-called bottom-up approach, which proceeds in much the same way that atomistic potentials are derived from quantum chemistry ab initio calculations: We compute the relevant microscopic properties of associated proteins from atomistic modeling and then fit the parameters of our CG model. This, we hope, will ensure the highest level of accuracy and transferability. Our final objective is to produce a CG residue-based force-field, which provides a good approximation of the potential of mean force (PMF) for protein−protein nonspecific interactions. Thus, we adopt a CG model where each amino acid residue is replaced by a van der Waals sphere, a charge, and a dipole. The corresponding parameters are fitted to reproduce relevant all-atom properties of the systems.

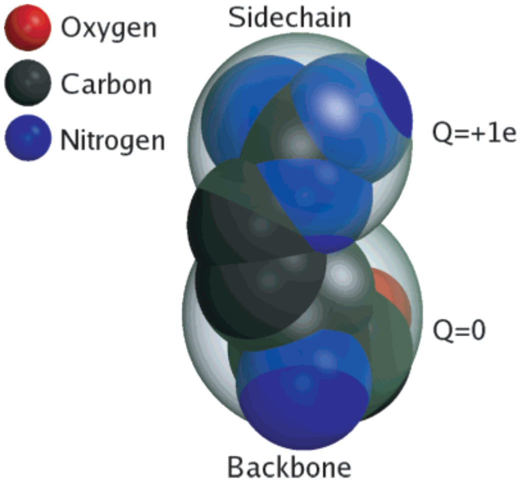


Figure 1. SASA of an arginine residue in its AA (red, black, and blue opaque balls) and CG (gray transparent balls) representations. Note that at pH 7 the arginine has a total net charge equal to −1e. Accordingly, this amino acid is represented in our CG approach by two CG spheres: The first one, labeled $Q = 0$, models the uncharged backbone part of the residue, whereas the second one, labeled $Q = +1e$, models side-chain atoms where the excess charge of −1e is distributed.

Our CG model is tested here on three widely studied proteins, Cu−Zn superoxide dismutase (SOD), lysozyme, and basic pancreatic trypsin inhibitor (BPTI), as well as on a small segment of the amyloid-forming peptide Aβ_7. The test consists of a comparison of the potential of mean force (PMF) for homodimer association, obtained using either a CG or AA representation with fixed internal geometries.

II. Materials and Methods

A. Coarse-Grained Model. We built the CG representation of a protein by associating one CG sphere to every uncharged residue and two CG spheres to every charged residue (Lys, Arg, Glu, and Asp residues and the protein terminals, which were considered to be charged at pH 7). For a neutral residue, the CG sphere was centered on the center of mass of the AA residue. For a charged residue, the first sphere was centered on the center of mass of the neutral part of the residue, whereas the second sphere was centered on the center of charge of the charged part (see Figure 1 for a representation of the CG model of an arginine residue). Every CG sphere was characterized by a radius, R_{CG}, a charge, Q_{CG}, and an electric dipole, $\mathbf{P}_{CG}$.

Coarse-Grained Radii. For any given protein, the R_{CG} values were chosen to procure a CG solvent-accessible surface area for each CG residue of the protein (SASA$_{CG}$)

Table 1. CG Parameters for the Aβ7 Peptide[a]

	residue	SASA$_{AA}$ (Å^2)	SASA$_{CG}$ (Å^2)	R_{CG} (Å)	p_{AA} (e Å)	P_{CG} (e Å)
1	Lys0	246.9	235.0	4.0	0.79	0.77
2	Lysq	95.7	121.1	2.4	0.00	0.00
3	Leu	139.0	104.9	3.6	0.66	0.68
4	Val	131.2	118.0	3.4	1.03	1.41
5	Phe	209.6	211.1	4.0	0.54	1.30
6	Phe	157.6	149.0	4.0	0.28	1.58
7	Ala	88.4	75.6	3.0	1.81	0.93
8	Glu0	123.1	127.7	3.2	0.98	0.89
9	Gluq	115.9	127.4	2.7	0.00	0.00

[a] The table presents the 9 CG radii (R_{CG}) and the absolute values of the 9 optimized CG dipoles (P_{CG}) for the Aβ_7 peptide. Note that Lys1 and Glu9, two charged residues, are represented by two CG spheres each, Lys0 and Glu0, with a dipole in their center and (Lysq and Gluq) with only a charge in their center. The AA and CG SASAs, SASA$_{AA}$ and SASA$_{CG}$, respectively, and the absolute values of the AA dipoles are also reported.

equal to its corresponding AA SASA$_{AA}$. The latter is defined as the SASA of the residue in the actual conformation it has in the native structure of the given protein. In particular, the SASA$_{AA}$ of a given residue was computed after the atomic radii of all remaining protein residues are switched off. Throughout this paper, SASA is defined as the surface traced by the center of a rolling probe sphere of a radius of 1.4 Å over the van der Waals surface of the solute molecule itself.[14] From the SASA$_{AA}$, we extract the R_{CG} through the relation

$$R_{CG} = \sqrt{\frac{\mathrm{SASA}_{AA}}{4\pi}} - 1.4 \qquad (1)$$

where the radii are expressed in Å.

For the test proteins considered in this work, the obtained R_{CG} values in a range from 2 to 4 Å. As an example, the radii distribution for the Aβ_7 peptide are reported in Table 1.

Coarse-Grained Charges and Dipoles. The set of CG charges, $\{Q_{CGi}\}$, and CG dipoles, $\{\mathbf{P}_{CGi}\}$, was obtained by optimization of the electrostatic potential generated by the CG systems around the protein to reproduce, most accurately, the electrostatic potential generated by the AA system. More specifically a mean square deviation function, $\Psi(Q_{CG}, \mathbf{P}_{CG})$, depending on the set of the $\{Q_{CGi}\}$ and $\{\mathbf{P}_{CGi}\}$, was defined as

$$\psi(Q_{CG}, \mathbf{P}_{CG}) = (\sum_{i=1}^{N_{res}} \sum_{k=1}^{N_{grid}} \phi_{CG}\,(Q_{CGi}, \mathbf{P}_{CGi}, \mathbf{x}_k) - \phi_{AA}(q_{AAi}, \mathbf{x}_k))^2 \qquad (2)$$

where the electrostatic potentials $\phi_{AA}(q_{AAi}, \mathbf{r})$ and $\phi_{CG}(Q_{CGi}, \mathbf{P}_{CGi}, \mathbf{r})$, generated by the AA and CG systems, respectively, are discretized over a grid and evaluated at every point $\mathbf{x}_i$ of the grid. N_{res} is the total number of residues and N_{grid} is the number of grid points. The discretization grid was built inside an orthorombic box containing the AA system. The dimensions of the box were chosen in such way that the AA system was contained in 75% of the box volume. The distance between the grid points was always between 0.5 and 1.0 Å, depending on the system considered. The deviation function $\Psi(Q_{CG}, \mathbf{P}_{CG})$ was then minimized with respect to the set $\mathbf{P}_{CGi}$ through a conjugated gradient algorithm. We point out that the CG charges were not optimized, but were constrained to their assigned values.

B. Interaction Free Energy Calculations. To evaluate the interaction free energy between biomolecules, we used an implicit solvent approach. As noted by Roux and Simonson,[15] the total solvation free energy of a molecule in water can be expressed as the sum of a polar term, resulting from electrostatic interactions, and a nonpolar term, from the Van der Waals and hydrophobic contributions: $\Delta G_{solv} = \Delta G_{elect} + \Delta G_{np}$.

From this relation, we can obtain the interaction free energy, G_{int}, between two molecules, A and B, at a distance r, as the change in ΔG_{solv} when the two molecules approach each other to a distance r from infinity plus the direct electrostatic interaction V_{int}. The resulting expression for the interaction free energy is

$$G_{int} = \Delta\Delta G_{solv} + V_{int} = \Delta\Delta G_{elect} + \Delta\Delta G_{np} + V_{int} \qquad (3)$$

Here, we have

$$\Delta\Delta G_{solv} = \Delta G^{A+B}_{solv}(r) - \Delta G^{A}_{solv} - \Delta G^{B}_{solv} \qquad (4)$$

$$\Delta\Delta G_{elect} = \Delta G^{A+B}_{elect}(r) - \Delta G^{A}_{elect} - \Delta G^{B}_{elect} \qquad (5)$$

$$\Delta\Delta G_{np} = \Delta G^{A+B}_{np}(r) - \Delta G^{A}_{np} - \Delta G^{B}_{np} \qquad (6)$$

where $\Delta G^{A+B}_{solv}(r)$ and $\Delta G^{A,B}_{solv}$ denotes the solvation free energy of a system composed of two molecules, A and B, when the distance between the two molecules is r and when they are isolated, respectively. The notation is the same for the electrostatic and nonpolar contributions, ΔG_{elect} and ΔG_{np}, respectively. It should be noted that, in principle, the interaction free energy definition of eq 3 does not account for the totality of the free energy changes caused by the interactions between the two molecules. Specifically, the following contributions from the interaction between the two molecules are neglected: the cost of structural reorganization, the loss of configurational reorganization entropy, and the loss of the translational and rotational degrees of freedom of the two molecules. In what follows, these contributions will not be taken into account, which is equivalent to considering the two interacting molecules to be rigid bodies. Thus, our calculations consider only contributions to the interaction free energy coming from direct molecule−molecule interactions and desolvation effects.

Polar Interaction Energy Contribution. The electrostatic term, G_{elect}, is defined as

$$G_{elect} = \Delta G^{A+B}_{elect}(r) - \Delta G^{A}_{elect} - \Delta G^{B}_{elect} + \frac{1}{\epsilon_p} \sum_{i\epsilon A, j\epsilon B} \frac{q_i q_j}{|x_i - x_j|} \qquad (7)$$

In this work, we calculated the first three terms in a continuum solvent context by numerically solving the Poisson−Boltzmann (PB) equation.[16] Following this approach, the water and the interior of molecules are treated as dielectric continuum of dielectric constant ϵ_s and ϵ_p, respectively. The volume occupied by the solute is defined by its molecular surface (obtained using a solvent radius of 1.4 Å). This volume contains, in full, the partial charges of the solute atoms.[17,18] The resulting electrostatic problem is treated numerically by solving the related Poisson equations with the ionic force set to zero. The last term in eq 7 is the pairwise Coulomb free energy between the two interacting molecules embedded in a dielectric medium of identical dielectric constant that their interior. The indexes i and j refers to molecules A and B, respectively; q and $\mathbf{x}$ are the charge and the position of the particles.

Non-Polar Interaction Energy Contribution. The non-polar term accounts for free energy differences resulting from both formation of cavities in the solvent and van der Waals interactions after the removal of the solvent from the interface between two interacting proteins. This term can be calculated according to eq 6. Every term on the right-hand side of this equation can be calculated using the well-established linear relation[19,20] that connects experimental hydration energies of small alkanes chains with their surfaces

$$\Delta G_{\mathrm{np}} = \gamma \mathrm{SASA} + b \tag{8}$$

where SASA is the solvent accessible surface of the molecule and γ and b are constants.

In summary, the final form of our interaction free energy is

$$G_{\mathrm{int}} = G_{\mathrm{elect}} + \Delta\Delta G_{\mathrm{np}} = G_{\mathrm{elect}} + \gamma \Delta \mathrm{SASA} \tag{9}$$

C. System Setups and Calculation Parameters. In this work, we have tested systematically our CG protein model by computing the interaction free energies between proteins using either the AA or CG representation. The systems chosen to test the accuracy of the CG model are small biomolecules, such as the 16−22th residue segment in the amyloid forming peptide Aβ_7 associated to the Alzheimer's disease,[21] and three globular proteins, Cu−Zn superoxide dismutase (SOD) from *Photobacterium leiognathi*,[22] egg-white lysozyme from *Gallus gallus* (Protein Data Bank 2lym),[23] and basic pancreatic trypsin inhibitor (BPTI) from *Bos taurus* (Protein Data Bank 1bpi).[24] For each molecule, we have generated a homodimer at close contact, and we have progressively increased the distance between the two monomers. The interaction free energy was calculated approximately every 0.5 Å until a maximum interdimer edge-to-edge distance of 25 Å. All degrees of freedom of the system, other than the dimer separation, were frozen. For technical reasons, the CG dipoles on every uncharged residue were represented by two identical charges of opposite sign located at the residue center of mass and placed at a distance of 1 Å from each other.

Among the chosen biomolecules, only SOD and the Aβ_7 peptide are known to form stable aggregates in solution. Aβ_7 organizes itself in antiparallel β-sheet fibrils,[21] whereas SOD yields a highly stable homodimer. In the case of SOD, we started our calculations from the structure of the SOD dimer presented in ref 22 that was equilibrated through an MD run. For the other proteins, an initial homodimer structure was built by superimposition of two identical AA single molecules configurations and then translation of one of them along the x-axis until the minimum distance between two atoms belonging to different halves was greater than 1.8 Å. Figure 2 displays the AA initial configuration of the SOD dimer, superimposed with the corresponding CG model. For lysozyme and BPTI, the AA geometry of the individual proteins was taken from a X-ray configuration from the PDB database. The molecule orientations with respect to the reference axis were the same as those in the PDB files. For the Aβ_7 peptide, the starting structure was obtained as a result of the equilibration of a MD run performed with the ORAC[25] program, in the NPT ensemble, using the CHARMM27[26] force field. The total length of the run was 500 ps, and the peptide was solvated by 543 tip3[27] explicit waters. The longitudinal axis of the obtained relaxed structure was then oriented parallel to the reference z-axis.

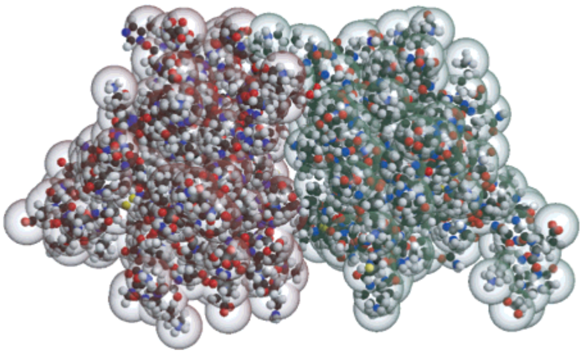


Figure 2. Pictorial representation of the coarse-grained SOD dimer. The AA SOD dimer (ball-and-stick) is superimposed to the CG representation (transparent, gray spheres).

Partial charges and radii for the AA systems were assigned according to the CHARMM27 force field.[26]

To eliminate numerical errors caused by the charge distribution interpolation used in the Poisson−Boltzmann solver, we computed the electrostatic contribution to the interaction energy in eq 5 through the relation

$$G_{\mathrm{elect}} = \Delta E^{\mathrm{A+B}}_{\mathrm{elect}} - \Delta E^{\mathrm{A}}_{\mathrm{elect}} - \Delta E^{\mathrm{B}}_{\mathrm{elect}} \tag{10}$$

where, for a fixed dimer configuration, the three terms correspond to the total electrostatic energy E (solvation energy, plus direct Coulomb interaction, plus grid self-energy) of A + B, A, or B, respectively. This procedure is more time-consuming than the direct application of eq 7 because it involves a recalculation of the A and B isolated contributions for each distance, but it eliminates numerical errors resulting from finite grid effects. The solution of the linearized PB equation was calculated through the finite differences version of the APBS[28] program.

The radius of gyration, R_{g}, of the considered considered molecules ranges from ~9 Å for the small Aβ_7 peptide to ~20 Å for SOD, whereas the volume spanned by the pairs of interacting molecules also depends on the distance between the two partners. In the Poisson−Boltzmann calculations, for every molecular pair and distance, the dimensions of our orthorombic discretization grid a, b, and c are chosen such that the two molecules are contained in 75% of the volume occupied by the grid. Moreover, the distance between consecutive grid points is always kept below 0.5 Å.

In the PB calculations, the dielectric constants of solute and solvent are of crucial importance. The latter was set to that of water, $\epsilon_s = 78.5$, while the former, ϵ_p, crucially depends on the model of solute used in the calculation.[29] While a solute dielectric constant of 2 is usually used to account for the electronic polarizability of the atoms, larger solute dielectric constants are employed in PB calculations to account for the atomic and orientational polarizability of the protein matrix.[16] In the past, ϵ_p values ranging from 4 to 20 have been used to study protein−protein interactions,[30] and $\epsilon_p = 17$[31] was used in molecular dynamics simulations of proteins. We remind the reader that the purpose of this work is to assess the validity of our CG description of biomolecules. This is done by comparison with the corresponding AA model under identical dielectric conditions. Thus, in all our PB calculations, the dielectric constant of the solute was set to $\epsilon_p = 2$ for both AA and CG systems.

As far as the calculation of the nonpolar terms was concerned, the value of the constant γ was set to 0.24 kJ/mol Å^2 for both the CG and AA systems, as previously done by several authors.[32,33] The ionic strength was set to 0 in all PB calculations.

All the calculations were performed on a 2.2 MHz Pentium processor. For SOD, the overall calculations performed for all distances took about 7 h of CPU time.

III. Results and Discussion

A. Bare Electrostatic Energy. The first step needed to build the CG model of a biomolecule is the calculation of the charge and dipole for every CG residue. Through the optimization process described in the Materials and Methods section, we have obtained the complete charges and dipoles distribution of the four test CG systems. We note that amino acid charges depends on the residue ionization state. In our calculations, the molecular ionization state was fixed to the average ionization state at pH 7. Thus, during the optimization procedure, CG charges were constrained to their AA original values at pH 7, and the residue dipoles were considered as the free variables.

As an example, in Table 1 we have reported, the modulus of the nonoptimized Aβ_7 peptide dipoles, p_{AA}, as they come directly from the point-charge distribution of the CHARMM27 force field, and the corresponding optimized quantities, P_{CG}. We notice that the optimization procedure has already a non-negligible effect on the electrostatic representation. After dipole optimization, we calculated the bare electrostatic interaction in a medium of uniform dielectric constant $\epsilon_s = 2$, that is, V_{int} in eq 3. In Figure 3, we compare the CG and AA models by displaying V_{int} as a function of the monomer−monomer distance, r in the picture, for our four biomolecules. The dimers for BPTI, lysozyme, and Aβ_7 were formed by pulling away one of the two monomers along the x-axis direction from an initial configuration where the two monomers were superposed. Given that the SOD dimer structure is known experimentally, the dimer structures were formed from the initial structure by translation along the intermonomer direction of one of the two monomers. Thus, the zero point for the distance in the SOD case corresponds to the experimental relaxed structure, whereas for the remaining proteins, $r = 0$ is chosen as the conformation for which the minimum distance between two atoms belonging to two distinct monomers is 1.8 Å.

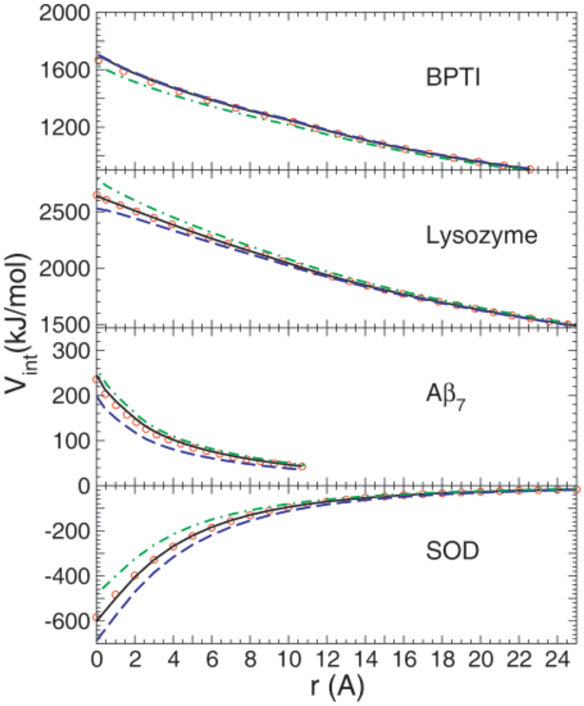


Figure 3. Dependence of the bare electrostatic interaction energy on the intermonomer distance, *r*. Starting from the initial reference configurations at *r* = 0 (see text), for all dimers, *r* was increased along a direction perpendicular to the dimer separation plane. In the pictures, the circles correspond to the AA systems and the solid curve to the full CG systems composed of charges and optimized dipoles. The other two curves corresponds to two models with nonoptimized dipoles (dashed line) and without dipoles (dashed-dotted line).

Figure 3 shows that the CG dipoles obtained by the optimization procedure described in section II are able to yield a very accurate representation of the bare electrostatic interaction energy. The same figure also shows the energy functions computed with nonoptimized dipoles and the ones obtained by setting all dipoles to zero. The latter corresponds to the electrostatic part of the coarse-grained models of proteins by Zacharias et al.[2] For the models including dipoles, we notice that both are able to yield the correct asymptotic behaviors, whereas the presence of the optimized dipoles is necessary to accurately reproduce the AA electrostatic fields at all distances. Curiously, the nonoptimized dipole curve comes out very close to the optimized one for BPTI, and the same observation is true for the results obtained with charge-only model for Aβ_7. Note also that the bare electrostatic interactions are repulsive for BPTI, lysozyme, and Aβ_7, whereas for SOD, they are quite attractive. Indeed, the SOD dimer is stabilized by many H-bonds and favorable electrostatic interactions between groups of opposite charge, located at the interface between the two SOD monomers.

B. Global Electrostatic Interactions. The next step in testing our CG models is to build solvated models of CG proteins and to calculate the overall electrostatic contribution to the interaction free energy as a function of the interdimer

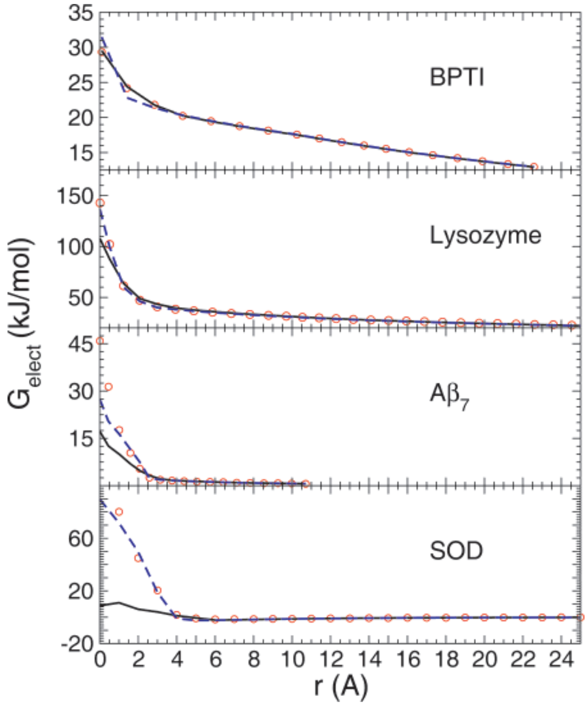


Figure 4. Total electrostatic interaction as a function of the intermonomer distance. We present results for the AA model (circles), the CG model (solid continuous line), and an intermediate model (dashed continuous line) with CG charges and dipoles, but with AA dielectric boundaries.

distances. To do this, it is necessary to add to the description of our CG systems an excluded volume component, in our case a van der Waals radius for every CG residue. This CG radii is computed as described in the Materials and Methods section. Typical radii are similar to those shown in Table 1 for the Aβ_7 case. For the other systems, the CG residue radii range from 2.3−2.5 Å for charged residues to 2.5−4.5 Å for non-charged residues.

With these excluded volume parameters to define the dielectric boundaries, the total electrostatic free energy, including solvation free energies and direct interactions as of eq 7, can be computed for the various dimer configuration using the PB solver. The resulting curves are reported in Figure 4. Two CG calculations are compared there to the AA reference (red circles). In one curve (dashed lines), the charge distribution is coarse-grained in terms of CG charges and dipoles, but the dielectric boundaries are those of the AA model. As expected, given the high accuracy of the CG charge distribution, the comparison is very favorable for all distances, even close to contact: the worst agreement involves the Aβ_7 peptide. The continuous curve accounts for the full CG model including the full coarse-graining of the electrostatic components and that of the molecular surface. Remarkably, this CG model reproduces very well the AA electrostatic interactions beyond a certain threshold distance. For the two proteins that are known to yield specific association, Aβ_7 and SOD, the threshold appears around 2 and 4 Å, respectively. For shorter distances, the CG curve underestimates the AA free energy. This discrepancy is caused by desolvation, which is accounted for in different ways in the AA and CG approaches. Even for identical electric fields, the different molecular surfaces leads to different surface polarization charges and thus different solvation free energies. For dimers where nonspecific interactions are dominant, such as BPTI and lysozyme, the accuracy of our CG model is larger at short distances. Indeed, for BPTI, there is a good agreement between AA and CG over the whole range of distances considered, whereas for lysozyme larger deviations are observed very close to contact.

Given the good agreement between the electrostatic energy of CG and AA in the homogeneous dielectric, the short-distance discrepancies, especially in the SOD dimer curve, are caused by desolvation effects. Thus, it is interesting to discuss those effects in more detail. When the distance between two atoms become smaller than the sum of the two atoms radii plus a solvent molecule diameter, the solvent molecules in the inner volume among the atoms start to be removed, and the atoms start to interact electrostatically through the low dielectric medium of the protein inner space. The change in the interaction free energy caused by this process can be separated into nonpolar and polar contributions. The nonpolar part, $\Delta\Delta G_{\mathrm{np}}$, will be treated in the next section. The polar part can be represented as the change in the electrostatic solvation energy of two molecules as they approach each other. Hence the desolvation polar part corresponds to the first three terms of eq 7.

A comparison between the curves in Figures 3 and 4, especially for SOD, shows that, even when the bare electrostatic interactions are favorable to the formation of the complex, the desolvation process can make the overall electrostatic interaction unfavorable. The net effect is more remarkable because the number of buried polar and charged groups located at the interface between the interacting molecules increases. Therefore, even when the flexibility of the protein matrix could rearrange the side-chains and the backbone to reduce the unfavorable electrostatic solute−solute interaction, behaviors like those in Figure 4 are to be expected: a sharp increase of the electrostatic effective interaction when two solvated molecules are closer than the desolvation threshold. In the case of SOD, the desolvation terms at $r = 0$ are equal to 396 and 310 kJ/mol for the AA and the CG models, respectively. These high energetic costs come from the desolvation of about 13% of the molecular surface in both models following the dimer formation. On the other hand, the direct Coulombic interaction energies with $\epsilon_s = 2$ amounts to −292 and −301 kJ/mol, respectively. Therefore, even though the CG gives a CG electrostatic field that differs by only 2% from the AA electrostatic field, the desolvation contributions are farther apart, by about 20%.

As we have seen earlier, using the CG charge distribution with the true AA dielectric boundaries does give excellent results, so that the deviation between AA and CG models is a direct consequence of the slight difference between the AA and CG molecular surface topologies. To increase the agreement between CG and AA, it is tempting to increase slightly the resolution and use a "finer" representation with, that is, 2−3 grains per residue instead of 1−2 in the present case. To test that proposition, we have used the finer CG residue definition of Zacharias[2] and have applied the same methodology as in section II.A to assign grain radii and grain

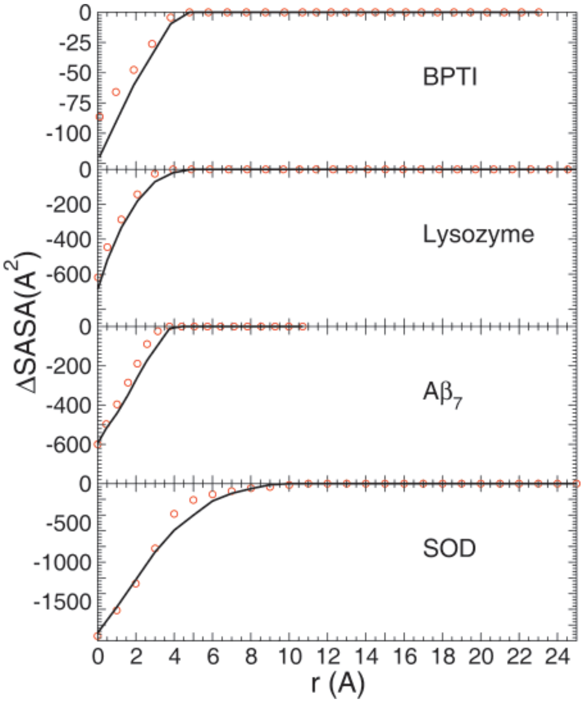


Figure 5. ΔSASAs as function of the intermonomer distance, obtained by taking the difference between the SASAs of dimers and the SASAs of the two isolated monomers. Results are reported for the AA (circles) and CG (continuous solid line) models.

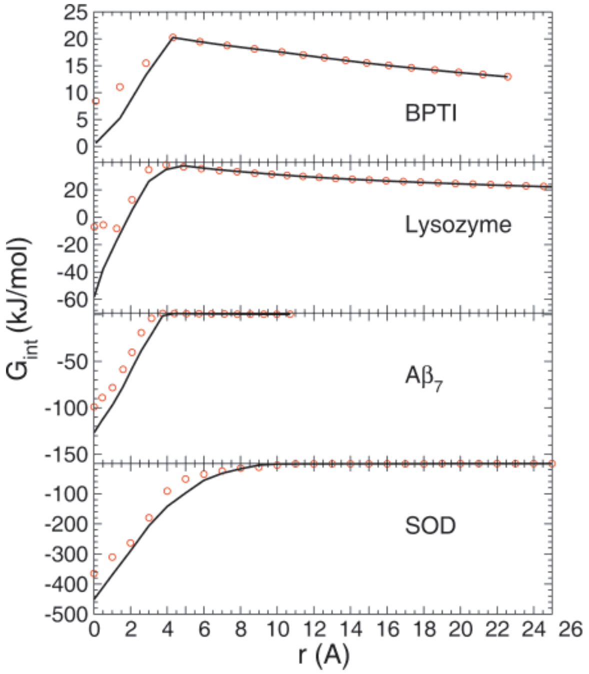


Figure 6. Total interaction free energy as a function of intermonomer distance. Circles are shown for the AA model, whereas the solid continuous lines are from the CG model.

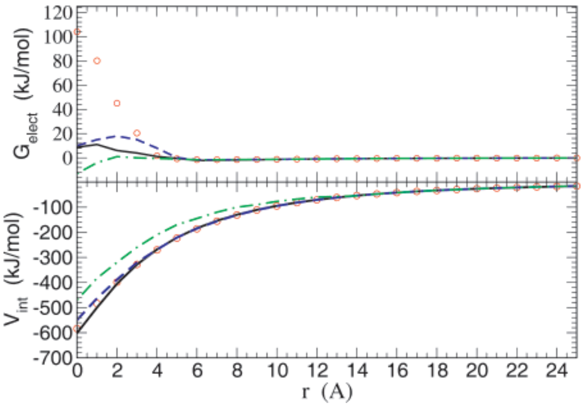


Figure 7. Bottom: Dependence of the bare electrostatic interaction energy on the intermonomer distance for the SOD dimer. The circles correspond to the AA system and the solid curve to the CG system composed of charges and optimized dipoles. The other two curves correspond to the two or three site Zacharias' model with natural charges (dot-dashed line) and optimized point charges (dotted line). Top: Same for the effective electrosatic interaction energy in solution.

partial charges. Every grain now carries a charge instead of a dipole, and all the charges are simultaneously optimized to best reproduce the AA electrostatic potential, as in eq 2, with the additional constraint of conserving the net charge of each residue (0 or ±1). This procedure improves upon the so-called *natural charge* representation of the original Zacharias' model, in which only charged residues carry one full charge on their extremal grain, all other grains staying neutral. We have carried out the same calculation described in Figures 3−6, comparing AA and CG dimer interaction energies as function of distance. Surprisingly, we find no significant improvement with respect to the 1−2-site model described before. For BPTI, lysozyme, and A$\beta_7$7, the results turn out more or less similar to those presented in Figures 3−6. The case of SOD is illustrated in Figure 7. For the vacuum electrostatics, it can be seen that the results are slightly deteriorated at short distances with respect to our initial one-dipole per residue representation. This is not surprising because a point-charge representation of the charge density is less flexible because it imposes the orientation of the residue dipoles. The electrostatic energy in solution, on the other hand, is slightly better, around 4−5 Å, as expected from the improvement of the cavity shape, but this improvement appears to no longer be true at shorter distances when interpenetration occurs. The natural charge representation of the original Zacharias' model does always give worse results, as can be expected from a less-accurate representation of the charge density.

The problem above clearly comes from the fact that the interaction between the two SOD monomers is highly specific and lateral chains from both partners become geometrically intricated at very short distances. A final possible refinement of our coarse-graining procedure is to stick to the AA representation at short residue−residue distances and switch to the CG model at only longer distances. To this end, given a certain threshold distance (for example 3 Å), we have computed the residue−residue interactions using the AA charge distribution and boundaries boundaries whenever the distance between any two atomic sites turns out below the threshold and the full CG representation otherwise. The results are presented in Figure 8 for a threshold at 3 and 6 Å. In the lowest panel, we have represented also the number of residues that are handled at

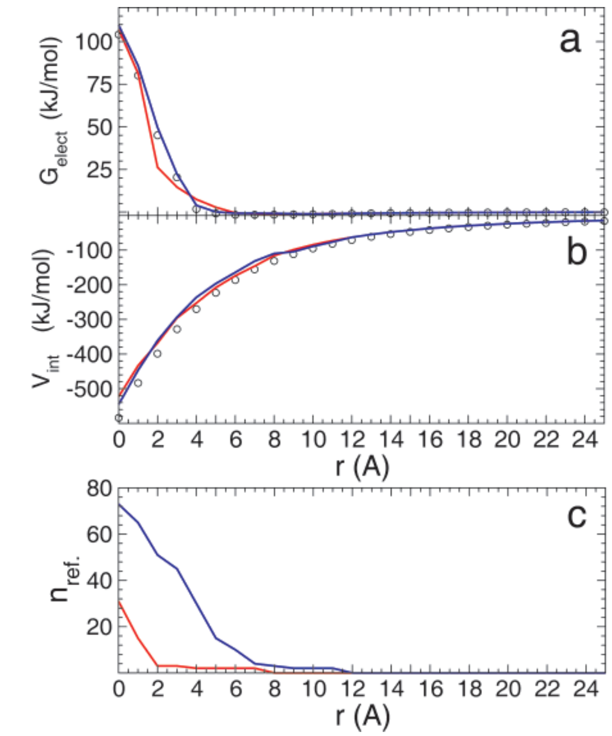


Figure 8. Electrostatic interaction energy in solution (a) and in vacuum (b) between the two monomers of the SOD dimer. Two short-ranged CG refiments are shown, with threshold at 3.0 Å (red curves) and at 6.0 Å (blue curves) and are compared to the AA results (black circles). In panel c, the number of refined residues, n_{ref}, as a function of distance are reported.

a AA level as a function of separation distance. It appears clearly that a threshold at 3 Å already gives a good description of the overall electrostatic interaction in solution, including desolvation, compared to the full AA calculation. The number of AA residues to consider remains relatively modest even at contact: about 30, that is, 15 per monomer (for a total number of 153). This number falls down to zero as soon as the intermonomer separation becomes greater than 2 Å.

C. SASA Term. For the four systems investigated, Figure 5 shows the variation of the differential solvent accessible surface area as a function of distance. As expected, when the minimum distance between the two interacting molecules is greater than a value $\mathrm{dist_{max}}$, the difference between the sum of the SASA of two isolated molecules and the SASA of the interacting pair is zero. This $\mathrm{dist_{max}}$ is the distance at which the solvent accessible surfaces of the two partners get in contact and the desolvation process starts. For lysozyme, BPTI, and Aβ_7 the ΔSASA term is vanishing for distances longer than ∼4 Å. For SOD, the threshold distance reaches ∼9 Å. From the visual analysis of the SOD solvent accessible surfaces, it is found that this early desolvation process is caused by the presence of specific residues, such as Lys25, Tyr26, and Ala112. These are placed at the border of the monomer separation surface and protrude out of one monomer toward the other, thus creating some steric hindrance between the two partners. Unexpectedly, the worst agreement between CG and AA is obtained for BPTI where no specific interaction between monomers is observed. Overall, the differential SASA are well reproduced by the coarse-grained model, which is certainly the result of our careful parametrization of the residue radii from their individual SASA. It is not yet clear why the CG model can provide such a good solvent-accessible surface but still somehow misses some of the fine electrostatic boundary effects.

D. Total Interaction Free Energy. From eq 3, we can obtain the interaction free energy profiles, corresponding to the potential of mean force (PMF) of the aggregate, by summing the electrostatic interaction free energy and the nonpolar term. The latter is obtained from the ΔSASA term and after multiplication with the γ proportionality constant introduced in eq 8. We have assigned to this proportionality factor a value of 0.24 kJ/mol Å^2, in agreement with previous studies of binding free energies from continuum dielectric methods.[32,35] It should be noted that a possible way to obtain the value of the γ parameter is to derive it from experimental results by fitting the calculated solvation free energies of small hydrophobic molecules to the experimental ones. This implies that the γ parameter takes different values if different sets of atomic radii and atomic charges are used. Furthermore, the γ parameter depends on the chosen solute and solvent dielectric constant ϵ_p and ϵ_s. The value of 0.24 kJ/mol for γ that we eventually chose was determined for the PARSE[34] parameter sets which reproduce the solvation free energies of proteins assuming $\epsilon_s = 2$. It should be stressed that the set of PARSE charges differ from the CHARMM27 set, used in this work for the AA models. For this reason, the sum of the electrostatic term G_{elect} and the nonpolar term $\Delta\Delta G_{\mathrm{np}}$ in eq 3 might not give fully meaningful results compared to experiments, and this applies to both the AA and CG systems. Notwithstanding, we remind to the reader that the purpose of this first study is to explore the validity of CG models of biomolecules and *not* (yet) to reproduce, through an appropriate CG parametrization, experimental results. Thus, for our purposes the most adequate choice of γ is the simplest, that is, an identical γ for CG and AA.

In Figure 6, the profiles of the interaction free energy for the four biomolecules are shown. Remarkably, above ∼4Å, the AA and CG free energy curves are in excellent agreement for Aβ_7, lysozyme, and BPTI. For SOD, deviations are noticeable only below 6 Å. For smaller molecule−molecule distances, the agreement is still quite good until 2−3 Å for all curves but degrades more strongly when the monomers are in closer contact.

This short distance discrepancy can be attributed to different contributions, the SASA term for BPTI, and desolvation effects in all other cases. It is noteworthy however that all the important physical features observed for the AA model are well reproduced by CG. BPTI and lysozyme are expected to have nonspecific interactions and to form very weak complexes if any. This is indeed what is observed, at least for the relative orientation for which the calculation were performed. At longer distances, the effective interaction potential is repulsive and dominated by like-charge repulsion; this part of the potential is perfectly reproduced by the CG description. The surface effects induce a rather shallow attraction at shorter distances, which is slightly overestimated by the CG model. For more specific interactions, Aβ_7, and SOD, the PMF shows a deep well at short distances and becomes completely flat afterward.

Bearing in mind the results in previous sections, we can conclude here that despute its not optimal parametrization, the CG model succeeds in reproducing the PMF of the AA systems. The agreement is semiquantitative when the two monomer are in contact and becomes quantitative at intermonomer distances larger than 4−6 Å.

IV. Conclusions

To efficiently characterize and model the interactions between biological macromolecules, it is necessary to consider a less-detailed description than that of the atomic scale. This can be done by modeling the system in term of elementary grains, coarse-grained description, such as the residues, the nucleotides, riboses, and even the higher molecular entities. Thus, this paper has developed a CG model of biomolecules that represents the amino acids in proteins and peptide chains as charged dipolar spheres. We have parametrized such a model based on the physical properties of individual amino acids and used it to study the interaction between solvated proteins and peptides. We have then computed the protein−protein potential of mean force for several selected systems and systematically compared results for the CG systems with those for the corresponding AA systems.

Despite the expected loss of atomic definition in the interactions, which also implies approximations in specific bonding such as hydrogen bonding, our CG model is capable of reproducing well the potential of mean force of the AA model until the intermonomer distance becomes too small. In particular, for conformations where specific interactions between monomers are unimportant, the CG interaction free energies are comparable to results from the AA model until the atom hard cores get into contact. This is the case of the lysozyme and BPTI homodimers. SOD and Aβ_7, instead, form stable aggregates in solution and presents specific electrostatic interactions. Hence, for the electrostatic interaction free energy, we obtain CG curves that deviate from those obtained for the AA models when the distance between the molecules are below the desolvation threshold. We also find that to get quantitative predictions at shorter distance a slight increase in the resolution of the model from 1−2 to 2−3 grains per residue is not sufficient. On the other hand, a mixed description where the interaction between grains at very short distances (<3 Å) is described at an atomic level is found to be quite accurate. In this latter approach, only a limited number of residues needs an atomistic representation, for example, 30 for the interaction between monomers in SOD, which will not degrade the computational efficiency of our coarse-grained approach too much.

To conclude, although the parametrization of our CG model is not optimal, the major result of this first investigation is that our CG model is very successful in reproducing the potential of mean force of its corresponding AA model. In a protein−protein interaction screening context,[12] our CG model will likely be useful to decide if two proteins can bind together or not, but it will not be sufficiently accurate to predict the association free energy. For our systems, this energy could deviate at full contact of 20% from the reference model.

We warn the reader that our protein−protein interaction picture is still missing some important contributions before a meaningful comparison with experimental data can be attempted. In particular, the steric repulsion effect should be considered via a (smooth) repulsive pairwise residue−residue potential. The induced dipole−induced dipole contribution emerging from both the electronic and atomic polarizability of the residues (the later being mainly due to the flexibility of the lateral chain) should also be considered. We point out that a set of atomic polarizabilities for all amino acids has been proposed recently by Song.[7] The underlying CG model, which involves one dipolar polarizable spherical site per residue, is the direct extension of the model explored in our study.

Finally, we point out that our CG model of proteins can be easily adapted coupled with Gaussian network models[36] to include a certain degree of interresidue flexibility.

Abbreviations: Coarse-grained (GC), all-atom (AA), molecular dynamics (MD), solvent accessible surface area (SASA), potential of mean force (PMF), Cu−Zn superoxide dismutase (SOD), basic pancreatic trypsin inhibitor (BPTI), Poisson−Boltzmann (PB).

Acknowledgment. The authors would like to thank Dr. L. Maragliano for providing the equilibrated structure of SOD homodimer. The authors also acknowledge support from the “Action Concertée Incitative IMPbio” project “Mésobio”.

References

(1) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

(2) Zacharias, M. Protein−protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **2003**, *12*, 1271.

(3) Brown, S.; Fawzi, N. J.; Head-Gordon, T. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10712.

(4) 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.

(5) Baker, N. A. Improving implicit solvent simulations: A Poisson-centric view. *Cur. Opin. Struct. Biol.* **2005**, *15*, 137.

(6) Feig, M.; Brooks, C. L., III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217.

(7) Song, X. An inhomogeneous model of protein dielectric properties: Intrinsic polarizabilities of amino acids. *J. Chem. Phys.* **2002**, *116*, 9359.

(8) Song, X. The extent of anisotropic interactions between protein molecules in electrolyte solutions. *Mol. Simul.* **2003**, *29*, 643.

(9) Sheinerman, F. B.; Norel, R.; Honig, B. Electrostatic aspects of protein−protein interactions. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153.

(10) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein−protein recognition sites. *J. Mol. Biol.* **1999**, *285*, 2177.

(11) Simonson, T.; Archontis, G.; Karplus, M. Free-energy simulations come of age: Protein−ligand recognition. *Acc. Chem. Res.* **2002**, *35*, 430.

(12) Janin, J. Welcome to CAPRI: A critical assessment of predicted interactions. *Proteins* **2003**, *47*, 257.

(13) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3−5 shows progress in docking procedures. *Proteins* **2005**, *60*, 150.

(14) Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379.

(15) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1.

(16) Sharp, K. A.; Honig, B. Electrostatic interactions in macromolecules. *Annu. Rev. Biophys. Chem.* **1990**, *19*, 301.

(17) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins Struct. Funct. Genet.* **1988**, *4*, 7.

(18) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144.

(19) Hermann, R. B. Theory of hydrophobic bonding II. The correlation of hydrocarbon. Solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1971**, *76*, 2754.

(20) Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1971**, *105*, 1.

(21) Balbach, J. J.; Ishii, Y.; Antzutkin, O. N.; Leapman, R. D.; Rizzo, N. W.; Dyda, F.; Reed, J.; Tycko, R. Amyloid fibril formation by Aβ_{16-22}, a seven-residue fragment of the Alzheimer's β-Amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* **2000**, *39*, 13748.

(22) Sergi, A.; Ciccotti, G.; Falconi, M.; Desideri, A.; Ferrario, M. Effective binding force calculation in a dimeric protein by molecular dynamics simulation. *J. Chem. Phys.* **2002**, *116*, 6329.

(23) Kundrot, C. E.; Richards, F. M. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. *J. Mol. Biol.* **1987**, *193*, 157.

(24) Parkin, S.; Rupp, B.; Hope, H. Structure of bovine pancreatic trypsin inhibitor at 125 K: Definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Crystallogr. D* **1996**, *52*, 18.

(25) Procacci, P.; Darden, T. A.; Paci, E.; Marchi, M. ORAC: A molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions. *J. Comput. Chem.* **1997**, *18*, 157.

(26) MacKerell, S., Jr.; Bashfor, D.; Bellotan, M.; Dunbrack Jrand, R. L.; Evansecand, J. D.; Fieland, M. J.; Fischeand, S.; Gaand, J.; Guand, H.; Hand, S.; Joseph-McCarthand, D.; Kuchniand, L.; Kuczerand, K.; Laand, F. T. K.; Mattoand, C.; Michnicand, S.; Ngo, T.; Nguyeand, D. T.; Prodhoand, B.; Reiher IIand, W. E.; Rouand, B.; Schlenkricand, M.; Smit, J. C.; Stotand, R.; Strauand, J.; Watanaband, M.; Wiorkiewicz-Kuczerand, J.; Yin, D.; Karplux, M. All-atom empirical potential for molecular modelling in dynamics studies of proteins. *J. Phys. Chem.* **1998**, *102*, 3586.

(27) 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.

(28) 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.

(29) Schutz, C. N.; Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct. Funct. Genet.* **2001**, *44*, 400.

(30) Dong, F.; Vijayakumar, M.; Zhou, H.-X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of Barnase and Barstar. *Biophys. J.* **2003**, *85*, 49.

(31) Lu, B. Z.; Chen, W. Z.; Wang, C. X.; Xu, X. J. Protein molecular dynamics with electrostatic force entirely determined by a single Poisson-Boltzmann calculation. *Proteins* **2002**, *48*, 497.

(32) Froloff, N.; Windemuth, A.; Honig, B. On the calculation of binding free energies using continuum methods: Application to MHC class I protein−peptide interactions. *Protein Sci.* **1997**, *6*, 1293.

(33) Nicholls, A.; Sharp, K. A.; Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **1991**, *11*, 281.

(34) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978.

(35) Trylska, J.; McCammon, J. A.; Brooks, C. L., III. Exploring assembly energetics of the 30S ribosomal subunit using an implicit solvent approach. *J. Am. Chem. Soc.* **2005**, *127*, 11125.

(36) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Folding Des.* **1997**, *2*, 173.

CT700121N