

JCTC

Journal of Chemical Theory and Computation

Using Selectively Applied Accelerated Molecular Dynamics to Enhance Free Energy Calculations

Jeff Wereszczynski* and J. Andrew McCammon

The Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, United States

Received June 11, 2010

Abstract: Accelerated molecular dynamics (aMD) has been shown to enhance conformational space sampling relative to classical molecular dynamics; however, the exponential reweighting of aMD trajectories, which is necessary for the calculation of free energies relating to the classical system, is oftentimes problematic, especially for systems larger than small poly peptides. Here, we propose a method of accelerating only the degrees of freedom most pertinent to sampling, thereby reducing the total acceleration added to the system and improving the convergence of calculated ensemble averages, which we term *selective aMD*. Its application is highlighted in two biomolecular cases. First, the model system alanine dipeptide is simulated with classical MD, all-dihedral aMD, and selective aMD, and these results are compared to the infinite sampling limit as calculated with metadynamics. We show that both forms of aMD enhance the convergence of the underlying free energy landscape by 5-fold relative to classical MD; however, selective aMD can produce improved statistics over all-dihedral aMD due to the improved reweighting. Then we focus on the pharmaceutically relevant case of computing the free energy of the decoupling of oseltamivir in the active site of neuraminidase. Results show that selective aMD greatly reduces the cost of this alchemical free energy transformation, whereas all-dihedral aMD produces unreliable free energy estimates.

1. Introduction

Molecular dynamics (MD) simulations have become a crucial theoretical tool in advancing our understanding of the function of biological macromolecules.¹ Advances in algorithms² and computing power^{3–5} continue to allow for simulations of increasingly larger systems on longer and longer time scales, permitting the direct observation of all-atom protein folding,^{6,7} the observation of ion permeation through a transmembrane channel,⁸ and the simulation of a complete virus.⁹ Despite the remarkable progress that has been made in the field, simulation times still often fall far short of the microsecond to millisecond time scales inherent in many biological processes. There have been several methodological advances which have aimed at simulating longer time scales within current computational power such as implicit solvation models,¹⁰ multiple time stepping algorithms,¹¹ and improved treatment of long-range electrostat-

ics.¹² Sampling of phase space may also be enhanced through the deformation of the underlying potential energy surface,¹³ as has been done in hyperdynamics,¹⁴ puddle jumping,¹⁵ conformational flooding,¹⁶ and the local boost method¹⁷ (to name only a few). Our group recently developed a method to enhance the crossing of barriers and the sampling of phase space termed accelerated molecular dynamics (aMD),¹⁸ which has been shown to enhance sampling of biomolecular systems as in the conformational switching of Ras,¹⁹ improve the agreement between experimental and calculated chemical shifts for IκBα,²⁰ and accelerate the calculation of pK_a values in lysozyme.²¹

An area of particular interest concerns using aMD simulations to calculate ensemble averages for physically relevant nonaccelerated systems. For systems in which the energy added to accelerate the system is low, trajectories may easily be reweighted; however, as the system size increases the boost energy required for significant acceleration increases, causing the exponential reweighting factor to produce

* Corresponding author e-mail: jmwerezsz@mccammon.ucsd.edu.

ensemble averages that are dominated by only a few configurations with high weight, thereby decreasing the precision of thermodynamic quantities such as free energy changes.²² Here, we propose limiting the acceleration to the degrees of freedom most responsible for conformational changes, thereby reducing the energy added to a system to enhance sampling and resulting in improved reweighting statistics. This work is an extension of a previous study in which aMD was limited to the dihedrals in the backbone of a peptide substrate bound to cyclophilin;²³ however, here we demonstrate that accelerated dihedrals may contain atoms which are also contained in nonaccelerated torsions (that is, individual molecules may contain both accelerated and nonaccelerated torsions). Two examples are highlighted. First, we show that even in the case of the model system alanine dipeptide, selectively targeting dihedrals in the molecule's backbone results in similar acceleration levels while reducing the amount of energy added to the system. Then we turn our attention to the larger problem of calculating the binding energy of the clinically approved drug oseltamivir (marketed as Tamiflu by Roche Pharmaceuticals, Basel, Switzerland²⁴) to the N1 flu protein neuraminidase.^{25,26} Using free energy perturbation with a modification to the Bennett acceptance ratio (to account for reweighting),²⁷ we show that the computational cost required to accurately calculate the binding energy may be reduced by as much as 70% while maintaining a similar level of precision.

2. Theory

To enhance phase space sampling, the original aMD applies an additional potential only when the potential energy, $V(\mathbf{r})$, is below a specified criterion E to produce dynamics on the artificial landscape $V^*(r)$ such that

$$V^*(r) = \begin{cases} V(\mathbf{r}) & \text{if } V(\mathbf{r}) \geq E \\ V(\mathbf{r}) + \Delta V(\mathbf{r}) & \text{if } V(\mathbf{r}) < E \end{cases} \quad (1)$$

The form of the “boost” potential $\Delta V(\mathbf{r})$ is defined as

$$\Delta V(\mathbf{r}) = \frac{(E - V(\mathbf{r}))^2}{\alpha + (E - V(\mathbf{r}))} \quad (2)$$

The formalism of this boost potential has several practical advantages: it produces a potential energy surface with a smooth first derivative, it does not require the definition of a “reaction coordinate” along which to enhance sampling, it reflects the shape of the original potential, and it is relatively simple with only two adjustable parameters (E and α).

Simulation results generated with aMD may be reweighted by the exponential of the boost potential, $\exp(\beta\Delta V(\mathbf{r}))$, to recover theoretically exact thermodynamic properties for the physically relevant unaccelerated system. In practice, however, the exponential dependence of the reweighting hinders convergence as trajectory averages become dominated by a smaller subset of their snapshots as the range of boost potentials increases. While this does not severely affect small systems such as alanine dipeptide, it does prevent the use of aMD in accurate free energy calculation in larger biomolecular systems. To improve the reweighting statistics and enhance sampling, several variants of aMD have been

developed including “barrier lowering” aMD,²⁸ replica-exchange aMD,²⁹ and adaptive aMD (personal communication P. Markwick). In this paper we discuss an extension to aMD which may be incorporated into other aMD implementations, which is to selectively accelerate a user-defined subset of dihedrals most pertinent to sampling the relevant degrees of freedom, which we refer to as *selective aMD*. Selective aMD has the advantage that by only accelerating the degrees of freedom most important to sampling, lower overall boosts may be utilized to achieve a similar acceleration level, thus resulting in improved reweighting statistics. The idea of enhancing sampling along a user-defined manifold has been previously shown to improve the calculation of time-correlation functions for kinetics of multidimensional systems.³⁰

2.1. Weighted Bennett Acceptance Ratio. Free energy perturbation (FEP) is a well-established technique used in free-energy calculations, specifically in the case of ligand binding and computational alchemy.^{31,13} In FEP, a nonphysical energy pathway is constructed between two physical end states, for example, a ligand bound in the active site of an enzyme (which we denote as $\lambda = 0$) and an active site without the ligand ($\lambda = 1$). The path between these two states is divided into a series of “windows” in which the Hamiltonian is transformed from state 0 to 1. Traditionally, free energy differences between successive windows are estimated by exponentially averaging the instantaneous work of going between the states, and the overall free energy is a sum of free energy differences between windows.³² Shirts et al. showed that the Bennett acceptance ratio (BAR) was superior to exponential averaging in producing asymptotically unbiased free energy estimates between two states that could improve precision by an order of magnitude.^{27,33}

For a series of work functions between two states in which individual works do not each have the same weight (as in aMD), the derivation of a weighted BAR follows that in Shirts et al. with the exception that their eqs 5 and 6, the probability of a single measurement of the work W_i for the forward and reverse work functions, are modified to

$$P(F|W) = \frac{\exp(\beta\Delta V(\mathbf{r}))}{1 + \exp[-\beta(M + W_i - \Delta F)]} \quad (3)$$

$$P(R|W) = \frac{\exp(\beta\Delta V(\mathbf{r}))}{1 + \exp[\beta(M + W_i - \Delta F)]} \quad (4)$$

where the constant M is redefined as

$$M = kT \ln \left(\frac{\sum_{i=0}^{n_F} \exp(\beta\Delta V(\mathbf{r}_i))}{\sum_{j=0}^{n_R} \exp(\beta\Delta V(\mathbf{r}_j))} \right) \quad (5)$$

Therefore, the value of ΔF that solves

$$\sum_{i=0}^{n_F} \frac{\exp(\beta\Delta V(\mathbf{r}_i))}{1 + \exp[-\beta(M + W_i - \Delta F)]} = \sum_{j=0}^{n_R} \frac{\exp(\beta\Delta V(\mathbf{r}_j))}{1 + \exp[-\beta(M + W_j - \Delta F)]} \quad (6)$$

is the optimal free energy estimate between adjacent windows. It has recently been shown that reweighting of states in BAR to account for non-Boltzmann sampling may have practical advantages outside of aMD simulations.³⁴

3. Computational Details

Molecular Dynamics Details. MD simulations were performed with the MD package Desmond (version 2.2) developed by D. E. Shaw Research.³⁵ Both systems were built, solvated, and ionized with Schrödinger's Maestro modeling suite such that there was a minimum of 12 Å of TIP3P³⁶ water buffer between the macromolecule and the periodic boundary and an ionic concentration of ~150 mM NaCl was present. The CHARMM22 force field with the CMAP correction was utilized (except where noted below in the neuraminidase calculations).³⁷ Following 10 000 steps of minimization, systems were continuously heated to 300 K over 1.5 ps. All simulations used the Martyna–Tobias–Klein constant pressure and temperature algorithms (a combination of Nosè–Hoover constant temperature and piston constant pressure algorithms)^{38,39} with a reference temperature and pressure of 300 K and 1.01425 bar, respectively. Short-range nonbonded interactions were truncated at 12 Å, while long-range electrostatics were calculated with a particle-mesh Ewald algorithm using a sixth-order B-spline for interpolation and a grid spacing of <1 Å in each dimension.⁴⁰ A time step of 2 fs was employed, and the M-SHAKE algorithm was used for constraining all hydrogen-containing bonds.⁴¹ A plugin was written for Desmond to perform aMD calculations on specified dihedrals.

Alanine Dipeptide. An alanine dipeptide molecule based on a model compound was solvated in a (27 Å)³ box using Maestro and equilibrated for 5 ns following heating. Two sets of 50 ns aMD trajectories were run, one in which all dihedrals were accelerated with aMD and one in which only the two dihedrals defined as ϕ and ψ were accelerated (selective aMD, see Figure 1). For each setup, 16 simulations spanning the parameter space of E and α were run to optimize these parameters. For the all-dihedral simulations, parameters of $E = 8$ and $\alpha = 4$ had an optimal fit to our metric χ (as discussed below), whereas for the selective aMD $E = 1$ and $\alpha = 0.75$ were optimal. Additionally, a 250 ns classical MD simulation was performed.

To determine the accuracy of the aMD results, well-tempered metadynamics was performed to calculate the underlying two-dimensional free energy landscape in ϕ/ψ space.⁴² In well-tempered metadynamics, the height of a Gaussian centered at position x is proportional to the Boltzmann weight of the metadynamics potential already present at x , that is the added Gaussian has a maximum value of $\omega_0 \cdot \exp(-V_i(x)/k\Delta T)$, with ω_0 being the initial Gaussian height, $-V_i(x)$ the metadynamics potential at x , and $k\Delta T$ a

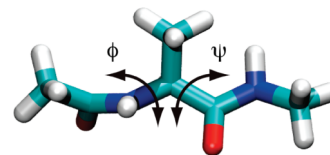


Figure 1. Conformation of the model system alanine dipeptide can be expressed by the angles of the two torsions ϕ and ψ .

user-defined energy which limits the explored energy range. We performed a 50 ns simulation in which Gaussians were added every 0.2 ps with a width of 0.1 radians and a height determined by $\omega_0 = 0.02$ kcal/mol and $k\Delta T = 2.4$ kcal/mol. To quantify convergence we define a metric χ as follows

$$\chi = \frac{1}{N \cdot M} \sum_{i=1}^N \sum_{U=6,7,\dots}^{3.0} \begin{cases} P_{i,U}^{\text{aMD}}/P_{i,U}^{\text{exact}} & \text{if } P_{i,U}^{\text{exact}} > P_{i,U}^{\text{aMD}} \\ P_{i,U}^{\text{exact}}/P_{i,U}^{\text{aMD}} & \text{if } P_{i,U}^{\text{exact}} < P_{i,U}^{\text{aMD}} \end{cases} \quad (7)$$

For a given well (defined as i in Figure 2f) this metric calculates the ratio of the population of states below the energy U ($P_{i,U}^{\text{aMD}}$) with that expected from the metadynamics results ($P_{i,U}^{\text{exact}}$) and averages this over M energy values of U from 0.6 to 3 kcal/mol (in increments of 0.1 kcal/mol) for each of the $N = 3$ wells. If a well has a population greater than that expected from the metadynamics result, the inverse of the ratio is taken to equally account for over and under sampling of the well. This metric has several advantageous features: by averaging over multiple energy levels it selects for smooth population densities (as are observed in cMD and metadynamics but may not result when the trajectory is reweighted with aMD), it treats over and under sampling a well as equally poor, and it equally weights all three of the main energy wells. $\chi = 1$ is considered an ideal reproduction of the exact population.

N1 Neuraminidase. A monomer of neuraminidase bound to the inhibitor oseltamivir taken from the 2HU0 crystal structure was solvated in an approximately (70 Å)³ box. Following the minimization and heating protocol outlined above, the system was equilibrated for 5 ns with the AMBER99SB force field⁴³ before protein and water parameters were changed to the CHARMM22 force field and further equilibrated for 1 ns. Parameters for oseltamivir were previously developed for use with the AMBER99SB force field⁴⁴ and maintained throughout the CHARMM simulations. The switch to the CHARMM force field was performed after testing of aMD free energy calculations revealed that increased acceleration levels in the AMBER force field tended to disturb the electrostatic components of the free energy calculation; therefore, this hybrid AMBER (for the ligand) and CHARMM (for the remainder of the system) force field was utilized. While the authors concede this may produce incorrect absolute binding energies, the goal of this paper is to study convergence of accelerated free energy calculations to results obtained from unaccelerated (and longer) calculations.

Alchemical free energy calculations for the decoupling of the ligand in the protein's active site were performed using 21 windows in which the electrostatics were decoupled over

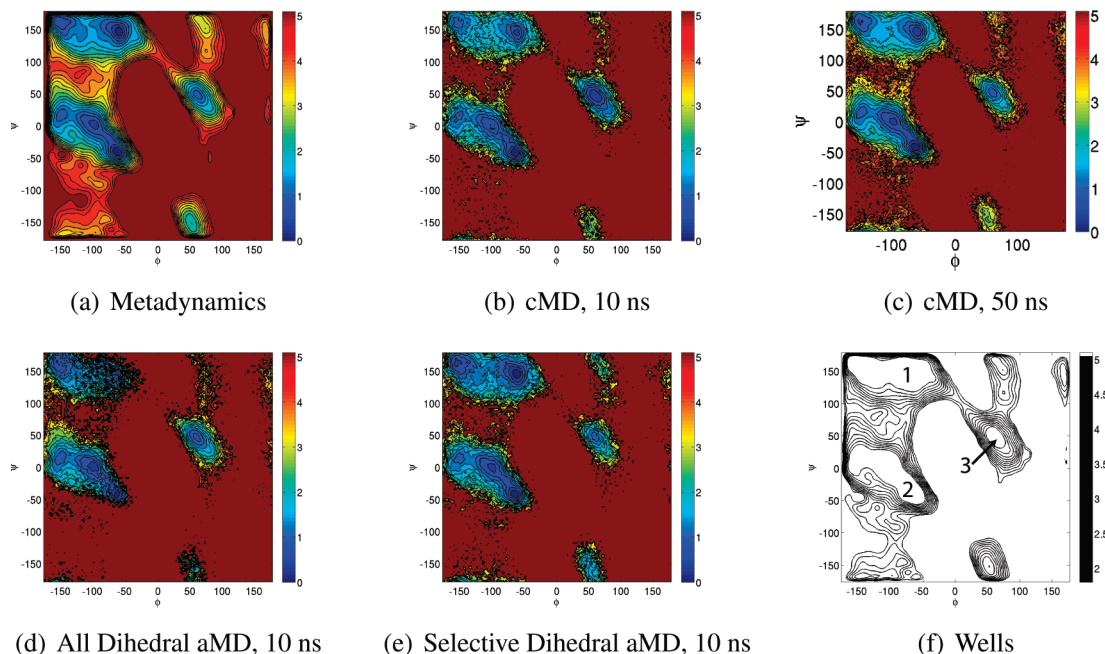


Figure 2. Free energy landscapes for alanine dipeptide. (a) Metadynamics results representing the expected results for infinite sampling of the system. Classical MD simulations visited all three energy wells on the 10 ns time scale (b); however, the statistics are poor (as shown in Table 1). Sampling is improved on the 50 ns time scale (c). (d) All dihedral aMD, showing that a sampling of all three energy wells is improved relative to the cMD results. (e) Selective dihedral aMD, showing the best agreement to the metadynamics results on the 10 ns time scale. (f) Wells labeled for discussion purposes.

Table 1. Free Energy Statistics for Alanine Dipeptide from cMD, All-Dihedral aMD, and Selective aMD Simulations Compared to the Metadynamics Results Which Effectively Represent the Infinite Sampling Limit^a

time	simulation type	well 1		well 2		well 3	
		energy	population	energy	population	energy	population
10 ns	metadynamics	0.065	0.406	0.000	0.429	1.010	0.049
	classical	0.485 (+.421)	0.249 (61%)	0.000 (+.000)	0.427 (99%)	0.301 (-0.706)	0.191 (392%)
	all dihedral aMD	0.000 (-0.065)	0.337 (83%)	0.085 (+.085)	0.415 (97%)	0.767 (-0.240)	0.078 (160%)
50 ns	selective aMD	0.008 (-0.057)	0.422 (104%)	0.000 (+.000)	0.369 (86%)	1.010 (-0.004)	0.041 (84%)
	classical	0.350 (+.285)	0.328 (81%)	0.000 (+.000)	0.475 (111%)	1.004 (-0.002)	0.058 (120%)
	all dihedral aMD	0.053 (-0.012)	0.372 (92%)	0.000 (+.000)	0.426 (99%)	1.003 (-0.004)	0.052 (106%)
250 ns	selective aMD	0.137 (+.072)	0.381 (94%)	0.000 (+.000)	0.422 (98%)	1.038 (+.033)	0.048 (98%)
	classical	0.232 (+.168)	0.357 (88%)	0.000 (+.000)	0.446 (104%)	0.994 (-0.012)	0.055 (112%)

^a For each of the three wells (Figure 2f) the minimum energy relative to the global minimum is calculated, as is the population of states within 1.8 kcal/mol of that minimum. Results show that both aMD forms converge on the order of five times as fast as the cMD simulations.

10 windows followed by the Lennard–Jones interactions decoupled over 10 windows with a softcore potential using $\alpha = 0.5^{45}$ (as in a previous study⁴⁶). Free energies were calculated using BAR, with the modified BAR formulas described above used in the aMD calculations. A positional restraint of .8 kcal/mol was placed on a central carbon atom in oseltamivir to prevent the ligand from sampling nonactive site portions of the simulation box,⁴⁷ and calculations were performed for cMD, all-dihedral aMD (with $E = 2600$ and $\alpha = 400$), and selective aMD (with $E = 13$ and $\alpha = 2$). For cMD calculations, three sets of windows were run (as has been shown to improve calculated free energies⁴⁸), each with the same initial coordinates but different velocities for 5 ns per window, whereas for the selective aMD the same three sets of windows were run with 200 ps of all-dihedral aMD (to quickly equilibrate the whole protein) followed by 1.5 ns of selective aMD per window. Note that the times indicated in the text include the simulation time spent in all-dihedral aMD; thus, a time of 500 ps represents 200 ps of

all-dihedral and 300 ps of selective aMD. One set of FEP calculations was performed for the all-dihedral case for 1.75 ns window to illustrate the futility of using standard aMD in large-scale biomolecular FEP calculations.

The choice of dihedrals to accelerate was based on previous work in which the tetramer was simulated for 100 ns.⁴⁶ Acceleration was applied to those dihedrals which contained only heavy atoms, were in residues that had a heavy atom within 5 Å of the oseltamivir in the crystal structure, and had a multivariate distribution for a total of 29 dihedrals.

Work functions were decorrelated based on the statistical inefficiency using code provided by Shirts and Chodera.⁴⁹ For each BAR calculation, a bootstrap analysis was performed (with 50 independent calculations) for an error σ_B , which was combined with the variance of the three means (σ_V) to calculate an overall error estimate for the free energy by

$$\sigma = \sqrt{\sigma_V + \sum_{i=1}^3 \sigma_B/3} \quad (8)$$

Results

Alanine Dipeptide. The free energy landscape of alanine dipeptide can be described by the rotation of two torsional angles, ϕ and ψ , making it an ideal model system for methodological development that has been extensively studied (Figure 1). The metadynamics, or “infinite sampling limit”, results (Figure 2a) show three distinct energy wells which we label for further discussion in Figure 2f. The energy barrier between wells 1 and 2 is relatively low, and classical MD (cMD) simulations sample both sets of configurations on the 10 ns time scale (Figure 2b). Well 3 is substantially oversampled, which we attribute to the system becoming trapped in this state due to the higher energy barrier, thereby discouraging transitions to and from well 3. With 50 ns of simulation the sampling of all three wells are improved (Figure 2c). Further detail is shown in Table 1, which compares the minimum energy of each well and probability of all states within 1.8 kcal/mol of that minimum to the theoretically exact answer derived from metadynamics. The 10 ns cMD shows good agreement for well 2 (it is identified as the global minimum and the population at 3kT is nearly identical to the metadynamics results); however, well 1 is undersampled by 39% whereas well 3 is oversampled by 292%, and the minimum energies are incorrect by .4 and .7 kcal/mol, respectively. With 50 ns of simulation time the sampling improves such that errors in the populations range from 11% to 20%, and by 250 ns of sampling the statistics agree much better with the metadynamics results for all three wells, although the populations of wells 1 and 3 are still off by >10%.

For comparison, the all-dihedral and selective aMD simulations sampled all three energy wells on the 10 ns time scale (Figure 2d and 2e). Free energy statistics indicate a maximum error in the minimum well energy estimate of 0.24 kcal/mol in the all-dihedral case and 0.06 kcal/mol in the selective aMD simulation, whereas the greatest disagreement in well populations was an undersampling of well 1 by 17% in the all-dihedral aMD and an undersampling of well 3 by 16% in the selective aMD. Extension of the simulations to 50 ns results in further improved statistics of the well populations, with well 1 only being undersampled by 8% in the all-dihedral simulation and by 6% in the selective aMD.

To further examine the convergence of the free energy statistics we defined the parameter χ to quantitate the difference in the two-dimensional energy profiles (as discussed in the methods) which has the property of a value of 1 representing ideal sampling of the wells as compared to the metadynamics results. In Figure 3 we compare the time course of this parameter between the cMD and the aMD simulations (note the different time scales for the two sets of simulations). The scores for both aMD are similar to those for cMD simulations of five times the length, with cMD simulations requiring 200 ns before consistently having values above 0.9, whereas the aMD simulations pass this value at 44 and 35 ns for all-dihedral and selective aMD, respectively.

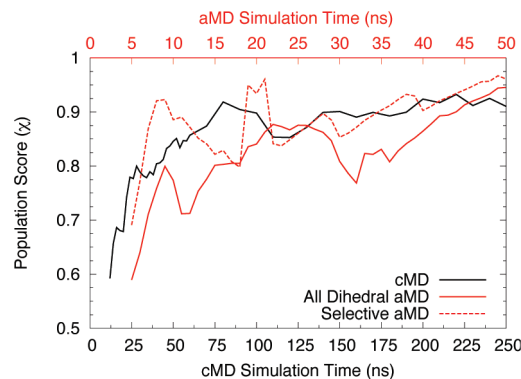


Figure 3. Comparison of the time evolution of our order parameter χ . Classical MD results (shown in black) correspond to the lower x axis, which is five times longer than that of the upper x axis, which corresponds to both aMD simulations. Despite this 5-fold difference in time scales, all three simulations show similar convergence patterns, with the lower weights in the selective dihedral slightly improving the statistics relative to the all-dihedral aMD.

A comparison of the boosts applied throughout the aMD simulations (Figure 4a) shows that higher boost potentials are applied throughout the all-dihedral simulation than in the selective aMD. For the all-dihedral case, ΔV varies from 0 to 4.2 kcal/mol, with an average value of 0.45 kcal/mol, whereas the selective aMD has a range of 0 to 0.54 kcal/mol with an average of 0.11 kcal/mol. This decrease in the applied boosts has a significant impact on reweighting as the maximum weight relative to an unaccelerated state is reduced from 1097 to 2.46. In Figure 4b the amount of total weight recovered from a simulation is plotted against the percentage of frames that contribute to that weight. In the case of all-dihedral aMD, very few frames contribute a substantial portion of the reweighting, 50% of the total weight comes from 4.8% of the trajectory, whereas 90% of the weight comes from 53.7% of the trajectory. This results in almost one-half the sampling (46.3%) contributing very little (less than 10%) to the calculated ensemble averages. In the selective aMD case the lower boosts result in more uniform weights, 50% of the weight comes from 37.9% of the trajectory and 90% from 87.6% of the trajectory (for comparison, in cMD configurations in the trajectory are uniformly weighted, so 50% of the weight comes from 50% of the trajectory). This increased reweighting efficiency improves the recovered statistics as high-boost configurations tend to dominate the ensemble average. For example, wells 1 and 2 appear significantly smoother in the energy landscape of the selective aMD relative to the all-dihedral aMD (Figure 2e and 2d) and χ is consistently higher in the selective case, both of which can be attributed to smoother statistics in the reweighted energy profiles.

N1 Neuraminidase. The binding of oseltamivir to neuraminidase is an example of how highly accurate free energy calculations may be employed in the study and development of novel pharmaceutical compounds. In order to validate our aMD simulations of the decoupling of oseltamivir in the N1 active site, we performed extensive sampling at each step of the alchemical transformation with 5 ns of cMD simulation for each of the 21 λ values, which

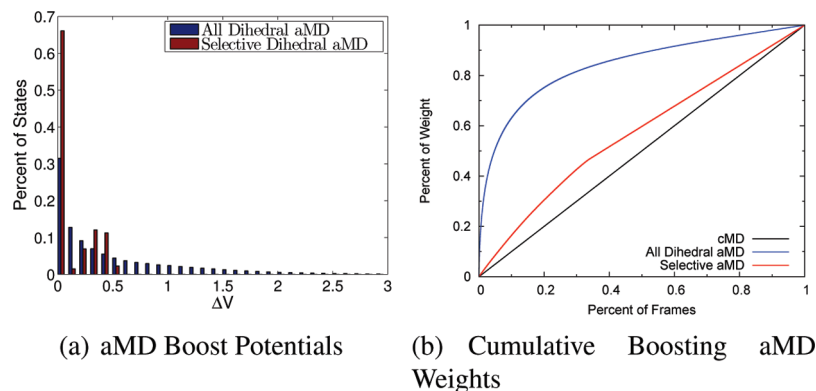


Figure 4. (a) Distribution of boost potentials applied during the 50 ns simulations, showing an increased range of ΔV values throughout the all-dihedral aMD simulation relative to the selective dihedral aMD. (b) Percent of the total weight from the simulations relative to the percent of frames that contributed this weight. The lower ranges of boosts in the selective aMD produce more uniform sampling relative to the all-dihedral aMD, which tends to be dominated by a small subset of the configurations.

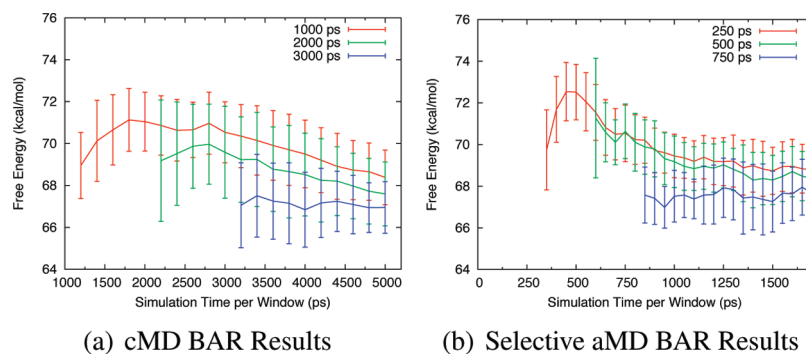


Figure 5. Free energies for the decoupling of oseltamivir from the neuraminidase active site as computed by alchemical transformation with classical MD (a) and selectively applied aMD (b). For each set of simulations three possible equilibration times were chosen; however, for only the longest equilibration period does the calculated free energy remain stable with increased sampling. The aMD results converged to values comparable to the cMD results with significantly less simulation time. The time scale on the x axis includes the equilibration time for which free energy values are not calculated.

was repeated three times (with different initial velocities). An equilibration period is typically discarded from the BAR calculations, the length of which is determined by several factors, including molecular rigidity, the slow motions of loops which are relevant to free energy differences, and the amount of computational time available.⁵⁰ In Figure 5a we show the time evolution of the mean free energies for three equilibration times, 1000, 2000, and 3000 ps, along with the associated errors. The 3000 ps of equilibration curve is the only one which does not have a mean that changes appreciably with increased sampling, suggesting that for this initial configuration of this complex a full 3 ns per window is required for each of the λ windows to equilibrate to their new Hamiltonians. The free energy of 66.9 ± 1.2 kcal/mol is calculated by using 3 ns for equilibration and the remaining 2 ns for sampling, as shown in Table 2. Note that this is not the free energy of binding in solution; rather it is only one leg of the thermodynamic cycle required for that calculation. Examination of individual BAR runs shows that increased sampling decreases both the variance between the three runs and the bootstrap errors associated with each (Figure S1, Supporting Information).

BAR results from simulations using selective aMD (with the first 200 ps utilizing all-dihedral aMD) are presented in Figure 5b. As in the cMD calculations, we have chosen three

Table 2. Free Energies for the Decoupling of Oseltamivir from the Neuraminidase Active Site As Calculated with cMD and Selective aMD for Varying Equilibration and Sampling Periods

simulation type	equilibration time	sampling time	free energy (kcal/mol)
cMD	1 ns	1 ns	71.0 ± 1.4
		2 ns	70.5 ± 1.5
		3 ns	69.5 ± 1.6
	2 ns	4 ns	68.4 ± 1.3
		1 ns	69.6 ± 2.2
		2 ns	68.5 ± 2.0
	3 ns	3 ns	67.6 ± 1.5
		1 ns	66.8 ± 1.8
		2 ns	66.9 ± 1.2
aMD	250 ps	500 ps	70.5 ± 1.4
		750 ps	69.4 ± 1.0
		1000 ps	69.2 ± 1.2
		1250 ps	68.7 ± 1.1
	500 ps	500 ps	69.2 ± 1.2
		750 ps	69.0 ± 1.1
		1000 ps	68.3 ± 1.2
		750 ps	67.9 ± 1.4
	750 ps	500 ps	67.9 ± 1.4
		750 ps	67.3 ± 1.5

equilibration times, 250, 500, and 750 ps, and the mean value of the free energy does not remain constant with increased sampling unless the longest of these equilibration times is discarded. With 750 ps of equilibration and 750 ps of sampling

we obtain a free energy value of 67.3 ± 1.5 kcal/mol, identical (within error) to that from the longer cMD calculations. Results from individual aMD runs are shown in Figure S2, Supporting Information. A comparison of the time evolutions of the BAR results shows similar behavior for the aMD and cMD free energies (with the aMD on shorter time scales) for the short, medium, and long equilibration times (Table 2 and Figure S3, Supporting Information). For short equilibration periods (cMD, 1000 ps; aMD, 250 ps) the free energy is initially overestimated, and while it approaches the values obtained for longer equilibration, the bias introduced in this nonequilibrated period results in free energies ~ 1.5 kcal/mol too high. Medium length equilibration periods (cMD, 200 ps; aMD, 500 ps) suffer from this effect as well but are not quite as biased, whereas for long equilibration times (cMD, 3000 ps; aMD, 750 ps) the calculated free energies remain stable (within error) with increased sampling.

As a comparison, we also performed BAR calculations on a single set of windows run for 1750 ps with all-dihedral aMD (Figure S4, Supporting Information). The much larger range of weights resulted in very few configurations contributing to the BAR results and poor free energy estimates. For example, with 250 ps of sampling and 1250 ps of equilibration a free energy of -15.0 ± 17.0 kcal/mol was computed, whereas 750 ps of equilibration and 750 ps of sampling resulted in a free energy of decoupling of 36.9 ± 17.3 kcal/mol. Both of these values are wildly inaccurate compared to the values of 66.9 ± 1.2 and 67.3 ± 1.5 kcal/mol obtained from cMD and selectively applied aMD.

5. Concluding Discussions

The method of accelerated molecular dynamics has been well established as a means of enhancing phase space sampling with minimal computational cost; however, the exponential reweighting required for the recovery of ensemble averages in the unaccelerated case introduces excessive noise such that it is often difficult, if not impossible, to recover accurate ensemble averages. Even in the case of the well-studied system alanine dipeptide this becomes evident. For example, in Figure 2d the free energy profile of well 1 appears discontinuous in the region of $(\phi, \psi) = (-50, 150)$, due in large part to the fact that several of the conformations visited have weights 1–3 orders of magnitude below that of the maximum weighted conformation. In contrast, by limiting the acceleration to only those dihedrals which are most pertinent to phase space sampling (in this case, ϕ and ψ) the maximum weight was reduced from 1097 to 2.46, which not only increased the smoothness of the recovered free energy profile (Figure 2e) but also moderately improved the agreement to the infinite sampling limit (as calculated with metadynamics) results in Table 1 and Figure 3. However, both aMD formalisms showed approximately a factor of 5-fold increase in efficiency relative to cMD for our order parameter χ .

Extension of this idea to the pharmaceutically relevant case of oseltamivir binding to neuraminidase shows the expensive free energy calculation of the decoupling of the ligand in the protein's active site may be reduced by up to 70% (from 5 to 1.5 ns/window) without a loss in precision. While this may not be crucial in the case of studying the binding of

only a single ligand to a protein, one could imagine that in the lead optimization stage of a drug-design effort, when highly accurate binding energies are necessary, the ability to examine three times the number of possible compounds at little extra computational cost may be highly desirable. In the case presented here, the accelerated dihedrals were chosen based upon extensive prior MD simulations; however, if this data were not available one could choose the accelerated dihedrals by residue type (accelerating dihedrals in residues with highly mobile side chains such as arginine and not accelerating dihedrals in aromatic rings), atom types (non-hydrogen containing), and proximity to the ligand. Additionally, in some cases where the ligand is bulky and has multiple torsions with high-energy barriers between local minima, one could accelerate dihedrals in the ligand molecules themselves, as was done in the case of cyclophilin.²³

Selective aMD may easily be incorporated into other free energy algorithms. For example, the methods of one-step perturbation and envelope distribution sampling provide techniques for effectively calculating the binding of several ligands to a protein with a single extended MD simulation.^{51,52} They have, however, not been extensively utilized due to the fact that, depending on the system being studied, they may require simulations on the microseconds time scale.⁵³ Therefore, a reduction of the computational cost of 3- to 5-fold (as observed in this study) could reduce the necessary simulation length from the highly expensive 1–2 μ s range into the manageable 200 ns time scale. These methods highlight that sampling of the partition function is, in general, a slow process requiring extensive calculations; therefore, methods such as selective aMD, which can enhance sampling of the relevant portions of phase space while not introducing excessive noise into the calculations, may prove useful in future applications.

Acknowledgment. We extend our thanks to M. Lawrenz, M. Fajer, P. Gasper, and Y. Wang for valuable discussion concerning the work presented here. The project described was supported by Award Number F32GM093581 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Additional support has been provided by the NSF, NIH, HHMI, CTBP, NBCR, and the NSF Supercomputer Centers.

Supporting Information Available: Plots of the free energies for oseltamivir decoupling as a function of sampling time for individual cMD and aMD simulations; comparisons between these cMD and aMD runs; and free energies calculated from a single all-dihedral aMD simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

- (2) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (3) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (4) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864–872.
- (5) Shaw, D. E.; et al. *Commun. ACM* **2008**, *51*, 91–97.
- (6) García, A. E.; Onuchic, J. N. *Proc. Nat. Acad. Sci. U.S.A.* **2003**, *100*, 13898–13903.
- (7) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806–816.
- (8) Jensen, M. O.; Borhani, D. W.; Lindorff-Larsen, K.; Maragakis, P.; Jogini, V.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *Proc. Nat. Acad. Sci. U.S.A.* **2010**, *107*, 5833–5838.
- (9) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. *Structure* **2006**, *14*, 437–449.
- (10) Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.
- (11) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990–2001.
- (12) Shan, Y. B.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2005**, *122*, 054101.
- (13) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2010**, *31*, 1569–1582.
- (14) Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908–3911.
- (15) Rahman, J. A.; Tully, J. C. *Chem. Phys.* **2002**, *285*, 277–287.
- (16) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (17) Wang, J. C.; Pal, S.; Fichthorn, K. A. *Phys. Rev. B* **2001**, *63*, 085403.
- (18) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
- (19) Grant, B. J.; Gorfe, A. A.; McCammon, J. A. *PLoS Comput. Biol.* **2009**, *5*, e100325.
- (20) Markwick, P. R. L.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 1220–1221.
- (21) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 560–568.
- (22) Shen, T. Y.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.
- (23) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2009**, *131*, 147–152.
- (24) Lew, W.; Chen, X. W.; Kim, C. U. *Curr. Med. Chem.* **2000**, *7*, 663–672.
- (25) Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. *Chem. Biol. Drug Des.* **2008**, *71*, 106–116.
- (26) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. *J. Med. Chem.* **2008**, *51*, 3878–3894.
- (27) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, *91*, 140601.
- (28) de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1516–1525.
- (29) Fajer, M.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1565–1569.
- (30) MacFadyen, J.; Wereszczynski, J.; Andricioaei, I. *J. Chem. Phys.* **2008**, *128*, 114112.
- (31) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (32) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (33) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 144107.
- (34) König, G.; Boresch, S. *J. Comput. Chem.* **2010**, in press.
- (35) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *SC'06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; ACM Press: New York, 2006; p 84.
- (36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (37) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (38) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.
- (39) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.
- (40) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (41) Krautler, V.; Van Gunsteren, W. F.; Hunenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501–508.
- (42) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.
- (43) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.
- (44) Amaro, R. E.; Minh, D. D. L.; Cheng, L. S.; Lindstrom, W. M.; Olson, A. J.; Lin, J. H.; Li, W. W.; McCammon, J. A. *J. Am. Chem. Soc.* **2007**, *129*, 7764–77677655.
- (45) Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; Vangunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.
- (46) Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2523–2532.
- (47) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- (48) Lawrenz, M.; Baron, R.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 1106–1116.
- (49) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
- (50) Yang, W.; Bitetti-Putzer, R.; Karplus, M. *J. Chem. Phys.* **2004**, *120*, 2618–2628.
- (51) Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. *J. Phys. Chem.* **1996**, *100*, 9485–9494.
- (52) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2008**, *128*, 174112.
- (53) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 6750–6754.

Simulating Lattice Spin Models on Graphics Processing Units

Tal Levy, Guy Cohen, and Eran Rabani*

*School of Chemistry, The Sackler Faculty of Exact Sciences, Tel Aviv University,
Tel Aviv 69978, Israel*

Received July 10, 2010

Abstract: Lattice spin models are useful for studying critical phenomena and allow the extraction of equilibrium and dynamical properties. Simulations of such systems are usually based on Monte Carlo (MC) techniques, and the main difficulty is often the large computational effort needed when approaching critical points. In this work, it is shown how such simulations can be accelerated with the use of NVIDIA graphics processing units (GPUs) using the CUDA programming architecture. We have developed two different algorithms for lattice spin models, the first useful for equilibrium properties near a second-order phase transition point and the second for dynamical slowing down near a glass transition. The algorithms are based on parallel MC techniques, and speedups from 70- to 150-fold over conventional single-threaded computer codes are obtained using consumer-grade hardware.

I. Introduction

In most common cases, computer programs are written serially: to solve a problem, an algorithm is constructed and implemented as a serial stream of coded instructions. These instructions are executed on a central processing unit (CPU) on one computer. Momentarily disregarding specific cases for which modern CPU hardware is optimized, only one instruction may be executed at a time; on its termination, the next instruction begins its execution. On the other hand, parallel computing involves the simultaneous use of multiple computational resources. In recent years, massively parallel computing has become a valuable tool, mainly due to the rapid development of the graphics processing unit (GPU), a highly parallel, multithreaded, many core processor. However, due to the specialized nature of the hardware, the technical difficulties involved in GPU programming for scientific use stymied progress in the direction of using GPUs as general-purpose parallel computing devices. This situation has begun to change more recently, as NVIDIA introduced CUDA, a general purpose parallel computing architecture with a new parallel programming model and instruction set architecture (similar technology is also available from ATI but was not studied by us). CUDA comes with a software environment that allows developers to use C (C++, CUDA

FORTRAN, OpenCL, and DirectCompute are now also supported¹) for program development, exposing an application programming interface and language extensions that allow access to the GPU without specialized assembler or graphics-oriented interfaces. Nowadays, there are quite a few scientific applications that are running on GPUs; this includes quantum chemistry applications,^{2,3} quantum Monte Carlo simulations,^{4,5} molecular dynamics,^{6–13} hydrodynamics,¹⁴ classical Monte Carlo simulations,^{15–17} and stochastic processes.^{18–20}

In this work, we revisit the problem of parallel Monte Carlo simulations of lattice spin models on GPUs. Two generic model systems were considered:

(1) The two-dimensional (2D) Ising model²¹ serves as a prototype model for describing critical phenomena and equilibrium phase transitions. Numerical analysis of critical phenomena is based on computer simulations combined with finite size scaling techniques.²² Near the critical point, these simulations require large-scale computations and long averaging, while current GPU algorithms are limited to small lattice sizes¹⁶ or to spin 1/2 systems.¹⁷ Thus, in three-dimensions (3D), the solution of lattice spin models still remains a challenge in statistical physics.

(2) The North-East model²³ serves as a prototype for describing the slowing of glassy systems. The computational challenge here is mainly to describe correlations and fluctua-

* Corresponding author e-mail: rabani@tau.ac.il.

Table 1. Summary of NVIDIA's Different Architectures

	G80	GT200	Fermi
maximum number of multiprocessors	16	30	16
scalar cores per multiprocessor	8	8	32
double precision capability	none	30 FMA ops/clock	256 FMA ops/clock
special function units	2	2	4
warp schedulers per multiprocessor	1	1	2
shared memory	16 kB	16 kB	48 kB
concurrent kernels	1	1	up to 16

tions on very long time scales, as the system approaches the glass temperature (concentration).²³ As far as we know, simulations of facilitated models on GPUs have not been explored so far, but their applications on CPU have been discussed extensively.^{24–27}

We develop two new algorithms useful to simulate these lattice spin models. For certain conditions, we obtain speedups of 2 orders of magnitude in comparison with serial CPU simulations. We would like to note that in many GPU applications, speedups are reported in Gflops, while in the present work, we report speedups of the actual running time of the full simulation. The latter provides a realistic estimate of the performance of the algorithms. This Article is organized as follows: Section II comprises a brief introduction to GPUs and summarizes their main features. Section III contains a short overview of lattice spin models and the algorithms developed. Section IV includes a short discussion of the random number generator used in this work. Section V presents the results and the comparisons between CPU and GPU performance. Finally, section VI concludes the Article.

II. GPU Architecture

To facilitate an understanding of GPU programming, we provide a sketch of NVIDIA's GPU device architecture. This is important for the development of the algorithms reported below, and in particular for understanding the logic and limitations behind our approach. By now, three generations of GPUs have been released by NVIDIA (G80, GT200, and the latest architecture codename "Fermi"). In Table 1 we highlight the main features and differences between generations.

A. Hardware Specifications. On a GPU device one finds a number of scalar multiprocessors (SMs). Each multiprocessor contains 8/32 (architecture dependent) scalar processor cores (SPs), a multithreaded instruction unit (MTIU), special function units for transcendental numbers and the execution of transcendental instructions such as sine, cosine, reciprocal, and square root, 32-bit registers, and the shared memory space. In general, unlike a CPU, a GPU is specialized for compute-intensive, highly parallel computation, exactly the task graphics rendering requires, and is therefore designed such that more transistors are devoted to data processing rather than data caching and flow control. This is schematically illustrated in Figure 1 and makes GPUs less general-purpose but highly effective for data-parallel computation with high arithmetic intensity. Specifically, they are opti-

mized for computations where the same instructions are executed on different data elements (often called single instruction multiple data, or SIMD) and where the ratio of arithmetic operations to memory operations is high. This puts a heavy restriction on the types of computations that optimally utilize the GPU, but in cases where the architecture is suitable to the task at hand it speeds up the calculations significantly.

B. Memory Architecture. NVIDIA's GPU memory model is highly hierarchical and is divided into several layers.

Registers. This is the fastest form of memory on the GPU. Usually automatic variables declared in a kernel reside in registers, which provide very fast access.

Local Memory. Local memory is a memory abstraction that implies "local" in the scope of each thread. It is not an actual hardware component of the SM. In fact, local memory resides in device memory allocated by the compiler and delivers the same performance as any other global memory region.

Shared Memory. This can be as fast as a register when there are no bank conflicts or when reading from the same address. It is located on the SM and is accessible by any thread of the block from which it was created. It has the lifetime of the block (which will be defined in the next subsection).

Global Memory. This is potentially about 400–600 times slower than register or shared memory. This module is accessible by all threads and blocks. Data transfers from and to the GPU are done from global memory.

Constant Memory. Constant memory is read only from kernels and is hardware optimized for the case when all threads read the same location (i.e., it is cached). If threads read from multiple locations, the accesses are serialized.

Texture Memory. Graphics processors provide texture memory to accelerate frequently performed operations. The texture memory space is cached. The texture cache is optimized for 2D spatial locality, so threads of the same warp that read texture addresses that are close together will achieve best performance. Constant and texture memory reside in device memory.

C. CUDA Programming Model. As has been mentioned, CUDA includes a software environment that allows developers to develop applications mostly in the C programming language. C for CUDA extends C by allowing the programmer to define C functions (kernels) that, when called, are executed N times in parallel by N different CUDA threads on the GPU scalar processors. Before invoking a kernel, the programmer needs to define the number of threads (N) to be created. Moreover, the programmer needs to decide into how many blocks these threads will be divided. When the kernel is invoked, blocks are distributed evenly to the different multiprocessors (hereby, having p multiprocessors requires a minimum of p blocks to achieve 100% utilization of the GPU). The blocks might be one-, two-, or three-dimensional with up to 512 threads per block. Blocks are organized into a one- or two-dimensional grid containing up to 65 535 blocks in each dimension. Each of the threads within a block that executes a kernel is given a unique thread ID. To differentiate between two threads from two different blocks,

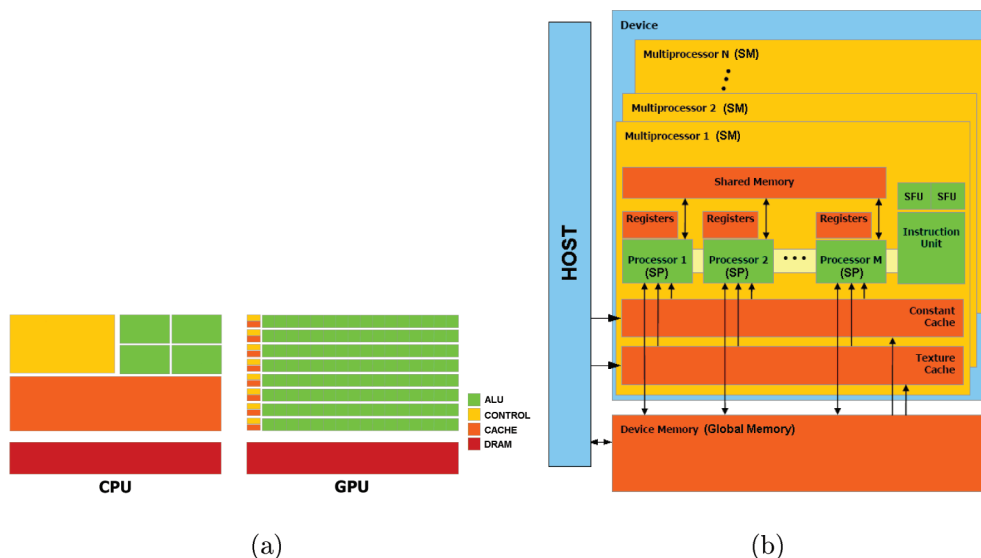


Figure 1. (a) A schematic delineation that illustrates the ratio of transistors devoted to data processing (ALUs) to those of data caching and flow control in CPUs and GPUs. This is a manifestation of the different philosophies between the “all purpose” CPU to the “dedicated purpose” GPU. (b) A visual hardware model that describes the memory architecture and hierarchy of a GPU device (courtesy of NVIDIA).

each of the blocks is given a unique ID as well. Threads within a block can cooperate among themselves by sharing data through shared memory. Synchronizing thread execution is possible within a block, but different blocks execute independently. Threads belonging to different blocks can execute on different multiprocessors and must exchange data through the global memory. There is no efficient way to synchronize block execution, that is, in which order or on which multiprocessor they will be processed; thus, an efficient algorithm should avoid communication between blocks as much as possible.

Much of the challenge in developing an efficient GPU algorithm involves determining an efficient mapping between computational tasks and the grid/block/thread hierarchy. The multiprocessor maps each thread to one scalar processor, and each thread executes independently with its own instruction address and register state. The multiprocessor SIMT (single instruction multiple threads) unit creates, manages, schedules, and executes threads in groups of 32 called warps. When a multiprocessor is given one or more thread blocks to execute, it splits them into warps that get scheduled by the SIMT unit. The SIMT unit selects a warp that is ready to execute and issues the next instruction to the active threads of the warp. A warp executes one common instruction at a time, so full efficiency is realized when all threads of a warp agree on their execution path (on the G80 and GT200 architectures, it takes 4 clock cycles for a warp to execute, while on the Fermi architecture it takes only 1 clock cycle). If threads of a warp diverge via a data-dependent conditional branch, the warp serially executes each branch path taken, disabling threads that are not on that path. When all paths terminate, the threads converge back to the same execution path. Branch divergence occurs only within a warp: different warps execute independently regardless of whether they are executing common or disjointed code paths. More information can be found in ref 1.

III. Lattice Spin Models

We consider a general lattice model of spins that are placed on a square lattice of dimensions $d = 1, 2, 3, \dots$. The spins may be of any dimension and may acquire continuous or discrete values. In the applications reported below, we focus, for simplicity, on the case where the spins at lattice site i take discrete values of $s_i = \pm 1$, but this can be easily extended to any spin dimension and value. The interactions between the spins are given by the Hamiltonian:

$$H = \sum_{i \neq j} J_{ij} s_i s_j + \sum_i B_i s_i \quad (1)$$

In the above equation, the first sum is usually carried over nearest neighbors only (which we will denote $\langle ij \rangle$). J_{ij} is the interaction parameter and may be constant, discrete, or continuous, and B_i is an external field. The above Hamiltonian can be used to study equilibrium properties as in the Ising model and spin glass models (Edwards–Anderson model,²⁸ Sherrington–Kirkpatrick model,²⁹ random orthogonal model,³⁰ etc.), or the dynamic behavior as in facilitated spin models (Fredrickson and Andersen,³¹ Jackle–Eisinger North-East model,^{23,32} etc.). Simulations of such systems are based on Monte Carlo techniques.³³ In most CPU implementations, the algorithms for equilibrium or dynamic simulations do not differ significantly. However, as will become clear below, they become very different when implemented on a GPU. Although the implementations described below are for two simple cases, the extension of our approach to the spin glass models mentioned above or to other facilitated spin models is straightforward.

A. Monte Carlo Simulation of the 2D Ising Model. The simplest 2D Ising model²² describes N magnetic dipoles (or spins) placed on a 2D square lattice with one spin per cell. We limit the discussion to the spin 1/2 case, where each spin has only two possible orientations, “up” and “down”. Each spin interacts with its nearest neighbors only, with a

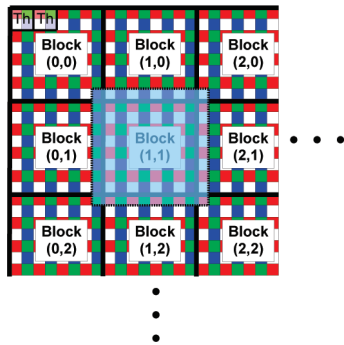


Figure 2. Illustration of the algorithm for an Ising model implemented on a GPU device. Every block handles a sublattice. Every thread is assigned 4 spin sites according to its ID and block. To improve memory access performance, each block copies its sublattice on to shared memory. Because a spin needs its 4 nearest neighbors to update, extra boundary sites are copied as well (blue-tinted square).

fixed interaction strength. In the absence of an external magnetic field, the Hamiltonian is given by

$$H = -J \sum_{\langle ij \rangle} s_i s_j \quad (2)$$

where $\langle ij \rangle$ represents a sum over nearest neighbors and J determines the energy scale. Monte Carlo simulation techniques based on the Metropolis algorithm³⁴ are perhaps the most popular route to obtain the thermodynamic properties of this model. In a CPU implementation of the Metropolis algorithm, a spin is selected at random, and an attempt to flip the spin is accepted with the Metropolis probability $P_{\text{acc}} = \min[1, \exp(-\Delta H/T)]$, where T is the temperature in units of energy. In the GPU algorithm developed here, we will take advantage of the fact that spins interact only with their nearest neighbors, and thus the problem can be divided into noninteracting domains. The generalization to the case of finite interacting regions is straightforward. The algorithm is as follows:

- (1) Randomly initialize the lattice (this is done on the CPU).
- (2) Copy lattice to the GPU.
- (3) Divide the lattice into Q sublattices, each with P spins.
- (4) A grid of Q thread-blocks is formed. Every thread-block contains $P/4$ threads.

(5) Every block copies a sublattice including its boundaries from the global memory to the shared memory. To avoid bank conflicts and save precious shared memory space, the short data type was used to form the lattice (Figure 2).

(6) Within the block, every thread is in charge of 4 spin sites (a sub block of 2×2). At first all red spins are updated (Figure 3a); that is, all threads are active. Once a thread finished updating the red spin, it continues to update its blue spin (Figure 3b). Because no native block synchronization exists, before updating the remaining spins, we make sure that all blocks finished the first two steps. This is done by recopying the data from the shared memory (boundaries excluded) onto the global memory and ending the kernel.

(7) Relaunch the kernel with the configuration generated in the previous step. Redo steps 5 and 6, only this time the green and white spins are being updated (Figure 3c).

(8) Recopy data to global memory.

This completes one Monte Carlo step (or one lattice sweep). In our implementation, we chose sublattices of 32×32 in size and blocks of 256 threads, as this choice turned out to be the most efficient. Using more threads does in fact reduce the time it takes the block to copy a sublattice to the shared memory, but then the ratio of arithmetic operations to memory operations is low and performance is poor. To obtain thermodynamic average properties, we use the fact the CPU and GPU can work in parallel and the lattice is copied from the device to the host from time to time, so averages can be calculated on the CPU while the GPU continues to sweep the lattice. We note in passing that similar algorithms have been proposed by Tobias et al.¹⁶ and Block et al.¹⁷ The former approach is restricted to lattices with up to 1024×1024 spins in 2D (assuming spins are stored as integer data type), whereas the algorithm presented in this work is applicable to larger systems, is designed for coalesced global memory access, and avoids bank conflicts, all of which make better use of the GPU architecture. The limit of system size in our approach is related to the size of the global memory on the GPU, which is typically on the order of 1–4 GB. This amounts to maximum system sizes of $20\,000 \times 20\,000$ to $40\,000 \times 40\,000$ spins. The algorithm of Block et al.¹⁷ is applicable to much larger systems and deals with the issue of multi-GPU programming, but is currently restricted to spin 1/2 systems, while the algorithm presented in this work is suitable for the more generalized Potts model³⁵ and gives approximately the same speedups in comparison to an equivalent CPU code.

To acquire the correct thermodynamic equilibrium state, the chosen set of Monte Carlo moves must satisfy either detailed balance or the weaker balance condition. On the CPU, detailed balance is rigorously satisfied when randomly selecting a spin within the Metropolis algorithm. For the GPU, however, the approach we developed breaks detailed balance, and only balance is satisfied.^{36,37} This is a sufficient condition to ensure that the algorithm correctly samples the Boltzmann distribution.

B. The North-East Model. The North-East model is based on the Ising model Hamiltonian with special constraints that are used to model facilitated dynamics.^{23,31,38} The Hamiltonian is given by:

$$H = J \sum_{\langle ij \rangle} s_i s_j + B \sum_i s_i \quad (3)$$

where s_i , J , B , and $\langle ij \rangle$ are described above. What makes this model different from the previous one is a constraint imposed on the transition probability to flip a spin. This probability is zero, unless the spin's upper (north) and right (east) neighbors point "up". In the latter case, one accepts a flip with the same Metropolis probability given by $P_{\text{acc}} = \min[1, \exp(-\Delta H/T)]$. Thus, the thermodynamics of the model is the same as the Ising model, but the Monte Carlo dynamics generated by the above rule is quite different. In most applications reported in the literature, one takes $J = 0$. For $J \neq 0$, the dynamics generated by this model are richer and show an interesting re-entrant transition.³⁹ When $J = 0$, the thermodynamics is trivial, and the coupling between neigh-

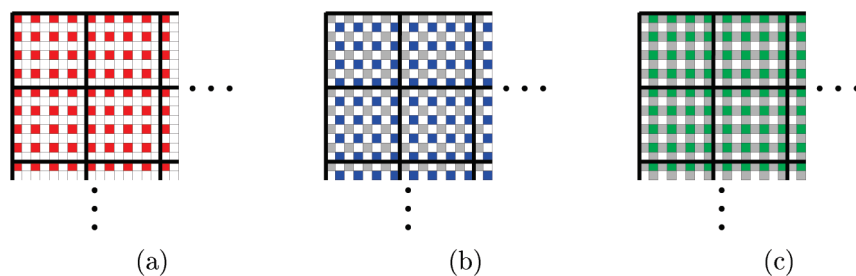


Figure 3. As native block synchronization is not available on a GPU, one must terminate the kernel's execution to achieve it. This is only necessary after the red and blue spins where updated (as can be clearly seen, they do not interact) and before moving on to the green and white spins, which in turn do interact with the aforementioned updated red and blue spins.

boring spins depends only on the aforementioned dynamical constraint. The lattice is initialized so spins point “up” with the probability c , which is also the equilibrium density of spins pointing “up”, and can be expressed as $c = 1/z e^{-1/T}$, where the partition function is $z = e^{1/T} + e^{-1/T}$ and T is the unitless temperature. Because of the dynamical constraints, if c is too low, one finds domains of spins that are stuck. For a high value of c on the other hand, all spins are flippable. As a consequence, there is a critical concentration c^* below which the system is not ergodic. This transition from ergodic to nonergodic behavior is modeled by the spin–spin autocorrelation function:

$$\Phi(t) = \sum_i \frac{\langle s_i(t) \cdot s_i(0) \rangle - \langle s_i \rangle^2}{1 - \langle s_i \rangle^2} \quad (4)$$

where $\langle s_i \rangle = 2c - 1$ is the average spin polarization and $s_i(t)$ is the spin polarization at Monte Carlo step t for site i . We expect the function to decay to zero for an initial concentration $c > c^*$ and to decay to a finite value f (which is the fraction of spins that are stuck) for $c < c^*$. The CPU implementation of the North-East model is identical to that described for the Ising model, with the additional constraint for the flipping probability.

The GPU implementation for this model, similar as it may seem to the Ising model, is a bit cannier. We found out that applying the checkerboard algorithm (described in section III.A) does not yield the same relaxation dynamics as the serial CPU implementation. This is understandable, because this model imitates a diffusion process: for a spin to be able to change its configuration, it is necessary that its north and east neighbors point up. If they do not, they in turn will also need their neighbors to point up to be able and change their configuration. Equilibration takes place by an up spin diffusion from “north-east” to “south-west”. By sequentially (instead of randomly) sweeping the lattice, we change the dynamics of this process. Such being the case, the corrected algorithm we developed is:

- (1) Initialize the lattice so spins point up with probability c (on the CPU).
- (2) Copy lattice to the GPU.
- (3) Divide the lattice into Q sublattices, each with P spins.
- (4) A grid of Q thread-blocks is formed. Every thread-block contains $P/4$ threads.
- (5) Blocks then randomly pick a sublattice (Figure 4) in such a way that two different blocks cannot pick the same

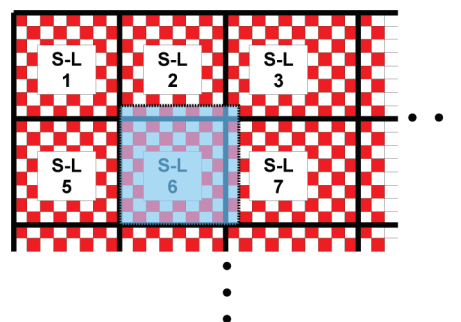


Figure 4. Although this figure looks similar to the Ising model implementation, in this case sublattices and spins are chosen randomly. This preserves the dynamical behavior in comparison to CPU algorithms. Note that when copying the data to the shared memory, only the upper and right boundaries are necessary.

sublattice. Every block copies a sublattice including its boundaries from the global memory to the shared memory.

(6) Within each block, λ threads concurrently pick λ spins from the sublattice randomly and update them. λ is chosen to be a small fraction of P .

(7) Synchronize the block.

(8) Reiterate steps 6 and 7 q times, such that $q \cdot \lambda = P/2$.

(9) Copy back data from shared memory to global memory (to allow block synchronization).

(10) Relaunch the kernel with the same configuration and redo steps 5–8.

(11) Recopy data to global memory.

This completes one lattice sweep. Again, we use the fact the CPU and GPU can work in parallel, and the lattice is copied from the device to the host from time to time to store the spin's configuration for the computation of the autocorrelation function.

The algorithm presented here does not preserve detailed balance nor the weaker balance condition, but because we are interested in its dynamics the given algorithm is correct. One should note that to obtain the correct dynamics an initialization of the parameter λ is needed. It is obvious that in the limit where $Q = \lambda = 1$ the algorithm is in fact serial and preserves detailed balance (results for this limit are given by the black line in Figure 5). We can now use this result as a reference and increase the number of threads and blocks that are working in parallel. There is an upper limit above which the results will diverge from the desired reference and the dynamics will no longer be correctly reproduced. Once

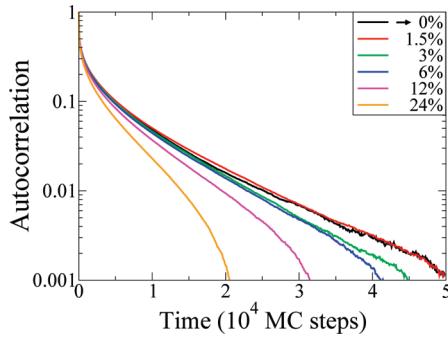


Figure 5. Convergence tests of the parameter λ for the North-East model at $c = 0.42$. The solid black line represents the reference value, which was obtained by running the simulation on a single thread—single block. This result is identical to the serial CPU result. The best performance is achieved for sublattices of 32×32 spins and for $\lambda = 16$ threads, that is, attempts to flip only $\sim 1.5\% \approx 16/(32 \times 32)$ of the spins inside a sublattice. One can see that for higher values of λ , the correlation diverges from the correct result.

λ has been evaluated, the simulation can be performed. In Figure 5 we provide a self-consistent test on the value of λ for the North-East model at $c = 0.42$.

It is possible to write an algorithm that maintains the weaker balance condition and at the same time preserves the dynamical relaxation of the model. The drawback of such an algorithm, however, is that it is $\sim 2.5\times$ slower than the one presented above. The only changes are in steps 6–9:

(6) Within each block, λ threads concurrently pick λ red spins (Figure 4) from the sublattice randomly and update them. λ is chosen to be a small fraction of P .

(7) Synchronize the block and pick λ white spins from the sublattice randomly and update them.

(8) Copy back data from shared memory to global memory (to allow block synchronization).

(9) Relaunch the kernel with the same configuration and redo steps 6–8 three more times (this will complete one MC step).

IV. Pseudo Random Numbers Generation

Monte Carlo simulations rely heavily on the availability of random or pseudorandom numbers. In this work, random

numbers were used to determine whether spin flips are accepted or rejected, in accordance with the Metropolis algorithm. As is widely known, the use of a poor quality PRNG (pseudo random numbers generator) may lead to inaccurate results.⁴⁰ In the course of this work, three different PRNGs were utilized:

(1) The first is L’Ecuyer with Bays–Durham shuffle and added safeguards.⁴¹ This PRNG has a long period (2×10^{18}) and is easy to implement on a CPU. Unfortunately, porting it to GPUs causes a dramatic decrease in performance: this is mostly due to the fact that the implementation of this algorithm requires too many registers.

(2) Second is the “minimal” random number generator of Park and Miller.⁴¹ This PRNG has a period of 2×10^9 and is portable to GPUs. In practice, however, it proved to work poorly, and the GPU simulation results were in poor agreement with reference values obtained with better PRNGs.

(3) Finally, we have the linear congruential random number generator (LCRNG).⁴¹ This PRNG has a period of 10^6 – 10^9 and provided good results even for very long runs. The LCRNG algorithm is very easy to port to the GPU and has the advantage of being very fast, requiring only a few operations per call.

In our implementations (section III), each thread used a separate LCRNG, creating its own sequence of pseudorandom numbers with a unique seed. The sequence (of thread i) is created as follows:

$$x_{j+1}^{\text{Th}_i} = (a \cdot x_j^{\text{Th}_i} + c) \bmod m \quad (5)$$

$$x_{j+1}^{\text{Th}_i} = \text{abs} \left(\frac{x_{j+1}^{\text{Th}_i}}{2^{31}} \right) \quad (6)$$

An appropriate choice of the coefficients is responsible for the quality of the LCRNG. We used⁴¹ $a = 1\,664\,525$, $c = 1\,013\,904\,223$, and $m = 2^{32}$. The different seeds were created per thread according to

$$x_0^{\text{Th}_{i+1}} = (16\,807 \cdot x_0^{\text{Th}_i}) \bmod m \quad (7)$$

with $x_0^{\text{Th}_0} = 1$. Similar PRNGs were used for other GPU applications.^{16,17}

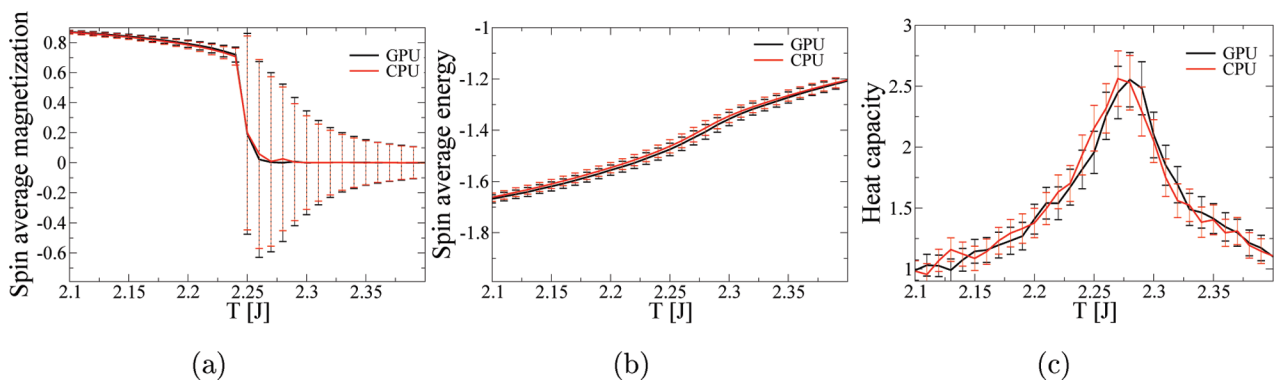


Figure 6. Comparison between the Monte Carlo simulation done on a CPU and on a GPU for a 128×128 spin–lattice. The temperature was stepwise reduced by 0.01 from $T = 2.4$ to $T = 2.1$. The critical temperature for this model is $T_c = 2/\log(1 + \sqrt{2})$. At each temperature, 2×10^7 sweeps through the lattice were performed, during which 10 000 different measurements were taken (after reaching equilibrium). In (a), we show the average spin magnetization, in (b) the average spin energy, and in (c) the heat capacity.

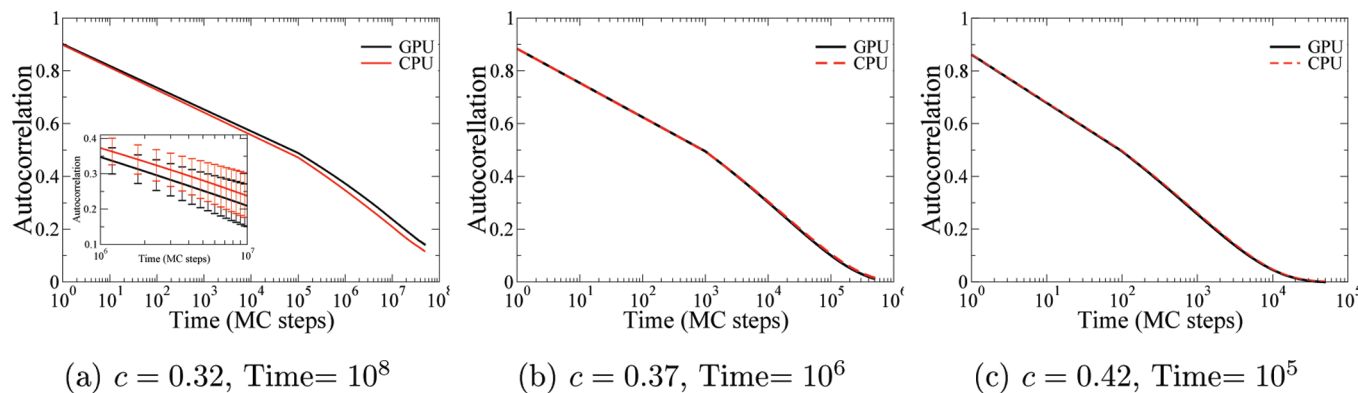


Figure 7. Comparison between spin–spin autocorrelation functions calculated from simulations done on a CPU and on a GPU for spin–lattices of different sizes and different initial concentrations. Results were obtained by averaging the autocorrelation functions of 30 different initial realization for every concentration. (a) Results for a 128×128 spin–lattice. As can be seen for low concentrations approaching c^* , relaxation time is very long (longer than the time we were willing to wait for the CPU results). Furthermore, at this concentration, more than 30 trajectories are required to average the autocorrelation function. (b and c) Results obtained from 512×512 spin–lattices.

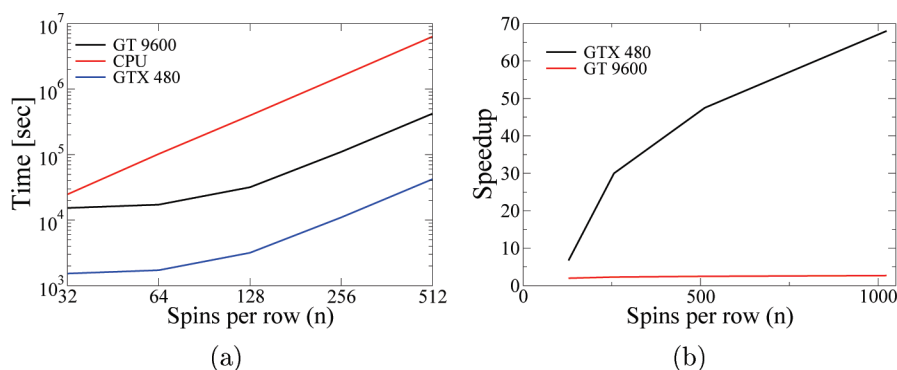


Figure 8. (a) Processing times for the 2D Ising model. Times are shown as a function of the number of spins n per row, which is related to the system size by $N = n^2$. A maximum acceleration factor of 15 was achieved (on the GT 9600GPU). We believe a higher value can be obtained for larger lattices. (b) Speedups of two different GPU cards versus the CPU serial implementation for the North-East model. Again, times are shown as a function of the number of spins n per row.

V. Results

For comparison, CPU codes were executed on a PC with Intel Core2 Duo E7400 @ 2.8GHz processor, Intel Raisin-City motherboard with Intel G41 chipset and Kingstone 2GB,800 MHz RAM (only a single core was used for the calculations). The operating system was CENTOS. Codes were compiled with an Intel C++ compiler (ICC) using all optimizations provided for best performance. Speedups reported here are given in terms of the complete application running time (from initialization until results are processed), rather than Gflops or spin updates per second. This choice is important, as it most closely describes the “real” gain in practical simulations by the use of a GPU rather than a CPU.

A. Ising Model. To verify the GPU implementation, we compared values of magnetization, energy, and heat capacity as a function of temperature (temperature was taken in energy units) between GPU and CPU versions. In Figure 6 are presented results from a 128×128 spin–lattice. We find the results to clearly agree. In terms of acceleration, we achieved a $15\times$ factor for lattices with 512×512 spins on the GT 9600 GPU (G80 architecture). Implementing the same code on the new GTX 480 GPU, we achieved a factor of $150\times$. This factor reduces to $2\times$ for small lattices ($32 \times$

32). The reason for this is that in small problems it becomes harder to hide memory access latencies, because there are not enough threads to execute between memory access operations. A theoretical analysis of our implementation shows that the GPU reaches full occupancy (a useful tool to check this is the “CUDA GPU Occupancy Calculator”, which can be freely downloaded from ref 42). The 10-fold factor obtained by the GTX 480 in comparison with the GT 9600 is easily understood when taking into consideration that the GTX 480 has 15 SMs instead of 8 (on the GT 9600), 4 blocks can be active simultaneously instead of 3, and a warp executes in one clock cycle instead of 4. Thus, the relative speedup is $15/8 \cdot 4/3 \cdot 4/1 = 10$. From Figure 8a it is obvious that the GPU reaches full occupancy only for lattices bigger than 128×128 . Note also that near the critical temperature, the fluctuations and noise increase. Because the present work is not concerned with determining the critical behavior, we have used the same number of Monte Carlo sweeps for all temperatures. Estimation of the critical behavior requires much longer runs, and perhaps also larger systems. In this respect, the GPU approach developed here provides the means to increase numerical accuracy by more than 1 order of magnitude (noise scales with the square root of the number

of MC steps), either by increasing the system size or by simulating longer Monte Carlo runs.

B. North-East Model. The North-East model has a critical concentration $c^* \approx 0.3$, below which the dynamics break ergodicity. In Figure 7 we show the results for spin–spin autocorrelation function for different concentrations above the critical value. Two lattice sizes were studied. Similar to the previous case, we find that the GPU results agree well with the CPU results, indicating that the proposed algorithm reproduces the correct dynamics. This is not trivial and depends on the value chosen for λ . Moreover, as pointed out above, the algorithm used for the Ising model fails to produce correct relaxation times. As the system approached the critical concentration, the dynamics become sluggish and the autocorrelation function decays slowly to zero.

In Figure 8b we compare the running times between the CPU and two GPU architectures. We achieved a $2\times$ factor on the GT 9600 GPU and a $\sim 70\times$ speedup running the same code on the new GTX 480 GPU. This factor reduces to $8\times$ for smaller lattices (128×128 and less). The GPU acceleration is nearly 2 orders of magnitude, implying that one can simulate the system closer to the critical density. This is important because the behavior of relaxation near the critical density is not necessarily universal, and thus extrapolations are often tricky.

VI. Conclusions and Summary

We have developed algorithms to simulate lattice spin models on consumer-grade GPUs. Two prototype models have been considered: the Ising model describing critical phenomena at equilibrium and the North-East model describing glassy dynamics. We showed that for equilibrium properties of lattice models an impressive speedup of $150\times$ can be achieved. To simulate the dynamics of such models, a more sophisticated approach was developed to preserve the dynamical rules and outcome. Our algorithm for the dynamic model reaches a $\sim 70\times$ factor in comparison to the serial CPU implementation for large system sizes. Although the algorithms were performed on specific models, we feel they can be easily extended to a larger class of similar systems.

Because the gain in computational power embodied by these results is in some cases 2 orders of magnitude, while the required GPU hardware is priced similarly to a CPU and can often be added to existing systems, the algorithms are certainly of interest. On the other hand, taking full advantage of it still requires savvy knowledge of the device and its capabilities and limitations, and the development of specific numerical algorithms. As the advantages of this useful technology become clear and knowledge about its implementation continues to build and spread, we hope it will become more accessible to a wide variety of computationally minded scientists.

Acknowledgment. We would like to thank Prof. Sivan Toledo for discussions. This work was supported by the Israel Science Foundation (grant no. 283/07). G.C. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. E.R. thanks the Miller Institute for Basic Research in

Science at UC Berkeley for partial financial support via a Visiting Miller Professorship.

References

- (1) NVIDIA CUDA, *Compute Unified Device Architecture Programming Guide* Version 3.0, http://www.nvidia.com/object/cuda_develop.html (accessed 7/2010).
- (2) Ufimtsev, I. S.; Martinez, T. J. *CiSE* **2008**, *10*, 26–34.
- (3) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619–2628.
- (4) Anderson, A. G.; III, W. A. G.; Schröder, P. *Comput. Phys.* **2007**, *177*, 298–306.
- (5) Meredith, J. S.; Alvarez, G.; Maier, T. A.; Schulthess, T. C.; Vetter, J. S. *Parallel Comput.* **2009**, *35*, 151–163.
- (6) van Meel, J. A.; Arnold, A.; Frenkel, D.; Portegies Zwart, S. F.; Belleman, R. G. *Mol. Simul.* **2008**, *34*, 259–266.
- (7) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (8) Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359.
- (9) Davis, J.; Ozsoy, A.; Patel, S.; Taufer, M. In *Bioinformatics and Computational Biology*; Rajasekaran, S., Ed.; Springer Berlin: Heidelberg, 2009, Vol. 5462.
- (10) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864–872.
- (11) Genovese, L.; Ospici, M.; Deutsch, T.; Mehaut, J.-F.; Neelov, A.; Goedecker, S. *J. Chem. Phys.* **2009**, *131*, 034103.
- (12) Demattè, L.; Prandi, D. *Briefings Bioinf.* **2010**, *11*, 323–333.
- (13) Eastman, P.; Pande, V. S. *J. Comput. Chem.* **2010**, *31*, 1268–1272.
- (14) Bernaschi, M.; Fatica, M.; Melchionna, S.; Succi, S.; Kaxiras, E. *Concurr. Comput.: Pract. Exper.* **2010**, *22*, 1–14.
- (15) Lee, A.; Yau, C.; Giles, M. B.; Doucet, A.; Holmes, C. *ArXiv e-prints 0905.2441*, 2009.
- (16) Preis, T.; Virnau, P.; Paul, W.; Schneider, J. J. *J. Comput. Phys.* **2009**, *228*, 4468–4477.
- (17) Block, B.; Virnau, P.; Preis, T. *Comput. Phys. Commun.* **2010**, *181*, 1549–1556.
- (18) Juba, D.; Varshney, A. *J. Mol. Graphics Modell.* **2008**, *27*, 82–87.
- (19) Januszewski, M.; Kostur, M. *Comput. Phys.* **2010**, *181*, 183–188.
- (20) Balijepalli, A.; LeBrun, T. W.; Gupta, S. K. *J. Comput. Inf. Sci. Eng.* **2010**, *10*, 011010.
- (21) Onsager, L. *Phys. Rev.* **1944**, *65*, 117–149.
- (22) Newman, M. E. J.; Barkema, G. T. *Monte Carlo Methods in Statistical Physics*; Oxford University Press: London, 1999.
- (23) Reiter, J.; Mauch, F.; Jäckle, J. *Physica A (Amsterdam)* **1992**, *184*, 493–498.
- (24) Garrahan, J. P.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9710–9714.
- (25) Jack, R. L.; Garrahan, J. P.; Chandler, D. *J. Chem. Phys.* **2006**, *125*, 184509.

- (26) Garrahan, J. P.; Jack, R. L.; Lecomte, V.; Pitard, E.; van Duijvendijk, K.; van Wijland, F. *Phys. Rev. Lett.* **2007**, *98*, 195702.
- (27) Chandler, D.; Garrahan, J. P. *Annu. Rev. Phys. Chem.* **2010**, *61*, 191–217.
- (28) Edwards, S. F.; Anderson, P. W. *J. Phys. F: Met. Phys.* **1975**, *5*, 965.
- (29) Sherrington, D.; Kirkpatrick, S. *Phys. Rev. Lett.* **1975**, *35*, 1792–1796.
- (30) Marinari, E.; Parisi, G.; Ritort, F. *J. Phys. A: Math. Gen.* **1994**, *27*, 7615.
- (31) Fredrickson, G.; Andersen, H. *Phys. Rev. Lett.* **1984**, *53*, 1244–1247.
- (32) Jäckle, J.; Eisinger, S. *Z. Phys. B: Condens. Matter* **1991**, *84*, 115–124.
- (33) Landau, D. P.; Binder, K. *A Guide to Monte Carlo Simulations in Statistical Physics*; Cambridge University Press: Cambridge, 2000.
- (34) Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, R. A.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1091.
- (35) Wu, F. Y. *Rev. Mod. Phys.* **1982**, *54*, 235–268.
- (36) Manousiouthakis, V. I.; Deem, M. W. *J. Chem. Phys.* **1999**, *110*, 2753–2756.
- (37) Ren, R.; Orkoulas, G. *J. Chem. Phys.* **2006**, *124*, 064109.
- (38) Nakanishi, H.; Takano, H. *Phys. Lett. A* **1986**, *115*, 117–121.
- (39) Geissler, P. L.; Reichman, D. R. *Phys. Rev. E* **2005**, *71*, 031206.
- (40) Ferrenberg, A. M.; Landau, D. P.; Wong, Y. J. *Phys. Rev. Lett.* **1992**, *69*, 3382–3384.
- (41) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.
- (42) *CUDA GPU Occupancy Calculator* version 3.1, http://developer.nvidia.com/object/cuda_3_1_downloads.html (accessed 7/2010).

CT100385B

Magnetizabilities at Self-Interaction-Corrected Density Functional Theory Level

Mikael P. Johansson^{*,†,‡} and Marcel Swart^{*,†}

Institut de Química Computacional, Universitat de Girona, Campus Montilivi, ES-17071 Girona, Spain and Laboratory for Instruction in Swedish, Department of Chemistry, University of Helsinki, FI-00014 Helsinki, Finland

Received May 6, 2010

Abstract: Using a recent high-quality ab initio coupled cluster benchmark set for magnetizabilities, we assess the performance of a set of density functionals, representing different levels of complexity, from the local density approximation (LDA), via generalized gradient approximations (GGA's) to kinetic energy density including meta-GGA's. The effect of self-interaction correction (SIC) is remarkable and, in most cases, leads to a significant error reduction, revealing the sensitivity of magnetizability toward a physically sound exchange-correlation potential.

1. Introduction

The static magnetizabilities are computed from the exchange-correlation potential as a second-order response of a molecule to an external magnetic field, as the second derivative of the energy with respect to the magnetic field, evaluated at zero field strength:

$$\xi = - \left. \frac{d^2 E}{d\mathbf{B}^2} \right|_{\mathbf{B}=0}$$

It is thus the magnetic-field equivalent of polarizability in an electric field. A negative value for the magnetizability implies a diamagnetic molecule, a class that encompasses most closed-shell, neutral species. The few exceptions include the paramagnetic ozone and boron monohydride,^{1–3} BH. Experimentally, measuring absolute magnetizabilities is something of a challenge. Vibrational effects are also large,^{4,5} which makes a direct comparison to computed values less straightforward. Recently, Lutnæs et al. published a careful benchmark set for magnetizabilities⁴ (and rotational g tensors), computed at the coupled-cluster singles, doubles, and perturbative triples, CCSD(T)⁶ level, using large, gauge-including basis sets. They compared their CCSD(T) values against several levels of density functional theory (DFT)^{7,8}

and concluded that all of the tested functionals show significant differences with respect to the ab initio values. Other studies have also shown that magnetizabilities are relatively insensitive to electron correlation effects, and already Hartree–Fock (HF) theory^{9,10} usually performs adequately. Therefore, it is somewhat awkward that most density functionals, which naturally capture more of the correlation than HF, actually perform worse in predicting this property.

Standard exchange-correlation functionals suffer from an unphysical self-interaction of the electrons. An immediate consequence of the Coulomb self-interaction error (SIE) is to raise the energies of occupied orbitals. As is well known, the highest occupied molecular orbital (HOMO) in exact DFT would correspond exactly to the negative of the ionization potential.¹¹ In approximate DFT, this is seldom the case. Thus, band gaps are usually much too small, and in some cases, it is even impossible to satisfy the aufbau principle. For anions, the HOMO energy often becomes positive, implying unbound electrons. In an attempt to minimize the self-interaction, orbitals become too diffuse, an especially problematic artifact for unpaired electrons¹² and dissociation of radicals.¹³ SIE has also been found to cause problems in the evaluation of polarizabilities, the electric field counterpart of magnetizabilities.¹⁴

Lutnæs et al. found that hybrid functionals that incorporate a portion of exact “HF” exchange improve the magnetizabilities. A further improvement was found by employing the optimized effective potential (OEP) approach.^{15–20} By

* To whom correspondence should be addressed. E-mail: mikael.johansson@iki.fi (M.P.P.), marcel.swart@icrea.es (M.S.).

[†] Universitat de Girona.

[‡] University of Helsinki.

treating the exchange part in this manner, an orbital-dependent nonlocality is introduced. One effect of this is to reduce the self-interaction error of the functional.²¹ Correcting for self-interaction has previously been shown to be beneficial for computing nuclear magnetic resonance (NMR) parameters^{22–24} and magnetization.²⁵ In the field of computational NMR, the Malkin correction²⁶ seems to mimic self-interaction correction to a degree, often improving the results.^{22,27} Also, calculation of polarizabilities benefits from self-interaction correction, especially within the OEP-SIC formalism.^{28,29} Therefore, it is of interest to see how explicitly self-interaction-corrected DFT (SIC-DFT) performs for magnetizabilities.

Here, we assess the performance of several density functionals, representing various levels of complexity, starting from the local density approximation³⁰ (LDA). Of the generalized gradient approximation³¹ (GGA) family, we include the Swart–Solà–Bickelhaupt³² (SSB-D) functional, recently presented as a good all-round functional, well suited for, among other properties, reliable spin state energetics. Thus, it is of interest to see how well this functional performs for the somewhat related property of magnetizability. SSB-D is to a large part based on the nonempirical Perdew–Burke–Ernzerhof (PBE) functional³³ and its modification OPBE, which uses Handy–Cohen optimized exchange, OPTX.³⁴ PW91,^{35,36} closely related to PBE, PBEsol, a reformulation of PBE which satisfies exact conditions more relevant for solids than molecules³⁷ (while still being nonempirical), for magnetic properties the tuned Keal–Tozer KT2,³⁸ and the popular BP86^{39,40} functionals are also benchmarked.

The meta-GGA functionals go beyond the GGA approximation by also taking the orbital-dependent kinetic energy density and/or the Laplacian of the electron density as an input parameter. This allows for a more complete emulation of the “true” functional, for example, by satisfaction of a fuller range of known physical properties. Potentially, ingredients important also for magnetizabilities would be included. Here, we test two common representatives of this level of theory, the nonempirical Tao–Perdew–Staroverov–Scuseria⁴¹ (TPSS) and the semiempirical Minnesota group functional M06-L⁴² (which has been found to work well for chemical shieldings⁴³) noting that neither uses the Laplacian in their functional form.

Model density functional potentials have been designed for providing a correct asymptotic behavior, another feature most traditional functionals fail to accomplish. Here, we use three of these exchange–correlation potentials: the van Leeuwen and Baerends⁴⁴ (LB94), the more recent statistical average of orbital potentials⁴⁵ (SAOP), and the gradient-regulated asymptotic correction⁴⁶ (GRAC-LB) methodologies. The potentials have been shown to improve upon the description of response properties, especially excitation energies and polarizabilities.^{47–49} We note that SAOP in particular has been suggested to perform reasonably well for NMR chemical shifts.⁵⁰

2. Computational Details

All computations were performed with the Amsterdam Density Functional (ADF) code,^{51–53} version 2009.01 (snapshot r21439); the self-interaction-corrected calculations were performed with a modified local version. The polarized quadruple- ζ QZ4P Slater-type orbital (STO) basis set⁵⁴ was used in all cases unless otherwise noted, with extra diffuse fit functions added (via the AddDiffuseFit keyword). For a further improvement in basis set convergence, the gauge-including atomic orbitals (GIAO) formalism,^{55–62} was employed. This is currently not exploitable at the meta-GGA level within the code used. Therefore, the GIAO corrections for TPSS and M06-L were taken from the PBE calculations, simply using the ratio of standard and gauge-origin-corrected magnetizabilities:

$$\xi(\text{mGGA, GIAO}) = \xi(\text{mGGA, standard}) \times \frac{\xi(\text{PBE, GIAO})}{\xi(\text{PBE, standard})}$$

The reader should be aware that the same gauge origin dependence for PBE and the kinetic energy density incorporating meta-GGA's is only a for the purpose necessary assumption. We note that in general, gauge-invariance can be satisfied also within the meta-GGA framework.⁶³

The magnetizabilities were computed perturbationally, neglecting the current density, using the standard implementation in ADF^{22,23,64} for LDA and GGA's; for details, we refer the reader to the original papers. We note that here we only report the mean magnetizabilities from the magnetizability tensor, and thus, the direction of the magnetic field does not affect the results. The self-interaction corrections were computed self-consistently, using the Perdew–Zunger⁶⁵ (PZ) scheme with the Krieger–Li–Iafrate^{66,67} (KLI) approximation of the exact optimized potential^{15–20} (OEP). The PZ SIC removes one-electron self-interaction completely and many-electron self-interaction efficiently.¹³ The orbitals were localized using the Boys–Foster procedure,^{68,69} which has been found to be suitable in this context.⁷⁰ GRAC-LB requires an estimate of the ionization potential; these were obtained from the National Institute of Standards and Technology (NIST) Chemistry WebBook. The molecular geometries were taken from ref 4. Default convergence and accuracy parameters were used, with the following tighter exceptions: The integration accuracy parameter was set to 7.5; the self-consistent field (SCF) equations were converged to a threshold of 10^{-8} au.

3. Results and Discussion

3.1. Basis Set Convergence. Magnetic properties show a notoriously slow convergence with respect to basis set size, due to the gauge-origin issue.⁷¹ The use of gauge-including atomic orbitals (GIAO's), also known as London atomic orbitals (LAO's), alleviates this problem significantly. In this section, we perform an initial assessment of the basis set convergence, using both standard Slater-type basis sets and the GIAO formalism. The computations have been performed with the SSB-D functional, and four representative molecules from the benchmark set were chosen: CH₂O (formaldehyde),

Table 1. Magnetizabilities (in SI units, 10^{-30} J T $^{-2}$) Using Different Basis Sets for Four Molecules, Computed at the SSB-D Level^a

	bf's for C	CH ₂ O		H ₂ S		LiF		OCS	
		normal	GIAO	normal	GIAO	normal	GIAO	normal	GIAO
DZ	4s2p	-395.2	-124.4	-531.4	-466.8	-394.7	-212.6	-1974.3	-598.6
DZP	4s2p1d	-196.1	-117.0	-459.3	-464.0	-343.9	-209.0	-1314.4	-591.6
TZP	5s3p1d	-147.4	-112.8	-455.6	-452.3	-262.1	-195.3	-1133.3	-572.0
TZ2P	5s3p1d1f	-144.5	-112.1	-454.5	-452.5	-260.5	-195.1	-1115.2	-571.5
<i>frozen core</i>	<i>4s3p1d1f</i>	<i>-188.2</i>	<i>-88.1</i>	<i>-454.7</i>	<i>-421.9</i>	<i>-337.5</i>	<i>-181.3</i>	<i>-1495.0</i>	<i>-538.7</i>
QZ4P	7s4p2d2f	-110.7	-108.3	-455.3	-456.3	-198.9	-196.5	-640.7	-569.6

^a The second column shows the size of the basis set as the number of basis functions for carbon (bf's for C). Both normal and GIAO calculations are reported.

H₂S, LiF, and OCS. The standard basis sets in ADF,⁵⁴ of increasing size, were employed: DZ, DZP, TZP, TZ2P, ZORA-QZ4P. The results are summarized in Table 1.

As can be seen from Table 1, the convergences of the GIAO calculations are, as expected, much better than for the standard procedure. For all molecules, the GIAO calculations appear reasonably converged at the singly polarized triple- ζ , TZP level. We also tested the effect of using a frozen-core approach, where the lowest energy orbitals are not allowed to change during the SCF. This approach is unusable when computing magnetizabilities, showing the importance of core orbitals for this property. With the TZ2P basis set, the GIAO results for the frozen core approach differ from the all-electron calculations by 14–33 $\times 10^{-30}$ J T $^{-2}$ for the four molecules above.

We also assessed the even-tempered ET-QZ+5P basis set (8s6p3d2f bf's for C) but experienced problematic linear dependency issues for some molecules. In general, for molecules without issues, the GIAO results compared to QZ4P are very small. The following calculations have been performed using the QZ4P basis set in connection with the GIAO formalism. We estimate that the error compared to the complete basis set limit is roughly $\pm 2 \times 10^{-30}$ J T $^{-2}$.

3.2. Performance of the Standard Functionals. Lutnæs et al. included the following molecules in their benchmark set:⁴ hydrogen fluoride (HF), carbon monoxide (CO), nitrogen (N₂), water (H₂O), hydrogen cyanide (HCN), hyperfluorous acid (HOF), ozone (O₃), ammonia (NH₃), formaldehyde (CH₂O), methane (CH₄), ethylene (C₂H₄), aluminum monofluoride (AlF), methyl fluoride (CH₃F), cyclopropene (C₃H₄), fluoroacetylene (FCCH), cyanogen fluoride (FCN), hydrogen sulfide (H₂S), methinophosphide (HCP), formyl fluoride (HFCO), ketene (H₂C₂O), lithium fluoride (LiF), lithium hydride (LiH), nitrous oxide (N₂O), carbonyl sulfide (OCS), difluoride monoxide (OF₂), ethylene oxide (H₄C₂O), phosphorus mononitride (PN), and sulfur dioxide (SO₂). Here, we study the same molecules, using the original CCSD(T)-optimized structures.

Tables 2 and 3 show the magnetizabilities, as computed with the different functionals benchmarked here, relative to the CCSD(T) data. Table 4 summarizes the error statistics; following Lutnæs et al., we excluded ozone (O₃) from the statistics, due to its complex multireference character, which renders even the CCSD(T) value of 121.5×10^{-30} J T $^{-2}$ somewhat unreliable; at the MCSCF level, the magnetizability has been found to be appreciably lower,⁷² 97.8×10^{-30} J T $^{-2}$.

The values for LDA, PBE, and KT2 agree well with the corresponding values computed by Lutnæs et al., which shows the expected result that close to the (GIAO) basis set limit; there is no real difference in performance between Gaussian- and Slater-type orbitals. Further, we used the Perdew–Wang³⁰ (PW92) formulation of LDA correlation, instead of the Vosko–Wilk–Nusair⁷³ (VWN) formulation; the differences are negligible.

We start by giving an overview of the combined error statistics for the functionals, after which a more detailed analysis of representative molecules is provided.

Let us first consider the nonempirical functionals at each of the three represented rungs of the DFT ladder.⁷⁴ The performance of LDA is unsatisfactory, with a mean absolute relative error (MARE) of 3.7% and a standard deviation from the benchmark data of 11.5×10^{-30} J T $^{-2}$. Surprisingly, PBE, at the next rung, improves the results very weakly, if at all, as already commented upon in ref 4. PBEsol, tuned for solids, is actually just as good (or bad). Only at the third ladder do we see an improvement; TPSS performs quite well overall, having the smallest errors of all the functionals without self-interaction correction. We also note the smallest error, compared to CCSD(T), for ozone. Thus, the added flexibility of the meta-GGA formulation compared to GGA's does seem to capture more of the physics relevant also for magnetizabilities. This is encouraging, adding to the repertoire of the wide range of chemistry and physics, this fully nonempirical functional has previously been found to perform well for.^{75–82}

Replacing the standard PBE exchange by OPTX exchange notably degrades the magnetizabilities; all statistical error measures of OPBE are worse than those of PBE. The SSB-D functional, which largely is an interpolation of PBE and OPBE, shows for magnetizabilities a performance between those of PBE and OPBE. The MARE of 4.0% is closer to that of OPBE, while the standard deviation of 9.8×10^{-30} J T $^{-2}$ is closer to that of PBE. SSB-D, which has been shown to harvest the “best of both worlds” for other properties, the worlds being PBE and OPBE, is thus not a clear improvement over either when it comes to magnetizabilities. It is, however, not worse than its major constituents either, showing that its parametrization is sound. The performance of PW91 is very close to that of PBE, as expected due to their similarity. The magnetizabilities obtained with BP86 are also very similar. Of the GGA's, KT2 is still the best performer and together with TPSS the only functional that exceeds Hartree–Fock quality.

Table 2. Magnetizabilities (in SI units, $10^{-30} \text{ J T}^{-2}$) for the Lutnæs et al. Benchmark Set⁴ Using Hartree–Fock (from ref 4), LDA, and Different GGA Functionals^a

molecule	CCSD(T)	HF	LDA	PBE	PBEsol	OPBE	SSB-D	KT2	PW91	BP86
HF	-176.4	3.7	-4.9	-3.3	-3.7	0.8	0.1	-0.2	-3.1	-2.7
CO	-209.5	5.0	3.0	4.4	4.4	10.2	8.8	0.8	4.2	4.9
N ₂	-205.2	2.4	4.2	6.1	5.9	13.0	11.8	1.0	6.2	6.5
H ₂ O	-235.1	3.8	-6.2	-3.2	-4.1	1.7	1.1	0.0	-3.3	-2.6
HCN	-271.8	-8.3	6.3	7.6	7.9	17.7	13.6	1.4	7.4	7.6
HOF	-235.4	-9.2	6.8	8.6	7.1	11.0	12.2	8.2	9.2	8.2
O ₃	121.5	459.0	75.0	64.6	67.1	66.8	54.9	19.5	65.9	62.1
NH ₃	-290.3	2.9	-7.9	-3.0	-4.8	1.3	0.5	-0.4	-2.7	-1.5
CH ₂ O	-127.4	-12.1	31.6	22.5	26.3	20.7	19.1	9.3	22.4	19.3
CH ₄	-316.9	3.2	-12.8	-3.6	-7.8	0.3	1.5	0.5	-3.1	-1.5
C ₂ H ₄	-345.6	-9.1	14.2	14.8	14.7	20.4	18.9	9.8	15.9	15.0
AlF	-394.5	-4.9	-1.4	-2.1	-1.1	1.3	-1.2	1.8	-1.7	0.4
CH ₃ F	-315.7	-2.3	0.1	4.4	1.9	5.6	7.9	7.8	5.0	4.1
C ₃ H ₄	-478.9	0.9	14.2	19.1	16.3	23.0	25.0	21.2	21.7	19.2
FCCH	-441.6	-10.6	2.9	4.3	3.3	11.9	9.3	1.4	5.0	3.9
FCN	-370.0	-8.0	4.4	5.4	5.0	10.4	8.1	2.7	5.8	5.5
H ₂ S	-455.1	2.3	-11.9	-4.4	-7.8	1.3	-1.1	-1.8	-3.3	-2.1
HCP	-492.8	-18.7	14.7	13.3	14.9	22.5	19.4	4.3	13.0	13.3
HFCO	-307.2	-4.3	10.2	10.2	10.6	13.6	11.5	7.7	10.4	10.9
H ₂ C ₂ O	-423.9	-8.7	-4.1	2.2	-1.3	7.1	8.1	1.7	4.4	3.0
LiF	-195.5	4.6	-1.1	-0.9	-0.9	2.0	-1.0	-1.2	0.2	-1.9
LiH	-127.2	1.9	-8.9	-7.9	-8.5	-8.2	-8.4	-10.0	-8.7	-6.1
N ₂ O	-339.1	-3.7	4.4	7.3	-6.3	12.7	11.5	8.5	8.2	6.8
OCS	-584.1	-13.4	7.1	9.2	8.4	16.5	14.5	8.0	10.1	9.3
OF ₂	-247.1	-24.5	27.6	26.8	25.8	27.4	29.2	21.8	27.8	26.1
H ₄ C ₂ O	-535.2	-9.6	5.4	11.9	7.9	15.1	16.4	14.2	13.8	12.4
PN	-308.2	4.4	23.2	23.7	24.9	33.9	27.7	11.1	23.0	23.4
SO ₂	-314.3	12.5	19.0	20.4	-4.4	30.3	24.0	17.3	20.5	22.0

^a DFT values computed with the QZ4P basis set; values shown are relative to the CCSD(T)/aug-cc-pCV[TQ]Z reference data.**Table 3.** Magnetizabilities (in SI units, $10^{-30} \text{ J T}^{-2}$) for the Lutnæs et al. Benchmark Set⁴ Using Different meta-GGA's and Model Potentials, Computed with the QZ4P Basis Set^a

molecule	CCSD(T)	TPSS	M06-L	LB94	SAOP	GRAC
HF	-176.4	-1.9	2.4	2.3	2.8	-1.1
CO	-209.5	-0.4	2.3	25.6	5.3	6.2
N ₂	-205.2	2.0	6.5	29.7	7.6	8.0
H ₂ O	-235.1	-1.1	3.9	2.0	2.6	-0.5
HCN	-271.8	-0.3	7.0	29.2	9.0	8.9
HOF	-235.4	-0.4	-9.2	18.5	12.0	8.4
O ₃	121.5	4.5	-25.1	187.6	48.4	59.2
NH ₃	-290.3	0.1	5.3	-1.3	1.0	-1.8
CH ₂ O	-127.4	-5.1	-15.7	46.0	5.8	19.1
CH ₄	-316.9	0.5	3.2	-11.3	-5.1	-3.4
C ₂ H ₄	-345.6	-4.8	-5.9	33.9	16.1	14.4
AlF	-394.5	-10.2	-15.3	17.5	-8.6	0.7
CH ₃ F	-315.7	-8.6	-14.1	1.0	4.8	4.2
C ₃ H ₄	-478.9	-7.9	-22.0	24.8	18.2	18.4
FCCH	-441.6	-16.1	-16.1	21.9	6.5	5.0
FCN	-370.0	-9.7	-25.5	23.4	7.3	5.9
H ₂ S	-455.1	0.3	0.0	0.5	-5.9	-0.5
HCP	-492.8	-2.6	6.0	50.2	8.4	14.1
HFCO	-307.2	-8.5	-22.6	27.2	7.3	11.0
H ₂ C ₂ O	-423.9	-19.1	-41.1	17.1	0.9	2.1
LiF	-195.5	-2.2	-0.4	13.1	2.4	1.5
LiH	-127.2	-7.0	-6.2	13.8	-5.0	-1.7
N ₂ O	-339.1	-5.4	-22.4	21.4	9.9	6.7
OCS	-584.1	-12.0	-42.0	28.2	4.8	8.9
OF ₂	-247.1	5.4	-18.7	44.6	25.3	26.1
H ₄ C ₂ O	-535.2	-10.8	-18.8	10.6	12.5	12.0
PN	-308.2	8.9	12.0	62.5	11.5	24.8
SO ₂	-314.3	0.2	-26.4	50.0	9.7	22.5

^a The values are shown relative to the CCSD(T)/aug-cc-pCV[TQ]Z reference data.

The other meta-GGA studied, the empirical M06-L, performs quite poorly for magnetizabilities. Of the normal

Table 4. Statistical Error Analysis for the Density Functional Methods Studied^a

method	ME	MAE	MaxE	MRE	MARE	SD
Hartree–Fock	-3.7	7.2	-24.5	1.1%	2.5%	8.4
LDA	5.2	9.6	31.6	-2.0%	3.7%	11.5
PBE	7.2	9.3	26.8	-2.4%	3.4%	9.5
PBEsol	5.0	8.7	26.3	-1.8%	3.3%	10.3
OPBE	12.0	12.6	33.9	-3.9%	4.4%	10.4
SSB-D	10.7	11.5	29.2	-3.4%	4.0%	9.8
KT2	5.4	6.4	21.8	-1.6%	2.3%	7.3
PW91	7.7	9.6	27.8	-2.5%	3.5%	9.6
BP86	7.5	8.9	26.1	-2.5%	3.2%	8.9
TPSS	-4.3	5.6	-19.1	1.3%	1.7%	6.3
M06-L	-10.1	13.7	-42.0	2.9%	4.2%	14.8
LB94	22.3	23.2	62.5	-8.1%	8.4%	17.9
SAOP	6.2	8.0	25.3	-2.0%	2.7%	7.5
GRAC	8.1	8.8	26.1	-2.8%	3.1%	8.5

^a Hartree–Fock results from ref 4. Mean error (ME), mean absolute error (MAE), maximum error (MaxE), mean relative error (MRE), mean absolute relative error (MARE), and standard deviation (SD) are shown. Except for the relative errors, units are $10^{-30} \text{ J T}^{-2}$.

functionals tested here and in ref 4, it is actually the worst, with the largest maximum error, largest spread in errors, and largest mean absolute errors. The parametrization for the functional thus appears unsuitable for this property. One can note, however, that for the highly multiconfigurational ozone, M06-L performs reasonably, being on par with TPSS and KT2. This suggests the possibility that the true nonlocality entering, indirectly, via the orbital-dependent kinetic energy density does capture more of the relevant correlation also for difficult multireference species. A data set of one molecule is naturally too small for any conclusions; a more rigorous study should be performed.

Table 5. Magnetizabilities (in SI units, 10^{-30} J T $^{-2}$) for the Lutnæs et al. Benchmark Set⁴ Using SIC-DFT, Computed with the QZ4P Basis Set^a

molecule	LDA	PBE	PBEsol	OPBE	SSB-D	KT2	PW91	BP86	LB94	SAOP	GRAC
HF	6.9	4.8	5.3	5.6	3.0	-1.1	5.0	4.3	2.8	8.2	0.7
CO	8.4	2.9	3.9	1.0	0.9	-6.6	3.6	-0.4	1.8	-2.2	-10.9
N ₂	7.2	1.9	2.8	0.5	1.1	-7.5	2.7	-1.3	-0.7	-3.0	-9.9
H ₂ O	9.3	7.0	7.9	7.9	5.2	1.1	8.2	8.6	5.8	9.2	0.5
HCN	9.4	4.3	5.4	4.2	5.0	-3.6	0.8	5.0	2.4	-0.8	-6.5
HOF	12.2	5.8	7.1	3.5	0.2	-8.0	6.3	4.8	5.3	7.8	4.6
O ₃	-48.2	-66.4	-66.1	-83.5	-75.9	-89.1	-62.1	-78.7	-95.3	-85.7	-61.9
NH ₃	10.3	8.4	9.1	9.4	5.4	2.4	8.9	8.3	9.0	7.3	-2.4
CH ₂ O	3.9	-3.2	-1.7	-8.1	-4.7	-9.9	-2.6	-8.0	-9.5	-17.1	-1.8
CH ₄	4.0	3.3	3.0	3.9	2.6	0.2	3.6	4.7	6.7	-2.8	-9.7
C ₂ H ₄	17.7	12.1	13.2	10.2	12.0	6.5	13.1	11.1	13.5	8.3	4.3
AlF	21.6	12.3	15.6	9.4	8.2	7.8	13.4	11.6	19.4	3.3	0.5
CH ₃ F	5.2	2.7	3.0	1.7	0.6	-1.6	2.9	3.2	4.8	3.3	4.5
C ₃ H ₄	17.8	12.1	13.1	10.6	11.9	8.0	13.5	12.1	15.5	8.4	13.8
FCCH	11.9	4.8	6.3	2.7	2.2	-2.8	5.7	2.9	8.5	2.4	0.0
FCN	9.4	2.8	4.1	0.9	-0.5	-6.8	-8.4	-0.1	4.4	-0.9	-5.0
H ₂ S	10.5	6.6	7.3	7.0	2.4	-2.0	7.4	6.6	12.4	-2.0	-2.7
HCP	17.3	6.9	9.5	4.2	5.2	-1.9	8.6	1.3	10.5	-7.4	-13.0
HFCO	9.4	3.3	4.6	1.1	-1.4	-7.6	4.3	0.1	4.8	-4.5	-7.2
H ₂ C ₂ O	4.9	-1.6	-1.1	-5.2	-4.3	-9.5	1.1	-2.2	6.2	-9.2	-6.3
LiF	8.0	5.7	6.4	6.3	3.1	-1.3	6.2	4.2	6.9	8.1	3.3
LiH	9.9	8.9	8.8	7.7	9.2	8.5	9.2	8.0	13.3	2.6	-6.0
N ₂ O	1.5	-4.0	-3.4	-6.0	-6.1	-10.5	-3.0	-6.4	-1.8	-4.2	-3.2
OCS	4.3	-5.3	-3.3	-6.7	-8.5	-15.5	-3.5	-9.6	-1.0	-19.9	-13.7
OF ₂	23.6	12.1	14.2	5.9	1.0	-11.4	13.2	9.1	7.5	8.6	9.3
H ₄ C ₂ O	16.3	10.9	11.5	9.5	6.7	1.3	12.2	11.3	14.5	8.8	12.0
PN	6.7	-1.9	-0.8	-6.1	-5.8	-18.7	-0.5	-9.2	-3.8	-22.8	-30.3
SO ₂	4.3	-5.3	-3.5	-8.1	-9.6	-17.0	-2.5	-11.4	-3.3	-23.5	-23.1

^a Values are shown relative to the CCSD(T)/aug-cc-pCV[TQ]Z reference data.

Of the three model potentials studied, the older LB94 is surprisingly weak. Of all DFT methods studied, it shows the largest deviations from the reference data. The GRAC-LB and especially SAOP potentials, on the other hand, perform quite well. There seems to be, however, no advantage of using SAOP or GRAC-LB compared to well-performing standard functionals. Indirectly, it shows that a correct asymptotic behavior of the potential is not, as such, sufficient for obtaining good magnetizabilities; this has previously been noted also for excitation energies.⁸³ It should be noted that in ref 4 the Coulomb attenuated CAM-B3LYP,⁸⁴ with improved asymptotic behavior over the original B3LYP,^{85,86} did improve magnetizabilities.

The molecules where correlation is most important, measured as the largest relative errors of HF compared to CCSD(T), are, in decreasing order, O₃ (378%), OF₂ (9.9%), CH₂O (9.5%), SO₂ (-4.0%), HOF (3.9%), and HCP (3.8%). One could expect that for these, at least, DFT methods would outperform HF. We first consider LDA and the GGA's as well as the three model potentials for which this is *not* the case. For ozone, there is a clear improvement but the errors in general are still very large at the DFT level. For OF₂, the absolute error is of the same magnitude. It is however consistently of opposite sign, thus indicating an overcorrection. This overcorrection is even more serious in the case of CH₂O, where the absolute errors are significantly larger than for Hartree-Fock. For LDA and all the GGA's, these two molecules are, after O₃, the most problematic of the moderately correlated molecules. For all of these, the triple excitations are also significant. Coupled cluster with single and double excitations, CCSD,⁸⁷ is still in error by 3.3% (OF₂), 3.4% (CH₂O), and 39% (O₃).⁴

Nevertheless, problems at the DFT level are also observed for the molecules where triple excitations are not important. For SO₂, most functionals actually increase the absolute error. HOF and HCP still exhibit the same trend, with the DFT functionals providing roughly the same absolute error as HF but of opposite sign.

The situation changes when investigating the above subset of molecules at the TPSS level. The very good agreement between TPSS and CCSD(T) for ozone was already commented on. For the other molecules with an important correlation contribution, TPSS provides an improvement over Hartree-Fock that brings the magnetizabilities much closer to the CCSD(T) reference, with relative errors ranging from -0.1% (SO₂) to 4.0% (CH₂O). For all of these molecules, TPSS also clearly outperforms KT2, which again suggests that the added completeness in the description of exchange and correlation at the meta-GGA level is important.

The subset of molecules where the correlation contribution to magnetizability is very small, 1% or less, consists of C₃H₄ (-0.2%), H₂S (-0.5%), CH₃F (0.7%), CH₄ (-1.0%), and NH₃ (-1.0%). With the exception of C₃H₄, all of these are described quite well by DFT, which suggests that where exchange only is important DFT does not overshoot the correlation correction and thus works well. The case of C₃H₄, where the HF value is closest to the CCSD(T) reference data, is special, with DFT errors being quite large, between -3.0% for LDA up to -5.2% for the GGA's and model potentials. TPSS again performs best but is still in error by 1.7%. Studying the convergence for this molecule in the series HF → CCSD → CCSD(T) reveals that correlation effects *are* present but hidden by the fact that going from CCSD to CCSD(T) brings the value back toward the HF value. For

Table 6. Statistical Error Analysis for the Self-Interaction-Corrected Density Functional Methods Studied, Computed with the QZ4P Basis Set^a

method	ME	MAE	MaxE	MRE	MARE	SD
SIC-LDA	10.1	10.1	23.6	-3.4%	3.4%	5.7
SIC-PBE	4.4	6.0	12.3	-1.5%	2.1%	5.4
SIC-PBESol	5.5	6.5	15.6	-1.9%	2.2%	5.5
SIC-OPBE	2.7	6.3	10.6	-0.9%	2.0%	5.9
SIC-SSB-D	1.7	4.7	12.0	-0.6%	1.6%	5.7
SIC-KT2	-4.0	6.6	-18.7	1.4%	2.4%	7.2
SIC-PW91	4.8	4.8	13.5	-1.7%	2.2%	5.9
SIC-BP86	2.5	6.1	12.1	-0.8%	2.2%	6.9
SIC-LB94	5.8	7.3	19.4	-1.7%	2.5%	6.7
SIC-SAOP	-1.3	7.7	-23.5	0.4%	2.8%	10.0
SIC-GRAC	-3.6	7.6	-30.3	1.3%	2.5%	9.8

^a Mean error (ME), mean absolute error (MAE), maximum error (MaxE), mean relative error (MRE), mean absolute relative error (MARE), and standard deviation (SD) are shown. Except for the relative errors, units are $10^{-30} \text{ J T}^{-2}$.

the other “low correlation” molecules, the magnetization consistently becomes (slightly) more negative when correlation is described more completely in the ab initio series. Correlation is thus present in cyclopropene (still to a relatively small degree), but its various effects cancel.

Another anomaly that is worth mentioning is LiH. For this four-electron molecule, the correlation effect is quite small, -1.5% , and the triples contribution negligible. Despite this, all functionals, save GRAC-LB, perform poorly. The reason could be that for LiH, the potential energy surface is very shallow⁷² and small changes in geometry, although energetically insignificant, could have a notable effect on magnetizability. Thus, the errors for this molecule could be somewhat artificial. We discuss this in more detail in the following section.

3.3. Effect of Self-Interaction Correction. Almost all contemporary functionals and all functionals studied in the previous section including the model potentials⁸⁸ suffer from an unphysical self-interaction of the electrons. Also, TPSS, even if it describes the hydrogen atom correctly in this regard,

suffers from many-electron self-interaction,⁷⁵ an effect more complicated to quantify than “simple” one-electron self-interaction.^{89,90} In this section, we study the influence of (approximately) removing self-interaction from the standard functionals considered in the previous section. Table 5 shows the magnetizabilities computed at the SIC-DFT level, and Table 6 summarizes the error statistics.

We again start with a general overview of the error statistics. For SIC-LDA, both the mean error and the mean absolute error actually increase compared to standard LDA. The mean absolute relative error is slightly decreased. The errors become much more systematic, however, with the standard deviation halved. For all molecules except ozone, SIC-LDA now provides consistently too positive magnetizabilities.

All of the GGA’s, with one exception, gain a substantial error reduction from SIC. The standard deviation drops below $6 \times 10^{-30} \text{ J T}^{-2}$ for most functionals and the mean absolute relative error to around 2%; SIC-SSB-D is now the most accurate functional, with a MRE of -0.6% and a MARE of 1.6%. The one exception is SIC-KT2, with a performance almost identical to that of standard KT2. SIC does affect the KT2 values too, but overall the correction is too large; the mean error changes from $+5.4 \times 10^{-30}$ to $-4.0 \times 10^{-30} \text{ J T}^{-2}$. As the parent functional was already explicitly tuned for magnetic properties, the increased accuracy observed for the other functionals is absent.

Of the model potentials, SIC-LB94 performs significantly better than its parent and is now almost on par with the GGA’s. Still, there is no advantage in using LB94 over normal functionals. Neither SIC-SAOP nor SIC-GRAC-LB represent improvements over their uncorrected versions.

Looking at ozone, we observe a dramatic change when doing the SIC correction. The magnetizability, at all SIC levels, is lowered significantly and instead of being much too positive compared to CCSD(T) becomes not nearly positive enough. This can be related to the effect that despite

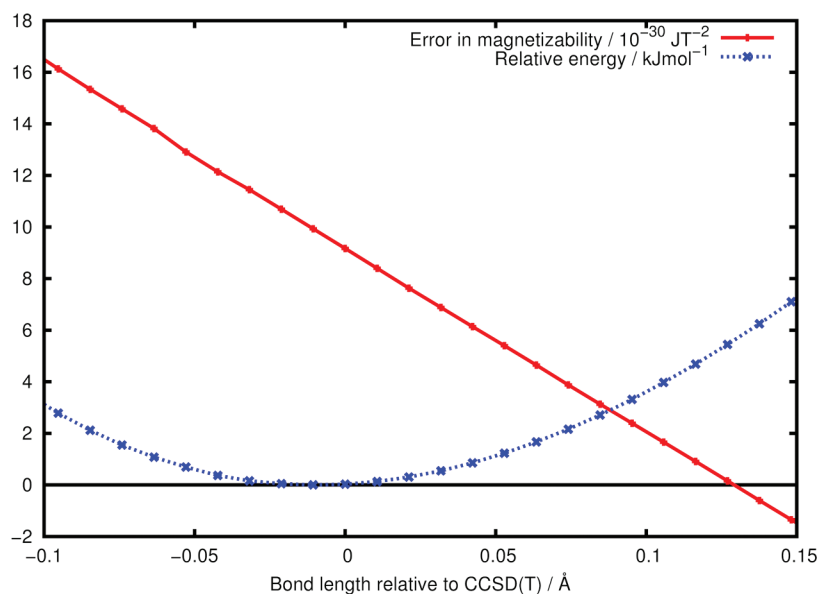


Figure 1. Magnetizability as a function of the bond length relative to the CCSD(T) value of 1.596 Å (red curve) and relative energy (blue curve), computed at the SIC-SSB-D/QZ4P level.

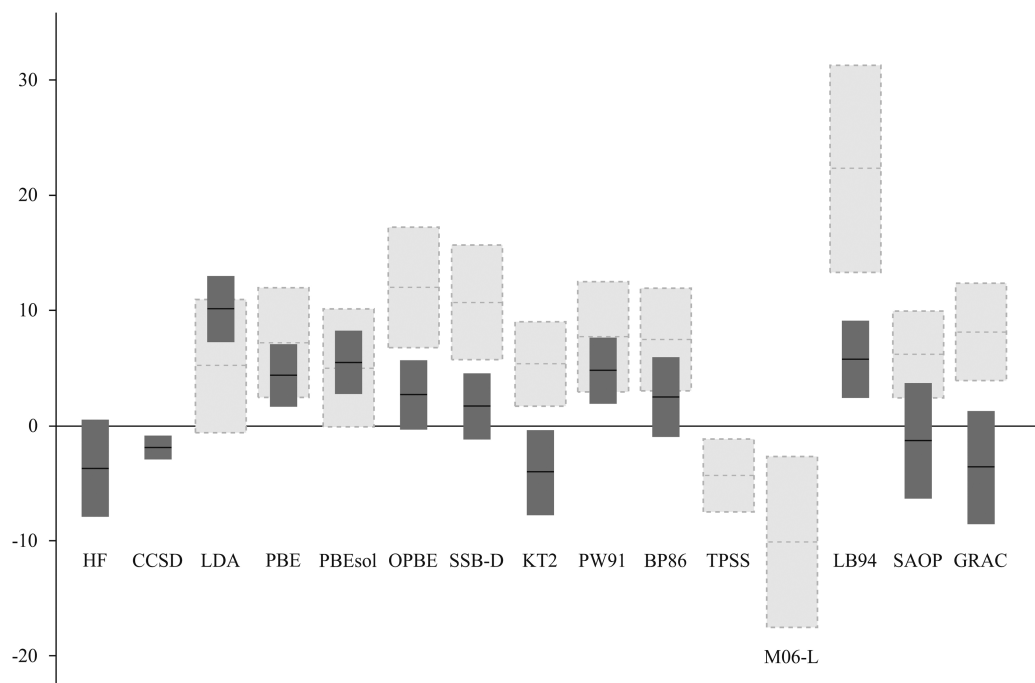


Figure 2. Mean errors and standard deviations of magnetizabilities (in SI units, $10^{-30} \text{ J T}^{-2}$) at different levels of theory. Light-gray boxes represent uncorrected, self-interaction including functionals; dark-gray boxes self-interaction-corrected (or self-interaction free) methods. The boxes are centered at the mean error of the method, and the vertical size of the boxes correspond to the standard deviation, both relative to the CCSD(T)/aug-cc-pCV[TQ]Z reference data from ref 4.

its many drawbacks, the self-interaction error present in standard functionals mimics static correlation, with the effect that DFT usually works quite well for mild multireference species; with the SIE removed, this advantageous side effect also disappears.^{34,21,91–96}

For CH_2O and OF_2 , the two other molecules with a significant contribution from the triple excitations in wave function theory, SIC-DFT performs remarkably well, although the description of CH_2O is still somewhat lacking. The exception is again SIC-KT2, where, in effect, only the sign of the error changes.

For the other molecules in the benchmark set, SIC-DFT almost consistently reduces the error at the GGA level. A somewhat surprising exception here is water, for which SIC-DFT in all cases performs notably worse than uncorrected DFT. The error for LiH also increases slightly but as discussed in the previous section could be artificial. Figure 1 shows how the error in magnetizability, compared to the CCSD(T) reference, evolves with a change of the bond length, computed at the SIC-SSB-D level, along with the relative energy. The sensitivity of the magnetizability toward the bond length is striking: with distortions between -0.05 and $+0.05 \text{ \AA}$ around the equilibrium bond length computed at CCSD(T) level, the energy difference is minute but the error in magnetizability drops from 13 to $5.6 \times 10^{-30} \text{ J T}^{-2}$.

Thus, using nonhybrid SIC-DFT, reliable magnetizabilities are obtainable. SIC-SSB-D provides the best overall performance. For self-interaction-corrected condensed matter studies,^{97,98} SIC-PBEsol could be a good choice. It would also be interesting to see how meta-GGA's would perform with SIC applied to them. The possibility of a further error reduction for, e.g., TPSS is appealing.

4. Summary and Conclusions

We studied the performance of several functionals and model exchange-correlation potentials for magnetizabilities using the recent high-quality reference set by Lutnæs et al. Of the standard functionals, without self-interaction correction, we find that the nonempirical TPSS meta-GGA performs very well, having a mean relative error (MRE) of only 1.3% and a standard deviation (SD) of $6.3 \times 10^{-30} \text{ J T}^{-2}$ compared to the reference CCSD(T) set with a mean value of magnetizability of $-324 \times 10^{-30} \text{ J T}^{-2}$. TPSS correlation is self-interaction free, and the exchange part is notably more physically correct in this respect compared to LDA and GGA's,⁷⁵ which with the results of the SIC functionals presented in this work suggests a reason for the good performance of this meta-GGA. Pending a rigorous implementation of gauge-independent magnetizabilities at the meta-GGA level, the reader should consider the values reported here somewhat preliminary though. Of the GGA's, KT2 performs best. The SSB-D functional exhibits a performance between that of its major constituents, PBE and OPBE. The model potentials LB94, SAOP, and GRAC-LB, designed for providing a correct asymptotic behavior, do not outperform standard functionals; LB94 performs quite poorly, and SAOP and GRAC-LB perform no better than normal functionals.

Correcting for the electron self-interaction present in the functionals leads to a significant reduction of the errors in almost all cases. With self-interaction correction, the best GGA studied here, SIC-SSB-D, provides a MRE of only -0.6% , a MAE of $4.7 \times 10^{-30} \text{ J T}^{-2}$, and a SD of $5.7 \times 10^{-30} \text{ J T}^{-2}$. Further, other GGA's closely follow. Notably, SIC leads to a significant reduction of the SD at LDA and

most GGA levels. Figure 2 summarizes the presented results, with the mean errors and standard deviations of both the standard and the self-interaction corrected methods depicted graphically.

Overall, DFT methods *can* be used to get reliable magnetizabilities. To achieve this level of quality, in addition to using a suitable functional, a minimum requirement is an all-electron polarized triple- ζ quality basis set, employed within a gauge-independent formalism.

It is perhaps fair to say that magnetizability is a second-order property not only mathematically, even with its direct relation to bulk permeabilities and susceptibilities, and to the spontaneous ordering of molecules in a magnetic field. This underappreciation is probably partly due to the experimental difficulties associated with the measurement of magnetizabilities. This could, of course, be seen as an excellent opportunity for chemical theory and computation to excel, especially as reliable computations can be performed at the “pure” DFT level, without costly Hartree–Fock exchange, as shown here. The good performance of the nonempirical meta-GGA TPSS and the SIC including GGA’s and conversely the poor performance of standard DFT methods further suggest that magnetizability could be used as a guiding property when designing new density functionals as it seems to be a quite sensitive measure of a correct description of the underlying physics and chemistry.

Acknowledgment. We thank Prof. Juha Vaara for enlightening discussions. The work was supported by the Ministerio de Ciencia e Innovación (MICINN, project CTQ2008-06532/BQU), the DIUE of the Generalitat de Catalunya (project 2009SGR528), a MICINN Juan de la Cierva research grant (MPJ), and the Academy of Finland (project 136079). The Finnish IT Centre for Science provided ample computational resources.

References

- (1) Stevens, R. M.; Lipscomb, W. N. Perturbed Hartree-Fock Calculations. V. Magnetic Properties of the BH Molecule. *J. Chem. Phys.* **1965**, *42*, 3666–3669.
- (2) Sauer, S. P. A.; Enevoldsen, T.; Oddershede, J. Paramagnetism of closed shell diatomic hydrides with six valence electrons. *J. Chem. Phys.* **1993**, *98*, 9748–9757.
- (3) Ruud, K.; Helgaker, T.; Bak, K. L.; Jørgensen, P.; Olsen, J. Accurate magnetizabilities of the isoelectronic series BeH⁻, BH, and CH⁺. The MCSCF-GIAO approach. *Chem. Phys.* **1995**, *195*, 157–169.
- (4) Lutnæs, O. B.; Teale, A. M.; Helgaker, T.; Tozer, D. J.; Ruud, K.; Gauss, J. Benchmarking density-functional-theory calculations of rotational *g* tensors and magnetizabilities using accurate coupled-cluster calculations. *J. Chem. Phys.* **2009**, *131*, 144104–16.
- (5) Ruud, K.; Åstrand, P.; Taylor, P. R. Molecular Magnetizabilities: Zero-Point Vibrational Effects and the Breakdown of Pascal’s Rule. *J. Phys. Chem. A* **2001**, *105*, 9926–9930.
- (6) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (7) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.
- (8) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (9) Hartree, D. R. The wave mechanics of an atom with a non-coulomb central field. I. Theory and methods. *Proc. Cambridge Philos. Soc.* **1928**, *25*, 89–110.
- (10) Fock, V. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.* **1930**, *61*, 126–148.
- (11) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (12) Zhang, Y.; Yang, W. A challenge for density functionals: Self-interaction error increases for systems with a noninteger number of electrons. *J. Chem. Phys.* **1998**, *109*, 2604–2608.
- (13) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. Density functionals that are one- and two- are not always many-electron self-interaction-free, as shown for H₂⁺, He₂⁺, LiH⁺, and Ne₂⁺. *J. Chem. Phys.* **2007**, *126*, 104102–8.
- (14) Mori-Sanchez, P.; Wu, Q.; Yang, W. Accurate polymer polarizabilities with exact exchange density-functional theory. *J. Chem. Phys.* **2003**, *119*, 11001–11004.
- (15) Grabo, T.; Kreibich, T.; Gross, E. Optimized Effective Potential for Atoms and Molecules. *Mol. Eng.* **1997**, *7*, 27–50.
- (16) Sharp, R.; Horton, G. A Variational Approach to the Unipotential Many-Electron Problem. *Phys. Rev.* **1953**, *90*, 317–317.
- (17) Norman, M.; Koelling, D. Towards a Kohn-Sham potential via the optimized effective-potential method. *Phys. Rev. B* **1984**, *30*, 5530–5540.
- (18) Talman, J.; Shadwick, W. Optimized effective atomic central potential. *Phys. Rev. A* **1976**, *14*, 36–40.
- (19) Kümmel, S.; Perdew, J. Simple Iterative Construction of the Optimized Effective Potential for Orbital Functionals, Including Exact Exchange. *Phys. Rev. Lett.* **2003**, *90*, 043004.
- (20) Engel, E.; Dreizler, R. M. From explicit to implicit density functionals. *J. Comput. Chem.* **1999**, *20*, 31–50.
- (21) Polo, V.; Kraka, E.; Cremer, D. Electron correlation and the self-interaction error of density functional theory. *Mol. Phys.* **2002**, *100*, 1771–1790.
- (22) Patchkovskii, S.; Autschbach, J.; Ziegler, T. Curing difficult cases in magnetic properties prediction with self-interaction corrected density functional theory. *J. Chem. Phys.* **2001**, *115*, 26–42.
- (23) Patchkovskii, S.; Ziegler, T. Phosphorus NMR Chemical Shifts with Self-Interaction Free, Gradient-Corrected DFT. *J. Phys. Chem. A* **2002**, *106*, 1088–1099.
- (24) Vaara, J. Theory and computation of nuclear magnetic resonance parameters. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399–5418.
- (25) Filippetti, A.; Fiorentini, V. Self-interaction-free density-functional band theory for magnetic cuprates. *J. Magn. Magn. Mater.* **2007**, *310*, 1648–1650.
- (26) Malkin, V. G.; Malkina, O. L.; Casida, M. E.; Salahub, D. R. Nuclear Magnetic Resonance Shielding Tensors Calculated with a Sum-over-States Density Functional Perturbation Theory. *J. Am. Chem. Soc.* **1994**, *116*, 5898–5908.
- (27) Arbuznikov, A. V.; Vaara, J.; Kaupp, M. Relativistic spin-orbit effects on hyperfine coupling tensors by density-functional theory. *J. Chem. Phys.* **2004**, *120*, 2127–2139.
- (28) Pemmaraju, C.; Sanvito, S.; Burke, K. Polarizability of molecular chains: A self-interaction correction approach. *Phys. Rev. B* **2008**, *77*.

- (29) Körzdörfer, T.; Mundt, M.; Kümmel, S. Electrical Response of Molecular Systems: The Power of Self-Interaction Corrected Kohn-Sham Theory. *Phys. Rev. Lett.* **2008**, *100*, 133004.
- (30) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (31) Perdew, J. P.; Wang, Y. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B* **1986**, *33*, 8800–8802.
- (32) Swart, M.; Solà, M.; Bickelhaupt, F. M. A new all-round density functional based on spin states and S_N2 barriers. *J. Chem. Phys.* **2009**, *131*, 094103–9.
- (33) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (34) Handy, N.; Cohen, A. J. Left-right correlation energy. *Mol. Phys.* **2001**, *99*, 403–412.
- (35) Wang, Y.; Perdew, J. Correlation hole of the spin-polarized electron gas, with exact small-wave-vector and high-density scaling. *Phys. Rev. B* **1991**, *44*, 13298–13307.
- (36) Perdew, J. P.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (37) Perdew, J.; Ruzsinszky, A.; Csonka, G.; Vydrov, O.; Scuseria, G.; Constantin, L.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **2008**, *100*, 136406.
- (38) Keal, T. W.; Tozer, D. J. The exchange-correlation potential in Kohn–Sham nuclear magnetic resonance shielding calculations. *J. Chem. Phys.* **2003**, *119*, 3015–3024.
- (39) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (40) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (41) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (42) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101–18.
- (43) Zhao, Y.; Truhlar, D. G. Improved Description of Nuclear Magnetic Resonance Chemical Shielding Constants Using the M06-L Meta-Generalized-Gradient-Approximation Density Functional. *J. Phys. Chem. A* **2008**, *112*, 6794–6799.
- (44) van Leeuwen, R.; Baerends, E. Exchange-correlation potential with correct asymptotic behavior. *Phys. Rev. A* **1994**, *49*, 2421–2431.
- (45) Gritsenko, O. V.; Schipper, P. R. T.; Baerends, E. J. Approximation of the exchange-correlation Kohn-Sham potential with a statistical average of different orbital model potentials. *Chem. Phys. Lett.* **1999**, *302*, 199–207.
- (46) Grüning, M.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. Shape corrections to exchange-correlation potentials by gradient-regulated seamless connection of model potentials for inner and outer region. *J. Chem. Phys.* **2001**, *114*, 652–660.
- (47) Swart, M.; Snijders, J. G.; van Duijnen, Piet Th. Polarizabilities of amino acid residues. *J. Comput. Meth. Sci. Engrg.* **2004**, *4*, 419–425.
- (48) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. Molecular calculations of excitation energies and (hyper)polarizabilities with a statistical average of orbital model exchange-correlation potentials. *J. Chem. Phys.* **2000**, *112*, 1344–1352.
- (49) Swart, M.; van Duijnen, P. T.; Snijders, J. G. Mean polarizabilities of organic molecules. A comparison of Restricted Hartree Fock, Density Functional Theory and Direct Reaction Field results. *J. Mol. Struct. (THEOCHEM)* **1998**, *458*, 11–17.
- (50) Poater, J.; van Lenthe, E.; Baerends, E. J. Nuclear magnetic resonance chemical shifts with the statistical average of orbital-dependent model potentials in Kohn–Sham density functional theory. *J. Chem. Phys.* **2003**, *118*, 8584–8593.
- (51) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (52) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. Towards an order- N DFT method. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (53) Krykunov, M.; Autschbach, J. Calculation of static and dynamic linear magnetic response in approximate time-dependent density functional theory. *J. Chem. Phys.* **2007**, *126*, 024101–12.
- (54) van Lenthe, E.; Baerends, E. J. Optimized Slater-type basis sets for the elements 1–118. *J. Comput. Chem.* **2003**, *24*, 1142–1156.
- (55) London, F. Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *J. Phys. Radium* **1937**, *8*, 397–409.
- (56) Hameka, H. F. On the nuclear magnetic shielding in the hydrogen molecule. *Mol. Phys.* **1958**, *1*, 203–215.
- (57) Ditchfield, R. Self-consistent perturbation theory of diamagnetism - I. A gauge-invariant LCAO method for N.M.R. chemical shifts. *Mol. Phys.* **1974**, *27*, 789–807.
- (58) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.
- (59) Gauss, J.; Ruud, K.; Helgaker, T. Perturbation-dependent atomic orbitals for the calculation of spin-rotation constants and rotational g tensors. *J. Chem. Phys.* **1996**, *105*, 2804–2812.
- (60) Gauss, J.; Stanton, J. F. Electron-Correlated Approaches for the Calculation of NMR Chemical Shifts. In *Advances in Chemical Physics*, Vol. 123; Prigogine, I., Rice, S.A., Eds.; John Wiley & Sons: Hoboken, NJ, 2003; pp 355–422.
- (61) Krykunov, M.; Autschbach, J. Calculation of origin-independent optical rotation tensor components in approximate time-dependent density functional theory. *J. Chem. Phys.* **2006**, *125*, 034102–10.
- (62) Krykunov, M.; Autschbach, J. Calculation of optical rotation with time-periodic magnetic-field-dependent basis functions in approximate time-dependent density-functional theory. *J. Chem. Phys.* **2005**, *123*, 114103–10.
- (63) Maximoff, S. N.; Scuseria, G. E. Nuclear magnetic resonance shielding tensors calculated with kinetic energy density-dependent exchange-correlation functionals. *Chem. Phys. Lett.* **2004**, *390*, 408–412.

- (64) Patchkovskii, S.; Ziegler, T. Improving “difficult” reaction barriers with self-interaction corrected density functional theory. *J. Chem. Phys.* **2002**, *116*, 7806–7813.
- (65) Perdew, J. P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **1981**, *23*, 5048.
- (66) Krieger, J. B.; Li, Y.; Iafate, G. J. Construction and application of an accurate local spin-polarized Kohn-Sham potential with integer discontinuity: Exchange-only theory. *Phys. Rev. A* **1992**, *45*, 101.
- (67) Krieger, J. B.; Li, Y.; Iafate, G. J. Systematic approximations to the optimized effective potential: Application to orbital-density-functional theory. *Phys. Rev. A* **1992**, *46*, 5453.
- (68) Boys, S. F. Construction of Some Molecular Orbitals to Be Approximately Invariant for Changes from One Molecule to Another. *Rev. Mod. Phys.* **1960**, *32*, 296–299.
- (69) Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300–302.
- (70) Vydrov, O. A.; Scuseria, G. E. Effect of the Perdew-Zunger self-interaction correction on the thermochemical performance of approximate density functionals. *J. Chem. Phys.* **2004**, *121*, 8187–8193.
- (71) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons: Chichester, 1999.
- (72) Ruud, K.; Helgaker, T.; Jørgensen, P. The effect of correlation on molecular magnetizabilities and rotational g tensors. *J. Chem. Phys.* **1997**, *107*, 10599–10606.
- (73) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (74) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.* **2005**, *123*, 062201–9.
- (75) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119*, 12129–12137.
- (76) Zhao, Y.; Truhlar, D. G. Benchmark Databases for Nonbonded Interactions and Their Use To Test Density Functional Theory. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.
- (77) Furche, F.; Perdew, J. P. The performance of semilocal and hybrid density functionals in $3d$ transition-metal chemistry. *J. Chem. Phys.* **2006**, *124*, 044103.
- (78) Kanai, Y.; Wang, X.; Selloni, A.; Car, R. Testing the TPSS meta-generalized-gradient-approximation exchange-correlation functional in calculations of transition states and reaction barriers. *J. Chem. Phys.* **2006**, *125*, 234104.
- (79) Kossmann, S.; Kirchner, B.; Neese, F. Performance of modern density functional theory for the prediction of hyperfine structure: meta-GGA and double hybrid functionals. *Mol. Phys.* **2007**, *105*, 2049–2071.
- (80) Tao, J.; Tretiak, S.; Zhu, J. Performance of a nonempirical meta-generalized gradient approximation density functional for excitation energies. *J. Chem. Phys.* **2008**, *128*, 084110.
- (81) Johansson, M. P.; Lechtken, A.; Schooss, D.; Kappes, M. M.; Furche, F. The 2D–3D transition of gold cluster anions resolved. *Phys. Rev. A* **2008**, *77*, 053202.
- (82) Johansson, M. P.; Patzschke, M. Fixing the Chirality and Trapping the Transition State of Helicene with Atomic Metal Glue. *Chem.—Eur. J.* **2009**, *15*, 13210–13218.
- (83) Lehtonen, O.; Sundholm, D.; Send, R.; Johansson, M. P. Coupled-cluster and density functional theory studies of the electronic excitation spectra of *trans*-1,3-butadiene and *trans*-2-propeniminium. *J. Chem. Phys.* **2009**, *131*, 024301–13.
- (84) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (85) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (86) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.
- (87) Purvis, G. D., III; Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **1982**, *76*, 1910–1918.
- (88) Della Sala, F.; Görling, A. The asymptotic region of the Kohn-Sham exchange potential in molecules. *J. Chem. Phys.* **2002**, *116*, 5374–5388.
- (89) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Many-electron self-interaction error in approximate density functionals. *J. Chem. Phys.* **2006**, *125*, 201102–4.
- (90) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. Tests of functionals for systems with fractional electron number. *J. Chem. Phys.* **2007**, *126*, 154109–9.
- (91) Polo, V.; Kraka, E.; Cremer, D. Some thoughts about the stability and reliability of commonly used exchange-correlation functionals - coverage of dynamic and nondynamic correlation effects. *Theor. Chem. Acc.* **2002**, *107*, 291–303.
- (92) Cremer, D.; Filatov, M.; Polo, V.; Kraka, E.; Shaik, S. Implicit and Explicit Coverage of Multi-reference Effects by Density Functional Theory. *Int. J. Mol. Sci.* **2002**, *3*, 604–638.
- (93) Gräfenstein, J.; Kraka, E.; Cremer, D. The impact of the self-interaction error on the density functional theory description of dissociating radical cations: Ionic and covalent dissociation limits. *J. Chem. Phys.* **2004**, *120*, 524–539.
- (94) Allouti, F.; Manceron, L.; Alikhani, M. On the performance of the hybrid TPSS meta-GGA functional to study the singlet open-shell structures: A combined theoretical and experimental investigations of the Ni_2O_2 molecule. *J. Mol. Struct. (THEOCHEM)* **2009**, *903*, 4–10.
- (95) Johansson, M. P.; Sundholm, D. Spin and charge distribution in iron porphyrin models: A coupled cluster and density-functional study. *J. Chem. Phys.* **2004**, *120*, 3229–3236.
- (96) Gräfenstein, J.; Cremer, D. The self-interaction error and the description of non-dynamic electron correlation in density functional theory. *Theor. Chem. Acc.* **2009**, *123*, 171–182.
- (97) Filippetti, A.; Spaldin, N. Self-interaction-corrected pseudo-potential scheme for magnetic and strongly-correlated systems. *Phys. Rev. B* **2003**, *67*, 125109.
- (98) Van Aken, B. B.; Palstra, T. T.; Filippetti, A.; Spaldin, N. A. The origin of ferroelectricity in magnetoelectric YMnO_3 . *Nat. Mater.* **2004**, *3*, 164–170.

Transferability of Atomic Properties in Molecular Partitioning: A Comparison

Yu Zhang* and Adam Wasserman

Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, Indiana 47907

Received May 10, 2010

Abstract: For a given choice of fragmentation of a molecule, Partition Density Functional Theory (PDFT) provides fragment densities that add up to the correct molecular density, and produce the—in principle exact—molecular energy. Using a simple model system of a heteronuclear diatomic molecule, we investigate the transferability of the resulting PDFT fragment densities by examining how their shapes and dipoles are preserved as the environment changes, and compare with other partitioning schemes. Our results show that (1) the transferability of PDFT densities is about an order of magnitude higher than that of real-space partitioning schemes, and (2) the PDFT *dipoles* are about an order of magnitude more transferable than Hirshfeld dipoles in regions of chemical relevance.

1. Introduction

Partition Density Functional Theory (PDFT), introduced in refs 1, 2 and summarized in the next Section, is a method for calculating molecular properties from Kohn–Sham calculations on isolated fragments. This method was born from the union of the Partition Theory of ref 3 and Kohn–Sham Density Functional Theory.⁴ One of the outcomes of a molecular PDFT calculation is a set of fragment densities that can be taken as a convenient alternative to the various definitions of “atoms-in-molecules” (AIMs) proposed over the years by many authors.^{5–8} These are useful to define formal charges, local dipoles, and as conceptual tools to understand chemical reactivity.^{9,10} In addition, they are also useful *computational* tools. For example, as we explain in the next section, the first step in a PDFT calculation for the ground-state energy of a molecule requires knowledge of a set of densities that add up approximately to the correct molecular density. One could use isolated-atom densities, but the closer the sum of these densities is to the correct molecular density, the faster the convergence of the PDFT equations (at convergence, the resulting fragment densities add up in principle to the *exact* molecular density).

Like other fragment-based computational methods,¹¹ the appeal of PDFT lies in its promise to allow for electronic-

structure calculations that scale linearly with system size.¹² It is thus desirable that the fragment densities be as *transferable* as possible, in the sense that once obtained for a given system, they can be used effectively as the starting point for efficient electronic-structure calculations on other systems.

Our purpose in this work is to study the transferability of PDFT densities. We explore in detail a model system where all quantities can be obtained exactly for different partitioning schemes, compare them with each other, and reveal this way some of the key differences in transferability that are likely to appear in typical applications of AIMs. The model system we study is the same that was recently solved analytically in ref 13. At least in this model of a heteronuclear diatomic molecule, we find that the PDFT densities are significantly more transferable than those arising from other popular schemes. We also show that the PDFT dipoles are in a sense more chemically meaningful than those given by competing methods.

2. Self-Consistent Density Partitioning

An analogy between KS-DFT and PDFT provides a good way to explain the essence of PDFT.¹ Consider the simplest real-life example of interacting electrons, the He atom, and the simplest real-life example of interacting fragments, the H₂⁺ molecule. In KS-DFT, the ground-state density of He is expressed as the sum of two densities, i.e., those of the two

* To whom correspondence should be addressed. E-mail: zhang339@purdue.edu.

particles composing the fictitious Kohn–Sham system of *noninteracting electrons*. Similarly, in PDFT, the ground-state density of the H_2^+ molecule is given by the sum of two densities: those of the two fragments composing a fictitious system of *noninteracting fragments*. In KS-DFT, there is a global potential $v_s(\mathbf{r})$ shared by the two noninteracting electrons. In PDFT, there is a global *partition potential* $v_p(\mathbf{r})$ shared by the two noninteracting fragments.

The method works like this: First, we choose a way to fragment the system, that is, assign the nuclei into different fragments. Examples of *binary fragmentation* have been worked out,^{2,13} where each atom of a diatomic molecule constitutes in itself a fragment. But a fragment in a molecule could well be a set of atoms characterizing a functional group in organic molecules, or whole amino acids in proteins, or nucleotides in DNA, etc. The number of fragments N_f into which one wishes to partition the molecule plays a role that is analogous to that of the number of electrons N in DFT. Denoting the molecular 1-body potential as $v(\mathbf{r})$, it is decomposed as $v(\mathbf{r}) = \sum_{\alpha=1}^{N_f} v_{\alpha}(\mathbf{r})$, where $v_{\alpha}(\mathbf{r})$ is the α^{th} -fragment potential. Fragment α contains N_{α} electrons, not necessarily an integer (e.g., example of H_2^+ above, where $N_{\alpha} = 0.5$), but $N = \sum_{\alpha} N_{\alpha}$ is obviously an integer. To correctly treat noninteger numbers, we obtain the fragment energy \mathcal{E}_{α} via the two-component ensemble of Perdew, Parr, Levy, and Balduz¹⁴ (PPLB):

$$\mathcal{E}_{\alpha} = (1 - \nu_{\alpha})E_{\alpha}[n_{p_{\alpha}}] + \nu_{\alpha}E_{\alpha}[n_{p_{\alpha}+1}] \quad (1)$$

where p_{α} is the lower bordering integer of N_{α} , so: $N_{\alpha} = p_{\alpha} + \nu_{\alpha}$, and $0 < \nu_{\alpha} < 1$. The energy $E_{\alpha}[n_{p_{\alpha}}]$ is that of a system of p_{α} electrons in the presence of the external potential $v_{\alpha}(\mathbf{r})$ for fragment α .

The fragment densities $\{n_{\alpha}(\mathbf{r})\}$ are then obtained by minimizing the total fragment energy $E_f = \sum_{\alpha} \mathcal{E}_{\alpha}$ subject to the constraints that $N = \sum_{\alpha} N_{\alpha}$, and that $n(\mathbf{r}) = \sum_{\alpha} n_{\alpha}(\mathbf{r})$, the total density. Such constrained minimization needs two Lagrange multipliers: the chemical potential μ to guarantee satisfaction of the number constraint, and the *partition potential* $v_p(\mathbf{r})$ to guarantee satisfaction of the *density* constraint. Like the chemical potential, $v_p(\mathbf{r})$ is a global property of the system - the same for all fragments. Like the Kohn–Sham potential, it is a local potential for the electrons within each fragment, and it is “the” potential that yields the set of $\{n_{\alpha}\}$ adding to the true interacting molecular density. In fact, the fragment density $n_{\alpha}(\mathbf{r})$ is the ensemble ground-state density for N_{α} electrons in the external potential $v_{\alpha}(\mathbf{r}) + v_p(\mathbf{r})$ (for the important proof of uniqueness of $v_p(\mathbf{r})$, see ref 15 where $v_p(\mathbf{r})$ was termed a “reactivity potential”).

Since we do not know the total density $n(\mathbf{r})$ in advance, a practical scheme is needed. To that end, the *partition energy* E_p is defined as the difference between the exact energy E and the fragment energy E_f :

$$E_p[\{n_{\alpha}\}] = E[n] - E_f[\{n_{\alpha}\}] \quad (2)$$

This is analogous to defining the Hartree-exchange-correlation energy E_{HXC} in KS-DFT as the difference between the exact energy and the sum $T_s[n] + \int d\mathbf{r} v_{\text{ext}}(\mathbf{r})n(\mathbf{r})$ (non-interacting kinetic energy plus energy due to external

potential). The analogy continues: just like practical KS-DFT calculations rely on approximations to E_{HXC} as a functional of the ground-state density, the success of PDFT will ultimately rely on the quality of approximations of E_p as a functional of the set of fragment densities $\{n_{\alpha}\}$. The partition potential is given by $\delta E_p[\{n_{\alpha}\}]/\delta n_{\alpha}(\mathbf{r})$, for any fragment α . Thus, approximating $E_p[\{n_{\alpha}\}]$ as a functional of the set of fragment densities leads to a definite prediction for $v_p(\mathbf{r})$, which in turns yields a unique set of $\{n_{\alpha}(\mathbf{r})\}$, and therefore a prediction for the total molecular density and total fragment energy (from which the true ground state energy $E = E_p + E_f$ is deduced).

In principle, any electronic-structure method could be used in conjunction with the scheme above to solve for the fragment densities and energies. The connection with Kohn–Sham DFT was derived in ref 1. Here, since we deal with a model system¹³ whose partition potential and fragment densities can be obtained analytically, numerical self-consistent calculations are not needed. Our goal is to study the transferability of fragments obtained via *exact* PDFT, and compare it with that of other density-partitioning schemes.

In the next Section we define the transferability measures to be used, and in Section 4 we provide a direct comparison between the different methods.

3. Shape Transferability

Rigorous discussions on the mathematical foundation of transferability can be found in ref 16. Here we use an intuitive definition for the transferability of molecular properties. Given a recipe to calculate a molecular property P_M for a molecule M from the set of fragment densities $\{n_{\alpha}^M(\mathbf{r})\}$, and the set of their centers of gravity $\{\mathbf{r}_{\alpha}^M\}$, we say that a fragment β common to two different molecules M and M' is P -transferable from M to M' if using n_{β}^M in place of $n_{\beta}^{M'}$ leads via the same recipe to a value of P that is “close enough” to that of $P_{M'}$. For example, we discuss below the case of μ -transferability, where the recipe to calculate the dipole moment μ is as follows:

$$\mu_M = \sum_{\alpha} (\mu_{\alpha}^M - \mathbf{R}_{\alpha}^M N_{\alpha}^M) \quad (3)$$

The α^{th} -dipole is measured with respect to an origin placed at \mathbf{R}_{α}^M ,

$$\mu_{\alpha}^M = \int d\mathbf{r} n_{\alpha}^M(\mathbf{r} - \mathbf{R}_{\alpha}^M)\mathbf{r} \quad (4)$$

For reasons that are made clear in the first paragraph of Section 4.5, we focus here on *shape transferability*, where rather than testing the adequacy of $n_{\beta}^M(\mathbf{r})$ for predicting properties of M' , we test the adequacy of the shape function^{17,18}

$$g_{\beta}^M(\mathbf{r}) \equiv n_{\beta}^M(\mathbf{r})/N_{\beta}^M \quad (5)$$

where N_{β}^M is the number of electrons of fragment β in molecule M , not necessarily an integer. The degree of μ -transferability of fragment β is thus measured by the smallness of:

$$\Delta\mu_{M' \rightarrow M}^\beta = \int d\mathbf{r} N_\beta^M \mathbf{r} [g_\beta^{M'}(\mathbf{r} - \mathbf{R}_\beta^M) - g_\beta^M(\mathbf{r} - \mathbf{R}_\beta^M)] \quad (6)$$

We shall also compute

$$D_{M' \rightarrow M}^\beta = \int d\mathbf{r} (N_\beta^M)^2 [g_\beta^{M'}(\mathbf{r} - \mathbf{R}_\beta^M) - g_\beta^M(\mathbf{r} - \mathbf{R}_\beta^M)]^2 \quad (7)$$

which is a measure of the density transferability of fragment β from M' to M .

4. Comparison

Out of the many existing partitioning schemes, we choose four that do not rely on the use of basis functions. The first two split the density in real space, yielding *nonoverlapping* “atomic” densities (Bader⁷ and Voronoi⁸). The last two, by contrast, yield *overlapping* densities (Hirshfeld⁶ and PDFT¹). We apply the four schemes to a simple model system of a heteronuclear diatomic molecule where all quantities can be obtained analytically. Our model consists of two noninteracting electrons moving in one dimension in the presence of two attractive δ -function potentials of strengths $-Z_A$ and $-Z_B$, with $0 < Z_B < Z_A$, a distance $|R_A - R_B| \equiv R$ apart. This model was explored in detail in ref 13, where its relevance for understanding electron localization upon dissociation was discussed. When $Z_A = 1.05$ and $Z_B = 0.95$, the true ionization energy of lithium hydride (0.3 a.u.) is obtained for $R = 2$ a.u. Although the model is admittedly a caricature of the LiH molecule, it nevertheless retains some of the essential elements of real closed-shell heteronuclear diatomic molecules: 2 nuclei with different effective charges (one slightly higher than 1, the other slightly less than 1), 2 active electrons, a permanent dipole moment, density cusps at the nuclei, and exponential decay of the density at large distances, with a decay constant determined by the ionization energy of the most electronegative atom. We therefore expect our comparison to be meaningful for real heteronuclear diatomic molecules.

Since there are two electrons, the total molecular density is given by:

$$n(x) = 2|\psi(x)|^2 \quad (8)$$

where the molecular orbital $\psi(x)$ has the piecewise expression:

$$\psi(x) = \begin{cases} Ce^{\kappa(R/2+x)}, & x < -R/2 \\ De^{\kappa x} + Fe^{-\kappa x}, & -R/2 \leq x \leq R/2 \\ Ge^{\kappa(R/2-x)}, & x > R/2 \end{cases} \quad (9)$$

The wavenumber κ is given by the solutions of the transcendental equation:

$$(\kappa - Z_A)(\kappa - Z_B) = e^{-2\kappa R} Z_A Z_B \quad (10)$$

and the constants C , D , F , and G are given explicitly in the Appendix in terms of κ , Z_A , Z_B , and R .

We now propose one-dimensional versions that capture the main features of the four partitioning schemes, and apply them to our model system:

4.1. Bader. Bader’s Atoms-in-Molecules scheme⁷ uses zero-flux surfaces to partition the entire space. These are defined as the surfaces through which the divergence of the electron density $\nabla n(\mathbf{r})$ has no flux. In one-dimensional systems, the dividing “surfaces” collapse to points where $dn(x)/dx = 0$. Bader’s A-fragment density in our model system is thus identical to $n(x)$ for all x less than $x_{\text{Bader}} \equiv (1/2\kappa)\ln(F/D)$, and zero otherwise:

$$n_A^{\text{Bader}}(x) = \begin{cases} n(x), & x < (1/2\kappa)\ln(F/D) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

The left panel of Figure 1 compares this fragment density (solid line) with an isolated atom density when the choice of parameters is such that $n(x)$ is almost symmetric, having $F \approx D$ and therefore $x_{\text{Bader}} \approx 0$.

4.2. Voronoi. In this scheme, the atomic basins are bounded by planes perpendicular to the interatomic bonds. According to the original proposal of Voronoi,¹⁹ the dividing planes should be located at the midpoints of the interatomic bonds, but a modified version⁸ puts them at a distance $R_A = (X_A/(X_A + X_B))R$ from nucleus A, where R is the distance between nuclei A and B, and the $X_{A,B}$ ’s are the van der Waals radii of atoms A and B. In our one-dimensional model, since the nuclear radii are ill-defined, we locate the dividing “planes” at distances $R_A = (Z_B/(Z_A + Z_B))R$ instead, where Z_A and Z_B are the nuclear charges. Voronoi’s A-fragment density is then identical to $n(x)$ for all x less than $x_{\text{Voronoi}} \equiv R[Z_B/(Z_A + Z_B) - 1/2]$:

$$n_A^{\text{Voronoi}}(x) = \begin{cases} n(x), & x < R[Z_B/(Z_A + Z_B) - 1/2] \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The second panel of Figure 1 shows this density for a choice of parameters such that $Z_B/(Z_A + Z_B) \approx 1/2$, and therefore $x_{\text{Voronoi}} \approx x_{\text{Bader}}$.

4.3. Hirshfeld. The one-dimensional version does not require any modification of Hirshfeld’s original proposal.⁶ Here, the α^{th} -fragment density $n_\alpha(x)$ is found by multiplying the molecular density $n(x)$ by a weight function $w_\alpha(x)$ given by the ratio $n_\alpha^0(x)/n^0(x)$, where $n_\alpha^0(x)$ is the density of fragment α when truly isolated, and $n^0(x) = \sum_\alpha n_\alpha^0(x)$. For our model system, the truly isolated A-atom density is the simple exponential $n_A^0(x) = Z_A e^{-2Z_A|x+R/2|}$, so

$$n_A^{\text{Hirshfeld}}(x) = n(x) \left[1 + \frac{Z_B}{Z_A} \left(\frac{\exp(-2Z_A|x + \frac{R}{2}|)}{\exp(-2Z_B|x - \frac{R}{2}|)} \right) \right]^{-1} \quad (13)$$

The third panel in Figure 1 shows how $n_A^{\text{Hirshfeld}}$ extends over all space (solid line), and resembles $n_A^0(x)$ (dotted line).

4.4. PDFT. We applied the method as described in Section 2. The A-fragment density was found analytically in ref 13. It is conveniently expressed as follows:

$$n_A^{\text{PDFT}}(x) = n(x)(1 + \sin 2\beta(x))/2 \quad (14)$$

where $\beta(x)$ is an auxiliary polar-angle function given by:

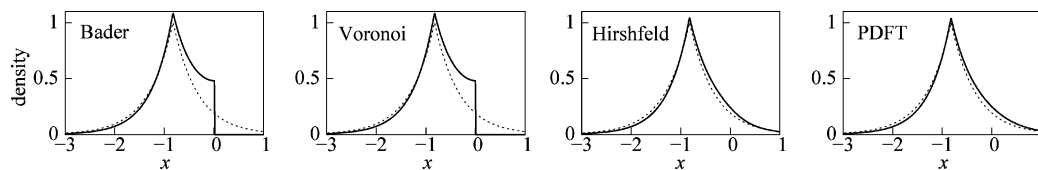


Figure 1. Comparison of A-fragment densities for four different density-partitioning schemes applied to the 1dAB model with $\Delta Z = 10^{-2}$, $Z = 1$, and $R = 1.65$. The dotted lines show the isolated-atom density: $Z_A e^{-2Z_A|x+R/2|}$.

$$\beta(x) = \begin{cases} \beta_A, & x < -\frac{R}{2} \\ \beta_A + \alpha_1 \left(\alpha_2 - \frac{e^{-2\kappa x}}{De^{(2\kappa x)} + F} \right), & -\frac{R}{2} \leq x \leq \frac{R}{2} \\ \beta_B, & x > \frac{R}{2} \end{cases} \quad (15)$$

The values for the constants β_A , β_B , α_1 , and α_2 are provided in the Appendix.

The right panel in Figure 1 shows $n_A^{\text{PDFT}}(x)$ when $Z_A = 1.005$, $Z_B = 0.995$, and $R = 1.65$. The distortion of the PDFT fragment density with respect to the isolated-atom density equals that of Hirshfeld's to first-order in $\Delta Z \equiv Z_A - Z_B$. However, to higher orders of ΔZ , the differences will prove important, as we discuss next.

4.5. "Atomic" Densities and Dipoles. It is clear from Figure 1 that Bader and Voronoi densities belong to a different family than those of Hirshfeld and PDFT. The overlapping densities (Hirshfeld and PDFT) resemble atomic densities, but differ from them in a meaningful way. It might seem at first sight from Figure 1 that the Hirshfeld and PDFT schemes yield identical densities, but careful examination shows that the PDFT density of the most electronegative atom (A in this case) exhibits a slower decay in the bonding region (range $0 < x < 1$ in Figure 1). Interestingly, since all atomic dipoles vanish as $R \rightarrow \infty$, these subtle differences are large enough to determine the *sign* of the dipole at large but finite R .

Figure 2 shows the electronic dipole moment for the A-fragment (1d-version of eq 4) as a function of inter-"nuclear" separation R . Chemically, it makes sense to expect the sign of μ_A to be positive for all R . Indeed, as the negative charge density on the A-atom is pulled by the unscreened positive charge on the B-atom, the fragment density for the A-atom becomes asymmetric. There is more negative charge on the right-hand side of the A-nucleus than on the left. So, a chemically reasonable fragment density for the A-atom should be expected to exhibit a positive dipole at all internuclear separations. PDFT is the only scheme that satisfies this (see upper panel in Figure 2). The slower decay of $n^{\text{PDFT}}(x)$ compared to $n^{\text{Hirshfeld}}(x)$ in the bonding region, hardly visible in Figure 1 (the effect is second order on ΔZ), is in fact crucial to determine the "correct" sign of the dipole moment at large R , where Hirshfeld's becomes negative. We use quotes for *correct* because the fragment dipole is of course not measurable at finite R . The Bader and Hirshfeld dipoles switch sign at $R \approx 1.4$ and 2.2 respectively, and the Voronoi dipole is always negative. The lower panels in Figure 2 show the actual A-fragment densities at small and large R , explaining the observed sign of μ_A .

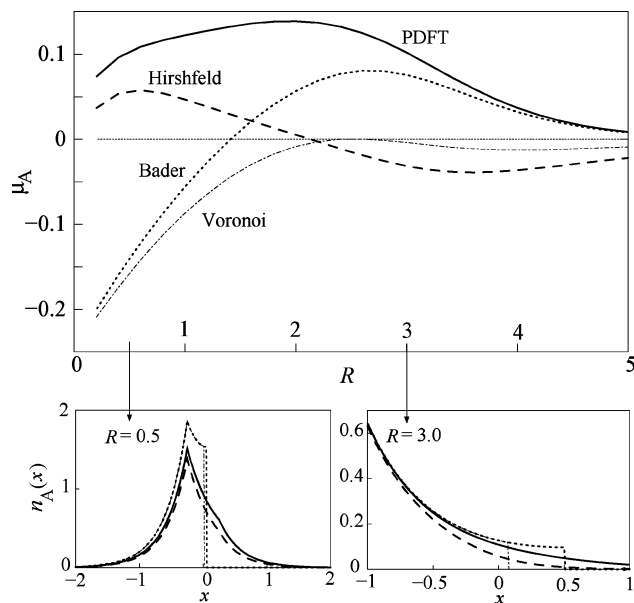


Figure 2. Top panel: Fragment dipole eq 4 for the A-atom as a function of inter-"nuclear" separation when $Z_A = 1.05$, and $Z_B = 0.95$. Bottom panels: Density of A-atom from the four different schemes when $R = 0.5$ (left) and $R = 3.0$ (right).

We make two more points based on Figure 2: (1) The densities from overlapping schemes preserve their atomic-like shapes even for very small R s. This is achieved partly because at small internuclear separations, the B-atom contributes significantly to the molecular density at the position of the A-nucleus; (2) For large values of R , because of its slower decay in the bonding region, the PDFT density for the A-atom runs closer than Hirshfeld's to the exact molecular density.

4.6. n-Transferability. When ΔZ is changed for fixed R , the fragment densities change in a way that can be understood by examining two separate effects: change of *shape*, and change of *number*. As $R \rightarrow \infty$, only the *number* effect is important (for A) because $n(x)$ becomes equal to the sum of two purely atomic densities containing noninteger numbers of electrons. At truly infinite separation, the 2 electrons sit close to the most electronegative atom (A), and the density becomes equal to twice the density of the 1-electron A-atom (since electrons are not interacting in our model). At small R , however, charge transfer is less important. For example, when ΔZ is changed from 0 to 10^{-4} the B-atom retains more than 99% of its single electron for all separations lower than $R = 5$. The main effect then is a change of *shape*. It is thus convenient to separate the two effects, as indicated in Section 3, eqs 6 and 7.

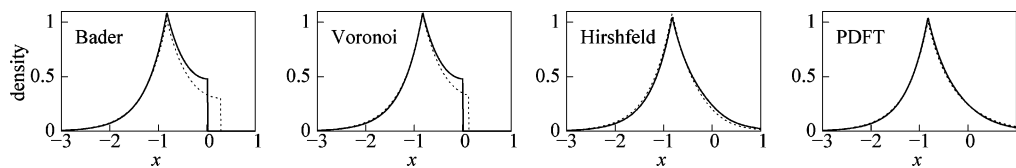


Figure 3. Fragment densities for the A-atom when $Z_A = 1.005$ and: *solid lines*: $Z_B = 0.995$, and $R = 1.65$; *dotted lines*: $Z_B = 0.895$, and $R = 1.80$. These have been shifted and renormalized to test shape transferability (see text).

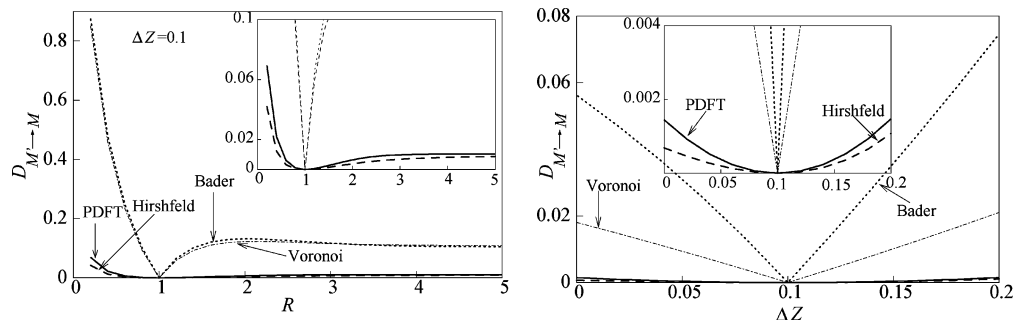


Figure 4. Density transferability as measured by $D_{M' \rightarrow M} = \sum_{\alpha} D_{M' \rightarrow M}^{\alpha}$, eq 7, for the four schemes, when: *Left*: the reference molecule M' corresponds to $R = 1$, with $\Delta Z = 0.1$, and the tested molecules M have the same value of ΔZ but different R 's; *Right*: the reference molecule has $R = 1$ and $\Delta Z = 1$, and the tested molecules M have the same value of R but different ΔZ 's.

To that end, we first take the same densities of Figure 1, extract their shapes (eq 5), and multiply them by the number of electrons that the A-atom would have had in a molecule defined by the same value of Z_A but different Z_B and R (this number obviously changes from scheme to scheme). Figure 3 compares the renormalized and shifted A-density of M' (the new ‘molecule’) with that of M (the original ‘molecule’). Clearly, for the set of parameters chosen in Figure 3, the PDFT and Hirshfeld densities are more transferable than those obtained by nonoverlapping schemes. We confirm this is true for *any* choice of parameters by plotting D (defined in eq 7) for fixed ΔZ as a function of R (left panel of Figure 4) and for fixed R as a function of ΔZ (right panel of Figure 4). Similar behavior was observed for all choices of R and ΔZ . We note that the slopes dD/dR and $dD/d\Delta Z$ around the reference point that defines the molecule M ($R = 1$ in the left panel of Figure 4) are more than an order of magnitude smaller for the overlapping than for the nonoverlapping schemes (better seen in the insets), proving the much higher transferability of Hirshfeld and PDFT densities over those based on real-space cutoff procedures. Interestingly, the PDFT scheme is superior to Hirshfeld’s in terms of *dipole* transferability, as we show below.

4.7. μ -Transferability. We now fix $Z_A = 1.05$, $Z_B = 0.95$, and $R = 2.0$, a chemically reasonable bond length for this model. The exact molecular dipole is -0.71073 a.u. Table 1 shows the values obtained under the 4 different schemes when Z_B is lowered to 0.9 (keeping Z_A fixed), and the shape of the new fragments is employed to estimate the dipole of the original molecule. The difference with the exact dipole is given by the 1-dimensional version of eq 6 (column 2 in Table 1). Clearly, for $R = 2$, the PDFT densities are the most μ -transferable. Lastly, Figure 5 plots $\Delta\mu_{M' \rightarrow M} = \sum_{\alpha} \Delta\mu_{M' \rightarrow M}^{\alpha}$ as a function of R . PDFT averages over the two nonoverlapping schemes, and does better than Hirshfeld’s for all separations larger than ~ 1 . In fact, in this window of

Table 1. Error for Molecular Dipole when $R = 2$ from the Four Partitioning Schemes^a

	μ_M	$\Delta\mu_{M' \rightarrow M}$	error (%)
Bader	-0.68315	0.02758	3.8
Voronoi	-0.73175	-0.02103	-2.9
Hirshfeld	-0.81283	-0.10210	-14.4
PDFT	-0.70912	0.00161	0.2

^a Z_A is fixed to 1.05 and Z_B lowered from 0.95 to 0.90 (see text).

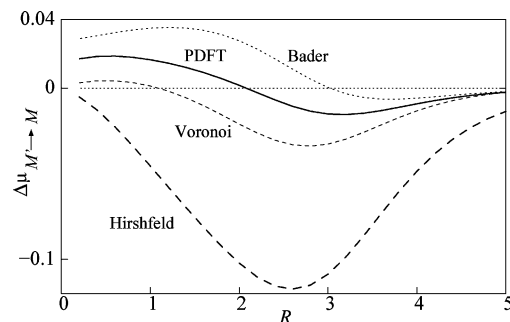


Figure 5. Transferability of the dipole moment as measured by $\Delta\mu_{M' \rightarrow M} = \sum_{\alpha} \Delta\mu_{M' \rightarrow M}^{\alpha}$, eq 6. Here, $Z_A = 1.05$ and Z_B changes from 0.95 for M to 0.90 for M' .

internuclear separations, both nonoverlapping schemes show better dipole-transferability than Hirshfeld’s. This can be understood in the following way. When decreasing Z_B for fixed Z_A , the change of the molecular density close to nucleus A is a change of *number* of electrons more than shape. This is exactly true when $R \rightarrow \infty$, and approximately so for finite R . The change of the molecular density close to B , however, is both of number *and* shape, since the decay of the density is governed by Z_A far from both nuclei. Not having the flexibility to reverse this change of shape upon resizing is the main source of error of nonoverlapping schemes. The error has a negative sign at large distances, indicating that

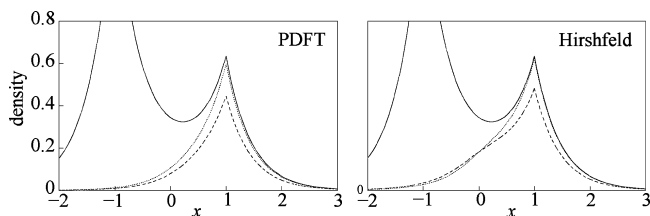


Figure 6. Comparison of the B-densities obtained by PDFT (left) and Hirshfeld (right) when $R = 2$, $Z_A = 1.05$ and Z_B is lowered from 0.95 (thin) to 0.90 (dashed). The solid line shows the full molecular density corresponding to the higher value of Z_B .

too much density leans toward the more electronegative nucleus A, understandably. Why is Hirshfeld's error so much larger? When multiplied by the actual molecular density corresponding to the lower value of Z_B , the isolated-atom density used as a weighting function in Hirshfeld's scheme shifts too much density toward A (right panel in Figure 6). After resizing (i.e., multiplying by a number larger than 1 to get the correct occupation of fragment B for the higher value of Z_B , while keeping the shape of the lower value), this leads to a molecular dipole that is much too negative, effectively squaring the error as compared to nonoverlapping schemes.

5. Conclusions

The results obtained for a simple model of a heteronuclear diatomic molecule highlight some of the advantages of the recently proposed Partition Density Functional Theory¹⁻³ over other density-partitioning schemes. First, the PDFT fragment-densities are more transferable than those of nonoverlapping schemes such as Bader's and Voronoi's. Second, the resulting PDFT dipoles are better adjusted to chemical intuition, at least in our model system.

We emphasize that PDFT is *not* simply a method to partition the density of a molecule. In fact, *all* molecular properties are encoded into each fragment density, and can be decoded from any of them. When connected to KS-DFT, refs 1 and 2 show how to extract the exact molecular ground-state energy via self-consistent calculations on the PDFT fragments. There is no obvious *a priori* reason to expect the PDFT fragment densities to be more transferable than those obtained by other density-partitioning methods. But the results reported here indicate that the *shapes* of the PDFT fragments are special. They are more transferable than the shapes obtained via other popular methods, either overlapping or nonoverlapping. It will be very interesting to see if the same conclusion holds for real molecules, so we are working on this. If the answer is positive (and we do not see why the picture should change qualitatively for real diatomics), then PDFT fragment densities have a bright future. On the one hand, they will be preferred as starting densities for improving convergence of fragment-based computational methods.¹¹ On the other hand, they will find good use in many branches of computational chemistry, from QM/MM applications to novel linear-scaling algorithms.

One of the slow steps of PDFT calculations as currently implemented is the simultaneous calculation of the occupa-

tion numbers $\{\nu_\alpha\}$ that minimize the partition energy, eq 2. These are of course an important output of a PDFT calculation. But when the main interest is not to compute formal charges, but to calculate molecular energies efficiently, an interesting possibility suggested here is that the approximate shape of the PDFT fragment densities can be found first for a fixed set of $\{\nu_\alpha\}$ (e.g., by imposing integer occupations), and then, simpler population analysis methods can provide occupation numbers to be used in conjunction with PDFT to yield approximate molecular energies.

Acknowledgment. Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund for support of this research under Grant No. 49599-DNI6. Y.Z. would like to thank Jing Zhu and Qi Wei for discussions on *Mathematica* programming.

Appendix

All calculations were accomplished with the *Mathematica* program.²⁰ To guarantee high precisions, analytical expressions or symbolic operations were involved when available.

The model system consists of two non-interacting electrons under the potential

$$v(x) = -Z_A\delta(x + R/2) - Z_B\delta(x - R/2) \quad (16)$$

where $Z_A > Z_B > 0$. The analytical solution of the 1-d Schrödinger equation with the potential above is shown in eq 9, where

$$\begin{aligned} D &= e^{\kappa R/2}(1 - Z_A/\kappa)C \\ F &= e^{-\kappa R/2}(Z_A/\kappa)C \\ G &= e^{\kappa R}\left(\frac{\kappa - Z_A}{Z_B}\right)C \\ C &= \sqrt{2\kappa}\left\{\frac{Z_A(\kappa - Z_A)}{Z_B(\kappa - Z_B)}\left(1 + \frac{Z_B^2}{\kappa^2}\right) - \frac{e^{-2\kappa R}Z_A^2}{\kappa^2} + \right. \\ &\quad \left. \frac{2Z_A}{\kappa}[1 + 2R(\kappa - Z_A)]\right\}^{-1/2} \end{aligned} \quad (17)$$

For convenience, an auxiliary polar-angle function $\beta(x)$ is introduced.¹³ In eq 15, β_A is the solution of the trigonometric equation

$$K_1 \cos(2\beta_A) - K_2 \cos[2(\beta_A + K_1 K_3 \cos(2\beta_A))] = 0 \quad (18)$$

and

$$\beta_B = \beta_A + K_1 K_3 \cos(2\beta_A) \quad (19)$$

where

$$\begin{aligned} K_1 &= -Z_A C^2 \\ K_2 &= -Z_B G^2 \\ K_3 &= \frac{1}{2\kappa D} \left(\frac{e^{\kappa R/2}}{C} - \frac{e^{-\kappa R/2}}{G} \right) \end{aligned} \quad (20)$$

and

$$\alpha_1 = K_1 \frac{\cos(2\beta_A)}{2\kappa D} \quad (21)$$

$$\alpha_2 = \frac{e^{\kappa R/2}}{C}$$

So the charge on atom B is

$$\nu = 1 - \int_{-\infty}^{\infty} |\psi(x)|^2 \sin[2\beta(x)] dx \quad (22)$$

where $\psi(x)$ is defined in eq 9. For all the details of the derivation of the equations above, and for explicit expressions for the partition potential $v_p(x)$, please see the Appendix of ref 13.

References

- (1) Elliott, P.; Burke, K.; Cohen, M. H.; Wasserman, A. *Phys. Rev. A* **2010**, *82*, 024501.
- (2) Elliott, P.; Cohen, M. H.; Wasserman, A.; Burke, K. *J. Chem. Theory Comput.* **2009**, *5*, 827–833.
- (3) Cohen, M. H.; Wasserman, A. *J. Phys. Chem. A* **2007**, *111*, 2229–2242.
- (4) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (5) (a) Löwdin, P.-O. *J. Chem. Phys.* **1953**, *21*, 374–375. (b) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840. (c) Politzer, P.; Harris, R. R. *J. Am. Chem. Soc.* **1970**, *92*, 6451–6454. (d) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1983**, *78*, 4066–4073. (e) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.
- (6) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129–138.
- (7) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: New York, 1990.
- (8) Rousseau, B.; Peeters, A.; van Alsenoy, C. *J. Mol. Struct.: THEOCHEM* **2001**, *538*, 235–238.
- (9) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1874.
- (10) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules; International Series of Monographs on Chemistry*; Oxford University Press, New York, 1989; pp 221–224.
- (11) (a) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050–8053. (b) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165–9169. (c) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701–706. (d) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791–11800. (e) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039–12041. (f) Li, S. H.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215–7226. (g) Huang, L. L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808–817. (h) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102. (i) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777–8785. (j) Řezáč, J.; Salahub, D. R. *J. Chem. Theory Comput.* **2010**, *6*, 91–99.
- (12) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.
- (13) Cohen, M. H.; Wasserman, A.; Car, R.; Burke, K. *J. Phys. Chem. A* **2009**, *113*, 2183–2192.
- (14) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (15) Cohen, M. H.; Wasserman, A. *J. Stat. Phys.* **2006**, *125*, 1125–1139.
- (16) Ayers, P. W. *J. Chem. Phys.* **2000**, *113*, 10886–10898.
- (17) Parr, R. G.; Bartolotti, L. J. *J. Phys. Chem.* **1983**, *87*, 2810–2815.
- (18) Ayers, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1959–1964.
- (19) Voronoi, G. *J. Reine Angew. Math.* **1907**, *133*, 97–178.
- (20) Wolfram Research, Inc., *Mathematica*, Version 6.0, Champaign, IL (2007).

CT100247Q

Investigation of Self-Interaction Corrections for an Exactly Solvable Model System: Orbital Dependence and Electron Localization

Daniel Vieira and K. Capelle*

*Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, Santo André,
09210-170 São Paulo, Brazil*

Received June 25, 2010

Abstract: A systematic investigation of two approximate self-interaction corrections (SICs), Perdew–Zunger SIC and Lundin–Eriksson SIC, and the local-density approximation (LDA) is performed for a model Hamiltonian whose exact many-body solution and exact LDA are known. Both SICs as well as LDA are applied in the calculation of ground-state energies, ground-state densities, energy gaps, and impurity densities of one-dimensional Hubbard chains differing in size, particle number, and interaction strength. The orbital-dependent potentials arising from either SIC are treated within the optimized-effective potential method, which we reformulate for the Hubbard model. The delocalization tendency of LDA is confronted with the localization tendency of SIC. A statistical analysis of the resulting data set sheds light on the role of SIC for weakly and strongly interacting particles and allows one to assess the performance of each methodology.

1. Introduction

Density-functional theory (DFT)^{1–3} is advancing at a rapid pace, driven by the demand of ever more accurate electronic-structure calculations. To meet this demand, ever more sophisticated density functionals are being constructed. A simultaneous, and often contradictory, demand is the applicability to larger and more complex systems, met by computational and methodological advances in the implementation of density functionals. In this context, the test, comparison, and implementation of different approximate density functionals in well-controlled environments is of crucial importance. This Article provides such a comparison for the case of orbital-dependent self-interaction corrections (SICs), using as theoretical laboratory an exactly solvable model system whose exact many-body ground-state energy and density are known, the one-dimensional Hubbard model.^{4,5} This model is known to display a crossover from weak to strong interactions, accompanied by increasing electron localization.^{6,7} We here specifically investigate the role self-interaction corrections play on each side of this crossover.

This Article is organized as follows: In section 2.1, we describe two different self-interaction corrections, Perdew–Zunger SIC and Lundin–Eriksson SIC. In section 2.2, we introduce the one-dimensional Hubbard model, and in section 2.3 we discuss the implementation of the orbital-dependent self-interaction corrected functionals by means of the optimized-effective potential (OEP) method for the Hubbard model. In section 3, we report results from a statistical analysis of hundreds of calculations of ground-state energies, ground-state densities, energy gaps, and impurity-site densities, always in comparison with the exact results for the many-body Hamiltonian. Section 4 contains our conclusions.

2. Background

2.1. Self-Interaction Error and Self-Interaction Corrections. One electron does not interact with itself. Simple as it is, this fact is at the heart of much trouble in many-body and electronic-structure theory. Traditional density functionals, such as the local density approximation (LDA), local spin-density approximation (LSDA), and generalized-gradient approximations (GGAs), as well as many more recently developed functionals, such as hybrids and meta GGAs, do have a spurious self-interaction error

* Corresponding author e-mail: klaus.capelle@ufabc.edu.br.

(SIE), and as a consequence predict a nonzero interaction energy even for a single electron.⁸ Consequences of the SIE of approximate functionals are multifarious and include wrong asymptotics of approximate exchange-correlation (XC) potentials, energy gaps, ionization energies, electron affinities, and transition-metal magnetic moments.

A very successful proposal for a self-interaction correction, made in 1981 by Perdew and Zunger (PZ),⁹ is

$$E_{xc}^{PZSIC}[n_{\uparrow}, n_{\downarrow}] = E_{xc}^{\text{approx}}[n_{\uparrow}, n_{\downarrow}] - \sum_k^{\text{occ}} \sum_{\sigma=\uparrow, \downarrow} (E_H[n_{k\sigma}] + E_{xc}^{\text{approx}}[n_{k\sigma}, 0]) \quad (1)$$

where $n_{\sigma}(\mathbf{r}) = \sum_k^{\text{occ}} n_{k\sigma}(\mathbf{r})$, $n_{k\sigma}(\mathbf{r}) = |\varphi_{k\sigma}(\mathbf{r})|^2$, and $\varphi_{k\sigma}(\mathbf{r})$ are single-particle orbitals. In PZSIC, the SIE is subtracted on a orbital by orbital basis. Evidently, for a one-electron density $E_{xc}^{PZSIC}[n^{(1)}] = -E_H[n^{(1)}]$, so that the one-electron SIE (1-SIE) is completely removed from the energy functional.

The PZSIC functional depends explicitly on the partial (orbital) densities $n_{k\sigma}(\mathbf{r})$ and, through these, on the orbitals $\varphi_{k\sigma}(\mathbf{r})$. Therefore, its minimization with respect to $n_{\sigma}(\mathbf{r})$ must employ the optimized-effective potential (OEP) method or one of its simplifications.^{10–14} In practice, however, the functional is usually minimized only with respect to the orbital densities $n_{k\sigma}(\mathbf{r})$, because the derivative of the PZ energy functional with respect to $n_{k\sigma}(\mathbf{r})$ can be calculated directly. The resulting orbital-specific (i.e., k -dependent) effective single-particle potential is

$$v_{s,k\sigma}^{PZSIC}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{xc,\sigma}^{\text{approx}}[n_{\sigma}, n_{\bar{\sigma}}](\mathbf{r}) - v_H[n_{k\sigma}](\mathbf{r}) - v_{xc,\sigma}^{\text{approx}}[n_{k\sigma}, 0](\mathbf{r}) \quad (2)$$

$$=: v_{\text{ext}}(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{xc,k\sigma}^{PZSIC}[n_{\sigma}, n_{\bar{\sigma}}](\mathbf{r}) \quad (3)$$

We note that this potential depends on the orbital it acts on, but is constructed from the density of all orbitals. The PZSIC approach, simplified by minimizing with respect to orbital densities instead of the spin densities, has been very successful in removing the one-electron SIE in solids.¹⁵ Results from proper minimization with respect to the densities, by means of the OEP, have also been reported for atomic^{16,17} and molecular systems.^{18–22} PZSIC has become so popular that frequently the abbreviation SIC is used as synonymous with PZSIC, but it has been repeatedly noted that the PZ proposal is not the only possibility for removing the SIE.^{23–27}

An innovative proposal for an alternative self-interaction correction was put forward in 2001 by Lundin and Eriksson (LE).²⁷ The LE proposal attempts to construct an effective potential that, when acting on the orbital of one electron, is constructed only from the density of the other orbitals. In practice, this is done by introducing an orbital-specific effective potential that is determined by subtracting from the full density $n(\mathbf{r})$ the partial density of the orbital the potential acts on:^{27–29}

$$v_{s,k\sigma}^{\text{LESIC}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_H[n - n_{k\sigma}](\mathbf{r}) + v_{xc}^{\text{approx}}[n_{\sigma} - n_{k\sigma}, n_{\bar{\sigma}}](\mathbf{r}) \quad (4)$$

$$= v_{\text{ext}}(\mathbf{r}) + v_H[n](\mathbf{r}) - v_H[n_{k\sigma}](\mathbf{r}) + v_{xc}^{\text{approx}}[n_{\sigma} - n_{k\sigma}, n_{\bar{\sigma}}](\mathbf{r}) \quad (5)$$

$$=: v_{\text{ext}}(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{xc,k\sigma}^{\text{LESIC}}[n_{\sigma}, n_{\bar{\sigma}}](\mathbf{r}) \quad (6)$$

where the step from eq 4 to eq 5 exploits the linearity of the Hartree potential, and eq 6 is the definition of $v_{xc,k\sigma}^{\text{LESIC}}$. In this way, the approximate effective potential acting on a given orbital, $v_{s,k\sigma}^{\text{LESIC}}(\mathbf{r})$, is constructed only from the density arising from the other orbitals. Clearly, for a one-particle system, this approach also correctly zeroes the one-electron SIE (1-SIE) contribution to the effective potential.

The LE proposal for a corrected effective potential is accompanied by a similar expression for the exchange-correlation energy:^{27,30}

$$E_{xc}^{\text{LESIC}}[n_{\uparrow}, n_{\downarrow}] = - \sum_k^{\text{occ}} \sum_{\sigma=\uparrow, \downarrow} E_H[n_{k\sigma}] + \sum_k^{\text{occ}} \sum_{\sigma=\uparrow, \downarrow} \int d\mathbf{r}^3 n_{k\sigma}(\mathbf{r}) e_{xc}^{\text{approx}}[n_{\sigma} - n_{k\sigma}, n_{\bar{\sigma}}](\mathbf{r}) \quad (7)$$

We note in passing that the change from LESIC to PZSIC can be affected by substituting

$$e_{xc}^{\text{approx}}[n_{\sigma} - n_{k\sigma}, n_{\bar{\sigma}}] \rightarrow e_{xc}^{\text{approx}}[n_{\sigma}, n_{\bar{\sigma}}] - e_{xc}^{\text{approx}}[n_{k\sigma}, 0] \quad (8)$$

and

$$v_{xc}^{\text{approx}}[n_{\sigma} - n_{k\sigma}, n_{\bar{\sigma}}] \rightarrow v_{xc,\sigma}^{\text{approx}}[n_{\sigma}, n_{\bar{\sigma}}] - v_{xc,\sigma}^{\text{approx}}[n_{k\sigma}, 0] \quad (9)$$

in the XC energy density and potential of former. For the Hartree potential, the substitution $v_H[n - n_{k\sigma}] \rightarrow v_H[n] - v_H[n_{k\sigma}]$ is an identity.

In the original LE work,²⁷ it was argued that the LE approach should be superior to the PZ approach because it removes the SIE of the potential, which, according to the original work, remains in the PZ approach. It has been objected³¹ that the LE proposal cannot be right because it would even correct the hypothetical exact functional, for which the PZ approach correctly reduces to zero. While we agree with this objection as a matter of principle, we feel that in practice the key issue is not only what the correction does to the hypothetical exact functional, but also what it does to the actually available approximate functionals. If such functionals were consistently improved by a correction, few workers would refrain from using this correction in practice, only because it overcorrects the hypothetical exact functional.

Interestingly, there are hints in the literature that approximate XC functionals do indeed benefit more from LESIC than from PZSIC. Novak et al.²⁸ compare the performance of LSDA, LSDA+PZSIC, and LSDA+LESIC in the calculation of hyperfine parameters and find that LESIC significantly improves agreement with experiment, relative to PZSIC. (Terra et al.²⁹ also reported a successful application of LESIC to hyperfine parameters.) In a separate

study, Friis et al.³² compared density distributions predicted for the Mg crystal with experimental electron diffraction data and observed that LESIC produces better core and valence densities than PZSIC.

The available data are too limited, however, to already conclude that LESIC is definitely superior, in practice, to PZSIC, in particular because, as a matter of principle, LESIC cannot be correct. It therefore becomes an important task to systematically investigate the performance of both approaches. This Article is a first step toward this task, providing a systematic analysis of the performance of PZSIC, LESIC, and LDA in a well-controlled model system whose exact solution is known, the one-dimensional Hubbard model.

2.2. The Hubbard Model. The Hubbard model^{4,5} (HM) is one of the most important models in quantum mechanics. In one dimension, and in second-quantized notation, the HM is defined as

$$\hat{H} = -t \sum_{i,\sigma}^L (c_{i\sigma}^\dagger c_{i+1,\sigma} + \text{H.c.}) + U \sum_i^L c_{i\uparrow}^\dagger c_{i\downarrow}^\dagger c_{i\downarrow} c_{i\uparrow} + \sum_{i,\sigma}^L V_i c_{i\sigma}^\dagger c_{i,\sigma} \quad (10)$$

where L is the number of sites, t represents the amplitude for hopping between neighboring sites, U is the local (on-site) interaction, and V_i is a local external potential acting on site i . Frequently, V_i is taken to be zero, but here we include this term explicitly to describe spatial inhomogeneities, such as impurities. Occupation of each site is limited to two particles, necessarily of opposite spin. Like-spin particles do not interact in the most common form of the model. The HM is a minimal model accounting for the competition between itineracy and localization of electrons. While it is widely used in solid-state physics to describe metallic and insulating phases of correlated solids, it may also be considered a special case of the PPP model occasionally used in quantum chemistry.³³

The basic Hohenberg–Kohn and Kohn–Sham theorems of DFT hold for the HM too, once the density $n(\mathbf{r})$ is replaced by the on-site occupation number.^{34,35}

$$n(\mathbf{r}) = \sum_{\sigma} n_{\sigma}(\mathbf{r}) \rightarrow n(i) = \sum_{\sigma} n_{\sigma}(i) = \sum_{\sigma} \langle c_{i\sigma}^\dagger c_{i\sigma} \rangle \quad (11)$$

In terms of this variable, local-density and spin-density approximations for Hubbard chains and rings have been constructed^{36–38} and can be employed in the usual way.

What makes this model attractive as a theoretical laboratory for DFT is that its exact many-body solution is known in the thermodynamic limit ($L \rightarrow \infty$, $n(i) = n = N/L = \text{const}$),³⁹ and an exact numerical solution is possible for small systems (up to $L \approx 14$ on a workstation, for arbitrary density distributions $n(i)$). As a consequence of the availability of the exact solution for the infinite homogeneous system, the exact LDA is known too.⁴⁰ Moreover, a complete basis consists of two spin–orbitals per site, so that we can always work at the basis-set limit. Another welcome feature, which we will exploit below, is that the model allows one to continuously vary the interaction strength both in the exact and in the DFT calculations, which simulates the behavior

of weakly and strongly interacting systems in ab initio calculations.

We stress that the Hubbard model, when used in this spirit, is a theoretical laboratory for many-body quantum mechanics, not a simplified description of a real system. In particular, the values of U and E and the corresponding wave functions need not be derived from the Wannier orbitals obtained from an ab initio calculation of a real system.

2.3. Optimized Effective Potential Method. Formally, the local orbital-independent XC potential corresponding to a given approximation to $E_{xc}[n_{\sigma}, n_{\bar{\sigma}}]$ is given by

$$v_{xc,\sigma}^{\text{approx}}[n_{\sigma}, n_{\bar{\sigma}}](i) = \frac{\delta E_{xc}^{\text{approx}}[n_{\sigma}, n_{\bar{\sigma}}]}{\delta n_{\sigma}(i)} \quad (12)$$

where the continuous spatial argument (\mathbf{r}) from section 2.1 has been replaced by the discrete site label i .

For orbital-dependent functionals $E_{xc}^{\text{orb}}[\{\varphi_{k\sigma|n}\}]$, a common multiplicative potential for all orbitals can be generated by means of the OEP method,¹⁰ whose equation for discrete Hubbard chains reads

$$v_{xc,\sigma}^{\text{OEP}}[n_{\sigma}, n_{\bar{\sigma}}](i) = \frac{1}{2n_{\sigma}(i)} \sum_k^{N_{\sigma}} \{n_{k\sigma}(i)[v_{xc,k\sigma}[n_{\sigma}, n_{\bar{\sigma}}](i) + \bar{v}_{xc,k\sigma}^{\text{OEP}}[n_{\sigma}, n_{\bar{\sigma}}] - \varphi_{k\sigma}(i)\hat{t}_s\Delta_{k\sigma}^*(i) + \Delta_{k\sigma}^*(i)\hat{t}_s\varphi_{k\sigma}(i)\} + \text{c.c.} \quad (13)$$

where

$$\Delta_{k\sigma}^*(i) = - \sum_{m(m \neq k)}^L \frac{\varphi_{m\sigma}^*(i)}{\varepsilon_{k\sigma} - \varepsilon_{m\sigma}} \sum_{j=1}^L \varphi_{k\sigma}^*(j)[v_{xc,k\sigma}[n_{\sigma}, n_{\bar{\sigma}}](j) - \bar{v}_{xc,\sigma}^{\text{OEP}}[n_{\sigma}, n_{\bar{\sigma}}](j)]\varphi_{m\sigma}(j) \quad (14)$$

and the barred quantities are defined as

$$\bar{v}_{xc,k\sigma}^{\text{OEP}}[n_{\sigma}, n_{\bar{\sigma}}] = \sum_{i=1}^L n_{k\sigma}(i)v_{xc,\sigma}^{\text{OEP}}[n_{\sigma}, n_{\bar{\sigma}}](i) \quad (15)$$

and

$$\bar{v}_{xc,k\sigma}[n_{\sigma}, n_{\bar{\sigma}}] = \sum_{i=1}^L n_{k\sigma}(i)v_{xc,k\sigma}[n_{\sigma}, n_{\bar{\sigma}}](i) \quad (16)$$

The terms \hat{t}_s , $\varepsilon_{k\sigma}$, and $\varphi_{k\sigma}(i)$ stand for the single-particle kinetic energy operator and the OEP-type Kohn–Sham (KS) eigenvalues and orbitals, respectively.

The preceding equation can be applied directly to the PZSIC functional to generate the corresponding KS potential. In the case of the LESIC functional, however, an additional ambiguity arises, because the LE energy expression (7) is constructed in analogy to eq 6, but its functional derivative with respect to either $n_{\sigma}(\mathbf{r})$ or $n_{k\sigma}(\mathbf{r})$ is not the LE XC potential in eq 6. Rather, the LE energy and potential are separate constructions.

Consequently, there are (at least) two possibilities to construct a local multiplicative Kohn–Sham potential for the LE approach. One is to apply the OEP to the orbital-dependent LE energy functional (7), in complete analogy to

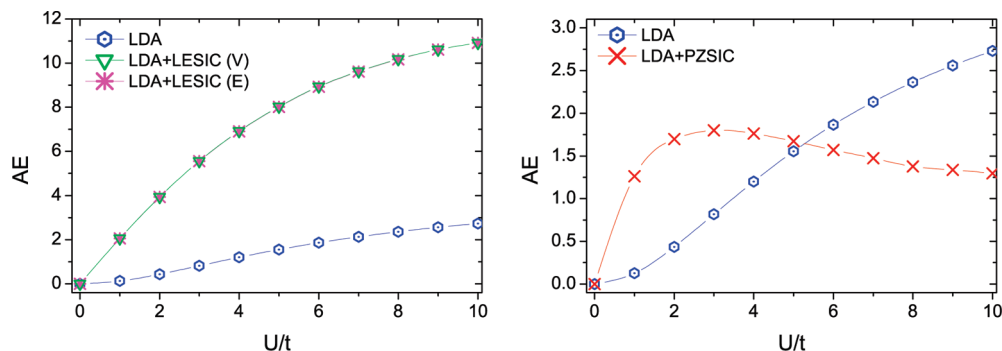


Figure 1. Average error in the ground-state energy, as defined in eq 17, for LDA and for LDA corrected by the Lundin–Eriksson (in its two flavors, V and E) and Perdew–Zunger self-interaction corrections, relative to the exact many-body ground-state energy. $N_S = 37$ systems are considered at each U/t .

what is done for the PZ energy functional (1). This requires differentiating the LE energy functional with respect to the partial densities to generate the corresponding orbital-dependent potential. We refer to this approach as LESIC(E) to emphasize that it makes use of the LE proposal for the energy and derives the XC potential from that.

Alternatively, we can formally introduce a second, hypothetical, LE energy functional, say $E_{xc}^{LESIC-h}[n_{\sigma}, n_{\bar{\sigma}}]$, such that its derivative with respect to the partial densities is the LE potential (6). In practice, this hypothetical energy functional need not be known explicitly, because all that is required to solve the KS equation is to substitute the orbital-specific potential (6) in the OEP equation to generate a common KS potential for all orbitals. This is the prescription followed, for example, in the study of discontinuities of the time-dependent XC potential via SIC calculations.⁴¹ Once the KS orbitals and densities have been obtained, we then follow the original LE proposal and calculate the energy correction from eq 7. This implementation is referred to as LESIC(V), to emphasize that it makes use of the LE proposal directly for the potential.

Either procedure formally eliminates the key feature of the LE correction that the potential acting on a given orbital is constructed only from the density of the other orbitals, because by construction the OEP delivers a common multiplicative potential for all orbitals. However, the feature is implicitly preserved by the OEP, because the resulting Kohn–Sham potential self-consistently produces the orbitals and densities that minimize the LE energy E_{xc}^{LESIC} or $E_{xc}^{LESIC-h}$, respectively.

3. Results

In this section, we compare the two self-interaction corrections LDA+PZSIC and LDA+LESIC (in its two flavors E and V), and the uncorrected LDA functional, using the OEP method of eq 13 to deal with the orbital-specific PZ and LE potentials. All calculations are done for finite chains of different lengths L , particle numbers N , and interaction strengths U . Variation of these parameters for each combination of methodology produces thousands of data, which we analyze statistically by reporting average errors, specifically defined for each observable. These observables are the ground-state energies (section 3.1), ground-state densities

Table 1. Average Error in the Ground-State Energy, As Defined in Eq 17, Separately for Weakly and for Strongly Interacting Systems (Columns 1 and 2) and for All Investigated Systems (Column 3)^a

N_S :	$U/t = 1...5$	$U/t = 6...10$	$U/t = 1...10$
LDA	0.8264	2.3301	1.5782
LDA+LESIC(V)	5.3011	10.0576	7.6794
LDA+LESIC(E)	5.2954	10.0508	7.6731
LDA+PZSIC	1.6388	1.4104	1.5246

^a Entries in boldface are the best in each column (interaction regime).

(section 3.2), energy gaps (section 3.3), and impurity-site densities (section 4).

3.1. Ground-State Energies. We start by evaluating the total ground-state energy. For each data set (see below), the accumulated deviation between approximate and exact results is measured by means of the percentage average error defined as

$$AE = 100 \frac{\sum_{j=1}^{N_S} |E_{\text{approx}}^j - E_{\text{exact}}^j|}{\sum_{j=1}^{N_S} |E_{\text{exact}}^j|} \quad (17)$$

where E_{exact}^j refers to numerically exact many-body energies for the same set of model parameters. N_S is the number of systems in the data set.

Our results are summarized in Figure 1 and Table 1. The figure compares the average error defined in eq 17 as a function of interaction strength U , ranging from the noninteracting system $U = 0$, over weakly $U \approx 2t$, moderately $U \approx 6t$, to strongly $U \approx 10t$ interacting systems. For each value of U , our sample consists of $N_S = 37$ chains of different sizes, particle number, and densities ($L = 3, 4, 5, \dots, 14$ and $N = 2, 4, 6, \dots, 10$, with $N/L < 0.85$). All systems have $V_i = 0$ (no impurities) but are spatially inhomogeneous due to the density modulations arising due to the boundaries.

First, comparing both SICs, we find that LDA+PZSIC consistently performs better than LDA+LESIC(V) and LDA+LESIC(E). To permit a better visualization, the scale on the vertical axis of the two figures is not the same, but the values of the average error are directly comparable and

leave little doubt about the superiority of PZSIC energies for these systems.

Next, we note that for weakly interacting systems (where the physics is dominated by the kinetic energy term, and the particles are delocalized), LDA alone is actually better than LDA+PZSIC and LDA+LESIC. Of course, both SICs remove the self-interaction error of the LDA also in this regime, but they are not exact themselves, and thus introduce other errors. For weak interactions, it does not pay to correct a small error by a correction introducing larger ones. For stronger interactions, the particles start to localize, the SIE becomes more important, and its removal by PZSIC improves the total energies considerably. The crossover between the error curves occurs around $U \approx 5t$. The separation in these two regimes is a rather gratifying result to obtain, as $U \approx 5t$ is known to mark the region where the Hubbard model crosses over from weakly to strongly interacting systems, and the electrons become increasingly localized.^{6,7} The performance of PZSIC is clearly related to this crossover, which is a signature of strong correlation. In fact, in retrospect, one could have predicted that the electrons in the Hubbard model start to localize around $U \approx 5t$ simply by comparing the relative performance of LDA and LDA+PZSIC as a function of U .

In this context, numerical evidence of electronic localization is obtained by calculation of the double occupation (D) factor, which, for any system j , is defined as the expectation value of the interaction energy divided by U :

$$D_j = \sum_{i=1}^{L_j} \langle c_{it}^\dagger c_{it} c_{it}^\dagger c_{it} \rangle = \frac{\partial E_{\text{approx}}^j}{\partial U} \quad (18)$$

where the second equality arises from the Hellman–Feynman theorem. In general, D_j measures the probability of a site to be doubly occupied as a function of U . For up to half-filled bands and any repulsive U , a ground-state $D > 0$ indicates that the particles are spread over various sites. Under these circumstances, D approaches zero as $U/t \rightarrow \infty$, indicating maximum localization. Considering our previous sample of $N_s = 37$ systems, we evaluate the average value of D_j :

$$AD = \frac{\sum_{j=1}^{N_s} D_j}{\sum_{j=1}^{N_s} 1} \quad (19)$$

Results are summarized in Figure 2.

First, the comparison of both SICs indicates that LESIC (in its two flavors) tends to overlocalize the electrons, reducing the average value of D . This trend to localize strongly agrees with similar observations in the original LE work.²⁷ The LDA functional, on the other hand, tends to enforce double occupation and delocalize the electrons, resulting in more accurate results for weak interactions. By contrast, the LDA+PZSIC approach overlocalizes particles up to $U \approx 3t$. For stronger values of U , the LDA+PZSIC functional crosses the exact curve and becomes superior to LDA, yielding localized states whose double occupation

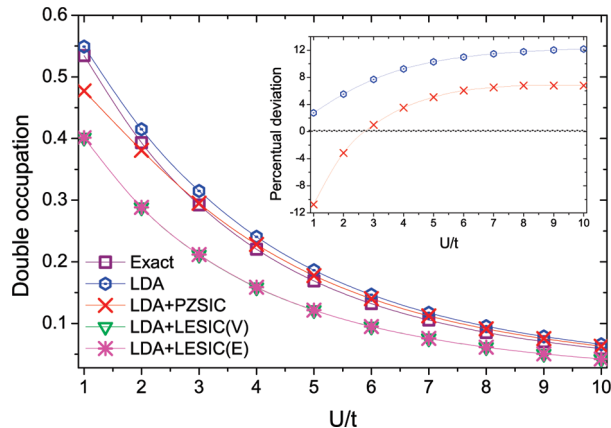


Figure 2. Average double occupation, as defined in eq 19, for exact diagonalization, LDA, and for LDA corrected by LESIC and PZSIC. $N_s = 37$ systems are considered at each U/t . Inset: Average percentual deviations with respect to the exact data.

curve is closer to the exact one. These conclusions are reinforced by the inset, which compares the percentual deviations of LDA and LDA+PZSIC. We note that localization is mainly driven by the onsite interaction, with electrons behaving like noninteracting spinless fermions as $U/t \rightarrow \infty$. The transition from LDA to LDA+PZSIC further localizes the electrons, reducing the spurious self-interaction, which favors delocalization.

To make this division in two distinct regimes clearer, we report in Table 1 the average error for each regime, that is, sum not only over systems with different size and particle number, but also over those with interaction strengths in the indicated range. The lowest average error in each regime is printed in boldface. Clearly, for weakly interacting systems, LDA does best. For strongly interacting systems, on the other hand, LDA+PZSIC wins. Overall, that is, considering the average error over all considered systems regardless of the value of U , LDA+PZSIC is the best combination among all methodologies tested here. Turning to a comparison of the different flavors of LESIC among each other, we note that the LDA+LESIC(E) and LDA+LESIC(V) approaches, despite their different starting points, give very similar results.

3.2. Ground-State Densities. Next, we turn to ground-state densities. To quantify the average error of these quantities, we sum for each system (labeled by j) and for all sites (labeled by i) the local absolute deviation between exact and approximate densities and express the resulting error as

$$AEn = 100 \frac{\sum_{j=1}^{N_s} \sum_{i=1}^{L_j} |n_{\text{approx}}^j(i) - n_{\text{exact}}^j(i)|}{\sum_{j=1}^{N_s} \sum_{i=1}^{L_j} |n_{\text{exact}}^j(i)|} \quad (20)$$

where L_j labels the total number of sites i composing each system j . Considering the same set of systems that we have considered for the total ground-state energies, the results for the average error in the ground-state densities are summarized in Figure 3 and in Table 2.

From Figure 3a we first note that uncorrected LDA performs better than any of the SICs and that LDA+PZSIC

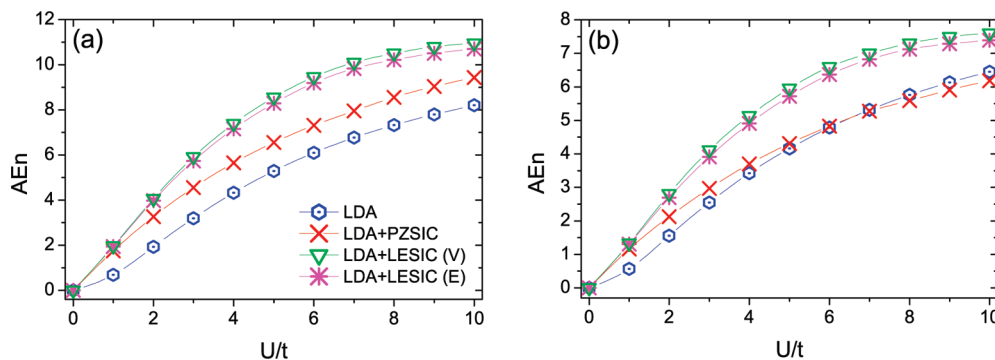


Figure 3. Average error in the ground-state density, as defined in eq 20, for LDA and for LDA corrected by LESIC and PZSIC, relative to the exact many-body ground-state density. (a) $N_s = 37$ systems are considered at each U/t ; (b) $N_s = 17$ systems are considered at each U/t .

Table 2. Average Error in the Ground-State Density, As Defined in Eq 20, Separately for Weakly and for Strongly Interacting Systems (Columns 1 and 2) and Jointly for All Investigated Systems (Column 3)^a

	$U/t = 1...5$	$U/t = 6...10$	$U/t = 1...10$
N_s :	185	259	444
LDA	3.0830	7.2373	5.1602
LDA+LESIC(V)	5.5745	10.36087	7.9676
LDA+LESIC(E)	5.4191	10.0873	7.7532
LDA+PZSIC	4.3526	8.4551	6.40381
N_s :	80	112	192
LDA	2.4536	5.6875	4.0706
LDA+LESIC(V)	3.8683	7.1997	5.5340
LDA+LESIC(E)	3.7070	6.9978	5.3524
LDA+PZSIC	2.8523	5.5571	4.2047

^a Entries in boldface are the best in each column. First group of data: complete statistics, including all sites. Second group of data: reduced statistics after exclusion of the border and central sites.

is better than LDA+LESIC(E) and LDA+LESIC(V). The relation between the various curves is rather different from what it was for the energies; in particular, there is no crossover between LDA and LDA+PZSIC.

Upon closer investigation of the exact and approximate density profiles, one finds that the LDA densities are closer to the exact ones than the SIC ones mostly at the borders of the chains, where the densities are reduced due to the presence of the boundary, and consequently the SIE is reduced too. At the central site (of systems with odd number of sites L , where a central site is well-defined), the SIC density is also substantially worse than the LDA one. Therefore, we additionally calculated the average errors after eliminating the boundary sites, as well as the central one. The result is shown in Figure 3b and clearly displays a crossover between LDA and LDA+PZSIC for sufficiently large values of U , just as was found for the total energies. We conclude that PZSIC densities are worse at the surface of the systems and at their geometric center, but that this local behavior has little effect on the global energies, which are dominated by the other sites.

For LDA+LESIC(V) and LDA+LESIC(E), on the other hand, even exclusion of these critical sites has no beneficial effect, and both continue to perform worse than uncorrected LDA for all U . We believe that this is an intrinsic problem of the LESIC (not due to the OEP-like implementation), because it is reproduced also in other implementations: in

unpublished calculations, we implemented LESIC via the GAM, Slater, and KLI approximations to the OEP equation,¹⁰ as well as through globally and locally scaled self-consistency.⁴² The behavior of LESIC is consistently worse than that of PZSIC, in any of these different modes of implementation. The same applies even for post-LDA calculations of total energies. The explanation is probably that LESIC overcorrects the LDA functional in the sense that it would even “correct” the exact functional.

Table 2 summarizes these discussions by displaying the average error separately for weak and strong interaction, as well as for all interactions, with and without elimination of boundary and central sites.

3.3. Energy Gaps. Besides the ground-state energy itself, the calculation of energy differences is also of interest because there is no guarantee that the trends we identified for energies will survive cancellation upon forming the difference. For this reason, we also evaluate the fundamental energy gap, defined as

$$E_g = E(N-1, L) - E(N, L) - [E(N, L) - E(N+1, L)] \quad (21)$$

where $E(N, L)$ labels the ground-state energy of a N electron system with L sites. Our data set comprises systems with $L = 4, 6, 7, \dots, 14$ sites, which for $N = 2, 4, 6, 8, 10$ (with $N/L < 0.85$) displays a band gap due to the system boundary, and for $U > 0$ and $N/L = 1.0$ with periodic boundary conditions (so that $n = 1$ on all sites) features a Mott gap. Previous work³⁷ already demonstrated that the Bethe–Ansatz LDA functional we use for the Hubbard model properly accounts for the Mott gap in homogeneous systems and provides a reasonable approximation to it in inhomogeneous systems. Here, we inquire whether the description of inhomogeneous systems is further improved by self-interaction corrections.

Similarly to the previous cases, we measure the average error in the gap by

$$AE_g = 100 \frac{\sum_{j=1}^{N_s} |E_{g, \text{approx}}^j - E_{g, \text{exact}}^j|}{\sum_{j=1}^{N_s} |E_{g, \text{exact}}^j|} \quad (22)$$

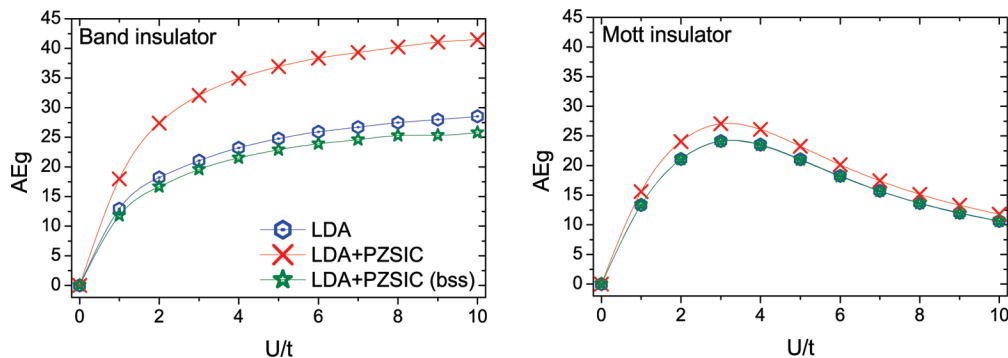


Figure 4. Average error in the energy gap, as defined in eq 22, for LDA and for LDA corrected by PZSIC (via post-LDA implementation), relative to the exact fundamental energy gap. The starred (green) data refer to calculations with broken spin symmetry (bss).

for the approximate and exact energy gap E_g^j of each system j . Results are shown in Figure 4. In this section, we employ post-LDA implementations of the SIC functionals instead of following the OEP procedure, which converges very badly when $N/L = 1.0$.

For both the band gap and the Mott gap, LDA+PZSIC is vastly superior over LDA+LESIC for all values of U . For simplicity, we therefore do not show LESIC data in the figures. Interestingly, the PZSIC results, although better than the LESIC ones, perform consistently worse than uncorrected LDA for all U . Unlike for total energies, there is no crossover between LDA and LDA+PZSIC curves even for large values of U . Because LDA+PZSIC yields the more accurate individual total energies, this improved performance for energy differences must be due to error cancellation upon subtraction, as frequently occurs within LDA.

Interestingly, comparison of both panels of Figure 4 clearly shows that the advantage LDA has over LDA+PZSIC is

much smaller for Mott insulators than for band insulators. This behavior is consistent with our previous findings, as in the Mott insulator the particles are more localized, LDA energies become worse, SIC energies become better, and the advantaged LDA derives from error cancellation is reduced.

In the LDA calculations, spin-symmetry is never broken. In the SIC calculations described so far, we explicitly enforced spin symmetry for systems with odd total particle number, by fractionally occupying the highest up and down orbitals. Alternatively, we can also permit the system to break spin symmetry, by placing the last electron completely in the highest orbital of one spin (say up). The resulting system displays an incorrect magnetization, but lower total energies and a much improved band and Mott gaps, as displayed in the starred (green) curves in Figure 4. For the band insulator, this improvement is enough to overcome the gain LDA obtained from error cancellation and transform PZSIC into the best performing functional, at the expense, however, of

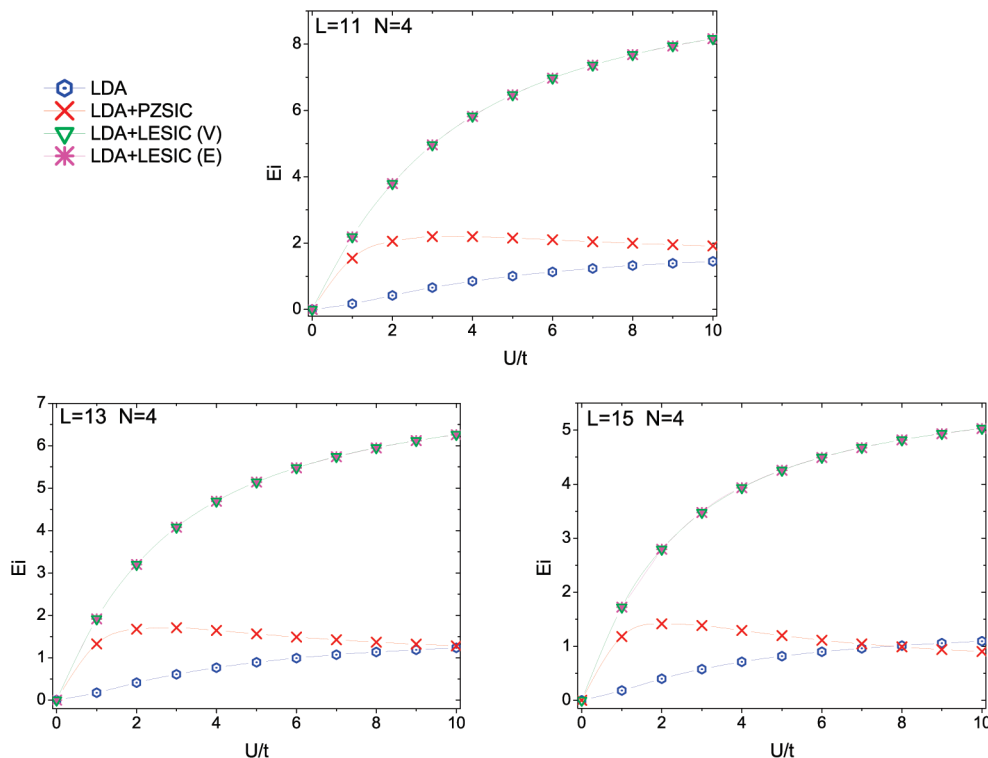


Figure 5. Error in the ground-state energy of the attractive impurity system, as defined in eq 23, relative to the exact many-body ground-state energy.

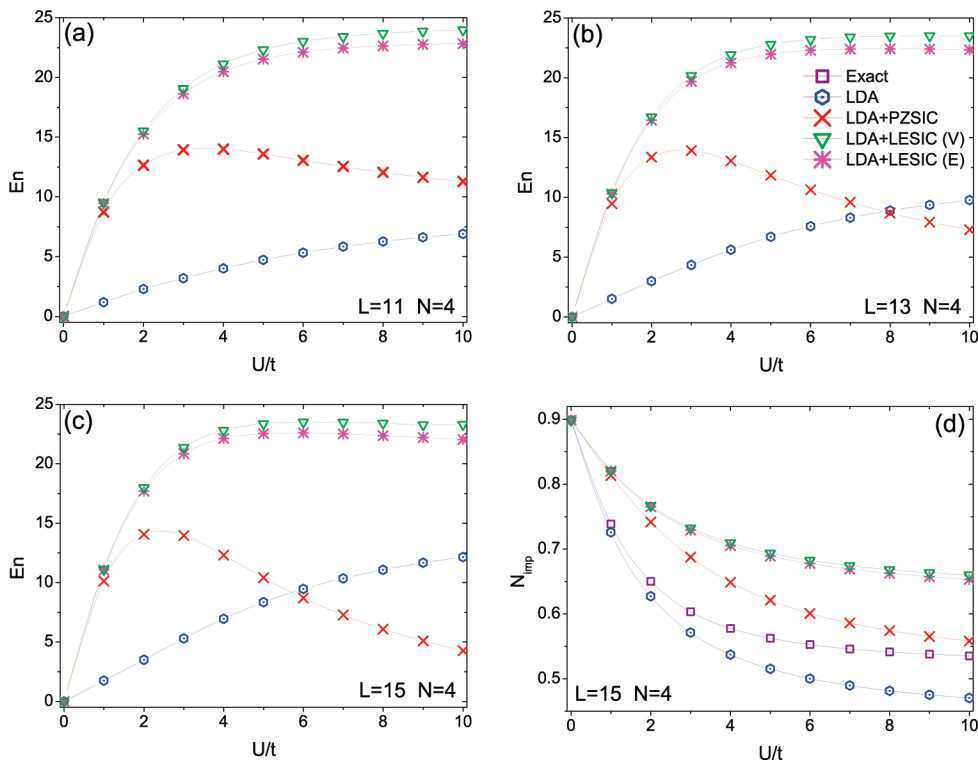


Figure 6. (a–c) Error in the ground-state density at the attractive impurity site, as defined in eq 24, relative to the exact many-body ground-state density. (d) Number of electrons N_{imp} at the impurity site.

wrong spin densities. This is the same tendency known from ab initio applications of PZSIC.⁹

3.4. Ground-State Energies and Densities in the Presence of a Local Impurity. So far, we have investigated chains without any additional external potentials. An interesting complementary analysis consists of considering systems with local impurities. Because of the very large number of possible variations of system parameters (which now include location, concentration and strength of the impurities), we here do not perform a statistical analysis, but rather a case study, limiting us to consider a few typical situations. Nevertheless, we consider different average densities N/L because the value of U that characterizes the crossover from weak to strong interactions can slightly depend on the average density.

As a function of the interaction U , the results are shown in Figures 5–8, in which we plot the percentual absolute error with respect to the exact results, defined as

$$E_i = 100 \frac{|E_{\text{approx}} - E_{\text{exact}}|}{|E_{\text{exact}}|} \quad (23)$$

for the ground-state energy and

$$E_n = 100 \frac{|n_{\text{approx}}(\text{imp}) - n_{\text{exact}}(\text{imp})|}{|n_{\text{exact}}(\text{imp})|} \quad (24)$$

for the density at the impurity site. As our sample, we take $N=4$, with $L=11, 13$, and 15 . First, we place an attractive impurity of magnitude $V_{\text{imp}} = -1.0t$ at the central site. This potential favors electron localization at the impurity site and competes with the local repulsive interaction U . In a separate

set of calculations, we consider a repulsive impurity of magnitude $V_{\text{imp}} = 1.0t$, also at the central site. Contrary to the attractive case, the repulsive impurity favors delocalization of electrons. By comparing both cases, we can thus analyze and assess the effects of the LDA and SIC on the delocalization and delocalization of electrons, respectively.

For the attractive impurity, considering either energies or densities, we see from Figures 5 and 6 a crossover between LDA and LDA+PZSIC whose precise location depends on the average density N/L . Again, no crossover is seen for LDA+LESIC(V) and LDA+LESIC(E). For our sample of systems, we find that the smaller is the average density, the smaller becomes the interaction U for which LDA+PZSIC starts to produce better results than uncorrected LDA. This is in accordance with the expectation that strong correlations are not only characterized by large values of U but also by low densities, and also with similar trends observed in calculations not using DFT.⁷

In Figure 6d, where we plot the density at the impurity site, we note the same tendency of electronic localization already seen in Figure 2: both LESIC approaches tend to strongly overlocalize electrons, PZSIC somewhat overlocalizes, while LDA overly delocalizes.

In the repulsive impurity case, depicted in Figure 7, our previous conclusions for the ground-state energies still hold. There is a clear crossover between LDA and LDA+PZSIC, which depends on the average density. On the other hand, for the densities at the impurity sites we see from Figure 8 that the uncorrected LDA surpasses all alternative approaches, for all values of U and N/L . This is because the repulsive impurity tends to delocalize the electrons, thus

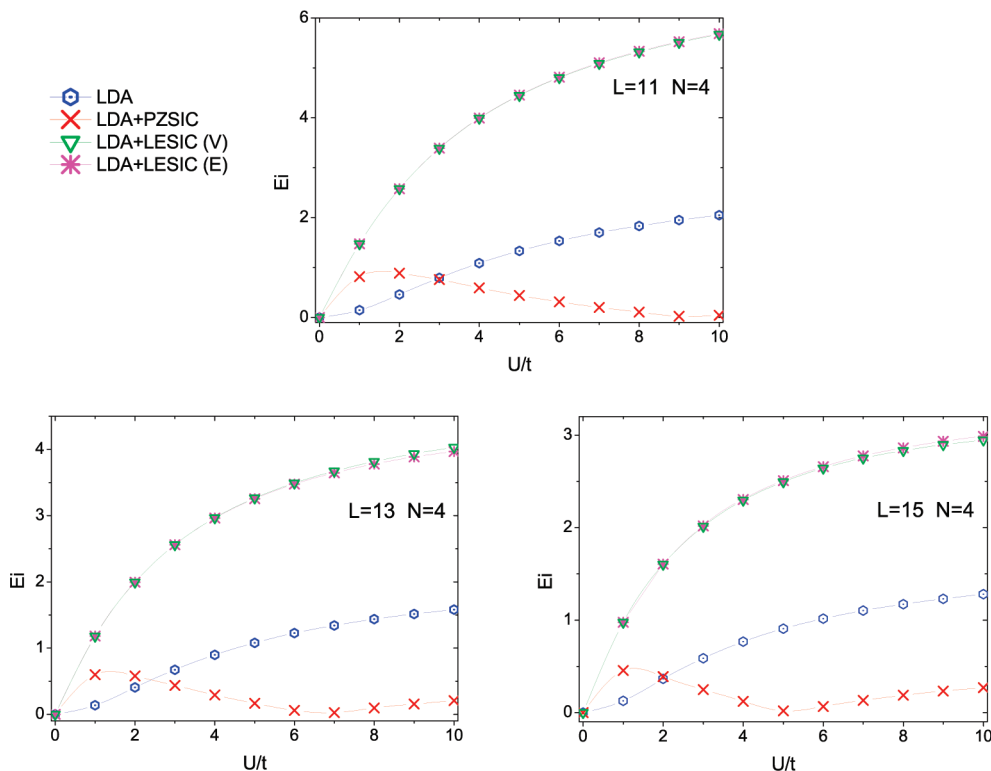


Figure 7. Error in the ground-state energy of the repulsive impurity system, as defined in eq 23, relative to the exact many-body ground-state energy.

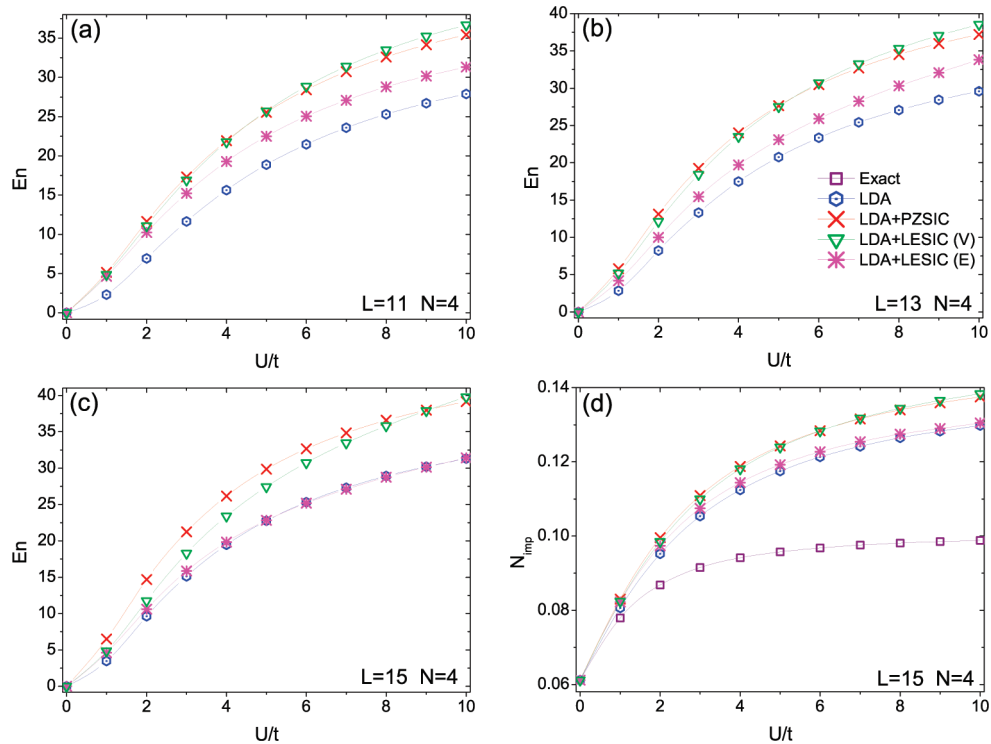


Figure 8. (a–c) Error in the ground-state density at the repulsive impurity site, as defined in eq 24, relative to the exact many-body ground-state density. (d) Number of electrons N_{imp} at the impurity site.

producing a situation in which the LDA is more reliable. Nevertheless, for ground-state energies, which are obtained from the densities of all sites of the Hubbard chain, LDA+PZSIC still delivers more accurate results in the strongly interacting regime.

4. Conclusions

We performed self-consistent calculations for the one-dimensional Hubbard model within the LDA and two orbital-dependent self-interaction corrections, in various implemen-

tations, and analyzed their performance by means of statistical comparison to the numerically exact many-body solution of the same model. Several trends emerge from this analysis:

(i) For the weakly interacting systems, the benefit of an approximate SIC is overcompensated by the intrinsic errors of the approximation. Uncorrected LDA here works best, for all investigated quantities and systems.

(ii) For more strongly interacting situations, correcting the SIE is important, but of the investigated corrections only PZSIC systematically improves on the LDA ground-state energies. Although it yields the most accurate total energies, PZSIC does not improve fundamental energy gaps, for which LDA benefits from substantial error cancellation, unless spin symmetry is allowed to break, in which case PZSIC becomes superior.

(iii) Despite their differences in the treatment of the potential, LESIC(E) and LESIC(V) produce quite similar results and are consistently worse than all other tested approaches, for energies and densities.

(iv) The crossover between the regimes where PZSIC worsens and improves on LDA is characterized by increasing localization. Its precise position as a function of interaction strength depends on the average density N/L . PZSIC turns out to be beneficial for lower densities and stronger interactions. The relative performance of LDA and LDA+PZSIC energies is found to be a clear indicator of the crossover between weak and strong correlations, and the concomitant localization, in the Hubbard model.

(v) Conclusions iii and iv also apply to ground-state densities, however, only if one ignores the densities right at the borders and at the central site of the systems, where PZSIC performs rather badly. If these sites are included, the accumulated (over all sites) error of the LDA densities is smaller than that of the LDA+PZSIC (and LDA+LESIC) densities for all investigated systems.

(vi) For impurity systems, the delocalization tendency of LDA sharply contrasts with the localization tendency of SIC. For attractive impurities, which favor electron localization, SIC performs better than LDA. On the other hand, for repulsive impurities, favoring electronic delocalization, LDA yields more accurate results than SIC. This shows that the performance of each functional is directly tied to the degree of localization, regardless of whether this localization is driven by particle–particle interactions or by external potentials.

Naturally, these conclusions may have been biased by the choice of the Hubbard model as theoretical laboratory. The disadvantages of such choice, on the other hand, are largely compensated by the availability of an exact many-body solution as a benchmark, of the exact LDA,⁴⁰ and by the absence of any problems due to finite basis sets or due to experimental uncertainties.

Other aspects of SIC, such as the asymptotic behavior of the potential of finite systems, cannot be addressed within the Hubbard model and require separate analysis. It remains to be seen if the superiority of PZSIC remains unchallenged by extension of these investigations to other corrections, implementations, and classes of systems, as well as in the study of systems of noninteger electron numbers. Similarly,

the important issue of the many-electron SIE^{43–46} requires separate investigation.

Acknowledgment. This work was supported by FAPESP and CNPq. K.C. thanks J. Terra and D. Guenzberger for bringing the LE approach to his attention. We thank V. L. Campo, Jr., for his exact diagonalization code for the Hubbard model.

References

- (1) Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253.
- (2) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory: An Approach to the Quantum Many-Body Problem*; Springer: Berlin, Germany, 1990.
- (3) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (4) Hubbard, J. *Proc. R. Soc. London, Ser. A* **1963**, *276*, 238.
- (5) Lieb, E. H.; Wu, F. Y. *Physica A* **2003**, *321*, 1.
- (6) White, S. R.; Affleck, I.; Scalapino, D. J. *Phys. Rev. B* **2002**, *65*, 165122.
- (7) Bedürftig, G.; Brendel, B.; Frahm, H.; Noack, R. M. *Phys. Rev. B* **1998**, *58*, 10225.
- (8) The exact exchange energy of one electron exactly cancels its Hartree energy, so that the Hartree + exchange energy of methods employing exact exchange is self-interaction free, but the correlation energy remains subject to self-interaction even in methods employing exact exchange.
- (9) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.
- (10) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3.
- (11) Yang, W.; Wu, Q. *Phys. Rev. Lett.* **2002**, *89*, 143002.
- (12) Kümmel, S.; Perdew, J. P. *Phys. Rev. B* **2003**, *68*, 035103.
- (13) Krieger, J. B.; Li, Y.; Iafate, G. J. *Phys. Rev. A* **1992**, *46*, 5453.
- (14) Ullrich, C. A.; Reinhard, P.-G.; Suraud, E. *Phys. Rev. A* **2000**, *62*, 053202.
- (15) Svane, A.; Petit, L.; Szotek, Z.; Temmerman, W. M. *Phys. Rev. B* **2007**, *76*, 115116. Lüders, M.; Ernst, A.; Dane, M.; Szotek, Z.; Svane, A.; Koderitzsch, D.; Hergert, W.; Gyroff, B. L.; Temmerman, W. M. *Phys. Rev. B* **2005**, *71*, 205109. Petit, L.; Svane, A.; Temmerman, W. M.; Szotek, Z. *Phys. Rev. B* **2001**, *63*, 165107. Strange, P.; Svane, A.; Temmerman, W. M.; Szotek, S.; Winter, H. *Nature* **1999**, *399*, 756.
- (16) Chen, J.; Krieger, J. B.; Li, Y.; Iafate, G. J. *Phys. Rev. A* **1996**, *54*, 3939.
- (17) Norman, M.; Koelling, D. D. *Phys. Rev. B* **1984**, *30*, 5530.
- (18) Körzdörfer, T.; Mundt, M.; Kümmel, S. *Phys. Rev. Lett.* **2008**, *100*, 133004.
- (19) Garza, J.; Nichols, J. A.; Dixon, D. A. *J. Chem. Phys.* **2000**, *112*, 7880.
- (20) Körzdörfer, T.; Kümmel, S.; Mundt, M. *J. Chem. Phys.* **2008**, *129*, 014110.
- (21) S. Patchkovskii, S.; Autschbach, J.; Ziegler, T. *J. Chem. Phys.* **2001**, *115*, 26.
- (22) Pemmaraju, C. D.; Sanvito, S.; Burke, K. *Phys. Rev. B* **2008**, *77*, 121204.
- (23) Fermi, E.; Amaldi, E. *Acad. Ital. Rome* **1934**, *6*, 117. Ayers, P. W.; Morrison, R. C.; Parr, R. G. *Mol. Phys.* **2005**, *103*, 2061.

- (24) Cortona, P. *Phys. Rev. A* **1986**, 34, 769.
- (25) Unger, H. J. *Phys. Lett. A* **2001**, 284, 124.
- (26) Vosko, S. H.; Wilk, L. J. *Phys. B: At. Mol. Phys.* **1983**, 16, 3687.
- (27) Lundin, U.; Eriksson, O. *Int. J. Quantum Chem.* **2001**, 81, 247.
- (28) Novak, P.; Kunes, J.; Pickett, W. E.; Ku, W.; Wagner, F. R. *Phys. Rev. B* **2003**, 67, 140403.
- (29) Guenzburger, D.; Ellis, D. E.; Terra, J. *Theor. Chem. Acc.* **2005**, 113, 191.
- (30) Seo, D. K. *J. Chem. Phys.* **2006**, 125, 154105.
- (31) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, 123, 062201.
- (32) Friis, J.; Maden, G. K. H.; Larsen, F. K.; Jiang, B.; Marthinsen, K.; Holmestad, R. *J. Chem. Phys.* **2003**, 119, 11359.
- (33) Pariser, R.; Parr, R. G. *J. Chem. Phys.* **1953**, 21, 767. Pople, J. A. *Trans. Faraday Soc.* **1953**, 49, 1375.
- (34) Gunnarsson, O.; Schönhammer, K. *Phys. Rev. Lett.* **1986**, 56, 1968.
- (35) Schönhammer, K.; Gunnarsson, O.; Noack, R. M. *Phys. Rev. B* **1995**, 52, 2504.
- (36) Lima, N. A.; Silva, M. F.; Oliveira, L. N.; Capelle, K. *Phys. Rev. Lett.* **2003**, 90, 146402.
- (37) Lima, N. A.; Oliveira, L. N.; Capelle, K. *Europhys. Lett.* **2002**, 60, 601.
- (38) Xianlong, G.; Polini, M.; Tosi, M. P.; Campo Jr., V. L.; Capelle, K.; Rigol, M. *Phys. Rev. B* **2006**, 73, 165120.
- (39) Lieb, E. H.; Wu, F. Y. *Phys. Rev. Lett.* **1968**, 20, 1445.
- (40) By exact LDA, we mean that there is no need for analytical interpolations such as the PZ81, VWN, or PW92 forms of the ab initio LDA, because the energy of the homogeneous reference system can be calculated numerically exactly for arbitrary densities.
- (41) Vieira, D.; Capelle, K.; Ullrich, C. A. *Phys. Chem. Chem. Phys.* **2009**, 11, 4647.
- (42) Lima, M. P.; Pedroza, L. S.; Silva, A. J. R.; Fazzio, A.; Vieira, D.; Freire, H. J. P.; Capelle, K. *J. Chem. Phys.* **2007**, 126, 144107.
- (43) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, 125, 194112.
- (44) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, 125, 201102.
- (45) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Science* **2008**, 321, 792.
- (46) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2008**, 100, 146401.

CT100352R

Including Charge Penetration Effects in Molecular Modeling

Bo Wang and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431, United States

Received July 11, 2010

Abstract: Electrostatic effects are often the dominant component of intermolecular interactions, but they are often modeled without accounting for charge penetration effects due to the finite extent of electronic orbitals. Here, we propose a new scheme to include charge penetration effects in electrostatic modeling, and we parametrize it and illustrate it by employing the electronically embedded combined quantum mechanical and molecular mechanical (QM/MM) method. It can also be extended to other molecular modeling approximations that include electrostatic effects. The method, which is based on introduction of a single parameter for each element, is simple in concept and implementation, modest in cost, and easily incorporated into existing codes. In the new scheme, the MM atomic charge density of an atom in a molecule is represented by a screened charge rather than by a point charge. The screened charge includes a point charge for the nucleus, core electrons, and inner valence electrons, and a smeared charge for the outer valence electron density, which is distributed in a Slater-type orbital representing the outer part of the atomic charge distribution such that the resulting pairwise interactions are still analytic central potentials. We optimize the exponential parameters of the Slater-type orbitals for 10 elements, in particular H, C, N, O, F, Si, P, S, Cl, and Br, to minimize the mean unsigned error (MUE) of the QM/MM electrostatic and induction energies with respect to the Hartree–Fock electrostatic energies and the sum of induction and induction-exchange energies calculated by symmetry-adapted perturbation theory (SAPT). The resulting optimized exponential parameters are very physical, which allows one to assign parameters to all nonmetal elements (except rare gases) with atomic number less than or equal to 35. For a test set of complexes, the improved description of MM charge densities reduces the error of electrostatic interactions between QM and MM regions in the QM/MM method from 8.1 to 2.8 kcal/mol and reduces the error of induction interactions from 1.9 to 1.4 kcal/mol.

1. Introduction

The combined quantum mechanical and molecular mechanical (QM/MM) method is a useful tool to model large and complex systems.^{1–5} It has been widely used in modeling complex molecules, condensed-phase chemistry, materials, and homogeneous, enzymatic, and heterogeneous catalysis. The electronically embedded QM/MM approach, in which interactions between QM electrons and MM partial charges are added as one-electron integrals into the QM Hamiltonian,

allows the polarization of the QM electron density by the MM environment; therefore, it is a more accurate embedding scheme than mechanical embedding, in which such polarization is neglected.⁶ The electrostatic interactions between the QM and MM regions are usually written as:

$$H_{\text{QM/MM}}^{\text{el}} = - \sum_{i,A} \frac{q_A}{r_{iA}} + \sum_{\alpha,A} \frac{Z_{\alpha} q_A}{R_{\alpha A}} \quad (1)$$

where q_A are the MM point charges; the indices i and α run over all QM electrons and nuclei, respectively; and r_{iA} and $R_{\alpha A}$ are the distances between the QM electrons and the MM

* Corresponding author e-mail: truhlar@umn.edu.

point charges and the distances between the QM nuclei and the MM point charges, respectively.

The use of MM partial atomic charges as point charges at the positions of the nuclei, as in eq 1, is a very popular way to parametrize the electrostatics in molecular modeling, and it is sometimes called the distributed monopole approximation. However, an MM point charge need not be a good model for the electron density of the MM subsystem. Four possible improvements can be considered, in particular, the addition of higher-order multipole contributions at each nuclear center,^{7–9} the use of off-nuclei point charges,^{10,11} the inclusion of penetration effects,^{12–14} and the treatment of additional quantum mechanical effects associated with the distributed charge distribution.^{15–27} In a general way, the first three approaches all account for the same effect, namely, that the actual electron density has more structure than a collection of point charges. The distributed multipole method accounts for the asymmetry of the charge distribution of an atom in a molecule, the use of off-nuclei charge centers accounts for both asymmetry and finite orbital extent, and the penetration modeling accounts for finite orbital extent. One can include both higher multipole moments and penetration effects,^{20,28–30} but in the present study, we concentrate solely on the penetration effects. This has the advantage that we retain the radial, pairwise functional form for the contribution of each atom to the molecular electrostatic potential. The fourth approach is beyond the present scope, but it can include, for example, exchange repulsion or the effects of embedding atoms on the matrix elements of the QM atoms.

The essence of penetration effects, which cannot be described by MM point charges or distributed multipoles, is that, when two atoms are close enough, their charge densities can overlap, and the shielding of the nuclear charge of each atom by its own electron density decreases. Various approaches have been suggested to include this effect in both MM calculations and QM/MM calculations.^{13,14,16,20,28–41} In the MM studies, the MM charge densities were represented by Gaussian multipoles,³² point charges with damping functions,^{14,20,30,41} a set of s-type Gaussian functions,^{29,35,36} spherical atomic charge densities using Hartree–Fock-limit wave functions,³⁷ and single Slater-type contracted Gaussian multipole charge densities.⁴⁰ The coefficients of Gaussian-type functions or damping functions were chosen to fit the electron density and electrostatic potential,^{20,30,32,35–37,39,40} the electrostatic energy,^{14,41} or liquid-state properties,³³ or they were based on the intermolecular overlap.²⁹ Similar strategies have also been used in QM/MM calculations in which charges with damping functions^{14,16} and a set of s-type Gaussian functions³⁸ have been employed to include the charge penetration effects. Gaussian-type charge densities were used to mimic the real charge distributions of MM atoms in the link atom and double link atom methods.^{13,34}

In the present work, we describe a simple scheme to include penetration effects in QM/MM calculations. Because the inclusion of distributed multipoles complicates the implementation in QM/MM calculations, and the MM point charges fitted to electrostatic-potential (ESP) can effectively include some contributions due to higher-order multipoles,³

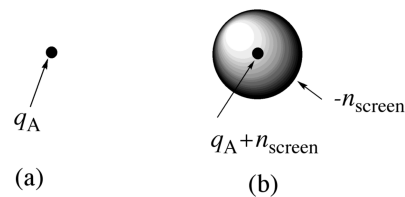


Figure 1. Comparison between (a) a point charge model and (b) a screened charge model of an MM atom A. The total smeared charge in model (b) is $-n_{\text{screen}}$, representing n_{screen} electrons.

we do not add multipole refinements to the MM charges in this study. Moreover, it was pointed out by Cisneros et al. that damping of atom-centered point charges is more important than using distributed multipoles.¹⁴ The basic idea of the method advanced here is using a charge screened by a Slater-type orbital⁴² (STO) to represent the outermost portion of the charge density of an MM atom. The parameter used to specify the spatial extent of the screened charge is optimized to give the best agreement with Hartree–Fock electrostatic energies and with induction energies computed by symmetry-adapted perturbation theory⁴³ (SAPT). In section 2, we will derive an expression for the screened MM charge. In section 3, we will present the method to optimize the exponential parameters of the STOs for different elements, and we will present the test suite and implementation details. Section 4 gives the optimized parameters and an analysis of the resulting accuracy. Finally, section 5 summarizes the main conclusions.

2. Theory

All variables and equations are in atomic units. We define a normalized STO by

$$\varphi = Ar^{n-1} \exp(-\zeta r) \quad (2)$$

where A is the normalization constant (normalizing φ^2 to integrate to unity), r is the distance of the electron from the nucleus, and n is the highest principal quantum number of the element. In particular, $n = 1$ for H and He, $n = 2$ for Li through Ne, $n = 3$ for Na through Ar, and $n = 4$ for K through Kr. The exponential parameter ζ is a parameter that depends on the atomic number.

We make the assumption that the charge density of an atom in the MM subsystem can be represented by two components: (i) a smeared charge Q distributed like electrons in the orbital φ and (ii) the rest of the charge, which is located at the nucleus. Since the net charge on an atom A in the MM subsystem is a parameter q_A , the charge at the nucleus is $q_A - Q$. Since the smeared charge represents the outer electrons, Q is negative and it is convenient to define $n_{\text{screen}} = -Q$. Then the charge density $\rho_A(r)$ of the smeared charge on atom A is given by

$$\rho_A(r) = -n_{\text{screen}}[Ar^{n-1} \exp(-\zeta r)]^2 \quad (3)$$

Figure 1 compares the point charge model and the screened charge model.

Table 1. Clementi–Raimondi Exponential Parameters for the Outermost Orbitals^a

atom parameter	H	C	N	O	F
atom parameter	1.00	1.57	1.92	2.23	2.55
atom parameter	Si	P	S	Cl	Br
atom parameter	1.43	1.63	1.83	2.04	2.26

^a Reference 44.

The exponential parameter ζ for the STO determines the spatial extent of the smeared charge distribution of the atom. It will be optimized for several elements (H, C, N, O, F, Si, P, S, Cl, and Br) in the present study. Our initial guess of the ζ values for these elements are the exponential parameters of their outermost orbitals (that is, 1s for H and He, 2s for Li through Be, 2p for B through Ne, 3s for Na through Mg, 3p for Al through Ar, 4s for K through Zn, and 4p for Ga through Kr) as optimized by Clementi and Raimondi,⁴⁴ and they are shown in Table 1 for the elements considered in this study. Another possible choice of initial guess would be the ζ values optimized by Cusachs et al. to reproduce the overlap integrals.⁴⁵

It can be shown that the electrostatic potential $U(r)$ at a distance r from the nucleus of an atom can be written as⁴⁶

$$U(r) = \frac{Z}{r} - 4\pi \left[\frac{1}{r} \int_0^r \rho(r') r'^2 dr' + \int_r^\infty \rho(r') r' dr' \right] \quad (4)$$

where $\rho(r)$ is the atomic electron density and Z is the nuclear charge. In our screened charge model, eq 4 becomes

$$U(r) = \frac{q_A + n_{\text{screen}}}{r} + 4\pi \left[\frac{1}{r} \int_0^r \rho_A(r') r'^2 dr' + \int_r^\infty \rho_A(r') r' dr' \right] \quad (5)$$

where ρ_A is given by eq 3. [Notice that $\rho(r)$, being an electron density, is positive, whereas $\rho_A(r)$, being a smeared charge density, is negative, just as Q is negative and n_{screen} is positive.] Integrating eq 5 yields

$$U(r) = \frac{q_A + n_{\text{screen}} f(\zeta r) \exp(-2\zeta r)}{r} \quad (6)$$

where the polynomial factor f is

$$\begin{aligned} f(\zeta r) &= 1 + \zeta r & n &= 1 \\ &= 1 + \frac{3}{2}\zeta r + (\zeta r)^2 + \frac{1}{3}(\zeta r)^3 & n &= 2 \\ &= 1 + \frac{5}{3}\zeta r + \frac{4}{3}(\zeta r)^2 + \frac{2}{3}(\zeta r)^3 + \frac{2}{9}(\zeta r)^4 + \frac{2}{45}(\zeta r)^5 & n &= 3 \\ &= 1 + \frac{7}{4}\zeta r + \frac{3}{2}(\zeta r)^2 + \frac{5}{6}(\zeta r)^3 + \frac{1}{3}(\zeta r)^4 + \frac{1}{10}(\zeta r)^5 + \frac{1}{45}(\zeta r)^6 + \frac{1}{315}(\zeta r)^7 & n &= 4 \end{aligned} \quad (7)$$

The electrostatic potential of a point charge is

$$U_A(r) = \frac{q_A}{r} \quad (8)$$

and the electrostatic potential of a screened charge can be written as

$$U_A(r) = \frac{q_A^*}{r} \quad (9)$$

where

$$q_A^* = q_A + n_{\text{screen}} f(\zeta r) \exp(-2\zeta r) \quad (10)$$

We substitute q_A by q_A^* as the MM charge in eq 1, and the first term of eq 1 enters into the QM Hamiltonian in our QM/MM calculations.

Our formula is similar to eq 3 in ref 14, but there are some differences. Their eq 3 can be rewritten as

$$q_A^* = q_A - (N_{\text{val}} - q_A) \exp(\Omega_{ij}(r)) \quad (11)$$

where N_{val} is the number of valence electrons, and $\Omega_{ij}(r)$ is a factor depending on the distance between atoms i and j , which satisfies

$$\Omega_{ij}(r) \rightarrow \begin{cases} 0 & \text{as } r \rightarrow 0 \\ -\infty & \text{as } r \rightarrow \infty \end{cases} \quad (12)$$

When r approaches infinity, both eqs 10 and 11 become $q_A^* = q_A$; however, when r approaches 0, q_A^* in eq 10 becomes more positive, but q_A^* in eq 11 becomes more negative. When the MM atom becomes close to a QM nucleus, the screening effect of the electrons of the MM atoms decreases, and the QM region experiences a more positive electrostatic potential; therefore, it is more physical to use eq 10. It seems likely that the first negative sign in eq 11 arose from an algebra error.

We explored two possible ways to parametrize eq 10. In one way, we assume that all N_{val} valence electrons are described by the STO, and we define n_{screen} as

$$n_{\text{screen}} = N_{\text{val}} - q_A \quad (13)$$

In the other way, since eq 10 will be parametrized to reproduce Hartree–Fock electrostatic and induction energies in the van der Waals region, we recognize that the parameter ζ will be suitable for describing the outermost fringe of an electron density, which may be appropriate for only one electron or a proper fraction of an electron. Moreover, the charge located at the MM nucleus equals $q_A + n_{\text{screen}}$, and it can be quite large if we use eq 13 (e.g., 6.0 for an oxygen). Since no exchange repulsion is added in the QM/MM self-consistent field (SCF) optimizations, this large charge at the MM nucleus can cause overpolarization of the QM region, especially for basis sets with a large number of diffuse functions. Therefore, we reduce the number of electrons that we treat as smeared to

$$n_{\text{screen}} = \min\{N_{\text{val}} - q_A, 1\} \quad (14)$$

In our tests, we found that the second choice is able to represent the penetration effects while avoiding overpolarization, so we only give detailed results for this choice. Looking ahead, we note that for nine of the 10 elements parametrized in this paper, eq 14 will yield $n_{\text{screen}} = 1$. The exception is hydrogen. Moreover, we found that the screening of metals (Na and Al in the test suite) does not systematically improve the results, which is possibly because the penetration effects are small for the positively charged cations. Therefore,

only the nonmetals are screened in the screened charge schemes and the metals are always treated as point charges, even when they appear in the same molecule in which nonmetal charges are screened.

3. Methods

3.1. Theoretical Framework. To calculate the accurate electrostatic energy and induction energy, we use symmetry-adapted perturbation theory⁴³ (SAPT) to partition the total interaction energy E_{int} into a sum of several physically distinct contributions. We will use three of these contributions: $E_{\text{elst}}^{(10)}$ represents the Hartree–Fock electrostatic interaction of the unperturbed monomers’ charge distributions; $E_{\text{ind}}^{(20)}$ is the Hartree–Fock induction energy and equals the sum of $E_{\text{ind}}^{(20)}(A \leftarrow B)$ and $E_{\text{ind}}^{(20)}(B \leftarrow A)$, where $E_{\text{ind}}^{(20)}(A \leftarrow B)$ and $E_{\text{ind}}^{(20)}(B \leftarrow A)$ are the Hartree–Fock induction energy of monomer A with the static field of the monomer B, and vice versa; and $E_{\text{ind-exch}}^{(20)}$ is the induction exchange energy. The sum of $E_{\text{ind}}^{(20)}$ and $E_{\text{ind-exch}}^{(20)}$ is called the damped-induction energy and is denoted by $E_{\text{damp-ind}}^{(20)}$ in this study.

The first question is how to compare these quantities to the components of a QM/MM calculation in which the QM method is a self-consistent-field (SCF) calculation. The SCF calculation can be either the Hartree–Fock approximation or a density functional calculation; we will use Hartree–Fock in the present paper. We consider a dimer comprised of two monomers, A and B. We can place either monomer A or monomer B in the QM region, and the other one is placed in the MM region. $E_{\text{elst}}^{\text{QM/MM}}$ is the interaction energy between the screened MM charges and the QM unpolarized electron density from a separate monomer calculation. $E_{\text{ind}}^{\text{QM/MM}}$ is derived in two steps. First, we place monomer A in the QM region and monomer B in the MM region, and we compare the interaction between the screened MM charges and the QM unpolarized electron density to the interaction between the screened MM charges and the QM electron density that is relaxed in the presence of the screened MM charges. The difference is $E_{\text{ind}}^{\text{QM/MM}}(A \leftarrow B)$. Second, we switch the QM and MM regions, and calculate $E_{\text{ind}}^{\text{QM/MM}}(B \leftarrow A)$ in the same way as $E_{\text{ind}}^{\text{QM/MM}}(A \leftarrow B)$. The induction energy $E_{\text{ind}}^{\text{QM/MM}}$ is the sum of these two terms.

The comparison of $E_{\text{elst}}^{\text{QM/MM}}$ and $E_{\text{ind}}^{\text{QM/MM}}$ from QM/MM calculations to $E_{\text{elst}}^{(10)}$, $E_{\text{ind}}^{(20)}$, and $E_{\text{ind-exch}}^{(20)}$ from SAPT calculations needs special attention. The $E_{\text{elst}}^{\text{QM/MM}}$ of the QM/MM calculation can be directly compared with $E_{\text{elst}}^{(10)}$ of the SAPT calculation. However, $E_{\text{ind}}^{\text{QM/MM}}$ in QM/MM calculations does not have the same meaning as $E_{\text{ind}}^{(20)}$ in SAPT. In the QM/MM calculations, the induction energy $E_{\text{ind}}^{\text{QM/MM}}$ includes all orders of the perturbation of the induction, while only the second-order perturbation energy is calculated in $E_{\text{ind}}^{(20)}$ in SAPT. Moreover, in SAPT calculations, the induction exchange energy $E_{\text{ind-exch}}^{(20)}$ is always evaluated and added to the induction energy $E_{\text{ind}}^{(20)}$ to give a reasonable estimation of the total interaction energy. In QM/MM calculations, it is hard to add this induction exchange energy in an empirical way, as may more readily be done for the static exchange energy, because the induction exchange energy is not even approximately pairwise additive. Therefore, we decide to

compare the QM/MM induction energy $E_{\text{ind}}^{\text{QM/MM}}$ with $E_{\text{damp-ind}}^{(20)}$, that is, with the sum of the induction $E_{\text{ind}}^{(20)}$ and induction-exchange $E_{\text{ind-exch}}^{(20)}$ energies of the SAPT calculations. The same strategy has sometimes been adopted in the development of polarizable MM force fields.^{36,47}

3.2. Basis Sets and MM Charges. Two Gaussian-type basis sets were used for the QM/MM and SAPT calculations, namely, the aug-cc-pVTZ basis set of Dunning and co-workers^{48,49} and the def2-TZVP basis set from TURBOMOLE.⁵⁰ Although in applications one might use a smaller number of diffuse functions than are present in the aug-cc-pVTZ basis set, it is important to use a large diffuse space during parametrization to be sure that the parametrized model is stable against overpolarization catastrophes.

The MM partial atomic charges are Hartree–Fock ChEIPG charges⁵¹ for the separated monomers using the same basis set as the QM method for each of the QM/MM calculations. For example, when the aug-cc-pVTZ basis set is used as the QM method in the QM/MM calculations, the MM charges are also derived using the aug-cc-pVTZ basis set.

3.3. Optimization Methods and Software. We optimize the ζ values for the 10 elements (H, C, N, O, F, Si, P, S, Cl, and Br) in our test suite in order to fit the QM/MM electrostatic energies to SAPT results. The error function is based on the difference between the QM/MM and SAPT Hartree–Fock electrostatic and damped-induction energies

$$\text{MUE} = \sum_{B=1}^2 \sum_{M=1}^{\text{molecules}} \sum_{G=1}^3 \left(\sum_{i=1}^2 |E_{\text{elst}}^{\text{QM/MM}}(\text{QM/MM}; B, M, G, i) - E_{\text{elst}}^{(10)}(\text{SAPT}; B, M, G)| + |E_{\text{ind}}^{\text{QM/MM}}(\text{QM/MM}; B, M, G) - E_{\text{damp-ind}}^{(20)}(\text{SAPT}; B, M, G)| \right) \quad (15)$$

where B labels the basis sets, M labels the molecules in the database (see below), G labels the geometries for each molecule in the database (see below), and i denotes which monomer is treated as QM. Equation 15 is minimized in the parametrization.

The SAPT calculations are performed with SAPT2008 program package⁵² interfaced to the version E.01 of *Gaussian 03*⁵³ integral and self-consistent-field package. All QM/MM calculations are carried out using our own QMMM program, which is based on the version D.01 of *Gaussian 03*⁵³ and TINKER⁵⁴ programs. We use a modified version of 1.3.5.⁵⁵ The optimized structures of some of the dimers in the test suite (see below) are acquired using a locally modified version of *Gaussian 03* (MN-GFM⁵⁶).

3.4. Test Suite. We included 40 dimers in our database, as shown in Figures 2–5. Because some molecules are too large for SAPT analysis using aug-cc-pVTZ, we do calculations with the aug-cc-pVTZ basis set for only 29 out of 40 molecules in the test suite; these 29 consist of those in Figures 3 and 5 plus the ones that are labeled by asterisks (*) in Figure 2. Thus the first two sums in eq 15 encompass 69 cases, not 80.

We considered three geometries for each of the dimers: the equilibrium geometry, a compressed geometry, and an extended geometry. The equilibrium geometry of all mol-

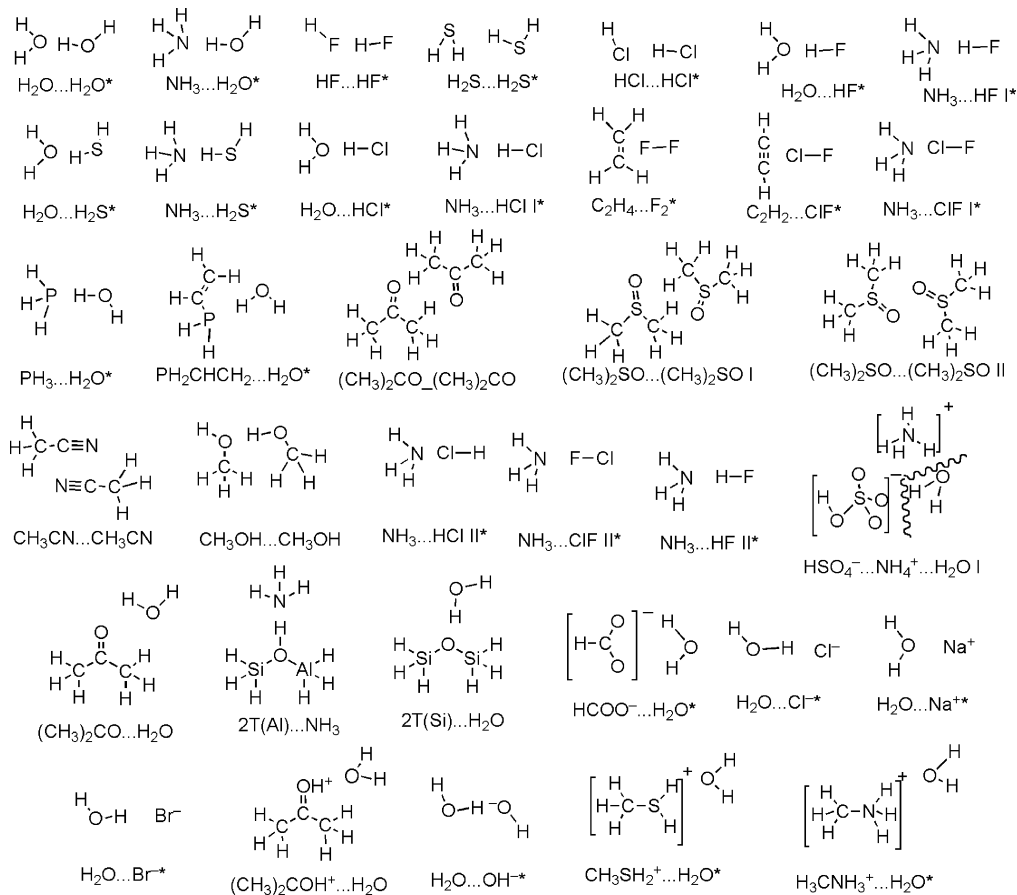


Figure 2. Thirty-six of 40 dimers in the test suite. We use * to label the molecules that are tested using both aug-cc-pVTZ and def2-TZVP basis sets.

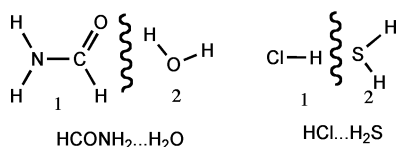


Figure 3. Geometry of the 37th and 38th dimers.

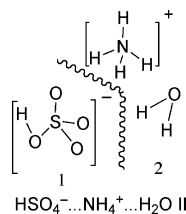


Figure 4. Geometry of the 39th dimer.

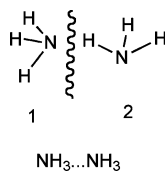


Figure 5. Geometry of the 40th dimer.

ecules is from several sources: (i) the HB6/04, CT7/04, and DI6/04 databases⁵⁷ are the first choice, when the dimer is present in one of them; (ii) $\text{H}_2\text{O}\cdots\text{OH}^-$ is the QCISD/MG3S-optimized structure;⁵⁸ and (iii) $\text{NH}_3\cdots\text{HCl}$ II, $\text{NH}_3\cdots\text{ClF}$ II, and $\text{NH}_3\cdots\text{HF}$ II are based on the geometry

of $\text{NH}_3\cdots\text{ClF}$ I in the CT7/04 database;⁵⁷ For the HCl, ClF, and HF monomers in the $\text{NH}_3\cdots\text{HCl}$ II, $\text{NH}_3\cdots\text{ClF}$ II, and $\text{NH}_3\cdots\text{HF}$ II dimers, the Cl, F, and H are placed at the same position as the Cl in $\text{NH}_3\cdots\text{ClF}$ I, and the Cl–H, F–Cl, and H–F are placed along the same direction as Cl–F in $\text{NH}_3\cdots\text{ClF}$ I with the bond lengths equal 1.282, 1.701, and 0.924 Å, respectively. (iv) Other dimers are M06-2X⁵⁹-optimized structures; the basis set we used for these optimizations is MG3S⁶⁰ for H through Cl, and 6-311+G(3d2f)⁶¹ for Br.

In the compressed geometry, we move the monomers closer, without changing their internal structures, along a line connecting their centers of mass until the distance between the centers of mass of the two monomers is 10% shorter than that in the equilibrium geometry. In the extended geometry, we move the monomers farther apart along a line connecting their centers of mass until the distance between the centers of mass of the two monomers is 10% longer than that in the equilibrium geometry.

3.5. Implementation in Gaussian 03. As is shown in eq 10, the screened MM charge contains a Slater-type function. To facilitate the implementation of the Slater-type function into the Hamiltonian of the QM/MM calculations, we expand a single Slater-type function in terms of three Gaussian-type functions, as follows:^{62,63}

$$\exp(-\lambda r) = \sum_{i=1}^3 C_i \exp(-\alpha_i \lambda^2 r^2) \quad (16)$$

Table 2. Contraction Coefficients and Exponential Parameters^a

	C_i	α_i
1	0.107 150	0.109 818
2	0.343 808	0.405 771
3	0.355 483	2.227 660

^a References 62 and 63.

The contraction coefficients C_i and exponential parameters α_i in eq 16 are listed in Table 2. The electrostatic potentials generated by the screened MM charges are incorporated in version D.01 of *Gaussian 03*⁵³ as pseudopotentials using the keyword “pseudo”. In the Gaussian implementation, the pseudopotentials do not interact with the QM nuclei. Therefore, in the QMMM program revised for this work, we calculate the interactions between the pseudopotentials and the QM nuclei, and we add them to the energy of the system.

4. Results and Discussion

4.1. SAPT Results. We computed averages over the SAPT results in order to give an indication of the typical values of the various terms. The averaged SAPT results of the electrostatic, induction, and damped-induction (induction plus induction-exchange) energies with two basis sets and three geometries are shown in Table 3. For the aug-cc-pVTZ basis set, they are averaged over 29 dimers. For the def2-TZVP basis set, they are averaged over 40 dimers. To compare the results for different basis sets, we also averaged the def2-TZVP results over only 29 dimers (the same molecules as those tested by the aug-cc-pVTZ basis set), with the results shown in Table 4. The results in Table 4 can be compared with the aug-cc-pVTZ results in Table 3 to see the energy differences when using different basis sets.

Table 3 shows that the average SAPT electrostatic, induction, and damped-induction energies over all geometries and basis sets are -19 , -13 , and -4 kcal/mol, respectively. Comparing the results from Tables 3 and 4, we found that for the electrostatic energy, the induction energy, and the damped-induction energy, the differences between two basis sets are less than 7% of the averaged values. The aug-cc-pVTZ basis set contains a larger number of diffuse functions than the def2-TZVP basis set. While these functions may be important for the dispersion interactions, the electrostatic and induction energies between the monomers can be described quite well without them. Using basis sets with smaller numbers of diffuse functions not only decreases the cost of the calculations but also it avoids the overpolarization in QM/MM calculations when large charges are placed near the boundary.

Comparing three different geometries, we found that both the electrostatic and induction energies become more negative when the two monomers are placed closer, while the exchange energy becomes more positive. The absolute value of the damped-induction energy is smaller than the induction energy, and it changes more slowly with geometry.

In the following discussion, we will use the electrostatic energy and damped-induction energy from SAPT calculations

as benchmarks to evaluate how well different charge schemes treat charge penetration effects in the QM/MM calculations.

4.2. QM/MM Results with Point Charge Scheme and Screened Charge Scheme. We define the number of electrons in the STO using eq 14. The ζ value for each element is optimized to minimize the MUE in eq 15. The optimized values are shown in Table 5. Note that the metal elements (Na and Al) are always treated as point charges (i.e., $\zeta = \infty$). Contrary to our initial guess that the optimized ζ values would be close to the Clementi–Raimondi exponential parameters for the outermost orbitals, they are instead similar (except for H) to half of the Strand and Bonham⁶⁴ exponential parameters for the outermost component (^a λ_1 for H–Ar and ^b λ_3 for K–Kr in their paper) of the density (the factor of 1/2 results from the fact that the screened charge used in this study has an exponential parameter of 2ζ because the orbital is squared, not ζ); these values are also shown in Table 5. The mean signed error (MSE) and mean unsigned error (MUE) of the electrostatic and damped-induction energies using MM point charges, MM screened charges with optimized parameters, and MM screened charges with modified Strand–Bonham (MSB) parameters (half of the Strand–Bonham parameters for all elements except H and the optimized parameter for H) are listed in Tables 6 and 7 respectively.

Table 6 shows that the MUE of the electrostatic energy using the screened charge scheme with optimized parameters is 2.8 kcal/mol, compared with 8.1 kcal/mol for the traditional point charge scheme. The screened charge scheme with the MSB parameters gives an MUE of 3.1 kcal/mol, which is quite close to the optimized result. Table 7 shows that the MUE of the QM/MM induction energy also decreases from 1.9 to 1.4 kcal/mol when using the screened charge schemes, although the improvement is not as significant as the electrostatic energy (a factor of 1.4 for the induction energy vs a factor of 2.9 for the electrostatic energy). The fact that our optimized STO exponential parameters are very close to the values we derived from the Strand–Bonham fits to atomic densities shows that the model parameters are very physical. The results clearly show that the inclusion of penetration effects improves the description of the MM electrostatic potential and its effect on the QM system. As the penetration effects decrease exponentially as a function of the distance between atoms, the improvement is most significant for the compressed geometry, in which the penetration effects are large. For the extended geometry, in which the monomers are placed farther from each other, the difference between the point charge scheme and the screened charge scheme is relatively small.

We also tested the screened charge scheme using eq 13, rather than eq 14, to define the number of screened electrons in the Slater-type orbital, and optimized the parameters to reproduce the SAPT electrostatic energies. The electrostatic energies can be well-reproduced. However, we found that the induction energy is greatly overestimated, especially when the aug-cc-pVTZ basis set is used. The error is hard to control; therefore, we abandoned eq 13.

One issue not so far considered is that in comprehensively parametrized force fields, one can, to some extent, make up

Table 3. Averaged SAPT Electrostatic, Induction, and Damped-Induction Energies (kcal/mol) Using Three Geometries and Two Basis Sets

	equilibrium/ acTZ ^a	equilibrium/ def2-TZVP	compressed/ acTZ	compressed/ def2-TZVP	extended/ acTZ	extended/ def2-TZVP	all
electrostatic	-14.50	-18.78	-25.58	-30.89	-9.20	-12.80	-18.98
induction	-9.61	-9.58	-24.86	-25.14	-4.14	-4.08	-12.91
damped-induction	-3.18	-3.49	-7.87	-8.58	-1.63	-1.77	-4.45

^a acTZ represents aug-cc-pVTZ.

Table 4. Averaged SAPT Electrostatic, Induction, and Damped-Induction Energies (kcal/mol) Using Three Geometries and the def2-TZVP Basis Set over Only 29 Molecules

	equilibrium/ def2-TZVP	compressed/ def2-TZVP	extended/ def2-TZVP
electrostatic	-15.17	-26.32	-9.80
induction	-9.55	-24.82	-4.08
damped-induction	-3.07	-7.73	-1.53

Table 5. ζ Values Used in the Slater-type Orbital

atom	H	C	N	O	F
optimized parameters	1.32	0.92	0.92	1.20	1.16
Strand and Bonham ^a	1.00	0.87	1.01	1.12	1.24
atom	Si	P	S	Cl	Br
optimized parameters	0.73	0.68	0.90	0.98	0.91
Strand and Bonham ^a	0.74	0.81	0.88	0.95	1.01

^a Half of the values in ref 64.

for errors in the electrostatics by parametrization of other MM parameters or by parametrization of the MM charges. However, neither approach is satisfactory. The first approach introduces systematic errors in the QM subsystem of combined QM/MM calculations because the electrostatic terms enter the QM Hamiltonian, but the other MM terms do not affect the quantum mechanical electronic Hamiltonian. The first approach is also unsatisfactory for MM calculations because electrostatic interactions have a different functional form than the usual Lennard-Jones or Buckingham forms used for other MM nonbonded terms and therefore cannot be completely mimicked by them. The second approach introduces systematic errors because the electrostatic potential due to an unscreened Coulomb interaction with a modified charge has a different dependence on geometry than a screened Coulomb interaction with the correct charge.

4.3. Case Studies. We use the HCONH₂⋯H₂O, HSO₄⁻⋯NH₄⁺⋯H₂O II, and HCl⋯H₂S dimers as three examples to illustrate the use of the new scheme. The equilibrium geometries of these three dimers are illustrated in Figures 3 and 4. In the calculations, either monomer 1 or 2 can be treated as the subsystem in the QM region. Tables 8–10 show the SAPT electrostatic and damped-induction energies and the QM/MM electrostatic and induction energies with the three charge schemes we considered, namely, unscreened point charges, screened charges with optimized STO parameters, and screened charges with MSB parameters for the STOs.

Table 8 shows the results for the HCONH₂⋯H₂O dimer. Including penetration effects in the screened charge scheme yields a much closer match to the SAPT results. The absolute error in electrostatic energy, averaged over three geometries and two basis sets, decreases from 5.4 to 1.3 kcal/mol. The

point charge scheme always underestimates the magnitude of the electrostatic and induction energies. Screening the point charge with the optimized parameters makes the electrostatic energy more negative by 9.3 kcal/mol for the compressed geometry, 2.8 kcal/mol for the equilibrium geometry, and 0.7 kcal/mol for the extended geometry, when averaged over two QM/MM calculations with different QM regions and two basis sets.

Although the SAPT damped-induction energies are similar for the aug-cc-pVTZ and def2-TZVP basis sets, the QM/MM induction energies are not. Taking the compressed geometry as an example, the SAPT damped-induction energies are -4.6 and -4.5 kcal/mol for the aug-cc-pVTZ and def2-TZVP basis sets, while the QM/MM induction energies using the screened charge scheme with the optimized parameters are -6.3 and -3.7 kcal/mol for aug-cc-pVTZ and def2-TZVP basis sets. The induction energies using the aug-cc-pVTZ basis set are more negative than those using the def2-TZVP basis set in the QM/MM calculations. As the aug-cc-pVTZ basis set contains more diffuse functions than the def2-TZVP basis set, the QM region is more prone to be polarized by MM point charges when the aug-cc-pVTZ basis set is used, which leads to the increase of the induction energy. This effect is more significant for the screened charge scheme than the point charge scheme, because the charge at the MM nucleus is larger in the screened charge scheme than that in the point charge scheme, which causes greater polarization. Despite this sensitivity, the average error in the induction energies is reduced by a about a factor of 2 when screening is included.

We also discuss a dimer composed of HSO₄⁻ and NH₄⁺⋯H₂O, with the results shown in Table 9. Only the def2-TZVP basis set is used. This is a very challenging test for the charge scheme because both of the monomers are charged. From the results, we can see that the screened charge scheme improves both the electrostatic and induction energies significantly.

The results for the HCl⋯H₂S dimer listed in Table 10 are not as good as those for the two examples that we have discussed. The improvement is not very significant when H₂S is in the MM region. The performance of the screened charge scheme for individual elements will be discussed in section 4.4.B.

In the examples singled out for illustration in this section, we found that the screened charge scheme with modified Strand–Bonham (MSB) parameters sometimes gives a better result than that with optimized parameters. This is because the optimized parameters are optimized for all the molecules in the test suite, rather than optimized for the specific dimers.

Table 6. MSE and MUE of Electrostatic Energies (kcal/mol) Using the QM/MM Method^a

	equilibrium/ acTZ		equilibrium/ def2-TZVP		compressed/ acTZ		compressed/ def2-TZVP		extended/ acTZ		extended/ def2-TZVP		all	
	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE
Pt charge ^b	6.43	6.43	6.43	6.43	14.80	14.80	15.10	15.10	3.02	3.02	2.98	2.98	8.13	8.13
Opt ^c	1.58	1.97	1.32	2.17	3.64	5.09	2.61	5.31	1.06	1.11	1.03	1.16	1.84	2.81
MSB ^d	1.59	2.22	1.19	2.29	3.55	5.39	2.15	5.90	1.07	1.25	1.00	1.24	1.71	3.06

^a Exact values are SAPT electrostatic energies. ^b Point charge scheme. ^c Screened charge scheme with optimized parameters. ^d Screened charge scheme with modified Strand-Bonham (MSB) parameters.

Table 7. MSE and MUE of the Induction Energies (kcal/mol) Using the QM/MM Method^a

	equilibrium/ acTZ		equilibrium/ def2-TZVP		compressed/ acTZ		compressed/ def2-TZVP		extended/ acTZ		extended/ def2-TZVP		all	
	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE	MSE	MUE
Pt charge ^b	0.67	1.09	1.12	1.18	3.46	3.98	4.34	4.46	0.12	0.45	0.35	0.38	1.72	1.94
Opt ^c	-0.74	1.00	0.69	0.82	-0.62	2.67	2.68	3.13	-0.34	0.45	0.25	0.30	0.46	1.40
MSB ^d	-0.79	1.02	0.68	0.80	-0.71	2.65	2.60	3.12	-0.36	0.46	0.25	0.29	0.42	1.39

^a Exact values are SAPT damped-induction energies. ^b Point charge scheme. ^c Screened charge scheme with optimized parameters. ^d Screened charge scheme with modified Strand-Bonham (MSB) parameters.

Table 8. Electrostatic and Induction Energies (kcal/mol) of HCONH₂⋯H₂O Dimer in QM/MM Calculations Compared with SAPT Results, and MUE (kcal/mol) of QM/MM Calculations over Three Geometries and Two Basis Sets

		equilibrium/ acTZ	equilibrium/ def2-TZVP	compressed/ acTZ	compressed/ def2-TZVP	extended/ acTZ	extended/ def2-TZVP	MUE
Electrostatic								
SAPT		-11.3	-11.6	-21.9	-22.3	-6.6	-6.9	
QM/MM	Pt charge	-7.4/-7.3 ^a	-8.0/-7.5	-10.5/-11.3	-11.4/-11.5	-5.1/-5.0	-5.5/-5.1	5.4
QM/MM	Opt	-10.5/-9.8	-10.8/-10.0	-22.0/-18.6	-22.5/-18.8	-5.9/-5.7	-6.2/-5.8	1.3
QM/MM	MSB	-10.8/-10.5	-11.2/-10.7	-23.4/-19.5	-23.8/-19.8	-6.0/-6.0	-6.3/-6.1	1.1
Induction								
SAPT (damped)		-1.7	-1.6	-4.6	-4.5	-0.7	-0.7	
QM/MM	Pt charge	-1.2	-1.1	-2.8	-2.5	-0.6	-0.5	0.9
QM/MM	Opt	-2.0	-1.3	-6.3	-3.7	-0.7	-0.5	0.5
QM/MM	MSB	-2.1	-1.3	-6.8	-3.9	-0.8	-0.5	0.6

^a x/y denotes that the electrostatic energy is x when HCONH₂ is the QM region and is y when H₂O is the QM region.

Table 9. Electrostatic and Induction Energies (kcal/mol) of HSO₄⁻⋯NH₄⁺⋯H₂O II in QM/MM Calculations Compared with SAPT Results, and MUE (kcal/mol) of QM/MM Calculations over Three Geometries

		equilibrium/def2-TZVP	compressed/def2-TZVP	extended/def2-TZVP	MUE
Electrostatic					
SAPT		-144.1	-176.4	-124.0	
QM/MM	Pt charge	-126.4/-130.2 ^a	-135.0/-147.6	-116.4/-117.5	19.3
QM/MM	Opt	-149.4/-138.7	-190.9/-161.5	-124.7/-121.6	7.2
QM/MM	MSB	-146.0/-139.4	-184.6/-162.2	-123.2/-122.3	5.2
Induction					
SAPT (damped)		-15.7	-30.6	-9.8	
QM/MM	Pt charge	-12.2	-17.4	-8.8	5.9
QM/MM	Opt	-15.7	-26.7	-9.9	1.4
QM/MM	MSB	-14.9	-25.4	-9.5	2.1

^a x/y denotes that the electrostatic energy is x when HSO₄⁻ is the QM region and is y when NH₄⁺ (H₂O) is the QM region.

4.4. Discussion of ζ Parameters. *A. ζ Value Effects on Electrostatic and Induction Energies.* In order to understand how the ζ parameters of the STOs affect the electrostatic and induction energies, we here examine the effect of varying ζ on the NH₃⋯NH₃ dimer with the equilibrium geometry and with the aug-cc-pVTZ basis set. The geometry is shown in Figure 5. The electrostatic and induction energies are most sensitive to the nearest MM atom, so we choose to vary the ζ parameter of the nearest MM atom. In Figure 6, we show the change of electrostatic energy with respect to the H ζ value with monomer 1 in the QM region and with the ζ value for N kept at its optimized value of 0.92. In

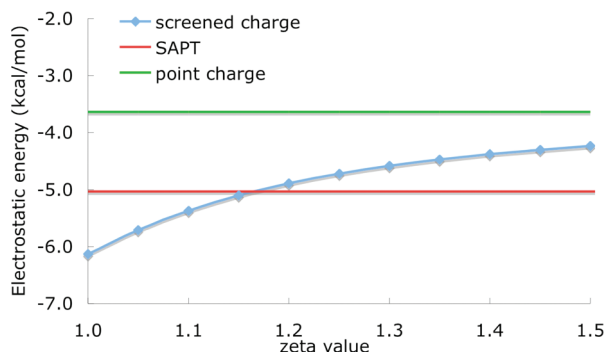
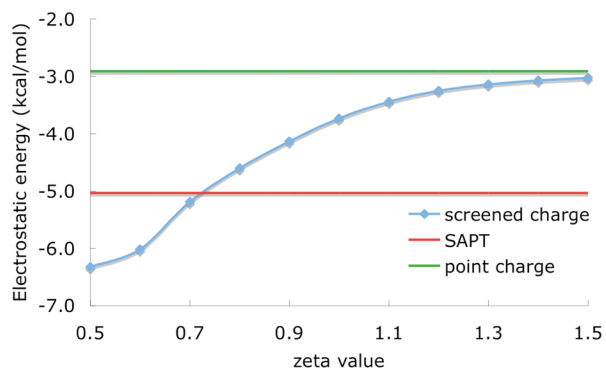
Figure 7, we show the change of electrostatic energy with respect to the N ζ value with monomer 2 in the QM region and with the ζ value for H kept at its optimized value of 1.32. To make the comparison, we draw two lines representing the SAPT result and the QM/MM result without screening H and N, respectively.

We can see that, when the ζ values become larger and larger, the electrostatic energy gets closer to the result from the point charge scheme, as is easy to understand both physically and mathematically. The figures show how optimum ζ values that best reproduce the electrostatic energy

Table 10. Electrostatic and Induction Energies (kcal/mol) of HCl \cdots H₂S Dimer in QM/MM Calculations Compared with SAPT Results, and MUE (kcal/mol) of QM/MM Calculations over Three Geometries and Two Basis Sets

		equilibrium/ acTZ	equilibrium/ def2-TZVP	compressed/ acTZ	compressed/ def2-TZVP	extended/ acTZ	extended/ def2-TZVP	MUE
		Electrostatic						
SAPT		-5.2	-5.4	-10.9	-11.1	-2.8	-3.0	
QM/MM	Pt charge	-1.0/-2.1 ^a	-1.1/-2.4	-1.7/-2.7	-1.8/-3.3	-0.6/-1.5	-0.7/-1.7	4.7
QM/MM	Opt	-2.4/-4.5	-2.5/-4.7	-4.0/-10.5	-4.0/-11.1	-1.2/-2.2	-1.2/-2.3	2.2
QM/MM	MSB	-2.5/-4.6	-2.5/-4.9	-3.9/-11.0	-3.9/-11.6	-1.2/-2.2	-1.3/-2.4	2.2
		Induction						
SAPT (damped)		-1.1	-1.0	-3.0	-3.0	-0.5	-0.4	
QM/MM	Pt charge	-0.4	-0.3	-0.8	-0.6	-0.2	-0.1	1.1
QM/MM	Opt	-1.5	-0.5	-5.1	-1.9	-0.4	-0.2	0.7
QM/MM	MSB	-1.5	-0.5	-5.4	-2.0	-0.5	-0.2	0.8

^a x/y denotes that the electrostatic energy is x when HCl is the QM region and is y when H₂S is the QM region.

**Figure 6.** QM/MM and SAPT electrostatic energies (kcal/mol) with respect to the H ζ value.**Figure 7.** QM/MM and SAPT electrostatic energies (kcal/mol) with respect to the N ζ value.

are located in a reasonable range. The same trend is found for other dimers.

B. Consideration of Individual Elements. In order to understand how the screened charge scheme works for individual elements, we divided the 40 dimers into several groups. Two division schemes have been used. In the first scheme, the dimers are divided on the basis of the elements included in the MM regions. For example, for the group corresponding to element O, we selected all the QM/MM electrostatic calculations in which one or more oxygen atom is included in the MM regions. Because an MM region always contains several different elements, each QM/MM calculation is included in several element groups. Then we calculated the MUE of electrostatic energies for different elements in their groups. The drawback of the first division scheme is that the screening effects of an MM element are not properly reflected in the QM/MM electrostatic energy

when the MM element is far from the QM region. Therefore, in the second scheme, only the MM atoms that are close to the QM region are considered. We calculated the distances between all QM and MM atom pairs and identified the shortest QM–MM distance. We define a close MM atom as one that has a distance from any QM atom that is less than the sum of 0.30 Å and the shortest QM–MM distance. Then we divided the 40 dimers on the basis of the elements among the close MM atoms. We carried out the division only for the equilibrium geometry, and we assigned the same group membership for the compressed and extended geometries as in the equilibrium geometry. Because the QM/MM electrostatic energy is most sensitive to the screening of the closest MM atoms, the second division scheme is more appropriate to test the contribution of individual elements. Note that if we change 0.30 Å in the criterion to define a MM atom close to infinity, the second division scheme becomes the first division scheme. We show the number of calculations included in each group and the MUE of the QM/MM electrostatic energy for both the two division schemes in Table 11. We define an improvement ratio (IR) as

$$\text{IR} = \frac{\text{MUE}(\text{point charge scheme})}{\text{MUE}(\text{screened charge scheme})} \quad (17)$$

IRs for individual elements using the screened charge scheme with the optimized parameters are also shown in Table 11, and it is gratifying that they are all greater than or equal to 1.7. The results of the two division schemes show similar trends (except for C, for which the first division scheme does not reflect the screening effect of the C atom).

For the point charge scheme, Cl and N have the largest MUEs. All of them are greater than 10 kcal/mol for both division schemes. The large error of the Si using the first division scheme is mainly due to other elements, as Si is not a close MM atom in any dimers.

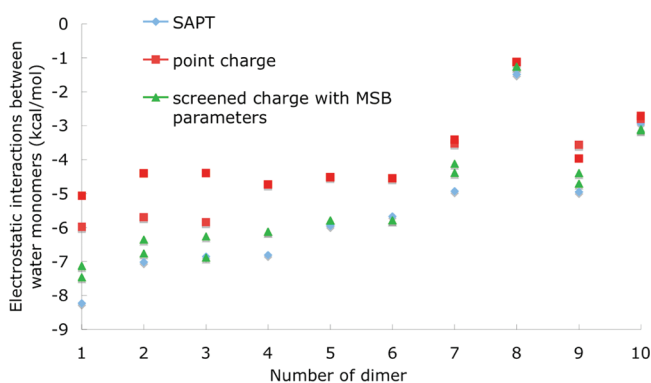
For the screened charge scheme with optimized parameters, N is the only element with an MUE of more than 3 kcal/mol by either division, whereas when screening is not employed all elements have an MUE greater than or equal to 3.6 kcal/mol for both divisions. We found that the large error of N is mainly due to the inclusion of NH₃ \cdots HF, NH₃ \cdots HCl, and 2T(Al) \cdots NH₃ dimers, in which the monomers are very close to each other.

Next, we compare the improvement ratios for individual elements. In the second division scheme, the elements that

Table 11. MUE (kcal/mol) of QM/MM Electrostatic Energies and Improvement Ratios (IR) for Individual Elements

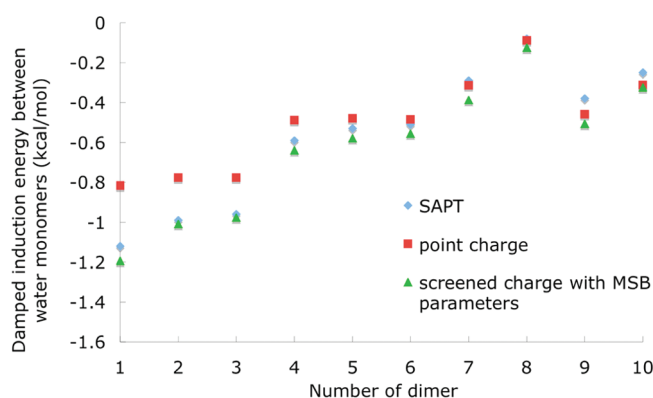
	H	C	N	O	F	Si	P	S	Cl	Br
First Division Scheme										
number of calculations ^a	372	78	87	171	54	6	12	54	60	6
Pt charge	7.9	6.9	10.2	7.5	9.6	11.0	4.9	7.3	13.7	3.6
Opt	2.9	2.4	5.1	2.6	1.6	1.7	2.9	2.5	2.1	1.4
MSB	3.1	2.7	5.5	2.7	2.0	2.2	3.1	2.6	2.6	1.6
IR (Opt) ^b	2.7	2.9	2.0	2.9	6.1	6.5	1.7	2.9	6.6	2.6
Second Division Scheme										
number of calculations ^a	213	12	66	96	18	0	12	18	30	6
Pt charge	6.9	5.2	11.1	6.1	5.9		4.9	4.0	18.6	3.6
Opt	1.8	0.7	5.7	2.8	1.2		2.9	2.3	2.7	1.4
MSB	1.8	1.2	6.6	3.0	2.1		3.1	2.2	3.1	1.6
IR (Opt)	3.8	7.7	1.9	2.1	4.9		1.7	1.7	6.9	2.6

^a This number includes all calculations using different geometries and different basis sets. ^b Improvement ratios (IRs) for individual elements using the screened charge scheme with optimized parameters.

**Figure 8.** QM/MM and SAPT electrostatic energies (kcal/mol) for 10 water dimers.

have the largest improvement ratios are C (7.7), Cl (6.9), and F (4.9), followed by H (3.8), Br (2.6), and O (2.1); N (1.9), P (1.7), S (1.7) have the smallest improvement, but even in these cases, the improvement is significant. These findings show that for all elements, electrons distributed in an STO can mimic the charge distribution of the MM atoms well enough to improve the results as compared to a point charge model.

4.5. An Application to 10 Water Dimers with a Variety of Geometries. To test the broader usefulness of the new method, we applied it to water dimers with 10 geometries tested by Reinhardt et al.⁶⁵ They have carried out SAPT calculations on these dimers using the aug-cc-pVTZ basis set. We carried out QM/MM calculations with the aug-cc-pVTZ basis set on their geometries using the point charge scheme and the screened charge scheme with MSB parameters, and we compared the results with their benchmark results of electrostatic and induction energies. The point charges of the water monomer are Hartree–Fock CHEIPG charges⁵¹ calculated with the aug-cc-pVTZ basis set. Figures 8 and 9 compare the electrostatic and induction energies for these 10 geometries. Because either monomer in the dimer can be treated as the QM region, we have two QM/MM electrostatic energies at each geometry. We found that using the screened charge scheme decreases the MUE of the electrostatic energy from 1.42 to 0.42 kcal/mol and the MUE of the induction energy from 0.11 to 0.06 kcal/mol. Note that the averaged electrostatic and damped-induction energies of the 10 water dimers are -5.48 and -0.57 kcal/mol,

**Figure 9.** QM/MM induction energies and SAPT damped-induction energies (kcal/mol) for 10 water dimers.

respectively. Therefore, the percent error drops from 26% to 8% and from 19% to 11% for the QM/MM electrostatic and induction energies, respectively. These findings show that including penetration effects is important for accurate modeling. We also carried out calculations using the screened charge scheme with optimized parameters, in which the ζ value for O is 1.20 rather than 1.12 as in the MSB parameters. The MUE of electrostatic energy is 0.57 kcal/mol, which is larger than that using MSB parameters, but still much better than that for point charges, and the MUE of the induction energy is 0.05 kcal/mol. We found that if the two monomers in the dimer are not very close (two monomers are very close in the case of $\text{H}_2\text{O}\cdots\text{OH}^-$ in the test suite for parametrization), the ζ value of 1.12 for O gives better results than the optimized value from Table 5.

4.6. Limitations of the Scheme.

We note five limitations. First, only a single STO with a fixed ζ value is used for each element in our scheme. However, a single STO may not describe the electron distribution well in all regions. When the QM electron goes closer to the MM atom, it should feel a charge distribution with a larger ζ value. This can be improved by double- ζ STOs, in which a more diffuse STO is used to describe the ultimate outer layer and a less diffuse one is used to describe the penultimate layer. Moreover, the ζ value for an element changes when the element has different partial charges and bonding environments, as pointed out before,⁶⁶ so the ζ value we optimize here is only an average value for an element in different environments. If one is willing to treat ζ as an MM parameter and optimize

it separately for different hybridization states (e.g., sp^2 O would have a different value than sp^3 O) or different functional groups or charge states, one could do even better.

Second, to avoid overpolarization, we use at most one electron in the STO. A small number of electrons in the STO may underestimate the charge penetration effect, especially for the compressed geometry. One may include more electrons in the STO if the overpolarization of the QM region can be avoided (e.g., when def2-TZVP basis set is used for the QM calculations); this would, for example, improve the results in Table 10 when H_2S is the MM region and S is the close MM atom.

Third, when multiply charged cations (such as Zn^{2+} and Mg^{2+}) are placed in the MM region near the QM–MM boundary, we found that the QM region may be greatly overpolarized by the MM region if a large number of diffuse functions are used for the QM region, such as that in the aug-cc-pVTZ basis set. This is due to the lack of exchange and orthogonality interactions between MM and QM regions in the QM/MM SCF optimizations. Although the problem is much less severe with a less diffuse basis set, our method cannot completely solve this problem. Ab initio model potentials^{67,68} (AIMP), smeared charges,^{13,34} and damped charges³⁸ can be applied to alleviate the overpolarization. One can also avoid the problem by not placing a QM–MM boundary next to a multiply charged MM cation.

Fourth, we have not tested the method for rare gases.

Fifth, we have so far developed the method only for the case where a QM–MM boundary does not pass through a covalent bond.

If desired, further studies can be carried out to remove limitations 1, 2, 4, and 5 mentioned above. Despite limitations 1–3, the method already provides a much more realistic treatment of the electrostatic and induction energies in the general case.

5. Conclusions

In this paper, we proposed a general screening scheme to include charge penetration effects in the treatment of electrostatic interactions in molecular modeling, and we parametrized it and applied it by using electronically embedded QM/MM calculations. Our scheme utilizes a Slater-type orbital to mimic the outer portion of the electron distribution around the MM atom. By including the charge penetration effects, we can greatly improve the description of electrostatic interactions in the QM/MM method. The parameters for the STOs of several common elements (H, C, N, O, F, Si, P, S, Cl, and Br) are optimized to reproduce the SAPT electrostatic and damped-induction energies. For the metal elements, we suggest keeping the point charge scheme, as no systematic improvement has been found by treating them as screened.

We found that the optimal exponential parameters are very close to the values describing the outermost layer of the electron density of atoms in Strand and Bonham's fits to atomic electron densities. This is extremely encouraging in that it shows that the method is very physical. The finding that screening does not offer systematic improvement for metal atoms is also physical, since metal atoms in molecules

usually have a partial positive charge and hence less diffuse electron density. These findings, combined with the availability of Strand and Bonham's fits for the first 36 elements, mean that parameters for the nonmetals (except rare gases) with $Z \leq 36$ that are not optimized here (B, Ge, As, and Se) can be obtained from their fits to electron densities.

Since the point charge model has been used in the overwhelming majority of MM parametrizations, but the present work shows that the point charge model leads to systematic errors in electrostatics and induction energies, we conclude that conventional MM parametrizations can only succeed by systematic cancellation of errors. Improving the electrostatics by including charge penetration effects can in principle lead to a new generation of more physical MM parameter sets.^{35,69} By evaluating the electrostatic energies more accurately, we can derive more accurate and physical empirical parameters for the exchange repulsion and dispersion. The present work shows how this can be done in a practical way. In particular, the bare Coulomb interaction is replaced by a central potential centered at the nuclei with only one additional parameter per element.

Although the formulation of the electrostatic interactions in the current work applies only to QM/MM calculations, the formula for MM–MM electrostatic energy calculations can also be derived on the basis of the screened charge model. One obtains

$$E = \int dr_1 \int dr_2 \frac{\rho_A(r_1) \rho_B(r_2)}{|r_1 - r_2|} \quad (18)$$

for the interaction energy between delocalized charge distributions at sites A and B. Evaluating the integrals of eq 18 in MM/MM electrostatic energy calculations would increase the computational cost; so for low-cost calculations, one should develop efficient schemes to evaluate or approximate the integrals; for example, these integrals can be replaced by point charge interactions for sites separated by more than a certain distance.

Because the optimized values of the parameters in our treatment of screened electrostatics are physical, they can be used without reoptimization in new force fields or in other methods that incorporate electrostatic effects, such as the electrostatically embedded many-body method,^{70,71} the electrostatically embedded multiconfiguration molecular mechanics method,^{72,73} or fragment methods for large molecular systems.⁷⁴

The present study only considered QM–MM boundaries that pass between nonbonded fragments. In future work, we will consider the use of this kind of scheme when QM–MM boundaries pass through covalent bonds.

Acknowledgment. We thank Hannah Leverentz for providing the geometry of several dimers. This work was supported in part by the National Science Foundation under grant no. CHE09-56776 and the Air Force Office of Scientific Research under grant no. FA9550-08-1-0183.

References

- (1) *Combined Quantum Mechanical and Molecular Mechanical Methods*; Gao, J.; Thompson, M. A., Eds.; ACS

- Symposium Series 712; American Chemical Society: Washington, DC, 1998.
- (2) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*, Grotendorst, J., Ed.; John von Neumann Institute for Computing: Jülich, 2000; p 285.
- (3) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.
- (4) Senn, H. M.; Thiel, W. *Angew. Chemie Int. Ed.* **2009**, *48*, 1198.
- (5) Bernstein, N.; Kermode, J. R.; Csányi, G. *Rep. Prog. Phys.* **2009**, *72*, 026501.
- (6) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
- (7) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233.
- (8) Sokalski, W. A.; Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86.
- (9) Leverentz, H.; Gao, J.; Truhlar, D. G. *Theor. Chem. Acc.*, in press.
- (10) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (11) Hancock, G. C.; Truhlar, D. G.; Dykstra, C. E. *J. Chem. Phys.* **1988**, *88*, 1786.
- (12) Stone, A. J. In *The Theory of Intermolecular Forces*; Oxford University Press Inc.: New York, 1996; p 94.
- (13) Das, D.; Eurenus, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. *J. Chem. Phys.* **2002**, *117*, 10534.
- (14) Cisneros, G. A.; Tholander, S. N.-I.; Parisel, O.; Darden, T. A.; Elking, D.; Perera, L.; Piquemal, J. P. *Int. J. Quantum Chem.* **2008**, *108*, 1905.
- (15) Bredow, T.; Geudtner, G.; Jug, K. *J. Chem. Phys.* **1996**, *105*, 6395.
- (16) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.
- (17) Yudanov, I. V.; Nasluzov, V. A.; Neyman, K. M.; Rösch, N. *Int. J. Quantum Chem.* **1997**, *65*, 975.
- (18) López, N.; Illas, F. *J. Phys. Chem. B* **1998**, *102*, 1430.
- (19) Pacchioni, G.; Ferrari, A. M. *Catal. Today* **1999**, *50*, 533.
- (20) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300.
- (21) Soave, R.; Pacchioni, G. *Chem. Phys. Lett.* **2000**, *320*, 345.
- (22) Gomes, J. R. B.; Illas, F.; Hernández, N. C.; Márquez, A.; Sanz, J. F. *Phys. Rev. B* **2002**, *65*, 125414.
- (23) Gomes, J. R. B.; Illas, F.; Hernández, N. C.; Sanz, J. F.; Wander, A.; Harrison, N. M. *J. Chem. Phys.* **2002**, *116*, 1684.
- (24) Gomes, J. R. B.; Lodziana, Z.; Illas, F. *J. Phys. Chem. B* **2003**, *107*, 6411.
- (25) Cinquini, F.; Valentin, C. D.; Finazzi, E.; Giordano, L.; Pacchioni, G. *Theor. Chem. Acc.* **2007**, *117*, 827.
- (26) Valero, R.; Gomes, J. R. B.; Truhlar, D. G.; Illas, F. *J. Chem. Phys.* **2008**, *129*, 124710.
- (27) Cembran, A.; Bao, P.; Wang, Y.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2010**, *6*, 2469.
- (28) Köster, A. M.; Kölle, C.; Jug, K. *J. Chem. Phys.* **1993**, *99*, 1224.
- (29) Kairys, V.; Jensen, J. H. *Chem. Phys. Lett.* **1999**, *315*, 140.
- (30) Piquemal, J. P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353.
- (31) Hall, G. G.; Smith, C. M. *Int. J. Quantum Chem.* **1992**, *42*, 1237.
- (32) Wheatley, R. J.; Mitchell, J. B. O. *J. Comput. Chem.* **1994**, *15*, 1187.
- (33) Guillot, B.; Guissani, Y. *J. Chem. Phys.* **2001**, *114*, 6720.
- (34) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43.
- (35) Piquemal, J. P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.
- (36) Torheyden, M.; Jansen, G. *Mol. Phys.* **2006**, *104*, 2101.
- (37) Spackman, M. A. *Chem. Phys. Lett.* **2006**, *418*, 158.
- (38) Cisneros, G. A.; Piquemal, J. P.; Darden, T. A. *J. Phys. Chem. B* **2006**, *110*, 13682.
- (39) Slipchenko, L. V.; Gordon, M. S. *J. Comput. Chem.* **2007**, *28*, 276.
- (40) Elking, D. M.; Cisneros, G. A.; Piquemal, J. P.; Darden, T. A.; Pedersen, L. G. *J. Chem. Theory Comput.* **2010**, *6*, 190.
- (41) Kumar, R.; Wang, F.-F.; Jenness, G. R.; Jordan, K. D. *J. Chem. Phys.* **2010**, *132*, 014309.
- (42) Slater, J. C. *Phys. Rev.* **1930**, *36*, 57.
- (43) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.
- (44) Clementi, E.; Raimondi, D. L. *J. Chem. Phys.* **1963**, *38*, 2686.
- (45) Cusachs, L. C.; Trus, B. L.; Carroll, D. G.; Mcglynn, S. P. *Int. J. Quantum Chem. Symp.* **1967**, *1*, 423.
- (46) Massey, H. S. W.; Burhop, E. H. S. In *The International Series of Monographs on Physics—Electronic and Ionic Impact Phenomena*; Marshall, W., Wilkinson, D. H., Eds.; University Press: Oxford, 1969; p 376.
- (47) Archambault, F.; Chipot, C.; Soteras, I.; Luque, F. J.; Schulten, K.; Dehez, F. *J. Chem. Theory Comput.* **2009**, *5*, 3022.
- (48) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (49) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (50) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (51) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.
- (52) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wheatley, R. J.; Wormer, P. E. S.; Zuchowski, P. S.; *SAPT2008: "An Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies" version, 2008.*
- (53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.;

- Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, version D.01 and E.01*, Gaussian, Inc: Wallingford, CT, 2004.
- (54) Ponder, J. W. *TINKER, version 4.2*; Washington University: St. Louis, MO, 2004.
- (55) Lin, H.; Zhang, Y.; Truhlar, D. G. *QMMM, version 1.3.5*, University of Minnesota: Minneapolis, 2007.
- (56) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module, version 4.1*; University of Minnesota: Minneapolis, 2009.
- (57) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 415.
- (58) Dahlke, E. E.; Orthmeyer, M. A.; Truhlar, D. G. *J. Phys. Chem. B* **2008**, *112*, 2372.
- (59) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (60) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (61) Curtiss, L. A.; McGrath, M. P.; Blaudeau, J. P.; Davis, N. E.; Binning, R. C.; Radom, L. *J. Chem. Phys.* **1995**, *103*, 6104.
- (62) Stewart, R. F. *J. Chem. Phys.* **1969**, *50*, 2485.
- (63) Szabo, A.; Ostlund, N. S. In *Modern Quantum Chemistry—Introduction to Advanced Electronic Structure Theory*; Dover Publications, Inc.: Mineola, NY, 1989; p 157.
- (64) Strand, T. G.; Bonham, R. A. *J. Chem. Phys.* **1964**, *40*, 1686.
- (65) Reinhardt, P.; Piquemal, J. P. *Int. J. Quantum Chem.* **2009**, *109*, 3259.
- (66) Gill, P. M. W. *J. Phys. Chem.* **1996**, *100*, 15421.
- (67) Barandiarán, Z.; Seijo, L. *J. Chem. Phys.* **1988**, *89*, 5739.
- (68) Nygren, M. A.; Pettersson, L. G. M.; Barandiarán, Z.; Seijo, L. *J. Chem. Phys.* **1994**, *100*, 2010.
- (69) Rotenberg, B.; Salanne, M.; Simon, C.; Vuilleumier, R. *Phys. Rev. Lett.* **2010**, *104*, 138301.
- (70) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.
- (71) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1.
- (72) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 790.
- (73) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 2925.
- (74) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646.

CT1003862

JCTC

Journal of Chemical Theory and Computation

Relative Weights of σ and π Ring Currents in a Few Simple Monocycles

Guglielmo Monaco,[†] Riccardo Zanasi,^{*†} Stefano Pelloni,[‡] and Paolo Lazzeretti[‡]

Dipartimento di Chimica, Università degli Studi di Salerno, via Ponte don Melillo, 84084 Fisciano (SA), Italy, and Dipartimento di Chimica, Università degli Studi di Modena e Reggio Emilia, via G. Campi 183, 41100 Modena, Italy

Received August 9, 2010

Abstract: By partitioning the bond current strength (current susceptibility) into plane symmetric and plane antisymmetric contributions, it is shown that 91% of the diatropic ring current of benzene is transported by the π electrons and the remaining non-negligible 9% is sustained by the σ electrons. In planar cyclooctatetraene 94% (6%) of the paratropic ring current is transported by the π (σ) electrons. In cyclopropane 95% (5%) of the diatropic ring current is transported by the σ (π -like) electrons. The 85% fraction of the diatropic ring current of Al_4^{2-} is transported by the σ valence electrons and 15% by the π valence electrons. In the nonaromatic borazine system the nitrogen-centered π electron circulations are surrounded by a weak diatropic “ring current” 6.5 times smaller than that of benzene.

1. Introduction

Within the orbital approximation, molecular properties result from a summation of orbital contributions, which can be conveniently classified according to their symmetry. If the properties under examination stem from the application of external fields, molecular orbitals are perturbed and their symmetry is not generally preserved. Orbital symmetry conservation is observed for plane symmetric molecules in the presence of a perpendicular magnetic field. In this case, plane symmetric and antisymmetric orbitals, hereafter referred to as σ and π orbitals, are only mixed among themselves and the first-order perturbed orbitals have the same symmetry as the unperturbed ones. Therefore, σ and π electrons give clearly distinguishable contributions to magnetic properties. This important feature, together with the assumption that σ electrons should give bond-localized circulations, has been exploited in past years to interpret the magnetic properties of a large class of (poly)cyclic conjugated hydrocarbons in terms of π ring current maps (see refs 1–4 for a few examples) and π contributions to the out-of-plane component of some molecular tensor properties, for

instance, magnetizability,⁵ magnetic shielding density,⁶ and magnetic shielding of either real⁷ or virtual nuclei. The latter properties are the target of the widespread computation of nucleus-independent chemical shift (NICS).^{8,9} Moreover, the delocalization of π ring currents is widely considered an indication of aromaticity.^{10,11} The dominant role of π electrons in determining the magnetic properties in polycyclic aromatic hydrocarbons (PAHs) is such a common tenet that it has been proposed even to recover *a posteriori* the strength of π currents from all-electron NICS grids.¹² This assumption is certainly questionable if both σ and π currents have appreciable magnitude. In this case, attention is readily paid to some quantification of the individual currents. The most used quantitative indicators are the just cited NICS and the “bond current strength” (current susceptibility).¹³ The latter is obtained by integrating the current density crossing a suitably chosen planar domain, which bisects a selected bond. When delocalization is present, a sizable net bond current flow is detected, whose strength can be used to make a direct comparison among different molecular systems. The retrieval of both contributions (σ and π) from the above all-electron indicators is certainly cumbersome. Indeed, a retrieval of σ and π currents in PAHs from an *a posteriori* analysis of NICS calculations had to introduce awkward saturated homologues,¹⁴ and the use of both NICS¹⁵ and bond current

* Corresponding author phone: +39089969590; fax: +39089969603; e-mail: rzanasi@unisa.it.

[†] Università degli Studi di Salerno.

[‡] Università degli Studi di Modena e Reggio Emilia.

strengths¹⁶ led to conclusions on the relative weight of σ and π currents in aluminum clusters that were shown to be incorrect.¹⁷

The introduction of orbital contributions to the bond current strengths to be discussed here would certainly help in the matter, allowing a quantitative clear-cut separation of σ and π currents. In a recent study, analyzing current profiles, Fliegl et al.¹⁸ reported that “calculations of the ring-current profile show that the ring current of benzene is not transported by the π electrons on both sides of the molecular ring” and that their “...calculations challenge the widespread notion that the ring current is transported by the π electrons...”. This is a disputable result which demands attention and needs to be carefully re-examined to understand whether a number of previous investigations make sense.

In the present paper we provide the partitioning of the current density bond cross section into π and σ contributions for typical aromatic, antiaromatic, and nonaromatic systems (benzene, planarized cyclooctatetraene, borazine), allowing for the same choice made to show the ring current signature in proton shielding density maps,⁶ with the aim to elucidate which electrons actually transport the ring current. Moreover, we consider cyclopropane and Al_4^{2-} , which have recently attracted considerable attention in the attempt to disclose the nature of their recognized ring current.^{15–20} According to Walsh,^{21,22} cyclopropane can be portrayed as a perturbed ethylene donating its π electrons to a methylene unit. The superposition of the three contributing structures leads to in-plane electron delocalization, which can only be plane symmetric. Alternatively, cyclopropane can be seen as formed by three bent C–C bonds (banana bonds).²³ Moreover, hyperconjugation can result from resonating structures involving the set of three π -like CH_2 group orbitals and this provides an out-of-plane electron delocalization which can be split into plane symmetric and plane antisymmetric contributions. As formerly suggested by Dewar,²⁴ the application of a perpendicular magnetic field switches on a ring current,¹⁹ which retains the symmetry of the electron delocalization from which it stems. By our approach, in-plane and out-of-plane circulations can be readily resolved and weighted. This will be discussed in a separate section hereafter. The aluminum cluster Al_4^{2-} is perhaps even more appealing, as the distinction between σ and π ring currents, recently resolved in favor of the former kind,¹⁷ can now be given on quantitative grounds, as shown in the following.

2. Method

Within the orbital approximation, the first-order current density induced by a magnetic field \mathbf{B} in the electrons of a molecule can be formally partitioned into orbital contributions:²⁵

$$\mathbf{J}^{\mathbf{B}}(\mathbf{r}) = \sum_j \mathbf{J}_j^{\mathbf{B}}(\mathbf{r}) \quad (1)$$

It is well-known that the total current density $\mathbf{J}^{\mathbf{B}} = \mathbf{J}^{\mathbf{B}}(\mathbf{r})$ is only exactly gauge invariant if the basis set is complete. In particular, the current density is independent of the choice of the gauge origin. This is not true, in general, for the individual orbital contributions, even in a complete basis set.

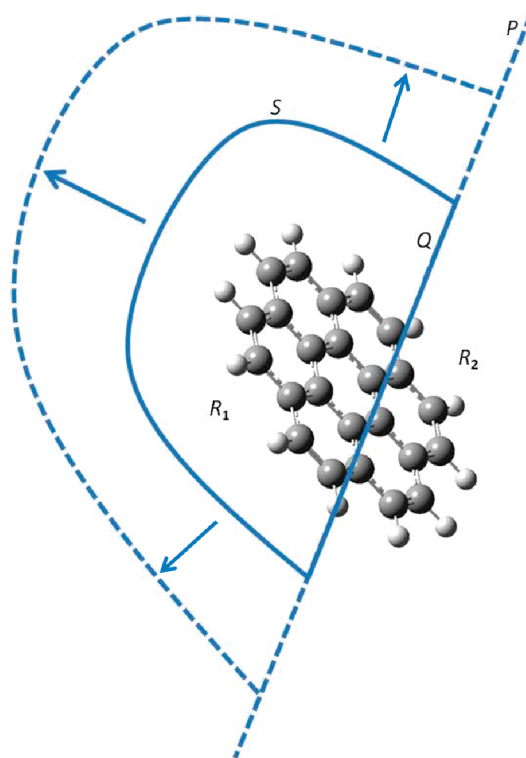


Figure 1

For practical applications with finite basis sets, methods have been devised to compute total origin-independent current densities. So far, mainly two strategies of calculation have been developed: one based on the continuous distribution of the origin of the current density^{26–29} and the other characterized by the use of gauge-including atomic orbitals (GIAO).^{13,30} The continuous transformation of the origin of the current density diamagnetic-zero (CTOCD-DZ) method, or *ipsocentric* method,^{31–33} is the simplest of the former kind.

In plane symmetric molecules in the presence of a perpendicular magnetic field, π and σ orbitals are symmetry distinct (see the Supporting Information for details). In this case the total current density is readily partitioned into π and σ contributions, regardless of the applied method. In particular, if the computed total current density is origin-independent, then also the π and σ current densities will be separately origin-independent.

Now let us consider the sketch in Figure 1, where a plane P , orthogonal to the plot plane, cuts the molecular space into two arbitrary regions, R_1 and R_2 . For any closed surface, formed by the portion Q of plane P and the surface S , enclosing the volume V of region R_1 , the Gauss theorem ensures that

$$\int_V \nabla \cdot \mathbf{J}^{\mathbf{B}} dv = \int_Q \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{q} + \int_S \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{s}$$

denoting by $d\mathbf{q}$ and $d\mathbf{s}$ the oriented element of area on Q and S , respectively. The intersection of plane P and the surface S with the plot plane is represented by a solid line in Figure 1. Since the divergence of the stationary current density vector vanishes everywhere, one finds that the sum of the two surface integrals on the right-hand side of the above relationship vanishes:

$$\int_Q \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{q} + \int_S \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{s} = 0$$

Expanding the volume V enclosing the region R_1 up to infinity, that is, taking an increasingly larger surface S and a wider portion Q of plane P , one has that $\lim_{Q \rightarrow P} \int_S \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{s} = 0$, a condition satisfied for well-behaved molecular wave functions going smoothly to zero at infinity. Consequently, the integral of the current density crossing *any* arbitrarily chosen plane cutting a molecular domain vanishes, i.e.

$$\int_P \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{q} = 0 \quad (2)$$

In practical calculations the field $\nabla \cdot \mathbf{J}^{\mathbf{B}}$ is not zero and eq 2 is not exactly fulfilled, except for symmetry reasons, as, for example, for every plane containing an even symmetry axis parallel to the inducing magnetic field \mathbf{B} , or when plane P is a symmetry plane perpendicular to \mathbf{B} . In all other cases the magnitude of the integral in eq 2 approaches zero on improving the quality of the calculation.

Let P be a plane bisecting at right angles a given bond between atoms K and L and let \mathbf{p} be a normal unitary vector

pointing from K to L. The cross section of the current density over P is given by

$$J_{\perp}^{\mathbf{B}} = \mathbf{J}^{\mathbf{B}}(\mathbf{r}) \cdot \mathbf{p} \quad (\mathbf{r} \in P) \quad (3)$$

The cross section $J_{\perp}^{\mathbf{B}}$ is a two-dimensional scalar field, having extremum points distributed around the center of the K–L bond. For each of these extrema a domain of integration of $J_{\perp}^{\mathbf{B}}$ can be defined as the area inside a contour line containing only that extremum point and no other point of maximum or minimum. The contour line value is set as close to zero as possible in agreement with the above condition. Then, if N is the number of domains, the net current strength for the selected bond is given by³⁴

$$\sum_{I=1}^N \int_I J_{\perp}^{\mathbf{B}} \cdot d\mathbf{p} \quad (4)$$

When the current flowing about a given bond is characterized by a substantially local vortical regime, the value of the integral in eq 4 for that bond will be vanishingly small.

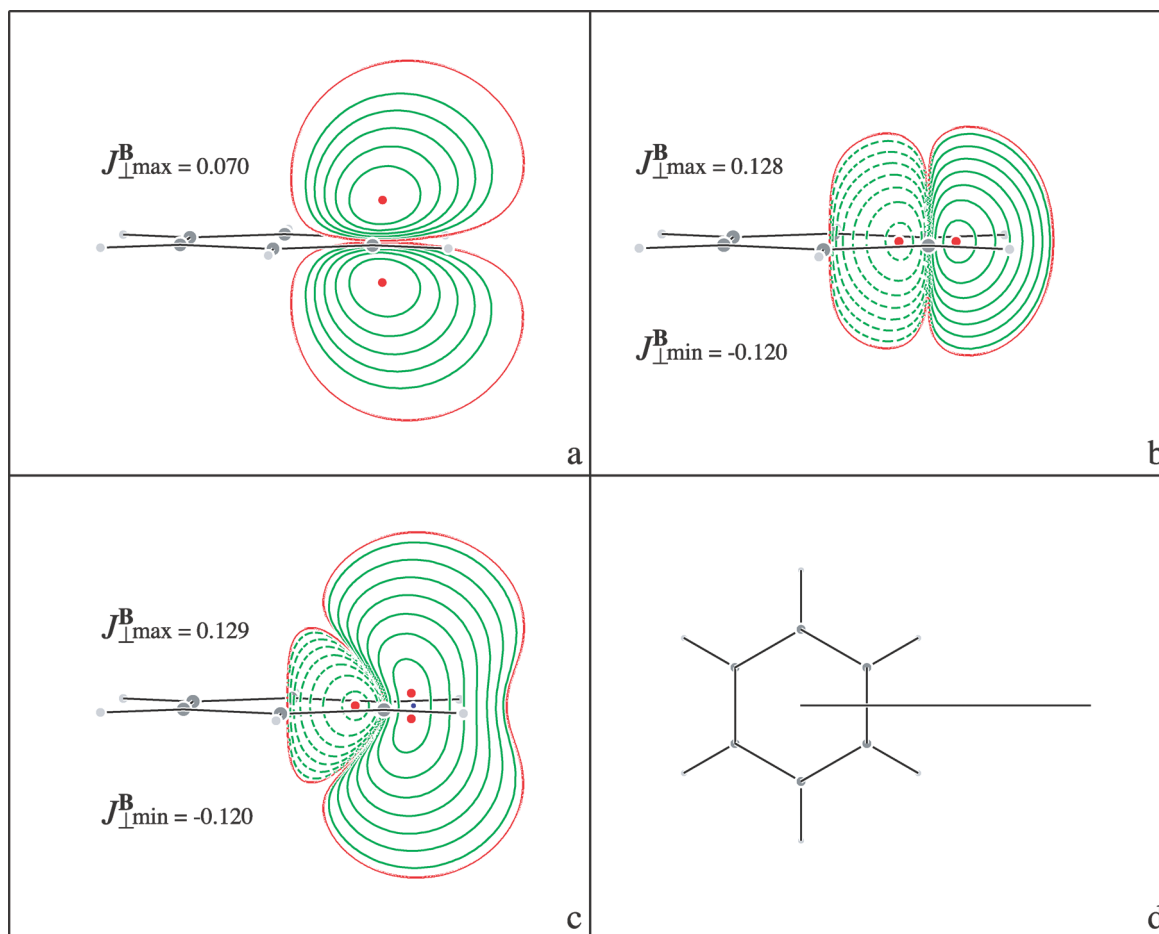


Figure 2. Current density cross sections for a plane perpendicular to a C–C bond in benzene and passing through the bond center: (a) π electron contribution; (b) σ electron contribution; (c) all electrons; (d) top view showing the profile of the cutting half-plane. The perturbing magnetic field of unitary magnitude is perpendicular to the molecular plane and in (a)–(c) is pointing from bottom to top. Solid (dashed) contour lines denote positive (negative) values of $J_{\perp}^{\mathbf{B}}$ corresponding to the outgoing (incoming) direction of the current density. Red dots locate a maximum of the cross section (within regions bound by solid contour lines) or a minimum (within regions bound by dashed contour lines), whose values are reported in atomic units. The small blue dot in (c) locates a saddle point in the outgoing current region. Green contour values are halved step by step, starting from the 80% inner extremum. The red contour at 10^{-3} au encloses the domain used to compute current strengths.

Table 1. Current Strengths (nA/T) Computed for Benzene^a

contribution	out	in	net
π	5.9 + 5.9	0	11.7
σ	7.9	-6.8	1.1
all-electron	17.7	-4.9	12.8

^a For each row, the sum of “out” and “in” terms matches the “net” result within numerical integration errors. This is also true for the sum of the σ and π contributions in the last column, but not for the columns labeled “out” and “in”, since the “out” and “in” integration domains in the all-electron cross section differ from those of the σ and π cross sections, as can be clearly observed in the figures.

Conversely, a value different from zero is a measure of current delocalization.

Calculations have been carried out for benzene, planarized cyclooctatetraene, borazine, cyclopropane, and Al_4^{2-} . For all the molecules but Al_4^{2-} , optimized geometries have been obtained at the RHF level using the (13s10p5d2f/8s4p1d) basis set of uncontracted Gaussian functions (see the Supporting Information for details). For Al_4^{2-} the optimal equilibrium geometry has been obtained at the RHF level using the 6-311+G(3df) basis set, as in ref 17. Origin-independent CTOCD-DZ current densities, computed at the coupled Hartree–Fock (CHF) level, using the same basis set adopted for the geometry optimization, have been

obtained and partitioned into π and σ contributions. Bond cross sections for π , σ , and total current densities have been computed via eq 3. Integration domains have been determined according to the above algorithm; π , σ , and total current strengths have been calculated via eq 4. Throughout this study, a magnetic field with a strength of 1 au is assumed within the linear approximation.

3. Results and Discussion

3.1. Benzene. Cross sections of the first-order current density induced by a unitary magnetic field perpendicular to the molecular plane of benzene are shown in Figure 2. These cross sections have been computed via eq 3 for the half-plane bound by the C_6 axis of symmetry, extending toward the tail molecular regions and bisecting one of the C–C bonds, as shown in panel (d).

As can be observed, the cross section of the total current density in panel c is in good agreement with the cross section reported in Figure 5 of ref 18. However, it should be remarked that (i) the total cross section is equal to the sum of the π and σ cross sections shown in panels (a) and (b), respectively, (ii) a substantial fraction of the outgoing total current returns through the region bound by the dashed contours as a result of the local diatropic vortex circulation

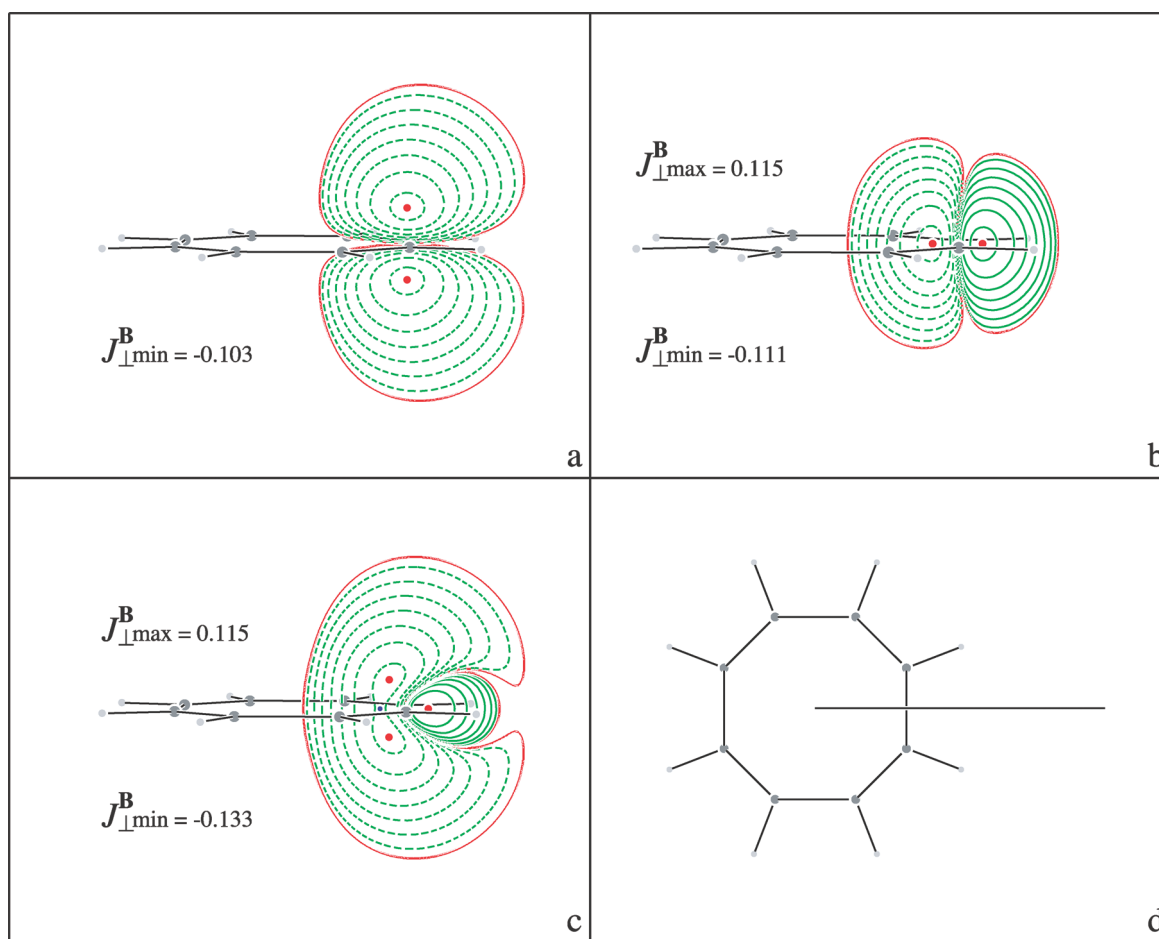


Figure 3. Current density cross sections for a plane perpendicular to a C–C single bond in planarized cyclooctatetraene and passing through the bond center: (a) π electron contribution; (b) σ electron contribution; (c) all electrons; (d) top view showing the profile of the cutting half-plane. See the Figure 2 caption for other details. The small blue dot in (c) locates a saddle point in the incoming current region.

Table 2. Current Strengths (nA/T) Computed for Planar Cyclooctatetraene^a

contribution	out	in	net
π	0	-9.0 - 9.0	-18.0
σ	6.5	-7.6	-1.1
all-electron	3.6	-22.8	-19.2

^a See footnote ^a of Table 1.

about the selected C–C bond (see Figure S1 of the Supporting Information), (iii) this localized regime of current is sustained only by the σ electrons, and (iv) the π current is fully delocalized as documented by the absence of incoming current in the cross section shown in panel (a).

In panel (c), the two red dots outside the carbon ring, on both sides of the molecular plane, correspond to maximum values of the total cross section. They result from the superposition of the two outgoing π current distributions, having maxima above and below the molecular plane, with the outgoing σ current distribution, having a single maximum over the molecular plane. Since the π current is vanishing over the molecular plane, the maximum value of the σ cross section becomes a *saddle point* in the total cross section, which is marked with a small blue dot in the figure. The superposition of the two outgoing π current distributions with the incoming σ current distribution inside the carbon ring does not alter the position of the minimum, which is the same in both σ and total cross sections.

In light of the above analysis, the claim made in ref 18 that “the strongest diatropic ring current flows on the outside of the ring and in the ring plane” cannot be justified by the current profile along a unique direction. Actually, starting from the maximum value of the σ cross section and moving perpendicularly to the molecular plane, the total cross section becomes greater, as required by the presence of the π ring current.

The current strengths collected in Table 1 provide a quantitative confirmation. They have been computed via eq 4, integrating the current density cross sections shown in Figure 2 over the domains enclosed within the red contour lines at 10^{-3} au. A smaller contour line value, i.e., larger integration domains, does not produce a significant effect on the computed current strengths. First of all, we observe a fairly good agreement between the all-electron current strengths and the results of ref 18 despite the different methods of calculation. Then we observe that the π contribution is fully delocalized, giving a net current strength of 11.7 nA/T, equally partitioned on both sides of the molecular plane.

The calculated σ current strengths are 7.9 nA/T for the outgoing current and -6.8 nA/T for the incoming current. A large fraction of these two current strengths is provided by the local diatropic vortex circulating about the C–C bond. Surely, the net current strength of 1.1 nA/T of the σ electrons

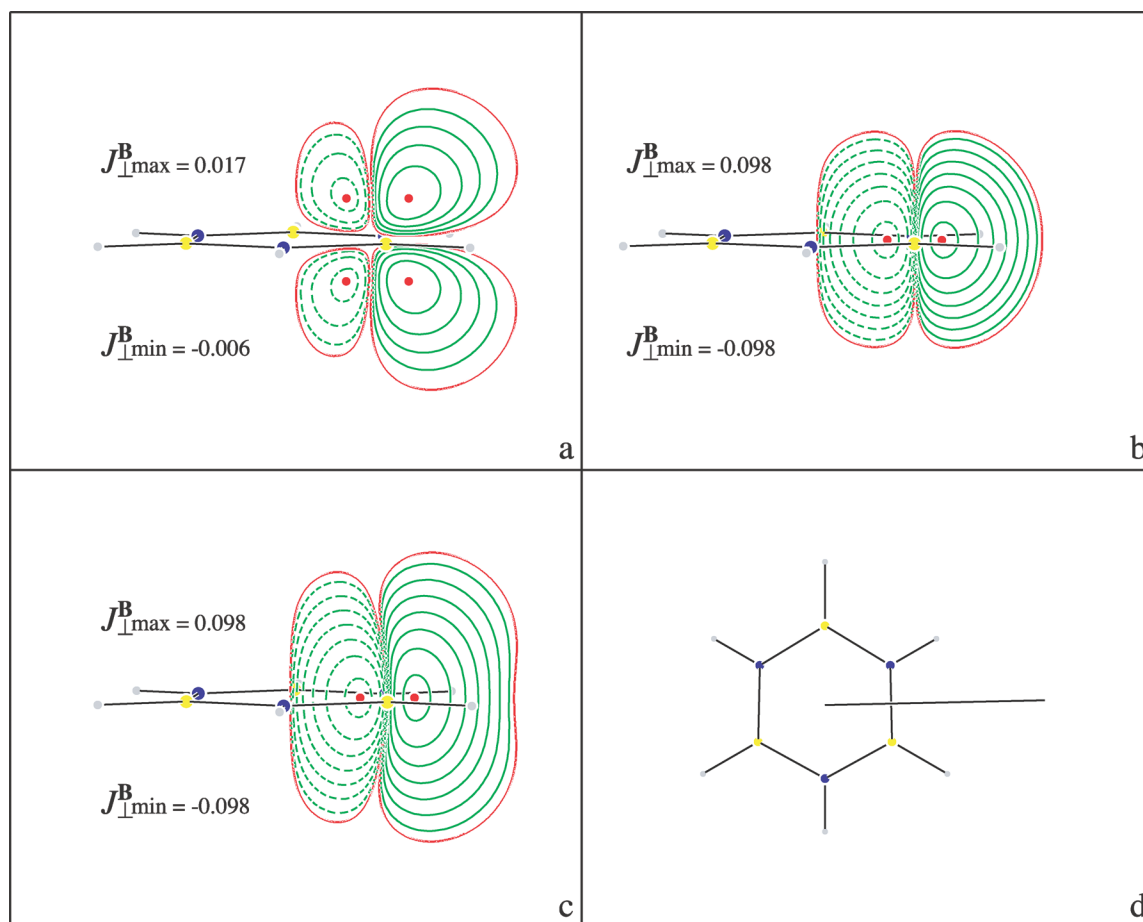


Figure 4. Current density cross sections for a plane perpendicular to a B–N bond in borazine and passing through the bond center: (a) π electron contribution; (b) σ electron contribution; (c) all electrons; (d) top view showing the profile of the cutting half-plane. B/N atoms are in yellow/blue. See the Figure 2 caption for other details.

Table 3. Current Strengths (nA/T) Computed for Borazine^a

contribution	out	in	net
π	1.2 + 1.2	-0.3 - 0.3	1.8
σ	6.2	-5.9	0.3
all-electron	8.5	-6.4	2.1

^a See footnote ^a of Table 1.

is delocalized, and therefore, it contributes to the diatropic ring current. The computed all-electron net current strength is 12.8 nA/T, which provides a quantitative measure of the diatropic ring current in benzene. From these figures, we can conclude that 91.4% of the diatropic ring current of benzene is transported by the π electrons and the remaining non-negligible 8.6% is sustained by the σ electrons.

3.2. Planarized Cyclooctatetraene. Cross sections of the first-order current density induced by a unitary magnetic field perpendicular to the molecular plane of planarized cyclooctatetraene are shown in Figure 3. These cross sections have been computed via eq 3 for the half-plane bound by the C_4 axis of symmetry, extending toward the tail molecular regions and bisecting one of the single C–C bonds, as shown in panel (d).

As can be observed in panel (a), the π electron cross section is reversed with respect to benzene and the absence of outgoing π current is consistent with a large paratropic

ring current. The σ cross section in panel (b) is similar to that in benzene, except that the incoming current seems to extend on a bit larger area than that of the outgoing current (see Figure S2 of the Supporting Information). The total cross section in panel (c) can be discussed much in the same way as in benzene. In this case the two red dots on both sides of the molecular plane appear inside the carbon ring and originate from the superposition of the incoming π and σ currents. The distance of these points from the molecular plane is larger than in benzene, documenting the larger effect of the paratropic π ring current.

Current strengths are reported in Table 2. Also in this case we observe a substantial agreement between the all-electron current strengths evaluated here and those of ref 18. The π contribution is fully delocalized, giving a net current strength of -18.0 nA/T, equally partitioned on both sides of the molecular plane.

The calculated σ current strengths are 6.5 nA/T for the outgoing current and -7.6 nA/T for the incoming current. In this case, the incoming current prevails over the outgoing current, giving a negative net current strength for the σ electrons of -1.1 nA/T, which contributes to the total paratropic ring current. The computed all-electron net current strength is -19.2 nA/T, which shows that the paratropic ring current of planar cyclooctatetraene is, in absolute value, 50%

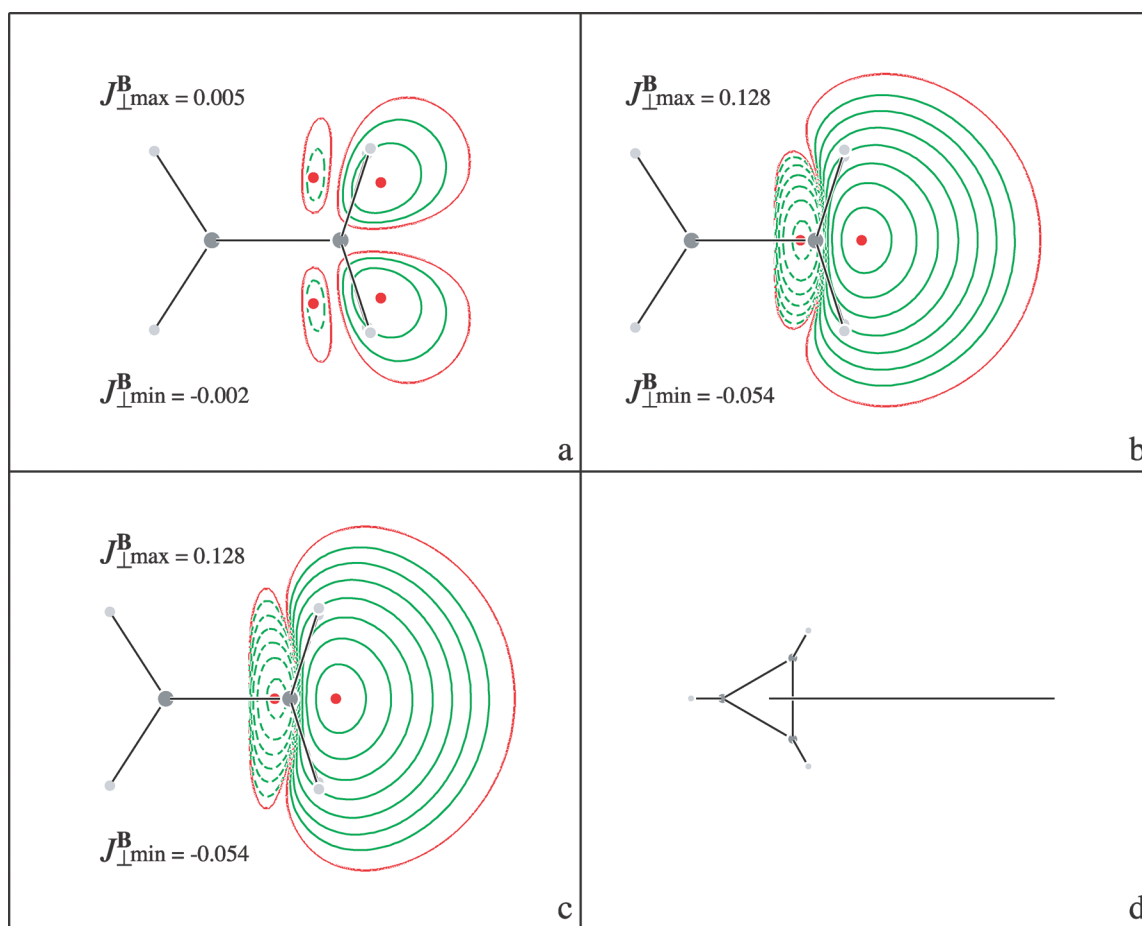


Figure 5. Current density cross sections for a plane perpendicular to a C–C bond in cyclopropane and passing through the bond center: (a) π -like contribution computed as the sum of plane antisymmetric current density orbital contributions; (b) σ contribution computed as the sum of plane symmetric current density orbital contributions; (c) all electrons; (d) top view showing the profile of the cutting half-plane. See the Figure 2 caption for other details.

Table 4. Current Strengths (nA/T) Computed for Cyclopropane^a

contribution	out	in	net
π -like	0.3 + 0.3	-0.0 - 0.0	0.5
σ	11.0	-1.3	9.7
all-electron	11.5	-1.4	10.2

^a See footnote ^a of Table 1.

stronger than in benzene. From these figures, we conclude that 94% of the paratropic ring current of planar cyclooctatetraene is transported by the π electrons and the remaining 6% is sustained by the σ electrons.

3.3. Borazine. Cross sections of the first-order current density induced by a unitary magnetic field perpendicular to the molecular plane of borazine are shown in Figure 4. These cross sections have been computed via eq 3 for the half-plane bound by the C_3 axis of symmetry, extending toward the tail molecular regions and bisecting one of the B–N bonds, as shown in panel (d).

Borazine represents an interesting example, as the non-aromatic character of the molecule is well documented by the π cross section shown in panel (a). In this case, a large fraction of the outgoing π current is connected to the incoming π current on both sides of the molecular plane, as a consequence of the substantially local vortical regime of

the π circulation. The σ cross section shown in panel (b) is almost identical to that of benzene. Owing to the much weaker π contribution, the total cross section in panel (c) does not show the red dot pair.

Current strengths are displayed in Table 3. As expected, the π contribution is small and partially reduced owing to the local character of the flow of the π current. Nonetheless, a net π current strength of 1.8 nA/T has been computed.

The calculated σ current strengths are 6.2 nA/T for the outgoing current and -5.9 nA/T for the incoming current. Therefore, the net current strength for the σ electrons is negligible, documenting an almost completely local regime of the current sustained by the σ electrons (see Figure S3 of the Supporting Information). The computed all-electron net current strength is 2.1 nA/T, which is typical of a weak diatropic ring current.

3.4. Cyclopropane. The nine valence molecular orbitals of cyclopropane can be split into six plane symmetric and three plane antisymmetric orbitals with respect to the σ_h plane of the carbon atoms. For a magnetic field perpendicular to that plane, the plane symmetric orbitals provide the σ contribution to the induced current density while the plane antisymmetric orbitals give a π -like contribution.

Cross sections of the first-order current density induced by a unitary magnetic field perpendicular to the molecular

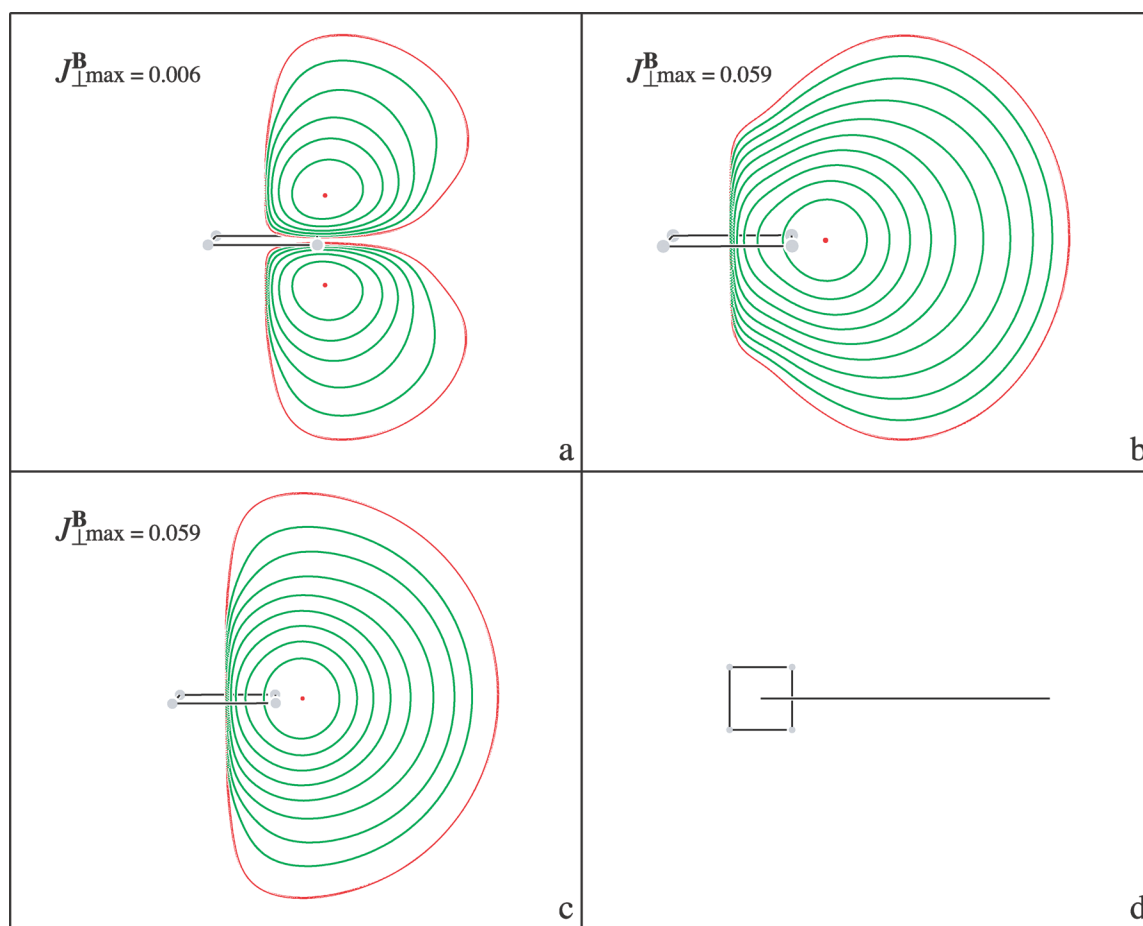


Figure 6. Current density cross sections for a plane perpendicular to an Al–Al bond in Al_4^{2-} and passing through the bond center: (a) contribution from the $2a_{2u}$ π HOMO; (b) contribution from the $5a_{1g} + 2b_{2g} + 4b_{1g} + 5e_u + 4a_{1g}$ σ valence orbitals; (c) all electrons; (d) top view showing the profile of the cutting half-plane. The red contour at 10^{-4} au encloses the domain used to compute current strengths. See the Figure 2 caption for other details.

Table 5. Current Strengths (nA/T) Computed for the Al_4^{2-} Cluster

contribution	out = net
π	2.1 + 2.1
σ	24.0
all-electron	28.4

plane of cyclopropane are shown in Figure 5. These cross sections have been computed via eq 3 for the half-plane bound by the C_3 axis of symmetry, extending toward the tail molecular regions and bisecting one of the C–C bonds, as shown in panel (d).

As can be observed in panel (a), the π -like cross section is weak and shows an outgoing current connected to a smaller incoming current on both sides of the molecular plane. The σ cross section displayed in panel (b) is much larger and, as in all previous cases, shows an outgoing current outside the carbon cycle and an incoming current inside the carbon cycle. Quite remarkably, for cyclopropane, incoming and outgoing σ currents are very different in regard to magnitude and extension, the external current largely prevailing over the internal current (see also Figure S4 of the Supporting Information). It should also be noted that the σ cross section is rather elongated on both sides of the σ_h plane of the carbon atoms up to the region about the hydrogen atoms. This fact, together with the bifurcation of the cyclopropane ring current in proximity of the methylene unit,^{20,35} suggests that both in-plane and out-of-plane circulations take place. The comparison of the total cross section in panel (c) with the σ cross section in panel (b) shows only minimal differences, thus revealing the very low importance of the π -like current.

Current strengths are reported in Table 4. As can be seen, the π -like current strength is negligible, its net value being as small as 0.5 nA/T. The calculated σ current strengths are 11.0 nA/T for the outgoing current and -1.3 nA/T for the incoming current. The large difference accounts for a net diatropic σ ring current whose strength (9.7 nA/T) is comparable to that of the π ring current of benzene.

The calculated all-electron current strengths given in the last row of Table 4 are in good agreement with the results of ref 18. Considering the all-electron net current strength of 10.2 nA/T, we can conclude that 95% of the diatropic ring current of cyclopropane is transported by the σ electrons and the remaining 5% is sustained by the π -like electron distribution. There is convincing evidence that the cyclopropane σ ring current is composed by in-plane and out-of-plane circulations.

3.5. Al_4^{2-} . The Al_4^{2-} subunit was initially indicated to be a π aromatic species, allowing for its detection by photoelectron spectroscopy.³⁶ On the other hand, direct calculation of the induced current density in Al_4^{2-} , using the ipsocentric method, confirmed the existence of a diatropic ring current, but almost entirely attributed to σ electrons.³⁷ This result was contradicted by CMO-NICS(0) calculations,¹⁵ which yield a π contribution to the NICS (-17.8) larger than the σ one (-11.1), and by bond current strength calculations, performed with the GIMIC method,¹⁶ which show substantially equal π and σ contributions. In a successive analysis,¹⁷ both the CMO-NICS(0) and the GIMIC conclusions were

shown to be largely incorrect and the prevailing role of σ electrons with respect to a (however) nonvanishing contribution from π electrons was reaffirmed. It is now interesting to check the validity of the latter result against the quantitative predictions provided by our approach.

Cross sections of the first-order current density induced by a unitary magnetic field perpendicular to the molecular plane of Al_4^{2-} are shown in Figure 6. These cross sections have been computed via eq 3 for the half plane bound by the C_4 axis of symmetry, extending toward the tail molecular regions and bisecting one of the Al–Al bonds, as shown in panel (d).

The absence of incoming currents in both π and σ cross sections, shown in panels (a) and (b), respectively, is consistent with a complete delocalization of the currents sustained by the valence electrons of Al_4^{2-} . However, the σ cross section is much larger, by nearly 1 order of magnitude.

Current strengths are reported in Table 5. As can be seen, the π current strength is not negligible, its value being 4.2 nA/T. The calculated σ current strength is 24.0 nA/T. The total current strength of 28.4 nA/T is in rather nice agreement with the value reported in ref 16 of 28.1 nA/T, despite the different method and basis set.

From these figures, we conclude that 85% of the diatropic ring current of Al_4^{2-} is transported by the σ valence electrons and the remaining 15% is sustained by the two π valence electrons. This result is in good agreement with the conclusion reported in refs 37 and 17 depicting Al_4^{2-} as a σ aromatic molecule.

4. Conclusions

Origin-independent CTOCD-DZ current densities induced by a magnetic field perpendicular to the σ_h planes of benzene, planarized cyclooctatetraene, borazine, cyclopropane, and Al_4^{2-} have been partitioned into π and σ contributions. In benzene and planar cyclooctatetraene, current density bond cross sections show that the π electrons sustain fully delocalized currents whereas the σ electrons sustain largely localized currents. These results confirm the widespread notion that the ring current in aromatic benzene and antiaromatic cyclooctatetraene is transported by the π electrons on both sides of the ring. A small fraction of the total ring current is transported also by the σ electrons. In the nonaromatic borazine the π electron current is mainly localized around nitrogen atoms; however, a small fraction circulates all over the ring. The ring current of cyclopropane has a net σ character. The Al_4^{2-} cluster is confirmed to be mainly σ aromatic on the magnetic criterion.

In the present paper we show that within the orbital approximation the π and σ contributions to the current can be quantitatively estimated, which will be helpful in all cases in which the two contributions are of comparable magnitude.

Acknowledgment. Financial support provided by MURST (Ministero dell'Università e della Ricerca Scientifica e Tecnologica), via FARB and PRIN funds, is gratefully acknowledged.

Supporting Information Available: Proof of the clear-cut partition of the first-order magnetic field induced current

density distribution into π and σ contributions for systems containing a symmetry plane immersed in a perpendicular magnetic field, optimized Cartesian coordinates of the investigated molecules, sum rules for charge conservation, and current density maps on the molecular planes for a unitary and perpendicular magnetic field. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Ligabue, A.; Pincelli, U.; Lazzeretti, P.; Zanasi, R. *J. Am. Chem. Soc.* **1999**, *121*, 5513–5518.
- (2) Steiner, E.; Fowler, P. W.; Jenneskens, L. W. *Angew. Chem., Int. Ed.* **2001**, *40*, 362–366.
- (3) Steiner, E.; Fowler, P. W. *ChemPhysChem* **2002**, *3*, 114–116.
- (4) Monaco, G.; Fowler, P. W.; Lillington, M.; Zanasi, R. *Angew. Chem., Int. Ed.* **2007**, *46*, 1889–1892.
- (5) Steiner, E.; Fowler, P. W. *Phys. Chem. Chem. Phys.* **2004**, *6*, 261–272.
- (6) Soncini, A.; Fowler, P. W.; Lazzeretti, P.; Zanasi, R. *Chem. Phys. Lett.* **2005**, *401*, 164–169.
- (7) Viglione, R. G.; Zanasi, R.; Lazzeretti, P. *Org. Lett.* **2004**, *6*, 2265–2267.
- (8) Schleyer, P. v. R.; Jiao, H.; Hommes, N. J. R. v. E.; Malkin, V. G.; Malkina, O. L. *J. Am. Chem. Soc.* **1997**, *119*, 12669–12670.
- (9) Schleyer, P. v. R.; Manoharan, M.; Wang, Z.-X.; Kiran, B.; Jiao, H.; Puchta, R.; Hommes, N. J. R. v. E. *Org. Lett.* **2001**, *3*, 2465–2468.
- (10) Schleyer, P. v. R. *Chem. Rev.* **2001**, *101*, 1115–1117.
- (11) Gomes, J. A. N. F.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349–1383.
- (12) Zonta, C.; De Lucchi, O. *Eur. J. Org. Chem.* **2006**, 449–452.
- (13) Jusélius, J.; Sundholm, D.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 3952–3963.
- (14) Stanger, A. *J. Org. Chem.* **2010**, *75*, 2281–2288.
- (15) Chen, Z.; Corminboeuf, C.; Heine, T.; Bohmann, J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2003**, *125*, 13930–13931.
- (16) Lin, Y.-C.; Jusélius, J.; Sundholm, D. *J. Chem. Phys.* **2005**, *122*, 214308–214316.
- (17) Havenith, R. W.; Fowler, P. W. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3383–3386, and references therein.
- (18) Fliegl, H.; Sundholm, D.; Taubert, S.; Jusélius, J.; Klopper, W. *J. Phys. Chem. A* **2009**, *113*, 8668–8676.
- (19) Fowler, P. W.; Baker, J.; Lillington, M. *Theor. Chem. Acc.* **2007**, *118*, 123–127.
- (20) Pelloni, S.; Lazzeretti, P.; Zanasi, R. *J. Phys. Chem. A* **2007**, *111*, 8163–8169.
- (21) Walsh, A. D. *Nature* **1947**, *159*, 165–165.
- (22) Walsh, A. D. *Nature* **1947**, *159*, 712–713.
- (23) Coulson, C. A.; Moffit, W. E. *J. Chem. Phys.* **1947**, *15*, 151–151.
- (24) Dewar, M. J. S. *J. Am. Chem. Soc.* **1984**, *106*, 669–682.
- (25) Stevens, R. M.; Pitzer, R. M.; Lipscomb, W. N. *J. Chem. Phys.* **1963**, *38*, 550–560.
- (26) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, *210*, 223–231.
- (27) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Chem. Phys. Lett.* **1994**, *220*, 299–304.
- (28) Zanasi, R.; Lazzeretti, P.; Malagoli, M.; Piccinini, F. *J. Chem. Phys.* **1995**, *102*, 7150–7157.
- (29) Soncini, A.; Lazzeretti, P.; Zanasi, R. *Chem. Phys. Lett.* **2006**, *421*, 21–26.
- (30) Johansson, M.; Jusélius, J.; Sundholm, D. *Angew. Chem., Int. Ed.* **2005**, *44*, 1843–1846.
- (31) Steiner, E.; Fowler, P. W. *Chem. Commun.* **2001**, 2220–2221.
- (32) Steiner, E.; Fowler, P. W. *J. Phys. Chem. A* **2001**, *105*, 9553–9562.
- (33) Havenith, R. W. A.; Fowler, P. W. *Chem. Phys. Lett.* **2007**, *449*, 347–353.
- (34) Monaco, G.; Zanasi, R. *AIP Conf. Proc.* **2009**, *1148*, 425–428.
- (35) Carion, R.; Champagne, B.; Monaco, G.; Zanasi, R.; Pelloni, S.; Lazzeretti, P. *J. Chem. Theory Comput.* **2010**, *6*, 2002–2018.
- (36) Li, X.; Kuznetsov, E.; Zhang, H.-F.; Boldyrev, A. I.; Wang, L.-S. *Science* **2001**, *291*, 859–861.
- (37) Fowler, P. W.; Havenith, R. W. A.; Steiner, E. *Chem. Phys. Lett.* **2001**, *342*, 85–90.

CT100442J

Geometry Optimization with Multilayer Methods Using Least-Squares Minimization

Wenkel Liang,[†] Craig T. Chapman,[†] Michael J. Frisch,[‡] and Xiaosong Li^{*†}

Department of Chemistry, University of Washington, Seattle, Washington 98195-1700, United States and Gaussian Incorporated, 340 Quinnipiac Street, Building 40, Wallingford, Connecticut 06492, United States

Received August 13, 2010

Abstract: In this article, we introduce a least-squares minimization scheme for optimizing molecular structures with mixed quantum mechanics (QM) and molecular mechanics (MM) multilayer models. A mixed-coordinate optimization framework was developed. The QM and MM regions are modeled with redundant internal coordinates and Cartesian coordinates, respectively. Within this mixed-coordinate system, a least-squares minimization method using the quasi-Newton step as the evaluation of error is constructed. The couplings between layers are treated rigidly in accordance with the mechanical embedding approach, and the MM Hessian is approximated as a scalar constant of the root-mean-square QM Hessian eigenvalues. Both two-layer and three-layer models were tested. The performance of the method developed herein shows consistently stable and fast convergence.

I. Introduction

Computational science has been very successful in many research areas, including chemistry, physics, materials, and biology. Computational efforts are applied to large systems of ever increasing size, such as biomolecules, polymers, and nanostructures with hundreds to thousands of atoms and electrons. Molecular mechanics, which scales linearly with respect to the system size at the cost of computational accuracy, is often the method of choice for large-scale systems. On the other hand, as the accuracy becomes more and more important in computational research, there is a strong need for using first-principles based methods for studying electronic structures of large systems. However, the requirement of computational resources for this need has outpaced advances in computer hardware because *ab initio* calculations scale nonlinearly with respect to the system size. A method that strikes a better balance between computational cost and accuracy is the mixed quantum mechanics and molecular mechanics (QM/MM) model^{1–4} or the ONIOM model, which has shown enormous success in biochemical research fields (see refs 5–9 for recent reviews on QM/MM).

In the QM/MM model, first-principles methods are used to describe electronic structures of chemically important subspaces while the rest of system is modeled by molecular mechanics force field theories. The interaction between these two subspaces is treated with embedding methods, such as the mechanical and electrostatic embedding approaches.

While energetics calculations based on the QM/MM or ONIOM model have been well defined in the literature, geometry optimization methods for such models are less frequently addressed, due to a 2-fold difficulty. First, the MM subspace often consists of thousands of atoms. An optimization method must be able to handle such a large geometric space efficiently and at an affordable computational cost. Second, optimization of the QM subspace is often carried out in internal coordinates with redundancy,^{10–13} while the MM subspace is described by nonredundant Cartesian coordinates. Geometry optimization in such a mixed-coordinate space is generally difficult, and an algorithm often treats the two subspaces independently.^{14–18} For example, the microiteration algorithm assumes that the QM subspace is stationary when optimizing the MM region. A conventional quasi-Newton based optimization method is then taken to optimize the QM region while holding the MM region intact.¹⁹ Such an optimization scheme is rather successful and has become the standard for optimization of multilayer

* Corresponding author e-mail: li@chem.washington.edu.

[†] University of Washington.

[‡] Gaussian Incorporated.

mixed QM and MM models. A more recent effort by Vreven et al. computes explicit coupling between QM and MM layers in the Hessian matrix.²⁰ This approach is very useful in transition-state searches but is generally computationally expensive due to the explicit computation and handling of the full second-derivative matrix.

In this article, we introduce a different optimization concept for QM/MM or ONIOM models. We will present a geometry optimization algorithm that optimizes both QM and MM subspaces simultaneously. The underlying motivation is to develop a method that is generalized to be able to handle a mixed-coordinate system and is essentially independent of the construction or division of the system into layers. The following notations will be used in this article extensively: \mathbf{R} and \mathbf{g}_R are vector and energy gradients in the general coordinate subspace; \mathbf{Q} and \mathbf{g}_Q are vector and energy gradients in redundant internal coordinates for the quantum mechanical (QM) subspace; \mathbf{X} and \mathbf{g}_X are vector and energy gradients in Cartesian coordinates for the molecular mechanical (MM) subspace; $\{c_i\}$ are variational parameters for the least-squared minimization; \mathbf{H} is the Hessian (energy second derivatives).

II. Algorithm

In the search space defined by $\{\mathbf{R}_i\}$, an optimization algorithm can be designed to look for a new vector \mathbf{R}^* that is associated with the minimal error. One of the most popular forms of the desired new vector in geometry and wave function optimization is the linear combination of the search space

$$\mathbf{R}^* = \sum_i c_i \mathbf{R}_i \text{ and } \sum_i c_i = 1 \quad (1)$$

With the definition of the new vector in eq 1, an associated error function $f(\mathbf{R}^*)$ can be constructed. Minimizing the square of $f(\mathbf{R}^*)$ with respect to variational coefficients $\{c_i\}$ leads to a set of coupled equations, which can be solved for $\{c_i\}$ that minimize f^2 . Optimization algorithms using eq 1 as the new vector are often referred to as the direct inversion in the iterative subspace (DIIS) method.^{21–24}

The key to the success of the DIIS algorithm is the choice of error function. In the context of geometry optimization, the error function f can be a measure of energy ΔE ,^{25–27} gradient \mathbf{g} ,^{21,22} or distance $\Delta \mathbf{R}$ ^{23,24,27} to the minimum on the local potential energy surface (PES). In the current paper, the following error function is considered

$$f(\mathbf{R}^*) = \sum_{i=1}^k c_i \Delta \mathbf{R}_i \quad (2)$$

Equation 2 uses the distance $\Delta \mathbf{R}$ to the local minimum as an evaluation of the error. $\Delta \mathbf{R}$ can be calculated with the Newton–Raphson or quasi-Newton approach

$$\Delta \mathbf{R}_i^T = -\mathbf{g}_i^T \cdot \mathbf{H}^{-1} \quad (3)$$

The stability of the quasi-Newton approach can be enhanced by controlling the step size using the rational function optimization technique (RFO)^{28,29} or the trust radius model

(TRM).^{30–36} According to the nature of this method, we will refer to it as quasi-Newton-DIIS (QN-DIIS). Minimization of eq 2 can be performed in a least-squares sense, written in a linear algebraic equation in matrix form²²

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,k} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{k,1} & \dots & a_{k,k} & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_k \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad (4)$$

where $a_{ij} = (\mathbf{g}_i^T \cdot \mathbf{H}^{-1}) \cdot (\mathbf{g}_j^T \cdot \mathbf{H}^{-1})^T$. Solving eq 4 is computationally trivial, and the solution $\{c_i\}$ minimizes the square of the error function in eq 2. Equations 1–4 illustrate the least-squares minimization approach for geometry optimization with a uniform coordinate system. For a mixed-coordinate system, such as those described by the QM/MM or ONIOM model, the QM and MM subspaces are represented by the internal (\mathbf{Q}) and Cartesian (\mathbf{X}) coordinates, respectively. Internal coordinate space is known for its fast convergence, as it is a natural choice for molecular structures and vibrations. However, internal coordinate spaces are often associated with a large number of redundancies. Such redundancy is certainly not computationally cost effective for the large MM subspace, which often includes thousands of atoms. Therefore, a widely used approach is to represent the MM subspace using the Cartesian coordinates. We herein define a generalized QM/MM or ONIOM coordinate space

$$\mathbf{R} = (\mathbf{Q} \ \mathbf{X})^T \quad (5)$$

$$\mathbf{g} = (\mathbf{g}_Q \ \mathbf{g}_X)^T \quad (6)$$

The DIIS search space is then defined by $\{\mathbf{R}_i\}$ as usual.

The QN-DIIS approach as described in eqs 1–4 requires computation of the Hessian

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{QQ} & \mathbf{H}_{QX} \\ \mathbf{H}_{XQ} & \mathbf{H}_{XX} \end{pmatrix} \quad (7)$$

where \mathbf{H}_{QQ} and \mathbf{H}_{XX} are Hessian matrices for the QM and MM subspaces, respectively, and $\mathbf{H}_{QX} = \mathbf{H}_{XQ}$ are related to the couplings between these two subspaces. For the relatively small QM subspace, computationally expensive analytical evaluations of the second derivatives are replaced with a numerical Hessian update scheme, such as BFGS,^{37–40} SR1,⁴¹ and PSB,^{42,43} to obtain an accurate QM Hessian \mathbf{H}_{QQ} . However, due to its large dimension, calculation and manipulation of the MM Hessian \mathbf{H}_{XX} are not computationally practical. One simple approach is to approximate \mathbf{H}_{XX} as a scaled identity matrix $\alpha \cdot \mathbf{I}$. In this implementation, the MM weighting factor α takes on the following form

$$\alpha = \sqrt{\frac{\sum_i^M \lambda_{ii}^2}{M}} \quad (8)$$

where λ 's are the Hessian eigenvalues of the QM subspace and M is the number of nonredundant QM coordinates. Equation 8 is constructed in order to equally weigh both QM and MM subspaces in the DIIS equation.

In principle, the off-diagonal coupling term \mathbf{H}_{QX} in eq 7 needs to be computed explicitly as it gives rise to the interaction between the two subspaces. Within the framework of QM/MM or ONIOM, various embedding approaches are used to treat the two subspaces independently via linking atoms and/or electrostatic point charges. To some extent, such a treatment implicitly decouples the QM and MM subspaces. Therefore, one can conveniently choose $\mathbf{H}_{\text{QX}} = 0$ in eq 7.

With approximations described above, eq 7 can be rewritten as

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\text{QQ}} & 0 \\ 0 & \alpha \cdot \mathbf{I} \end{pmatrix} \quad (9)$$

The quasi-Newton step in the QM/MM space can be defined as

$$\Delta \mathbf{R}^T = -\mathbf{g}_{\text{Q}}^T \cdot \mathbf{H}_{\text{QQ}}^{-1} - \mathbf{g}_{\text{X}}^T \cdot \alpha^{-1} \quad (10)$$

With this definition, the QN-DIIS equation (eq 4) can be easily constructed and solved and an optimization step can be taken as

$$\mathbf{R}_{N+1} = \mathbf{R}^* + \Delta \mathbf{R}^* = \sum_{i=1}^N c_i \cdot \mathbf{R}_i - \sum_{i=1}^N c_i \cdot \mathbf{g}_i^T \cdot \mathbf{H}^{-1} \quad (11)$$

The performance of the optimization scheme using DIIS critically depends on the quality of the approximated Hessian. Negative or saddle point Hessians are generally difficult to correct using the DIIS method. In the GDIIS method previously developed by Farkas,²⁴ additional controls on the direction and step size of the DIIS step were implemented. These controls were used to correct for possible bad Hessian approximations and to ensure a downhill optimization. In this implementation, these controls were replaced by using the standard RFO approach to preoptimize the structure into a near quadratic potential well with a positive definite Hessian in the spirit of the hybrid optimization method.²⁶ Nevertheless, the QN-DIIS method implemented here is formally equivalent to the GDIIS method in ref 24. On the other hand, one could, in principle, use the energy-based linear squares minimization approach, such as the GEDIIS method.²⁶ However, the GEDIIS approach for mixed-coordinate systems is not ideal in this scenario mainly because changes of energy in the Cartesian coordinate system along various degrees of freedom are strongly coupled.

After the RFO approach preoptimizes the structure to near-quadratic well, the overall optimization algorithm for QM/MM proceeds as follows. Gradients for both QM and MM subspaces are computed in the Cartesian coordinates. The QM Cartesian gradients are transformed into the internal coordinate space. The QM Hessian is updated, and the MM scaling factor α is calculated (eqs 8 and 9). The quasi-Newton step of the whole QM/MM or ONIOM space is constructed for all saved points using the current Hessian (eq 10). The DIIS equation is solved, and a new optimization step is taken (eq 11). The QM internal coordinates are transformed back to the Cartesian coordinates for gradient calculations. MM microiterations are carried out so that MM

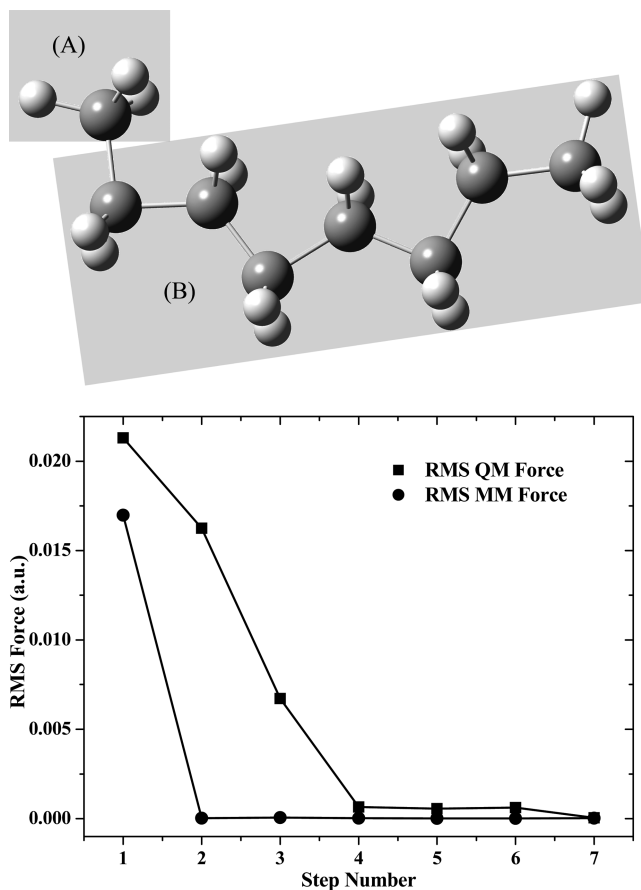


Figure 1. Force convergence profiles during optimizations of the two-layer ONIOM model of C_8H_{18} . Layer A is modeled with the HF/STO-3G level of theory, and layer B is modeled with the Amber⁴⁵ force field.

gradients are at the same convergence level as the QM gradients ($\mathbf{g}_{\text{X}} \leq \mathbf{g}_{\text{Q}}$).

III. Benchmarks and Discussion

Optimizations were carried out on a SGI Atlix 450 workstation (Intel dual core-Itanium 1.6 GHz with 48GB of RAM) using the development version of the GAUSSIAN⁴⁴ suite of programs with the implementation of the QM/MM optimization algorithm presented here. For all methods tested here, the optimization is considered converged when the maximum \mathbf{g} is less than 0.00045 au, the root-mean-square (rms) \mathbf{g} is less than 0.00030 au, the maximum $\Delta \mathbf{R}$ is less than 0.0018 au, and the rms $\Delta \mathbf{R}$ is less than 0.0020 au. The hybrid optimization scheme is used in this work. Simple RFO steps are taken to search for a good local potential well before the DIIS algorithm is used to optimize the system to the minimum (see refs 26 and 27 for details). Usually, the initial RFO search for a local potential well takes one-third to one-fifth of the total number of geometry optimization steps.

The first test case is a two-layer model of the C_8H_{18} molecule (Figure 1) using the ONIOM approach. The embedding method is mechanical with hydrogen as the linking atom. In this model, the ending methyl group is modeled with the HF/STO-3G level of theory, while the rest of system is modeled with Amber force field theory. Convergence profiles of rms \mathbf{g}_{Q} and \mathbf{g}_{X} are plotted in Figure

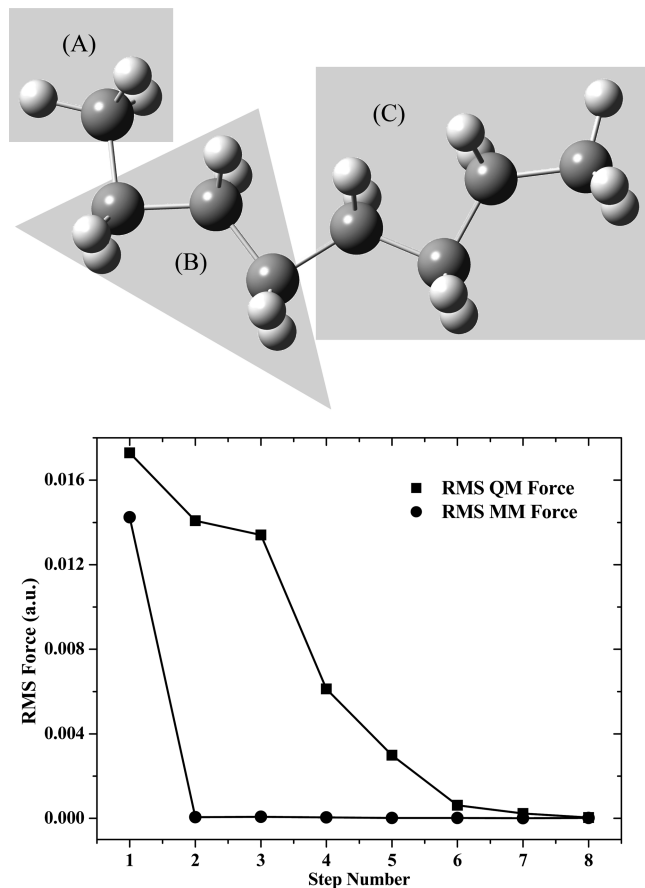


Figure 2. Force convergence profiles during optimizations of the three-layer ONIOM model of C_8H_{18} . Layer A is modeled with the B3LYP/STO-3G level of theory; layer B is modeled with the HF/STO-3G level of theory; layer C is modeled with the Amber⁴⁵ force field.

1. As described in the previous section, several MM microiterations using the conjugated gradient method are carried out in each optimization step. As a result, the rms \mathbf{g}_x converges faster than the rms \mathbf{g}_Q . The overall optimization profile is smooth and efficient for this two-layer test case. Optimizations for three-layer ONIOM systems are more difficult as the middle layer is often connected to both the higher and lower layers. Such a situation adds an implicit constraint and complexity to the optimization problem. Figure 2 depicts a three-layer ONIOM model for the C_8H_{18} molecule, where B3LYP/STO-3G, HF/STO-3G, and Amber force field theories are used to model the three layers. As an extension to the method described above, all QM layers are treated as independent subspaces, provided with the analytical first derivatives and updated Hessians for each layer. The couplings between QM layers are through the mechanical embedding method with hydrogen as the linking atom. The full gradient and Hessian for such a three-layer case can be written as

$$\mathbf{g} = (\mathbf{g}_{Q-I} \mathbf{g}_{Q-II} \mathbf{g}_X)^T \quad (12)$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{QQ-I} & 0 & 0 \\ 0 & \mathbf{H}_{QQ-II} & 0 \\ 0 & 0 & \alpha \mathbf{I} \end{pmatrix} \quad (13)$$

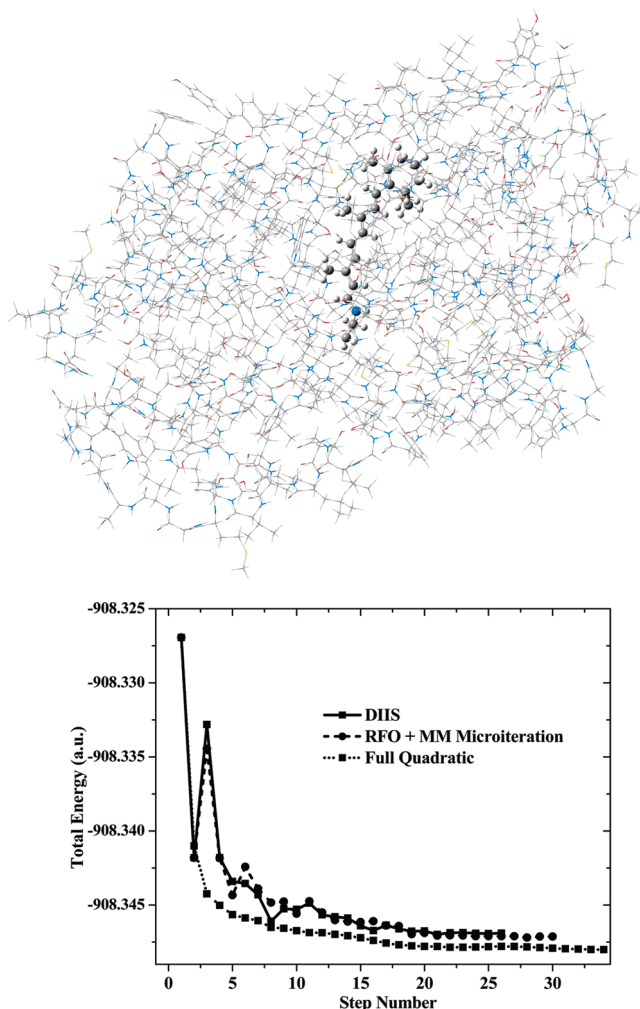


Figure 3. Comparison of optimization energy profiles using the multiplayer DIIS, conventional RFO, and full quadratic approaches for the two-layer ONIOM model of bacteriorhodopsin. The high-level layer (ball-and-stick) is modeled with the B3LYP/3-21G level of theory; the low-level layer (wire-frame) is modeled with the Amber⁴⁵ force field.

where \mathbf{H}_{QQ-I} and \mathbf{H}_{QQ-II} are updated Hessians for the two QM layers and α is computed using eq 8 with a summation over all QM variables. The performance of the optimization method introduced here for a three-layer ONIOM model is shown in Figure 2. With the three-layer ONIOM model, the optimization method also exhibits smooth and fast convergence behavior similar to that of a two-layer model. For both two-layer and three-layer models, the fast convergence of the MM force at the beginning of the optimization process is a result of the microiteration approach for the MM region. The importance of the DIIS approach comes into play at the later stage of the optimization when both QM and MM layers are in the near-quadratic potential well. For the test cases presented herein (Figures 1 and 2), the DIIS optimization scheme starts from the fourth step.

Figure 3 shows a two-layer ONIOM model for the bacteriorhodopsin. In this system, there are 3568 atoms modeled by Amber force field and 56 atoms by B3LYP/3-21G. The test is compared with the conventional RFO approach with microiterations to fully converge the MM layer at each step and the full quadratic method.²⁰ All three

methods considered here are able to converge such a large system within a reasonable number of steps (<34 geometry steps). Since the energy difference between the three optimized structures is <0.5 kcal/mol, they are considered to be at the same potential minimum. The optimization pathways are very similar in both RFO and DIIS test cases as they are quasi-Newton approach based methods. There are a number of unproductive oscillations at the beginning of the optimization process exhibited by both the RFO and the DIIS approaches. This is due to a lack of explicit coupling in the full Hessian between layers. On the other hand, with a proper treatment of the coupling, the full quadratic optimization is very smooth, however, at a much higher computational cost. The overall computational cost taken by the full quadratic optimization is more than three times that of the RFO or DIIS approach. While both RFO and DIIS methods take a similar number of steps to finish the test job and converge to a same minimum, the multilayer DIIS methods costs less than the conventional RFO method. In the multiplayer DIIS method, the MM region does not need to be converged at each optimization step because the MM geometry errors are taken into account in the least-squares minimization scheme. As a result, the overall computational savings using the method introduced here is about 25% compared to that of using the RFO method with microiterations.

IV. Conclusion

In this article, we introduced a method for optimizing molecular structures with multilayer models, such as QM/MM and ONIOM. A mixed-coordinate least-squares optimization framework was developed and generalized for QM and MM regions represented by redundant internal coordinates and Cartesian coordinates, respectively. Both two-layer (QM:MM) and three-layer (QM:QM:MM) models were tested. The performance of the method developed herein shows consistently smooth and fast convergence.

Acknowledgment. This work was supported by the U.S. National Science Foundation (CHE-CAREER 0844999 to X.L.). Additional support from Gaussian Inc. and the University of Washington Student Technology Fund is gratefully acknowledged.

References

- Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
- Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.
- Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.
- Gao, J. *Acc. Chem. Res.* **1996**, *29*, 298.
- Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.
- Shurki, A.; Warshel, A.; Valerie, D. Structure/Function Correlations of Proteins using MM, QM/MM, and Related Approaches: Methods, Concepts, Pitfalls, and Current Progress. In *Advances in Protein Chemistry*; Academic Press: New York, 2003; Vol. 66, pp 249.
- Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.
- Vreven, T.; Morokuma, K.; David, C. S. Hybrid Methods: ONIOM(QM:MM) and QM/MM. In *Annual Reports in Computational Chemistry*; Elsevier: New York, 2006; Vol. 2, p 35.
- Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.
- Fogarasi, G.; Zhou, X. F.; Taylor, P. W.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191.
- Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.
- Peng, C. Y.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.
- Liang, W.; Wang, H.; Hung, J.; Li, X.; Frisch, M. J. *J. Chem. Theory Comput.* **2010**, *6*, 2034.
- Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.
- Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959.
- Svensson, M.; Humbel, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 3654.
- Dapprich, S.; Komáromi, I. n.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. THEOCHEM* **1999**, *461–462*, 1.
- Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815.
- Vreven, T.; Morokuma, K.; Farkas, d.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.
- Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701.
- Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393.
- Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.
- Csaszar, P.; Pulay, P. *J. Mol. Struct.* **1984**, *114*, 31.
- Farkas, Ö.; Schlegel, H. B. *Phys. Chem. Chem. Phys.* **2002**, *4*, 11.
- Kudin, K. N.; Scuseria, G. E.; Cancès, E. *J. Chem. Phys.* **2002**, *116*, 8255.
- Li, X.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 835.
- Moss, C. L.; Li, X. *J. Chem. Phys.* **2008**, 129.
- Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.
- Simons, J.; Nichols, J. *Int. J. Quantum Chem.* **1990**, *24*, 263.
- Fletcher, R. *Practical Methods of Optimization*; Wiley: Chichester, 1981; p 95.
- Gill, P. E.; Murray, W.; Wright, M. H. *Practical Optimization*; Academic: New York, 1981; p 114.
- Powell, M. J. D. *Non-linear Optimization*; Academic: New York, 1982; p 29.
- Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; Prentice Hall: New York, 1983; p 129.
- Scales, L. E. *Introduction to Non-linear Optimization*; Springer-Verlag: New York, 1985; p 115.
- Schlegel, H. B.; Yarkony, D. R. Geometry optimization on potential energy surfaces. *Modern Electronic Structure Theory*; World Scientific: Singapore, 1995; p 459.

- (36) Schlegel, H. B.; Schleyer, P. v. R.; Allinger, N. L.; Kollman, P. A.; Clark, T.; Schaefer, H. F., III; Gasteiger, J.; Schreiner, P. R. Geometry optimization. In *Encyclopedia of Computational Chemistry*; Wiley: Chichester, 1998; Vol. 2, pp 1136.
- (37) Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 76.
- (38) Fletcher, R. *Comput. J. (Switzerland)* **1970**, *13*, 317.
- (39) Goldfarb, D. *Math. Comput.* **1970**, *24*, 23.
- (40) Shanno, D. F. *Math. Comput.* **1970**, *24*, 647.
- (41) Murtagh, B.; Sargent, R. W. H. *Comput. J. (Switzerland)* **1972**, *13*, 185.
- (42) Powell, M. J. D. *Nonlinear Programming*; Academic: New York, 1970; pp 31.
- (43) Powell, M. J. D. *Math. Program.* **1971**, *1*, 26.
- (44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development Version H.09+*; Gaussian, Inc.: Wallingford, CT, 2010.
- (45) 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.

CT100453X

Toward a General Formulation of Dispersion Effects for Solvation Continuum Models

Ville Weijo,[†] Benedetta Mennucci,^{†,‡} and Luca Frediani^{*,†}

Centre for Theoretical and Computational Chemistry, Department of Chemistry,
University of Tromsø, N-9037 Tromsø, Norway and Department of Chemistry,
University of Pisa, Via Risorgimento 35, 56126 Pisa, Italy

Received August 16, 2010

Abstract: We revised the quantum model of Amovilli and Mennucci (*J. Phys. Chem. B* **1997**, *101*, 1051) to include the dispersion contribution to the solvation free energy within the framework of continuum models. Our revised formulation makes use of a single adjustable solvent-dependent parameter, and it can be readily generalized to different quantum mechanical descriptions. In particular, we made use of DFT and applied the model to investigate dispersion effects on vertical excitation energies within a time-dependent DFT framework. Our findings show that dispersion effects constitute a significant component of the absolute solvent effect but when relative solvent–solvent shifts are considered a cancellation effect is observed.

1. Introduction

A faithful description of the molecular properties and processes has to take into account the possible presence of the environment surrounding the system under investigation. This clearly increases the modeling complexity as it is evident when, e.g., the number of interacting subsystems in a liquid solution (chemical processes are almost always involving liquid phases) is considered. In order to tackle such complexity, many strategies have been proposed to simplify the problem. The choice of the specific strategy is dictated by the problem under investigation. In the case of the modeling of electronic structure properties, it is necessary to adopt a quantum-mechanical description of the investigated system, whereas the remainder can be dealt with at a lower level of accuracy. Within this framework, a successful approach is constituted by continuum models, where the environment is described by a structureless medium characterized by a set of parameters in order to model its interactions with the molecular system.

Such interactions are then introduced in the quantum-mechanical description of the molecular system as additional terms in the molecular Hamiltonian operator.¹ The correct description is formally achieved by defining an idealized

solvation process, where all interactions are “switched on” from the ideal situation of two infinitely separated systems (the isolated solute and the pure solvent) to the fully interacting system (the solvated molecule). The first step of this ideal process is generally connected to the creation of a cavity in the solvent, where the solute molecule is accommodated. The energy involved in this process is commonly called the cavitation energy. The solute is then placed inside the cavity and allowed to interact electrostatically with the solvent. The electrostatic interaction is a purely classical term, and hence, its expression is well defined within continuum methods. Other nonelectrostatic interactions are intrinsically connected to the quantum nature of both solute and solvent, and it is therefore less obvious how to express them when the solvent is replaced by a continuum. One possible strategy comes from the theory of weakly interacting quantum systems where such interactions are dealt with as small perturbations to the Hamiltonians of the two isolated systems. The main terms in such treatment are called dispersion and repulsion, and they can loosely be connected to the attractive and repulsive terms of the Lennard–Jones (LJ) potential.

The free energy of solvation of a molecular solute can then be expressed as the sum of the above-mentioned contributions

$$G_{\text{sol}} = G_{\text{el}} + G_{\text{cav}} + G_{\text{disp}} + G_{\text{rep}} \quad (1)$$

where we neglected the contribution due to internal degrees of freedom.

* Corresponding author e-mail: luca.frediani@uit.no.

[†] University of Tromsø.

[‡] University of Pisa.

Historically, most continuum models have been developed to treat the electrostatic contribution to G_{sol} , but several proposals to include nonelectrostatic contributions have been formulated: the interested reader can find a review of the various methodologies developed so far in refs 1 and 2. Most such methodologies focused on reproduction of solvation energies only and they were therefore developed in a semiclassical framework.^{3–5}

The growing interest in the modeling of molecular properties that depend on the electronic structure, however, requires the development of a fully quantum mechanical approach to the determination of all solute–solvent interactions, including dispersion and repulsion. In this way both electrostatic and nonelectrostatic energy contributions can be fully integrated in the Hamiltonian operator. Practical expressions for dispersion and repulsion have been developed by Amovilli⁶ based on the theory of weak intermolecular forces⁷ and later reformulated within the polarizable continuum model^{1,2} (PCM) by Amovilli and Mennucci.⁸

This formulation allows for the dispersion and repulsion contributions being included in response property computations. The approach has been used in a coupled perturbed Hartree–Fock (CPHF) framework to compute electronic (hyper)polarizabilities⁹ and in a CASSCF framework to obtain vertical excitation energies.¹⁰

In the present work we will further develop the model by focusing on the dispersion term for which we propose a new parametrization that can easily be adapted to the chosen QM description of the electronic structure. By making use of density functional theory (DFT), we will then apply the method to the calculation of vertical excitation energies within a time-dependent DFT framework.

The paper is organized as follows: in section 2, a brief resume of the theoretical method will be given together with its extension to TDDFT equations. In section 3, the method will be parametrized according to new fitting strategy and it will be applied to study dispersion effects on excitation energies of various small- and medium-sized molecules in different solvents. Finally, in section 4, some conclusive remarks will be given.

2. Methods

2.1. Approximating the Dispersion Term to a Practical Form. Following refs 6 and 8, we begin our approximations by discarding the detailed molecular nature of the solvent and adopting a PCM-like description in terms of apparent surface charge distributions. Within this framework, the dispersion interaction contribution to the solvation Gibbs energy of eq 1 can be defined as

$$G_{\text{disp}} = \frac{1}{\pi} \sum_{ia} \int_0^\infty d\omega \frac{\Delta\epsilon_{ai}}{(\Delta\epsilon_{ai})^2 + \omega^2} \int d\mathbf{r} \int_\Gamma d\mathbf{s} \rho_{ai}(\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{s}|} \sigma(\mathbf{s}; \rho_{ia}, \epsilon(i\omega)) \quad (2)$$

where indexes a and i run over the virtual and occupied orbitals of the solute molecule, respectively, $\Delta\epsilon_{ai}$ is the orbital energy difference between orbitals a and i , Γ refers to the cavity surface, and $\sigma(\mathbf{s}; \rho_{ia}, \epsilon(i\omega))$ is the apparent surface

polarization charge at point \mathbf{s} due to the solute transition charge distribution ρ_{ia} and the solvent frequency-dependent dielectric constant $\epsilon(i\omega)$ evaluated at the imaginary frequency $i\omega$.

We further assume that in the case of dispersion, mutual interaction of the induced surface polarization charges does not play a significant role in the solvation energy. This assumption can be justified in light of the fact that dispersion interactions are mostly between two bodies (compare, e.g., to the two-body LJ potential), at least in nonmetallic materials. The surface charges can then be assumed to be proportional to the component of the electric field due to the charge distribution perpendicular to the cavity surface, E_{ia} ,⁸ so that

$$\sigma(\mathbf{s}; \rho_{ia}, \epsilon(i\omega)) = -\frac{\Omega^2}{\Omega^2 + \omega^2} \frac{(\eta^2 - 1)}{4\pi\eta^2} E_{ia}(\mathbf{s}) \quad (3)$$

where η is the refractive index of the continuum at optical frequencies and $\Omega = \eta I$, with I being the first ionization potential of the solvent.

Using eq 3 and denoting the Coulomb potential of the electronic charge distribution ρ_{ia} by V_{ia} , a simplified equation for the dispersion energy defined in eq 2 is obtained

$$G_{\text{disp}} = -\frac{(\eta^2 - 1)}{8\pi\eta^2} \sum_{ia} \frac{\Omega}{\Omega + \Delta\epsilon_{ai}} \int_\Gamma d\mathbf{s} V_{ia}(\mathbf{s}) E_{ai}(\mathbf{s}) \quad (4)$$

A further simplification, which is dictated by the need of achieving both an acceptable computational cost and an expression suitable to be used within the SCF framework, is obtained by introducing an averaged excitation energy $\Delta\epsilon_{\text{ave}}$ ⁸

$$G_{\text{disp}} = -\frac{(\eta^2 - 1)}{8\pi\eta^2} \frac{\Omega}{\Omega + \Delta\epsilon_{\text{ave}}} \int_\Gamma d\mathbf{s} \sum_{ia} V_{ia}(\mathbf{s}) E_{ai}(\mathbf{s}) \quad (5)$$

This expression can be written in terms of occupied and virtual density matrices \mathbf{P} and \mathbf{Q} , respectively, in the atomic orbital (AO) basis as

$$G_{\text{disp}} = -\frac{\beta}{2} \sum_{\mu\nu\gamma\delta} P_{\gamma\nu} Q_{\mu\delta} [\gamma\nu|\mu\delta] \quad (6)$$

where μ, ν, γ , and δ refer to AO indices and

$$P_{\gamma\nu} = 2 \sum_i^{\text{occ}} C_{\gamma i} C_{\nu i} \quad (7)$$

$$Q_{\gamma\nu} = \left(\sum_i^{\text{all}} C_{\gamma i} C_{\nu i} - \sum_i^{\text{occ}} C_{\gamma i} C_{\nu i} \right) = S_{\gamma\nu}^{-1} - \frac{1}{2} P_{\gamma\nu} \quad (8)$$

and the symmetrized form of the $V_{\mu\nu}(\mathbf{s})E_{\gamma\delta}(\mathbf{s})$ term is introduced as in the original formulation

$$[\gamma\nu|\mu\delta] = \frac{1}{2} \int_\Gamma d\mathbf{s} (V_{\gamma\nu}(\mathbf{s})E_{\mu\delta}(\mathbf{s}) + V_{\mu\delta}(\mathbf{s})E_{\gamma\nu}(\mathbf{s})) \quad (9)$$

The factor β is implicitly defined as

$$\beta = \frac{1}{4\pi} \frac{\eta^2 - 1}{\eta^2 \left(1 + \frac{\Delta\varepsilon_{\text{ave}}}{\Omega}\right)} \quad (10)$$

In the original formulation,⁸ the averaged excitation energy $\Delta\varepsilon_{\text{ave}}$ was obtained by considering a predefined set of occupied and virtual orbitals defined by a window of energies $[-a, a]$ and averaging for such a set. This choice is, however, strictly connected to the quantum mechanical method employed. In particular, it is well known that DFT and HF descriptions yield very different energies when occupied and virtual orbitals are compared, and the final value for β is strongly dependent on the chosen method.¹¹ In order to avoid this problem, we propose to use a value of $\Delta\varepsilon_{\text{ave}}$ that will yield the same dispersion energies predicted by the “exact” expression in eq 4.

In the original formulation the factor β was scaled by an adjustable parameter c_f replacing the constant $1/4\pi$ with the value $c_f = 0.036$. This value was empirically determined through a comparison with experimental data. In this work we followed a different strategy: we considered the semiclassical dispersion energy values as references and obtained c_f by linear interpolation between adjacent data points (see section 3 for more details).

2.2. SCF Formulation and Its Extension to a Linear Response Approach for Electronic Excitations. Within the SCF approach (here applied at the DFT level of theory), the solvent effect is represented in terms of additional contributions to the Fock matrix of the isolated molecule. In particular, such contributions can be computed by differentiating the corresponding free energy expressions with respect to the orbital parameters (e.g., the density-matrix elements). By adopting the same AO basis expansion used in eq 6, the expression for the dispersion contribution of a closed-shell molecule becomes

$$F_{\mu\nu} = \frac{G_{\text{disp}}}{\partial P_{\mu\nu}} = -\frac{\beta}{2} \sum_{\gamma\delta} ((S^{-1})_{\gamma\delta} - P_{\gamma\delta}) [\mu\gamma|\delta\nu] \quad (11)$$

The extension of dispersion interactions to excitation energies is achieved within the TDDFT linear response framework in a similar way as that already done for the electrostatic term.^{12–14} The excitation energies of a molecular system are determined by the poles of the linear response function to a time-dependent perturbation, and thus, they can be obtained from the solutions of the non-Hermitian eigensystem^{15–17}

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \omega \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \quad (12)$$

where the transition vectors (\mathbf{X} , \mathbf{Y}) correspond to collective eigenmodes of the density matrix with eigenfrequencies ω and the matrices \mathbf{A} and \mathbf{B} form the Hessian of the ground-state energy

$$\begin{aligned} A_{ai,bj} &= \delta_{ab}\delta_{ij}(\varepsilon_a - \varepsilon_i) + K_{ai,bj}^0 \\ B_{ai,bj} &= K_{ai,bj}^0 \end{aligned} \quad (13)$$

where indices i, j, \dots label occupied, a, b, \dots virtual, and s, t, \dots generic molecular orbitals and

$$K_{st,uv}^0 = \int d\mathbf{r} \int d\mathbf{r}' \psi_s^*(\mathbf{r}') \psi_t(\mathbf{r}') \left(\frac{1}{|\mathbf{r}' - \mathbf{r}|} + g_{xc}(\mathbf{r}', \mathbf{r}) \right) \psi_u^*(\mathbf{r}) \psi_v(\mathbf{r}) \quad (14)$$

Here g_{xc} is the exchange-correlation kernel within the adiabatic approximation.

If now we introduce dispersion effects, both \mathbf{A} and \mathbf{B} change as follows

$$\begin{aligned} A_{ai,bj} &\Rightarrow A_{ai,bj} + B_{ai,bj}^{\text{disp}} \\ B_{ai,bj} &\Rightarrow B_{ai,bj} + B_{ai,bj}^{\text{disp}} \end{aligned} \quad (15)$$

$B_{ai,bj}^{\text{disp}}$ is the dispersion contribution

$$B_{st,uv}^{\text{disp}} = \frac{\beta}{2} [st|uv] \quad (16)$$

which can be obtained from the Fock-operator contribution eq 11. We assume that β remains unchanged with respect to ground-state calculations.

In eq 15 we focused on the dispersion term, but all calculations performed also include the standard PCM electrostatic contribution and the repulsion contribution in both SCF and linear response equations. The interested reader can find detailed derivations of the electrostatic terms from the reference papers.^{8,12–14}

3. Results and Discussion

3.1. Computational Details. We computed dispersion energy contributions to solvation and excitation energies of a set of chromophores, including acetone, acrolein, 4-(*N,N*-dimethylamino)benzonitrile (DMABN), *p*-nitroaniline (PNA), and the three diazines (pyrazine, pyrimidine, and pyridazine) immersed in cyclohexane and acetonitrile. Molecular geometries were optimized for each solvent and vacuum using the B3LYP^{18–20} functional and 6-311G(d,p) basis set.²¹ Solvent contributions to geometry optimizations included only electrostatic terms. All calculations were carried out with a local version of the Gaussian 03 program package.²²

In order to compute dispersion effects on solvation free energies and excitation energies, the DFT and TDDFT approaches have been used in combination with two functionals, the generalized gradient approximation functional PBE²³ and the hybrid B3LYP.^{18–20} The dispersion contribution has been shown to be sensitive to basis set quality as discussed in ref 11. On the basis of the results of the reference, we have chosen to use augmented correlation-consistent double- ζ (aug-cc-pVDZ) basis set^{24,25} as a good compromise between the size of the set and the accuracy.

The solvent contribution to the dispersion energy contains two solvent-dependent parameters, namely, the refractive index η and the first ionization potential I . We used values of $\eta = 1.424$ and $I = 0.364$ au for cyclohexane and $\eta = 1.344$ and $I = 0.448$ au for acetonitrile.²⁶

For all systems, the molecular cavity was obtained as the envelope of spheres centered on selected atoms of the solute, and thus, it is determined by the solute geometry. In the present case, we used cavities constructed by applying the united atom topological model and the atomic radii of the UFF²⁷ force field as implemented in the Gaussian

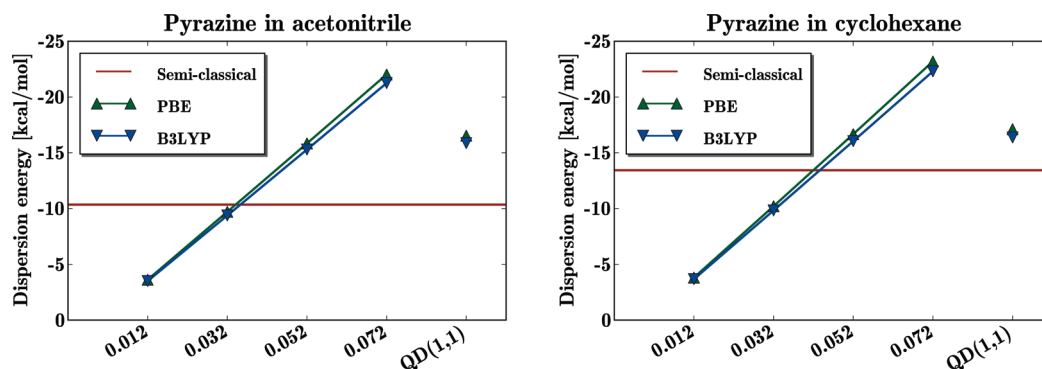


Figure 1. Illustration of the dependence of the dispersion energy on different parametrizations. All values refer to PBE and B3LYP calculations with the aug-cc-pVDZ basis set, 0.012, 0.032, 0.052, and 0.072 refer to different values of c_f (see text for details), and “QD(1,1)” is the value recommended by a previous model. The reference value (represented as an horizontal line) is obtained using a semiclassical method.

code. According to this model, a sphere is associated to each atom (excluding the hydrogens) with a radius defined according to the type of the atom and its bonds. In particular, we have $R(\text{CH}) = 2.125$, $R(\text{C}) = 1.925$, $R(\text{N}) = 1.83$, $R(\text{CH}_3) = 2.525$, $R(\text{O}) = 1.75$ (all values are in Angstroms). Semiclassical dispersion energy contributions were parametrized using cavities based on bare Bondi radii, and thus, evaluations of those contributions were made using smaller cavities, where the above-defined radii were divided by 1.2.

3.2. Parameterization. As reported in the Methods section, a proper adjustable parameter was originally introduced by Amovilli and Mennucci in the definition of the factor β , namely

$$\beta = c_f \frac{\eta^2 - 1}{\eta^2 \left(1 + \frac{\Delta \epsilon_{\text{ave}}}{\Omega} \right)} \quad (17)$$

with c_f originally set to 0.036.

In order to give an idea of the sensitivity of the method with respect to that parameter, two examples of the behavior of the dispersion energy as a function of parameter c_f are provided in Figure 1, where a pyrazine molecule is placed in two different solvents and its dispersion energy is evaluated. We recall that for each different c_f , the corresponding β factor is obtained using an averaged energy $\Delta \epsilon_{\text{ave}}$ that yields the same dispersion energy as the “exact” semiclassical value as reported in eq 4. This allows us to eliminate the arbitrary choice of orbitals used in the averaging procedure described in the original formulation.

In both cases, the DFT curves (“PBE” and “B3LYP”) are very close to linear, and they have a crossover point with the “semiclassical” reference between the c_f values of 0.032 and 0.052. “QD(1,1)” results, which are based on the previous parametrization of the model ($c_f = 0.036$ and $\Delta \epsilon_{\text{ave}}$ obtained using the mentioned window of orbital energies for the averaging), are off from the semiclassical values by roughly a factor two. This is most likely because the model was parametrized using the HF method and small basis sets, which are known to produce contributions that are too small when applied to DFT.¹¹ Results in the figures are illustrative in a sense that the c_f value which makes the dispersion energy equal to the semiclassical value lies between 0.032 and 0.052,

Table 1. Optimal Values for c_f^a

	acetonitrile		cyclohexane	
	PBE	B3LYP	PBE	B3LYP
acetone	0.04016	0.03407	0.04375	0.04076
acrolein	0.03838	0.03687	0.04724	0.04526
DMABN	0.03725	0.03261	0.04118	0.03688
PNA	0.03690	0.03390	0.04401	0.04132
pyrazine	0.03528	0.03409	0.04356	0.04194
pyridazine	0.03673	0.03546	0.04585	0.04191
pyrimidine	0.03534	0.03419	0.04359	0.04090
optimized values for each solvent and functional				
optimized	0.03690	0.03409	0.04375	0.04132
max. abs. error	0.912	0.775	1.305	2.497
avg. abs. error	0.342	0.275	0.423	0.587
max. rel. error	12.03%	7.62%	7.45%	12.21%
avg. rel. error	3.79%	2.44%	2.84%	3.79%
optimized values for each solvent				
optimized c_f	0.03535		0.04194	
max. abs. error	1.436		2.845	
avg. abs. error	0.539		0.696	
max. rel. error	17.72%		13.92%	
avg. rel. error	4.89%		4.80%	
optimized value (optimized with all the data)				
optimized c_f	0.03726			
max. abs. error	3.071			
avg. abs. error	1.231			
max. rel. error	21.33%			
avg. rel. error	9.53%			

^a Maximum and average absolute errors are in kcal/mol.

and that the variation of the dispersion energy with respect to c_f is very close to linear. On the basis of these observations, we optimized the value of c_f . The values obtained for different combinations of DFT functionals, solvents, and molecules are shown in Table 1.

Optimal c_f values for each molecule/solvent/functional combination (highest quadrant of the table) were obtained by taking the semiclassical value and interpolating c_f linearly to match the two numbers, and the optimal c_f 's are in the range from 0.0326 to 0.04724. Differences in optimal c_f 's within a column can be as large as 0.008, which translates to the smallest value being almost one-fifth smaller than the largest one. Such variations of c_f 's might seem to be large at first, but it is not evident that those will imply large errors in dispersion energy contributions; thus, we investigated the

possibility of reducing the number of free parameters in our model (from one optimal c_f for each molecule, solvent, and functional).

c_f values for combined data sets were optimized with respect to the sum of relative errors (as compared to the semiclassical values) of individual dispersion contributions. Optimizing the values with respect to the sum of absolute errors has only a minor effect on the optimized values, and thus, these results are not shown here. In addition to the optimized c_f values, we report maximum and average absolute and relative errors obtained when using an optimized c_f in Table 1. Optimized c_f 's for solvent/functional combinations are shown in the second quadrant of the table, where we can see that both functionals have optimal c_f 's close to each other for a given solvent, optimized values being at most 0.003 apart from each other. The largest absolute errors range from 0.8 to 2.5 kcal/mol, and the average absolute errors range from 0.3 to 0.6 kcal/mol. The largest relative errors are approximately 12%, while the average relative errors are less than 4%. Errors of these magnitudes are certainly acceptable, but it will be interesting to see if we can get rid of the functional specificity of the parametrization in order to be more in line with the idea of a method-independent continuum of PCM.

We optimized c_f values for both solvents by including contributions from both functionals (third quadrant of the table). As expected, errors are now slightly larger than before. Average errors are still very reasonable (0.6–0.7 kcal/mol and 5% for absolute and relative errors, respectively), and maximum absolute errors are 1.4 and 2.8 kcal/mol in acetonitrile and cyclohexane, respectively, which are slightly larger than maximum absolute errors from the previous level of parameter reduction (c_f for each functional/solvent). The parameter reduction can be taken a step further still by optimizing a single c_f value against all data. The results are shown in the fourth quadrant of the table. Using a single c_f value does not change the maximum absolute error much, but average absolute and relative errors roughly double to 1.2 kcal/mol and 10%, respectively. Maximum relative error also increases slightly from the acetonitrile-specific value of 18% to 21%.

As expected, the errors become larger when the number of parameters is reduced. From the point of view of average errors, we conclude that choosing a solvent specific c_f is the most reasonable choice in terms of accuracy and the number of free parameters. This choice fits well within a continuum solvent framework, where the solvent is described by a collection of parameters which determine its interactions with the solute. In order to have a more direct appreciation of the quality of the new solvent-specific parametrization with respect to the semiclassical reference values, we collect both the DFT dispersion energies obtained with solvent-optimized c_f and the corresponding semiclassical values in Table 2.

3.3. Excitation Energies. In this section we extend the analysis of dispersion contributions to excitation energies. In Table 3 we report the values obtained for selected excited states of the same set of molecules used for the previous analysis on solvation free energies. Excitations considered here are the following: the lowest $n-\pi^*$ electronic transition

Table 2. Dispersion Energies (in kcal/mol) with Solvent-Optimized c_f 's and Semiclassical (S-C) Reference Values

	acetonitrile		cyclohexane	
	DFT	S-C	DFT	S-C
B3LYP				
acetone	-10.37	-10.35	-12.92	-13.42
acrolein	-10.45	-10.86	-13.04	-14.27
DMABN	-10.36	-10.36	-12.92	-13.43
PNA	-7.32	-7.95	-9.15	-10.32
pyrazine	-6.24	-7.58	-9.38	-9.80
pyridazine	-15.65	-16.65	-20.82	-20.44
pyrimidine	-14.40	-15.03	-18.57	-19.51
PBE				
acetone	-10.74	-10.35	-13.42	-13.42
acrolein	-10.83	-10.86	-14.28	-14.27
DMABN	-10.72	-10.36	-13.81	-13.43
PNA	-7.62	-7.95	-9.56	-10.32
pyrazine	-7.88	-7.58	-10.09	-9.80
pyridazine	-18.09	-16.65	-23.28	-20.44
pyrimidine	-15.69	-15.03	-19.80	-19.51

is examined for diazines and acetone, for acrolein both $n-\pi^*$ and $\pi-\pi^*$ excitations are considered, and for PNA and DMABN, charge-transfer excitations (of A symmetry) are studied. In the case of DMABN, the locally excited B2 excitation results are also presented. For each molecule the symmetry of the corresponding excitation is indicated in the tables.

To have a more direct evaluation of the solvatochromic effects obtained by switching on the electrostatic and the dispersion interactions, we report gas-phase electronic excitation energies together with vacuum–solvent energy. There, c_f1 refers to the solvent-optimized value, whereas c_f2 is the globally-optimized coefficient. The QD(1.1) column contains results obtained by using the window parameter of refs 8 and 11, and the ES results are obtained by considering only the electrostatic contribution.

At a first glance, the magnitudes and signs of the solvation shifts might seem to be erratic. However, if the dispersion-included values are compared to the ES values, one sees that dispersion always gives a negative contribution to the corresponding excitation energy. The contributions from our c_f -parametrized models range roughly from -90 to -180 meV, whereas QD(1.1) contributions can be almost twice as large as compared to the corresponding c_f -parametrized shifts.

Depending on the magnitude of the original ES value, dispersion contributions can change the sign of the solvent shift. Magnitudes of dispersion contributions do not change considerably between different solvents, and the c_f1 and c_f2 results are close to each other, at least when comparing to the corresponding ES results. Unfortunately, the accuracy of dispersion-corrected excitation energy shifts is difficult to evaluate as reliable experimental vacuum–solvent excitation energy shift data are difficult to obtain. Experimental solvent–solvent shifts are, however, more readily available. Solvent–solvent excitation energy shifts along with some experimental results are shown in Table 4.

We will first comment on the QD(1.1) and c_f2 values (optimized against all solvent and functional combinations),

Table 3. Computed Vacuum Excitation Energies and Vacuum–Solvent Shifts (all values in eV)^a

molecule	vac.	solvation shift vac–ACN				solvation shift vac–CH			
		c_1	c_2	QD(1.1)	ES	c_1	c_2	QD(1.1)	ES
B3LYP									
pyrazine B3U	3.928	-0.094	-0.102	-0.172	0.049	-0.158	-0.137	-0.208	0.020
pyrimidine B1	4.245	0.038	0.030	-0.040	0.181	-0.110	-0.090	-0.159	0.067
pyridazine B1	3.547	0.178	0.169	0.092	0.327	-0.076	-0.056	-0.136	0.107
acetone A	4.415	-0.018	-0.030	-0.098	0.111	-0.161	-0.137	-0.179	0.036
acrolein A''	3.412	-0.026	-0.034	-0.114	0.121	-0.141	-0.120	-0.199	0.043
acrolein A'	5.692	-0.323	-0.331	-0.410	-0.205	-0.314	-0.293	-0.370	-0.161
DMABN B2	4.349	-0.128	-0.135	-0.186	-0.035	-0.164	-0.147	-0.188	-0.033
DMABN A1	4.623	-0.377	-0.385	-0.437	-0.279	-0.314	-0.295	-0.338	-0.175
PNA A	3.892	-0.583	-0.589	-0.660	-0.490	-0.410	-0.393	-0.462	-0.280
PBE									
pyrazine B3U	3.521	-0.102	-0.110	-0.186	0.049	-0.178	-0.156	-0.232	0.011
pyrimidine B1	3.751	0.021	0.013	-0.061	0.171	-0.124	-0.099	-0.174	0.068
pyridazine B1	3.082	0.186	0.177	0.096	0.343	-0.088	-0.063	-0.140	0.121
acetone A	4.245	-0.066	-0.075	-0.119	0.093	-0.175	-0.150	-0.187	0.029
acrolein A''	3.011	-0.037	-0.046	-0.133	0.116	-0.152	-0.130	-0.216	0.042
acrolein A'	5.456	-0.242	-0.251	-0.338	-0.117	-0.240	-0.218	-0.304	-0.085
DMABN B2	3.961	-0.163	-0.171	-0.215	-0.046	-0.200	-0.179	-0.216	-0.040
DMABN A1	4.314	-0.404	-0.412	-0.458	-0.280	-0.347	-0.325	-0.364	-0.179
PNA A	3.512	-0.456	-0.463	-0.539	-0.348	-0.368	-0.349	-0.423	-0.222

^a ACN marks acetonitrile and CH cyclohexane.**Table 4.** Computed Acetonitrile–Cyclohexane Solvent Shifts with Experimental Results (all values in eV)

molecule	func.	c_1	c_2	QD(1.1)	ES	exp.
pyrazine B3U	B3LYP	0.064	0.035	0.036	0.029	0.040 ^a
	PBE	0.076	0.046	0.046	0.038	
pyrimidine B1	B3LYP	0.148	0.119	0.119	0.114	0.120 ^a
	PBE	0.145	0.112	0.113	0.103	
pyridazine B1	B3LYP	0.254	0.225	0.228	0.220	0.260 ^a
	PBE	0.274	0.240	0.236	0.222	
acetone A	B3LYP	0.142	0.107	0.081	0.075	0.06 ^b
	PBE	0.109	0.074	0.069	0.064	
acrolein A''	B3LYP	0.115	0.085	0.085	0.078	
	PBE	0.115	0.084	0.084	0.075	
acrolein A'	B3LYP	-0.009	-0.039	-0.039	-0.045	
	PBE	-0.002	-0.033	-0.034	-0.032	
DMABN B2	B3LYP	0.036	0.012	0.002	-0.003	
	PBE	0.036	0.008	0.001	-0.005	
DMABN A1	B3LYP	-0.064	-0.090	-0.099	-0.104	-0.20 ^c
	PBE	-0.058	-0.087	-0.094	-0.100	
PNA A	B3LYP	-0.173	-0.196	-0.197	-0.210	-0.13 ^d
	PBE	-0.088	-0.114	-0.116	-0.125	

^a Acetonitrile–isooctane, ref 28. ^b Acetonitrile–*n*-hexane, ref 29. ^c Reference 30. ^d Acetonitrile–dichloromethane, ref 31.

which are reasonably close to each other in most cases, acetone's B3LYP value being a small exception. The similarity of the QD(1.1) and c_2 results comes from the fact that they use the same prefactor for both solvents, and the difference between the solvents is only obtained from the variations of the parameters of the permittivity at imaginary frequencies (see eq 3). For both solvents, the parameters η and I are almost the same, and consequently, without different prefactors, the excitation energy shift should be more or less the same. This implies that the solvent–solvent shifts should be small for QD(1.1) and c_2 results, and indeed, this can be observed from Table 4, where inclusion of dispersion contributions through QD(1.1) and c_2 creates a positive solvent–solvent shift of roughly 0.01 eV or less, which can be considered small as compared to the magnitude of ES solvent–solvent shifts in most cases. As expected by the differences in prefactors, c_1 solvent–solvent shifts are

larger than those obtained with c_2 and the differential effect is, in most cases, a positive further shift of approximately 0.03 eV.

Experimental data on Table 4 are scattered as they were not available for all chromophore/solvent combinations, and in some cases we reported data obtained in similar solvents. For all available data, differences between dispersion-corrected and ES shifts are quite small and they both are in relatively good agreement with experiments.

Finally, from the combined analysis of the two tables we observe that dispersion effects on relative solvatochromic shifts are small for all excitations and systems studied. This is in agreement with previous studies based on a different methodology.³² It is therefore not possible to make a final assessment of the different parametrizations. On the other hand, larger differences are observed when absolute shifts (with respect to vacuum) are considered: the old parametrization when applied to TDDFT calculations seems to overestimate the dispersion contribution, while the new proposed scheme yields corrections to the electrostatic values which are more balanced. This confirms that the combined use of “optimal” averaged energy $\Delta\varepsilon_{\text{ave}}$ and c_f parameter is a robust approach to extend the dispersion model to different QM levels of description.

4. Conclusions

We revised the quantum-mechanical formulation of the dispersion contribution to the solvation energy within the context of the polarizable continuum model originally proposed by Amovilli and Mennucci.⁸ The parametrization of the model, which avoids the approximated estimate of the averaged energy $\Delta\varepsilon_{\text{ave}}$ entering in the definition of the dispersion energy, is now achieved through a single parameter c_f used to “scale” the overall dispersion contribution. For the present work, c_f has been fitted to reproduce the dispersion energy values obtained with a highly parametrized semiclassical model. It turns out that the optimal strategy is

to select a different c_f for each solvent. Moreover, this new parametrization scheme does not show the large sensitivity to the QM level of calculation found in the original formulation, and therefore, it represents a promising method to extend the dispersion formulation to other QM descriptions.

We also employed optimized c_f values to compute the dispersion contribution to electronic excitation energies. The obtained net effect due to dispersion is a red shift for all the investigated systems in the two solvents employed. Our findings show that dispersion is a significant contribution to the overall solvent effect, but unlike the electrostatic contribution, it does not depend strongly on the specific solvent. As a result, if one is interested in solvatochromic shifts, i.e., changes of excitation energies passing from one solvent to the other, results obtained with the electrostatic model often yield satisfactory results in comparison to experimental shifts due to the almost exact cancellation of the dispersion contributions.

Acknowledgment. The support of the IT department of the University of Tromsø for computational resources is greatly acknowledged. L.F. and V.W. acknowledge the Tromsø Forskningsstiftelse for financial support.

References

- (1) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.
- (2) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.
- (3) Floris, F. M.; Tomasi, J.; Ahuir, J. L. P. *J. Comput. Chem.* **1991**, *12*, 784–791.
- (4) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760–768.
- (5) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. *J. Mol. Struct.:THEOCHEM* **2005**, *727*, 29–40.
- (6) Amovilli, C. *Chem. Phys. Lett.* **1994**, *229*, 244–249.
- (7) Amovilli, C.; McWeeny, R. *Chem. Phys. Lett.* **1986**, *128*, 11–17.
- (8) Amovilli, C.; Mennucci, B. *J. Phys. Chem. B* **1997**, *101*, 1051–1057.
- (9) Mennucci, B.; Amovilli, C.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 221–225.
- (10) Cossi, M.; Barone, V. *J. Chem. Phys.* **2000**, *112*, 2427–2435.
- (11) Curutchet, C.; Orozco, M.; Luque, F. J.; Mennucci, B.; Tomasi, J. *J. Comput. Chem.* **2006**, *27*, 1769.
- (12) Cammi, R.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2000**, *104*, 5631.
- (13) Cammi, R.; Mennucci, B. *J. Chem. Phys.* **1999**, *110*, 9877.
- (14) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708.
- (15) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- (16) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- (17) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291.
- (18) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (19) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (20) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (21) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
- (22) 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*; Gaussian, Inc.: Wallingford, CT, 2004.
- (23) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 2865.
- (24) T. H. Dunning, J. *J. Chem. Phys.* **1989**, *90*, 1007.
- (25) Kendall, R. A.; T. H. Dunning, J.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (26) In *Handbook of Chemistry and Physics*, 91st ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2009.
- (27) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (28) Baba, H.; Goodman, L.; Valenti, P. C. *J. Am. Chem. Soc.* **1966**, *88*, 5410.
- (29) Renge, I. *J. Phys. Chem. A* **2009**, *113*, 10678–10686.
- (30) Lipinski, J.; Chojnacki, H.; Grabowski, Z. R.; Rotkiewicz, K. *Chem. Phys. Lett.* **1980**, *70*, 449–453.
- (31) Moran, A. M.; Kelley, A. M. *J. Chem. Phys.* **2001**, *115*, 912–924.
- (32) Li, J.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **2000**, *77*, 264–280.

JCTC

Journal of Chemical Theory and Computation

On the Electronic Structure of H–Ng–Ng–F (Ng = Ar, Kr, Xe) and the Nonlinear Optical Properties of HXe₂F

A. Avramopoulos,^{*,†,||} L. Serrano-Andrés,^{*,‡} J. Li,[§] and M. G. Papadopoulos^{*,†}

Institute of Organic and Pharmaceutical Chemistry, National Hellenic Research Foundation, 48 Vas. Constantinou Ave., Athens 116 35, Greece, Molecular Sciences Institute, Universitat de València, Apartado 22085, València ES-46071, Spain, Accelrys Inc., Telesis Court, San Diego, California 92121, United States, and Department of Informatics and Computer Technology, Lamia Institute of Technology, Third Km Old National Road, 35100, Lamia, Greece

Received August 23, 2010

Abstract: The electronic ground state of H–Ng–Ng–F (Ng = Ar, Kr, Xe) has been studied theoretically by employing the ab initio complete active space valence bond (CASVB) and multi-state complete active space perturbation theory (MS-CASPT2) methods. Both levels of theory confirm the diradicaloid character (DC) of the HNg₂F ground state, increasing in the order Ar > Kr > Xe. The very significant effect of the first and, even more, the second Xe atom on the (hyper)polarizabilities has been shown and interpreted. Thus, the present results demonstrate a mechanism for producing very large (hyper)polarizabilities.

I. Introduction

There is currently a great need for novel materials for photonic applications (e.g., optical communications, data storage, and signal processing). Such materials must have a series of properties, prominently among which is the large magnitude of their nonlinear optical (NLO) properties. Various mechanisms and phenomena have been investigated, which lead to such large NLO properties, for instance, conjugation or charge transfer.^{1–3} Some time ago, we reported that insertion of a noble gas atom into a chemical bond, for example, Ar into HF leading to HArF⁴ or Xe into HC₂H leading to HXeC₂H,⁵ dramatically enhances the NLO properties.

Pauling predicted that stable bonds could be formed by heavy noble gas atoms, due to the reduced stability of the outer electrons, caused by the strong screening of the inner electrons.⁷ The first Xe derivative, XePtF₆, was reported by Bartlett in 1962,⁸ and soon after that, Claassen et al. reported

that xenon and fluorine can react to form XeF₄ (solid), which is stable at room temperature.⁹ After Bartlett's work, a large number of Xe derivatives were synthesized, and as Christie noted, this discovery "triggered a world-wide frenzy in this area."¹⁰

Various groups have inserted noble gas atoms in the bonds A–B (e.g., H–C, H–O, H–S, and C–C), leading to A–Ng–B and thus producing a large variety of compounds.^{11–14} Among those, one notes HNgY, where Y is an electronegative element or group (e.g., OH, CN, SH), which has been studied both experimentally and theoretically.^{15–17} These derivatives are metastable, with a global minimum HY + Ng. The bonding in HNgY involves the structure HNg⁺Y[–], where HNg⁺ is covalently bound and linked to Y[–] by a Coulombic attraction.¹⁸ It has been found that the inserted noble gas atom has a considerable positive charge.^{5,19} In general, there is a large energy barrier which prevents these molecules from dissociation.²⁰ A notable derivative in this series is HArF, which was synthesized by Khriachtchev et al.²¹ from the photolysis of hydrogen fluoride in a solid argon matrix. This is the first experimentally observed covalent neutral condensed phase argon derivative.²²

Among the most remarkable properties of the derivatives resulting by insertion of a noble gas atom are the linear and nonlinear optical (L&NLO) properties. It has been found that,

* Corresponding authors. E-mail: aavram@ie.gr (A.A.), Luis.Serrano@uv.es (L.S.-A.), mpapad@ie.gr (M.G.P.).

† National Hellenic Research Foundation.

|| Lamia Institute of Technology.

‡ Universitat de València.

§ Accelrys Inc.

for example, the insertion of Ar into HF (HArF) leads to a very large increase of the NLO properties.⁴ This has been confirmed by inserting Xe into H–C_n–H (e.g., H–Xe–C₂–H, H–Xe–C₄–H, and H–C₂–Xe–C₂–H)⁵ and AuF leading to Au–Xe–F.²³

Derivatives involving the group Xe–Xe have also been considered. For example, Stein et al.²⁴ reported the formation of Xe₂⁺ from the reduction of XeF⁺Sb₂F₁₁[–]. On the other hand, Drews and Seppelt²⁵ observed the formation of Xe₂⁺Sb₄F₂₁[–] at –30 °C, displaying a surprising, long Xe–Xe bond (3.087 Å). It was noted that this bond is not surpassed in length by any other element–element bond in main group chemistry.²⁵ Very recently, Frenking et al.⁶ reported the first quantum-chemical calculations on the stability of derivatives involving the Ng–Ng bond, where Ng = Ar, Kr, and Xe. The specific derivatives they investigated have the formula H–Ng–Ng–F.

Recently, Frenking et al.⁶ studied the insertion of *two* noble gas atoms in HF, leading to H–Ng–Ng–F, where Ng = Ar, Kr, and Xe. The present study reports on two questions related with H–Ng–Ng–F:

- Which is the electronic structure of the ground state of these derivatives?
- What is the effect of the second Xe atom on the linear and NLO properties of the resulting derivative?

Of considerable importance for this work is the analysis of the diradical character of HNg₂F. The literature connected with this feature and the NLO properties has been recently reviewed.²⁶

For this study, we have used a series of state-of-the-art quantum-chemical techniques (e.g., CCSD(T), CASPT2, and CASVB), and we have proved the need for a multiconfigurational description of the electronic structure of the mentioned systems, which turned out to have a large diradicaloid character in their ground state.

II. Computational Methods

II.1. Electronic Structure Computations. The electronic structure of H–Ng–Ng–F has been investigated by employing valence bond (VB), complete active space (CAS), and CASPT2 methods.

The VB calculations were performed using the code VB2000,^{27–29} integrated in GAMESS. VB2000 is an *ab initio* electronic structure package based on an Algranbrant algorithm.³⁰ A unique feature in VB2000 is its general implementation of McWeeny's group function theory (GFT)³¹ and the combination of GFT and VB methods. The program allows a user to specify multiple groups of electrons, and each one of them can be treated with correlated methods, such as the complete active space valence bond (CASVB) method. In this way, more electrons can be included in VB calculations in an affordable way. The electrons in the molecules HNgNgF (assuming that all atoms are placed on the *z* axis) can be divided into four groups as follows:

- six π electrons in the three valence p_x orbitals of the three heavy atoms
- six π electrons in the three valence p_y orbitals of the three heavy atoms

- six σ electrons, which are shared among the four atoms
- the remaining valence and the core electrons

All 18 valence electrons of the first three groups have the potential to describe the weak bonding in the molecules. Therefore, for a rigorous treatment of the derivatives without discrimination, all 18 electrons were included in VB calculations, and the remaining electrons were treated as a Hartree–Fock core. The orbital symmetry of the valence electrons provides a natural way of grouping the electrons, and each group is strongly orthogonal to the others. In this study, all groups were treated by the CASVB method. Each of the two π -electron groups has six electrons in six orbitals (each heavy atom contributes two p orbitals). Such a treatment can be denoted as CASVB(6,6)•CASVB(6,6)•CASVB(6,4), according to the “dot” notation, introduced in VB2000.^{28,29}

The above multigroup VB method is a natural extension of the conventional CASSCF and generalized VB (GVB) methods, both of which can be considered as special cases of a group function treatment. In the conventional complete active space self-consistent field (CASSCF) method, only one group is treated with correlation, and this group can have various numbers of electrons and orbitals. On the other hand, GVB can be considered as a multigroup VB treatment in which each VB group has two electrons in two orbitals. VB2000 has the most general implementation of the group function method: multiple groups can be included, and each group can have different numbers of electrons and orbitals and can be treated with different methods, such as VBSCF or CASVB. The wave functions are obtained by energy minimization under the strong orthogonality constraints between groups.

The electronic structure CASSCF/CASPT2^{32,33} computations were performed by using the ANO-RCC relativistic basis set³⁴ contracted to Xe [7s6p4d], Kr [6s5p3d], Ar [5s4p1d], F [5s3p1d], and H [3s1p]. The Douglas–Kroll procedure was included to consider relativistic effects.³⁴ The calculations were performed in the C_{2v} point group, where the C₂ axis coincides with the *z* molecular axis for all considered (linear) geometries. For these computations, one needs to take into account the existence of two electronic singlet states in A₁ symmetry (¹ Σ^+ in C_{∞v} symmetry) mainly involving the σ and σ^* orbitals. The strong interaction between the states makes necessary the use of a two-state model. One may notice the different levels of calculation used in the present work. First, one can run a single- or multiple-root CASSCF calculation. In the former case, this is named state average (SA) CASSCF and provides orthogonal states and wave functions for the averaged states, unlike what happens for single-root CASSCF, in which the states are not orthogonal with the other states of the same spin and spatial symmetry. On top of the CASSCF wave function (single-root or SA), one can perform a state-specific (SS) CASPT2 calculation, providing also nonorthogonal states. Using the SS-CASPT2 states as a reference, it is possible to apply the multistate (MS) CASPT2 technique, a procedure which makes the SS-CASPT2 states interact, leading to orthogonal states and wave functions, the latter named perturbatively modified complete active space configuration

interaction (PMCAS-CI) wave functions.^{35–37} As a summary, we have two sets of results available: CASSCF wave functions and CASPT2 state energies and PMCAS-CI wave functions and MS-CASPT2 state energies. We have applied here different procedures to establish the most accurate results. In the final calculations, we run a state-average (SA) CASSCF calculation considering the two states (one of them being the ground state) and then the multistate (MS) CASPT2 technique. The lowest $^1\Pi$ state was also computed. Different active spaces were employed to ensure the convergence and accuracy of the results, as described later. Except otherwise mentioned, the calculations are reported using the 4440/12 active space ($a_1b_1b_2a_2$ /active electrons in the C_{2v} symmetry) corresponding to four σ - and four π -type orbitals. In the σ space, the HOMO–1 and HOMO orbitals have been included (four electrons in this space) as well as two correlating ones. The occupied orbitals involve the valence orbital p_z of F and the p_z of Rg. The π space involves two strongly occupied (eight electrons) and two correlating π orbitals. The former are the π valence orbitals of Rg. The CASPT2 calculations were performed using the MOLCAS suite of programs.³⁴

II.2. Calculation of the L&NLO Properties. When a molecule is set in a uniform static electric field F , its energy, E , may be expanded as follows:³⁸

$$E = E^0 - \mu_i F_i - (1/2)\alpha_{ij} F_i F_j - (1/6)\beta_{ijk} F_i F_j F_k - (1/24)\gamma_{ijkl} F_i F_j F_k F_l - \dots \quad (1)$$

where E^0 is the field free energy of the atom or the molecular system; F_i , F_j , F_k , and F_l are the field components; μ_i , α_{ij} , β_{ijk} , and γ_{ijkl} are the tensor components of the dipole moment, linear dipole polarizability, and first and second hyperpolarizability, respectively. Summation over repeated indices is implied. A finite field approach was used to compute the longitudinal components of the polarizability (α_{zz}) and the first (β_{zzz}) and second hyperpolarizability tensors (γ_{zzzz}), since these are dominant among the other tensor components and are sufficient to demonstrate the effect of the Xe atoms on the NLO properties.

The nonlinear optical properties of interest were computed by employing a series of methods including (i) the CCSD(T), which involves the iterative calculation of single and double excitation amplitudes as well as a perturbative treatment of triple excitations, and (ii) CASSCF/CASPT2. CCSD(T) is known to provide a satisfactory estimate of the electron correlation contribution.³⁹ CCSD(T), in connection with an appropriate basis set, is known to be one of the most accurate methods for the computation of the L&NLO properties. In addition, MS-CASPT2 is particularly suitable for systems with diradical character like HXe_2F . Thus, a comparative study of the L&NLO properties of HXe_2F , involving both CCSD(T) and MS-CASPT2, is expected to provide reliable L&NLO properties.

All of the multiconfigurational NLO property computations, corrected with CASPT2,³⁴ were performed by using the PolX⁴⁰ basis set, developed by Sadlej, where X denotes the element symbol. For all computations, the C_{2v} point group was used, where the C_2 axis, as noted above, coincides with the z molecular axis. In order to check the effect and the

Table 1. The Structure Weights of HAr_2F from CASVB(6,6)·CASVB(6,6)·CASVB(6,4) Calculations with Different Basis Sets

structure	weight [3-21G*]	weight [cc-pVTZ]	weight [MCP-TZP]	type
I. F Ar Ar H	0.88	0.87	0.88	diradical, F and H
II. F ⁻ Ar ⁺ Ar H	0.01	0.01	0.01	diradical, Ar ⁺ and H
III. F Ar Ar ⁺ H ⁻	0.10	0.11	0.10	diradical, F, Ar ⁺
IV. F ⁻ Ar Ar ⁺ H	0.0	0.0	0.0	
V. F ⁻ Ar ⁺ Ar ⁺ H ⁻	0.0	0.0	0.0	
VI. F ⁻ ArAr H ⁺	0.0	0.0	0.0	
VII. FAr ⁺ ArH ⁻	0.0	0.0	0.0	

adequacy of the active space on the computed (hyper)polarizabilities, a number of different spaces and correlating electrons were considered. All of the multiconfigurational computations were performed by using the finite-field approach employing a number of field strengths along the C_2 axis. For the finite-field CASPT2 computations, one needs to take into account the near degeneracy of the two low-lying states of $^1\Sigma^+$ symmetry. These two states will be coupled by the electric field of A_1 symmetry (z direction). In order to account properly for this effect, we use first the (SA) CASSCF, considering the ground and the first excited state and then the multistate (MS) CASPT2^{35,36} technique, which leads to orthogonal states and smooth convergence with the finite field procedure.

The aug-cc-pVDZ series was also employed. For Xe, a small core (28 electrons), energy consistent, relativistic pseudopotential has been used.⁴¹ The 4spd outer-core shell is treated explicitly together with 5sp valence-orbitals. The usefulness and reliability of the pseudopotentials for the computation of the (hyper)polarizabilities has been documented.⁴²

All of the property values were computed by employing the Romberg approach,⁴³ in order to safeguard the numerical stability of our results and to remove higher order contaminations. A number of field strengths of the magnitude $2^m F$, where $m = 1, 2, 3$, and 4 and base field (F) is 0.0008 au, were used. The Gaussian 03 software⁴⁴ has been employed for all of the CCSD(T) calculations, while for the CASSCF/CASPT2 ones, the MOLCAS suite of programs was used.³⁴

III. Results

We shall first investigate the electronic ground state of HNg_2F , where Ng = Ar, Kr, Xe by employing CASVB, CASSCF/CASPT2 and MS-CASPT2 methods and subsequently we shall discuss the L&NLO properties of HXe_2F . All the computations were performed with geometries taken from ref.⁶

III.1. The Ground State of HNg_2F by CASVB Calculations. HAr_2F and HKr_2F . Tables 1 and 2 show the structure weights for HAr_2F and HKr_2F , computed with the Chirgwin–Coulson method⁴⁵ and by employing three different basis sets in order to check the sensitivity of the results. It was found that all three basis sets give quite similar results. It is observed that the total weight of the resonance structures, with diradical characteristics, is, approximately, 99% for HAr_2F and 97–98% for HKr_2F (depending on the basis set). The structure weights computed with different basis sets are

Table 2. The Structure Weights of HKr₂F from CASVB(6,6)·CASVB(6,6)·CASVB(6,4) Calculations with Different Basis Sets

structure	weight [3-21G ⁺]	weight [cc-pVTZ]	weight [MCP-TZP]	type
I. F Kr Kr H	0.81	0.81	0.81	diradical, F and H
II. F ⁻ Kr ⁺ Kr H	0.04	0.05	0.05	diradical, Kr ⁺ and H
III. F Kr Kr ⁺ H ⁻	0.13	0.12	0.11	diradical, F, Kr ⁺
IV. F ⁻ Kr Kr ⁺ H	0.01	0.01	0.01	
V. F ⁻ Kr ⁺ Kr ⁺ H ⁻	0.00	0.00	0.0	
VI. F ⁻ KrKr H ⁺	0.00	0.01	0.01	
VII. FKr ⁺ KrH ⁻	0.01	0.01	0.01	

Table 3. The Structure Weights of HXe₂F from CASVB(6,6)·CASVB(6,6)·CASVB(6,4) Calculations (Basis Set: MCP-TZP)

structure	weight	type
I. F Xe Xe H	0.53	diradical, F and H
II. F ⁻ Xe ⁺ Xe H	0.17	diradical, Xe ⁺ and H
III. F Xe Xe ⁺ H ⁻	0.14	diradical, F, Xe ⁺
IV. F ⁻ Xe Xe ⁺ H	0.06	
V. F ⁻ Xe ⁺ Xe ⁺ H ⁻	0.04	
VI. F ⁻ XeXe H ⁺	0.02	
VII. FXe ⁺ XeH ⁻	0.01	

Table 4. Atomic Partial Charge of HAR₂F from CASVB(6,6)·CASVB(6,6)·CASVB(6,4) Calculations (Basis Set: cc-pVTZ)

atomic charge	VB (a)	VB (b)	MP2/NBO ^a
H	0.34	-0.27	0.42
Ar (1)	0.58	0.27	0.51
Ar (2)	0.01	0.01	0.05
F	-0.93	-0.01	-0.98

^a Method of computation: MP2/def2-TZVPP//MP2/Def2-TZVPP.⁶

Table 5. The Structure Weights of the Two CASVB Solutions of HAR₂F (Basis Set: cc-pVTZ)

structure	state a	state b
I. F Ar Ar H	0.00	0.87
II. F ⁻ Ar ⁺ Ar H	0.02	0.01
III. F ArAr ⁺ H ⁻	0.00	0.11
IV. F ⁻ Ar Ar ⁺ H	0.57	0.00
V. F ⁻ Ar ⁺ Ar ⁺ H ⁻	0.01	0.00
VI. F ⁻ ArAr H ⁺	0.38	0.00
VII. FAR ⁺ ArH ⁻	0.00	0.00

remarkably similar; thus any possible uncertainty due to basis set choice is eliminated.

HXe₂F. Table 3 shows the structure weights for HXe₂F. A major difference of HXe₂F from the other two molecules is that the total weight for the nondiradical structures is, approximately, 13% for HXe₂F, while for HAR₂F and HKr₂F, it is 1% and 2–3% (depending on the basis set), respectively.

Diradical vs Charge Transfer State. Two VB wave functions with close total energies were obtained for HAR₂F. They differ by ~30 kcal/mol (lower is b, Table 4). The above solutions have very different charge distributions. The atomic charges obtained from the Mulliken population analysis of the two VB wave functions are shown in Table 4. The structure weights of states a and b show a dramatic difference (Table 5). Structures I–III have strong diradical characteristics [state b], while structures IV and VI have charge-

transfer features [state a]. We can label state a as a charge transfer state and b as a diradical. Similar results were obtained for HKr₂F. For HXe₂F, the VB calculations show only one diradical state. *Therefore, we believe that the derivatives HNg₂F have dominant diradical characteristics.*

III.2. The Ground State of HNg₂F by MS-CASSCF/CASPT2 Calculations. As the multiconfigurational calculations show, the HNg₂F systems display a number of close-lying excited states, in particular two of ¹Σ⁺ and one of ¹Π symmetry. Especially, the two ¹Σ⁺ states, related to the electronic configurations σ² and σσ* and belonging to the A₁ symmetry in the C_{2v} point group, strongly interact and require a large amount of electronic correlation to be properly described. This is mainly shown by the different descriptions obtained at different levels of theory. In particular, it will be shown that remarkable differences are observed between single vs multiple (two) root (state average, SA) CASSCF descriptions, and followed by MS-CASPT2 computations.

HAr₂F and HKr₂F. The electronic ground states of HAR₂F and HKr₂F have largely distinct descriptions at different levels of theory. Single reference methods like HF, MP2, or CCSD(T) give a closed-shell picture, characterizing the state as ¹Σ⁺(σ²), displaying a bound structure,⁶ a description at odds with that provided at the CASVB level in the previous section. When performing a single-root CASSCF/CASPT2 calculation, the resulting wave function has a predominantly closed-shell character. We have found, for instance for HAR₂F employing the 4440/12 active space, a weight of 70% of the (σ²) configuration, although the diradical-type (σσ*) configuration also contributes at 1%. However, the previous description remarkably changes a lot when performing a two-root state average SA (2) CASSCF calculation, followed by state-specific SS-CASPT2, which allows describing the state as ¹Σ⁺(σσ*), due to the predominance of the open-shell singlet diradical structure. For HAR₂F, the ground state wave function is composed of the following configurations: 24% (σ²) + 71% (σσ*), a description that does not change if we modify the number of averaged roots and that changes moderately when we apply the multi-state (MS) CASPT2 approach, which provides a PMCAS-CI description as 38% (σ²) + 56% (σσ*). The HOMO/LUMO orbitals are the same at the SA-CASSCF and PMCAS-CI levels. The HOMO σ orbital is mainly composed of the p_z of F, and the LUMO σ* orbital is mainly formed by the s of H with small contributions of the noble gas atoms. On the other hand, the low-lying excited state of the same symmetry (2¹Σ⁺), placed 0.036 au (0.98 eV) above the ground state, is given by 56% (σ²) + 38% (σσ*).

The picture of HKr₂F is similar to that of HAR₂F. Although the nonorthogonal single-state CASSCF/CASPT2 result provides a closed shell character for the ground state [75% (σ²)], the more accurate two-state MS-CASPT2 procedure yields a ground state description given by 40% (σ²) + 53% (σσ*). The 2¹Σ⁺ excited state lies 0.072 au (1.959 eV) above the ground state and is given by 53% (σ²) + 39% (σσ*) at the MS-CASPT2 level. These values were computed employing the 4440/12 space.

So far, it seems clear that the ground states of HAR₂F and HKr₂F can be characterized as ¹Σ⁺(σσ*), displaying a

predominantly diradicaloid and dissociative nature, as proved by CASVB and the most complete approach SA-CASSCF/CASPT2 and PMCAS-CI/MS(2)-CASPT2 multiconfigurational calculations. The most striking aspect is undoubtedly the dramatic change of the wave function composition displayed from the predominantly closed-shell single-root SR-CASSCF description to the two-root SA-CASSCF and PMCAS-CI characterizations, with prevalent open-shell singlet diradical structures. The latter is not entirely unexpected. Although both approaches display a multiconfigurational character, a single-root CASSCF wave function is nonorthogonal with respect to the other states of the same spin and spatial symmetry. If the single-root state strongly interacts with neighboring states like here, the obtained wave function can be incorrectly mixed or collapse, favoring one of the possible solutions, the closed-shell singlet here. This is common in cases of near degenerate states like conical intersections or avoided crossings.³⁶ Therefore, at least a two-state model needs to be used. SA-CASSCF wave functions are orthogonal for the computed states ($1^1\Sigma^+$ and $2^1\Sigma^+$ here), although the following second-order perturbation theory SS-CASPT2 solution actually displays nonorthogonal CASPT2 states.^{36,37} In cases like the present one, it is therefore compulsory to provide a PMCAS-CI/MS-CASPT2 multistate solution that finally yields fully orthogonal states and wave functions at the highest correlated level.

HXe₂F. A single-root CASSCF/CASPT2 calculation on HXe₂F gives a ground $1^1\Sigma^+$ state of predominantly closed-shell character [90% (σ^2)]. Different from the cases of the HAr₂F and HKr₂F molecules, the two-root state average CASSCF level (followed by state-specific (SS) CASPT2) provides a description where the closed-shell character slightly predominates, 49% (σ^2) + 44% ($\sigma\sigma^*$), a dominance that increases to 58% (σ^2) + 35% ($\sigma\sigma^*$) at the final PMCAS-CI/MS(2)-CASPT2 level. The orbital description is similar to that mentioned above for the other systems. The $2^1\Sigma^+$ excited state lies 0.114 au (3.099 eV) above the ground state and is described by 43% (σ^2) + 49% ($\sigma\sigma^*$) and 34% (σ^2) + 58% ($\sigma\sigma^*$) at the two-state SA-CASSCF/CASPT2 and PMCAS-CI/MS(2)-CASPT2 levels of theory, respectively.

The above results demonstrate that the multistate CASPT2 is the correct theory for the description of the ground state of HNg₂F and illustrate how the weight of the closed-shell σ^2 configuration increases with the atomic number of the noble gas atom, while the weight of $\sigma\sigma^*$ decreases.

The sequence of lowest-lying singlet excited states in the three molecules also illustrates the differences in their electronic structure. The MS-CASPT2 results are, for HAr₂F, $1^1\Sigma^+(\sigma\sigma^*) = 0.00$ eV, $1^1\Pi(\pi\sigma^*) = 0.60$ eV, and $2^1\Sigma^+(\sigma^2) = 0.98$ eV; for HKr₂F, $1^1\Sigma^+(\sigma\sigma^*) = 0.00$ eV, $1^1\Pi(\pi\sigma^*) = 1.55$ eV, and $2^1\Sigma^+(\sigma^2) = 1.96$ eV; and for HXe₂F, $1^1\Sigma^+(\sigma^2) = 0.00$ eV, $2^1\Sigma^+(\sigma\sigma^*) = 3.09$ eV, and $1^1\Pi(\pi\sigma^*) = 3.18$ eV. The location and properties of these low-lying excited states are basic features of HNg₂F, which lead to large NLO properties.²⁶ This will be discussed more extensively in section III.5.

It is known⁴⁷ that the lowest singlet–triplet gap provides an indication for the diradical character of a system. The lowest triplet state $3^3\Sigma^+$ has been computed for all of the

Table 6. The Static Dipole Moment and (Hyper)Polarizabilities of HXe₂F Computed Using Different Methods

property (au)	method			
	HF	MP2	CCSD	CCSD(T)
μ_z	6.454 ^a	4.855 ^a	4.929 ^a	3.788 ^a
	6.458 ^b	4.972 ^b	4.990 ^b	3.856 ^b
α_{zz}	206.96 ^a	349.05 ^a	312.63 ^a	420.43 ^a
	207.61 ^b	343.46 ^b	316.08 ^b	423.99 ^b
β_{zzz}	−9633 ^a	−19900 ^a	−11040 ^a	−11040 ^a
	−9630 ^b	−19010 ^b	−10900 ^b	−10700 ^b
$(\gamma_{zzzz} \times 10^{-3})$	720 ^a	853 ^a	−510 ^a	−4000 ^a
	700 ^b	745 ^b	−625 ^b	−4500 ^b

^a aug-cc-pVDZ basis set. For Xe, an effective core potential was used.⁴¹ ^b PolX basis set.

studied HNg₂F derivatives. It has been found that the triplet state has a strong diradical character (96% $\sigma\sigma^*$). The natural occupation numbers are 1.0 for both HOMO and LUMO. At the MS-CASPT2/ANO level, the triplet state lies 0.2 eV (4.7 kcal/mol), 0.64 eV (14.7 kcal/mol), and 1.25 eV (28.7 kcal/mol) higher in energy than the singlet ground state for the Ar, Kr, and Xe derivatives, respectively. Wirz⁴⁸ suggested that a diradical is “a molecular entity whose lowest singlet and triplet state energies do not differ by much more than kT, say 2 kcal mol^{−1}. The expression ‘biradicaloid’ would then extend this range to say 24 kcal mol^{−1}”. According to this definition, all three species are diradicaloids. The diradicaloid character depends on Ng and decreases as the atomic number of the noble gas atom increases.

As a conclusion, both final sets of results, computed with the CASVB and MS-CASPT2 methods, confirm a strong diradical element in the ground state of H–Ng–Ng–F, which follows the sequence Ar > Kr > Xe.

For comparison we note that the electronic ground state of HXeF is described by (σ^2), at the MS-CASPT2 level of theory (4550/16 active space). This contributes 88% to the ground state configuration, implying that the state is dominated by a strong closed shell character. In the lowest-lying excited state, which lies 7.07 eV above the ground state and thus very weakly interacting with it, the contribution of $\sigma\sigma^*$ is 88%. A similar picture emerges at the CASSCF level. Thus, the diradical character is clearly induced by the second Xe atom.

III.3. The L&NLO Properties. The computed property values for HXe₂F are presented in Tables 6 and 7. This is the system with the largest σ^2 character in its ground state. We will focus our attention on the longitudinal components, since the other ones are much smaller. For example, at the CCSD(T)/PolX level, we found that α_{xx} is 9.75 times smaller than α_{zz} . The PolX results are in satisfactory agreement with those computed with the aug-cc-pVDZ set (Table 6). It is observed that the dynamic electron correlation effect is quite strong for all of the considered properties and especially for the second hyperpolarizability, where at the CC level of theory a change of sign for γ_{zzzz} has been found, in comparison to that observed with the HF and MP2 methods. At the HF level of theory, the dipole moment of the system is overestimated, while a significant reduction of μ_z is observed by taking into account correlation. The very large effect of triples on γ_{zzzz} is noted.

Table 7. MS-CASPT2 Dipole Moment and (Hyper)Polarizability of HXe₂F. The PolX Basis Set was Used

property (au)	A ₁ B ₁ B ₂ A ₂ /act.elect.						
	2000/2	3000/4	2220/6	4220/8	2440/10	4440/12	4550/16
μ_z	3.347	3.841	4.938	4.454	5.007	4.290	4.719
α_{zz}	460.49	407.04	472.51	[3.564] ^a 372.74	[4.603] ^a 456.41	[3.926] ^a 398.94	383.20
β_{zzz}	-5460	-2450	-22640	[382.95] ^a -10200	[383.00] ^a -23990	[340.30] ^a -8300	-13800
$(\gamma_{zzzz} \times 10^{-3})$	-3670	-4130	-4120	[NC] ^b -3100	[-27000] ^a -4000	[-16100] ^a -3660	-3500
						[-15000] ^a	

^a Single root calculation. ^b Nonconvergence.

Table 8. The Dipole Moment and the (Hyper)Polarizabilities of HF, HXeF, and HXe₂F^a

property (au)	method					
	HF			CCSD(T)		
	HF ^b	HXeF ^c	HXe ₂ F ^c	HF ^b	HXeF ^c	HXe ₂ F ^c
μ_z	0.759	2.283	6.454	0.703	1.975	3.788
α_{zz}	5.59	51.25	206.96	6.19	59.59	420.43
β_{zzz}	-9.8	-436.7	-9633	-11.5	-582.1	-11040
$(\gamma_{zzzz} \times 10^{-3})$	0.219	16.9	720	0.284	22.7	-4000

^a The computations were performed with the aug-cc-pVDZ basis set. For Xe, an effective core potential of 28 was used.⁴¹ ^b The experimental geometry was used.⁵⁴ ^c The geometries were taken from ref 6.

We have employed the T_1 diagnostic in order to find an indication for the importance of nondynamical correlation effects. A T_1 value of 0.03 was found, another indication that the results from single reference methods are potentially unreliable.^{49–51}

The multiconfigurational computations are presented in Table 7, for which the MS-CASPT2/PolX method was employed. In order to check the reliability of the results, the effect of the enlargement of the active space is considered. Several spaces and a number of correlating electrons were used in order to fully support the adequacy of the multiconfigurational results. It is observed that inclusion of correlating π orbitals is quite important in order get reliable hyperpolarizability values (e.g., one may compare the results of the computations 2000/2 and 2220/6). This indicates the importance of the dynamic electron correlation effect on the considered properties. By further enlarging the 2220 space with two σ orbitals (one occupied, one correlated), a remarkable diminishment of the first and second order hyperpolarizability values is observed. By proper addition of π orbitals leading to either 4440/12 or 4550/16 spaces, a smaller dependence of the computed property values on the active space is observed. It is seen that SS-CASPT2 hyperpolarizability (β and γ) values differ remarkably in comparison with the MS-CASPT2 ones. This difference is particularly pronounced for γ_{zzzz} . The MS-CASPT2 property values, computed with the 4550/16 space, are in reasonable agreement with the CCSD(T) ones (Table 6).

III.4. The Effect of Xe Insertion. In this section, we shall consider the effect of inserting one and two Xe atoms in the HF system (leading to HXeF and HXe₂F) on the L&NLO properties (Tables 8 and 9).

Electron Correlation Effects. The electron correlation correction (ECC) for a property p , where $p = \mu_z, \alpha_{zz}, \beta_{zzz}$, or

Table 9. The Transition Dipole Moment (μ_{ge}), the Excitation Energy (ΔE), and the Dipole Moments of the Ground $1^1\Sigma^+$ (μ_g) and the Lowest-Lying $2^1\Sigma^+$ Excited State (μ_e)^a

property (au)	method					
	CCSD			MS-CASPT2 ^b		
	HF ^c	HXeF ^d	HXe ₂ F ^d	HF ^c	HXeF ^d	HXe ₂ F ^d
ΔE	0.681	0.259	0.117	0.521	0.256	0.107
$ \mu_{ge} $	0.463	1.647	3.948	0.502	1.761	4.199
μ_e				-0.513	-1.376	2.325
μ_g	0.709	2.051	4.990	0.650	1.956	4.719
$\Delta\mu_{eg}$				-1.163	-3.332	-2.394
β_{zzz}^e				-3.2	-473	-11070

^a All the properties were computed with the PolX basis set.

^b The 4550 space in C_{2v} point group was employed in these MS-CASPT2 computations. ^c The experimental geometry was used.⁵⁴ ^d The geometries were taken from ref 6. ^e Value computed by employing a two state model approximation.

γ_{zzzz} , is defined as $p^{ECC} = p[\text{CCSD(T)}] - p[\text{HF}]$ (HF: Hartree–Fock). It is observed that ECC decreases the value of μ_z and increases the values of α_{zz} , $|\beta_{zzz}|$, and $|\gamma_{zzzz}|$ (Table 6). The effect of ECC increases remarkably with the number of Xe atoms (Table 8). For example, insertion of one and two Xe atoms increases the value of γ_{zzzz} by 89.9 and 813.8 times, respectively.

Effect of Xe Insertion on the (Hyper)Polarizabilities. Comparing the results of HXe₂F with those of the HXeF and HF (hydrogen fluoride) systems, it is evident that Xe insertion significantly modifies all the presented properties, whereas its effect is quite enhanced, especially on β_{zzz} and γ_{zzzz} (Table 8). For example, the ratio $\gamma_{zzzz}[\text{HXeF}]/\gamma_{zzzz}[\text{HF}]$ is 80, while $\gamma_{zzzz}[\text{HXe}_2\text{F}]/\gamma_{zzzz}[\text{HXeF}]$ is -176. It is noted that the enhancement of the property, due to Xe, is multiplicative, that is,

$$p(\text{HXe}_2\text{F})/p(\text{HF}) = [p(\text{HXeF})/p(\text{HF})][p(\text{HXe}_2\text{F})/p(\text{HXeF})] \quad (\text{HF: hydrogen fluoride})$$

In order to further quantify the effect of the insertion of one and two Xe atoms in the HF system, we have computed the differential (hyper)polarizability, a property useful in the analysis of electron delocalization and bonding.⁵² These properties are defined as

$$\text{DP} = \alpha_{\text{diff}} = \alpha_{zz}(\text{HXe}_2\text{F}) - \alpha_{zz}(\text{HF}) - 2\alpha(\text{Xe}) \quad (1a)$$

$$\text{DHP} = \gamma_{\text{diff}} = \gamma_{zzzz}(\text{HXe}_2\text{F}) - \gamma_{zzzz}(\text{HF}) - 2\gamma(\text{Xe}) \quad (2)$$

We found, at the CCSD(T)/aug-cc-pVDZ level, $\text{DP} = 363.2$ au and $\text{DHP} = -3994.9 \times 10^3$ au. The corresponding

properties for the Xe atom are $\alpha = 27.87$ au and $\gamma = 20.03 \times 10^3$ au. *The dramatic effect on the (hyper)polarizabilities, induced by two Xe atoms, is quite clear.*

III.5. Interpretation of the Results. In this section, we will interpret the computed linear and nonlinear optical properties.

In particular, we would like to explain the very big change of β_{zzz} by inserting one and two Xe atoms in HF. We shall employ a two state-model (TSM), based on the sum over states (SOS) approach:⁵³

$$\beta_{zzz}(0)_{\infty} \approx \frac{3(\mu_e - \mu_g)\mu_{ge}^2}{(\Delta E)^2} \quad (3)$$

where μ_e and μ_g are the excited- and ground-state dipole moments, respectively, μ_{ge} is the transition dipole moment, and ΔE is the transition energy. The MS-CASPT2(4550/16) method with two roots and the PolX basis set were employed in order to compute the lowest-lying $2^1\Sigma^+$ excited state. Thus, for the implementation of eq 3, we have used μ_e , μ_g , and μ_{ge} values computed at the PMCAS-CI level, while the excitation energies were calculated by employing MS-CASPT2 theory.

Inserting one and, even more, two Xe atoms in HF leads to an increase of μ_{ge} and decrease of ΔE . Similarly $|\Delta\mu_{eg}| = |\mu_e - \mu_g|$ is larger for HXeF and HXeXeF, in comparison to that of HF. The above features explain the very big effect of one and, even more, two Xe atoms on the first hyperpolarizability. It is noted that β_{zzz} of HF, HXeF, and HXeXeF, computed by TSM (Table 9), is $-11\,070$ ($-11\,040$), -473 (-582.1), and -3.2 (-11.5) au, respectively. In parentheses are the corresponding CCSD(T) values (Table 8). Thus, the TSM model gives β_{zzz} values in qualitative agreement with those computed with the CCSD(T) method. A very good agreement between the TSM and the CCSD(T) values for the first hyperpolarizability of HXeXeF is observed, indicating the significant contribution of the $\sigma\sigma^*$ transition to the NLO response of this derivative.

Frenking et al.⁶ noted the following: “The substantial activation barrier, which is predicted at different levels of theory, suggests that HXeXeF could be observed in a low-temperature xenon matrix”. However, currently, HXeXeF is unlikely to be useful as an NLO material. The present NLO results are useful, because they demonstrate a mechanism (that is, insertion of noble gas atoms in a bond) for producing very large hyperpolarizabilities. Thus, although HXeXeF, currently, could not be considered as a material, it may show a way to produce one.

IV. Concluding Remarks

CASVB and CASPT2 computations have shown the diradicaloid character of the ground state of the HNg₂F systems, which decreases in the noble gas atom (Ng) order Ar > Kr > Xe. The diradicaloid character relies on the predominance of the $\sigma\sigma^*$ open-shell configuration in the ground state, the smallest in HXe₂F. The proper description of the state electronic structure requires the use of multiconfigurational approaches, but not only that. As the two low-lying states of the $1^1\Sigma^+$ symmetry strongly interact, a two-level treatment

like that provided by SA(2)–CASSCF/CASPT2 or, better, PMCAS–CI/MS(2)–CASPT2 is required, whereas single-root SR–CASSCF procedures may lead to nonorthogonal improper solutions. The decreasing diradicaloid character upon increasing the noble gas atomic number is confirmed by both the increasing singlet–triplet energy gap, which diminishes the diradical character of the ground state, and the singlet excited state structure, which evidences the stabilization of the closed-shell σ^2 configuration in the heaviest HXe₂F system. Both CCSD(T) and MS–CASPT2 computations have shown the great effect of the two Xe atoms on the NLO properties. Specifically, it has been found that insertion of one and two Xe atoms leads to an increase of β_{zzz} by 51 [= ($\beta_{zzz}\{\text{HXeF}\}/\beta_{zzz}\{\text{HF}\}$)] and 960 [= ($\beta_{zzz}\{\text{HXe}_2\text{F}\}/\beta_{zzz}\{\text{HF}\}$)] times, respectively, at the CCSD(T) level (Table 8). The corresponding increases of $|\gamma_{zzzz}|$ are 80 and 14 085 times, respectively.

Acknowledgment. The authors thank Professor G. Frenking for his useful comments. The research reported has been supported by the Spanish MICINN/FEDER projects CTQ2007-61260, CTQ2010-14892, and CSD2007-0010 Consolider-Ingenio in Molecular Nanoscience and by the Generalitat Valenciana. Grants in computing time from TerraGrid (USA) and DEISA (7th Framework Programme) are gratefully acknowledged.

References

- (1) Papadopoulos, M. G.; Waite, J. *J. Chem. Soc., Faraday Trans.* **1990**, *86*, 3525.
- (2) Waite, J.; Papadopoulos, M. G. *J. Phys. Chem.* **1990**, *94*, 6244.
- (3) Waite, J.; Papadopoulos, M. G. *J. Chem. Phys.* **1985**, *82*, 1427.
- (4) Avramopoulos, A.; Reis, H.; Li, J.; Papadopoulos, M. G. *J. Am. Chem. Soc.* **2004**, *126*, 6179.
- (5) Avramopoulos, A.; Serrano-Andrés, L.; Li, J.; Reis, H.; Papadopoulos, M. G. *J. Chem. Phys.* **2007**, *127*, 214102.
- (6) Jiménez-Halla, C. O. C.; Fernández, I.; Frenking, G. *Angew. Chem., Int. Ed.* **2009**, *48*, 366.
- (7) Pauling, L. *J. Am. Chem. Soc.* **1933**, *55*, 895.
- (8) Bartlett, N. *Proc. Chem. Soc. London* **1962**, 218.
- (9) Claassen, H. H.; Selig, H.; Malm, J. G. *J. Am. Chem. Soc.* **1962**, *84*, 3593.
- (10) Christie, K. O. *Angew. Chem., Int. Ed.* **2001**, *40*, 1419.
- (11) Berski, S.; Sivi, B.; Lundell, J.; Noury, S.; Latajka, Z. *New Trends in Quantum Systems in Chemistry and Physics*, Maruani, J., Minot, C., McWeeny, R., Smeyers, Y. G., Wilson, S., Eds.; Kluwer Academic Publishers: Norwell, MA, 1999; Vol. 1, p 259.
- (12) Brown, E.; Cohen, A.; Gerber, R. B. *J. Chem. Phys.* **2005**, *122*, 171101.
- (13) Khriachtchev, L.; Taskanen, H.; Lundell, J.; Pettersson, M.; Kiljunen, H.; Räsänen, M. *J. Am. Chem. Soc.* **2003**, *125*, 4696.
- (14) Papadopoulos, M. G.; Avramopoulos, A. *CP963, Computational Methods in Science and Engineering. Theory and Computation: Old Problems and New Challenges*. Maroulis, G., Simos, G., Eds.; American Institute of Physics: College Park, MD, 2007; p 316.

- (15) Ahokas, J.; Kunttu, H.; Khriachtchev, L.; Pettersson, M.; Räsänen, M. *J. Phys. Chem. A* **2002**, *106*, 7743.
- (16) Lundell, J.; Chaban, G. M.; Gerber, R. B. *J. Phys. Chem. A* **2000**, *104*, 7944.
- (17) Lundell, J.; Pettersson, M.; Khriachtchev, L.; Räsänen, M.; Chaban, G. M.; Gerber, R. B. *Chem. Phys. Lett.* **2000**, *322*, 389.
- (18) (a) Pettersson, M.; Lundell, J.; Räsänen, M. *Eur. J. Inorg. Chem.* **1999**, 729. (b) Lein, M.; Frunzke, J.; Frenking, G. *Struct. Bonding (Berlin)* **2004**, *106*, 181.
- (19) Lundell, J.; Chaban, G. M.; Gerber, R. B. *Chem. Phys. Lett.* **2000**, *331*, 308.
- (20) Ansbacher, T.; Gerber, R. B. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4175.
- (21) Khriachtchev, L.; Pettersson, M.; Runeberg, N.; Lundell, J.; Räsänen, M. *Nature* **2000**, *406*, 874.
- (22) Runeberg, N.; Pettersson, M.; Khriachtchev, L.; Lundell, J.; Räsänen, M. *J. Chem. Phys.* **2001**, *114*, 836.
- (23) Holka, F.; Avramopoulos, A.; Loboda, O.; Kellö, V.; Papadopoulos, M. G. *Chem. Phys. Lett.* **2009**, *472*, 185.
- (24) Stein, L.; Norris, J. R.; Downs, A. J.; Minihan, A. R. *J. Chem. Soc. Chem. Commun.* **1978**, 502.
- (25) Drews, T.; Seppelt, K. *Angew. Chem., Int. Ed.* **1997**, *36*, 273.
- (26) Serrano-Andrés, L.; Avramopoulos, A.; Li, J.; Labéquerie, P.; Bégué, D.; Kellö, V.; Papadopoulos, M. G. *J. Chem. Phys.* **2009**, *131*, 134312.
- (27) Li, J.; McWeeny, R. *Int. J. Quantum Chem.* **2002**, *89*, 208.
- (28) Li, J.; Duke, B. J.; Klapötke, T. M.; McWeeny, R. *J. Theor. Comput. Chem.* **2008**, *7*, 853.
- (29) Li, J.; Duke, B. J.; McWeeny, R. *VB2000*, version 2.1; SciNet Technologies: San Diego, CA, 2010. URL: <http://www.vb2000.net> (accessed Mar 2010).
- (30) Li, J.; Pauncz, R. *Int. J. Quantum Chem.* **1997**, *62*, 245.
- (31) McWeeny, R. *Proc. R. Soc. London, Ser. A* **1959**, *253*, 242.
- (32) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
- (33) Roos, B. O.; Andersson, K.; Fülcher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. *Adv. Chem. Phys.* **1996**, *93*, 219.
- (34) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-A.; Pedersen, T.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (35) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299.
- (36) Serrano-Andrés, L.; Merchán, M.; Lindh, R. *J. Chem. Phys.* **2005**, *122*, 104107.
- (37) Merchán, M.; Serrano-Andrés, L. *Ab Initio Methods for Excited States. Computational Photochemistry*; Olivucci, M., Ed.; Elsevier: Amsterdam, 2005.
- (38) Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107.
- (39) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley: New York, 2000.
- (40) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1998**, *53*, 1995. (b) <http://www.qch.fns.Uniba.sk/Baslib/POL> (accessed date March, 2010).
- (41) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113.
- (42) Jansik, B.; Schimmelpfennig, B.; Norman, P.; Mochizuki, Y.; Luo, Y.; Ågren, H. *J. Phys. Chem. A* **2002**, *106*, 395.
- (43) Davis, P. J.; Rabinowitz, P. *Numerical Intergration*; Blaisdell: London, 1967; p 166.
- (44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T., Jr.; 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, V.; 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, S.; Clifford, A. G.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
- (45) Chirgwin, B. H.; Coulson, C. A. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196.
- (46) (a) Sakai, V.; Miyoshi, E.; Klobukowski, M.; Huzinaga, S. *J. Chem. Phys.* **1997**, *106*, 8084. (b) Miyoshi, E.; Sakai, Y.; Tanaka, K.; Masamura, M. *THEOCHEM* **1998**, *451*, 41. (c) Noro, T.; Sekiya, M.; Koga, T. *Theor. Chem. Acc.* **1997**, *98*, 25. (d) Sekiya, M.; Noro, T.; Osanai, Y.; Koga, T. *Theor. Chem. Acc.* **2001**, *106*, 297.
- (47) Bachler, V.; Olbrich, G.; Neese, F.; Wieghardt, K. *Inorg. Chem.* **2002**, *41*, 4179.
- (48) Wirz, J. *Pure Appl. Chem.* **1984**, *56*, 1289.
- (49) Lee, T. J.; Rice, J. E.; Scuseria, G. E.; Schaefer, H. F., III. *Theor. Chim. Acta* **1989**, *75*, 81.
- (50) Lee, T. J.; Taylor, P. R. *Int. J. Quant. Chem. Symp.* **1989**, *23*, 199.
- (51) Lee, T. J.; Head-Gordon, M.; Rendell, A. P. *Chem. Phys. Lett.* **1995**, *243*, 402.
- (52) Maroulis, G. *J. Chem. Phys.* **2008**, *129*, 044314.
- (53) Oudar, L.; Chemla, D. S. *J. Chem. Phys.* **1977**, *66*, 2664.
- (54) Cadet, P. E.; Huo, W. M. *J. Chem. Phys.* **1967**, *47*, 614.

JCTC

Journal of Chemical Theory and Computation

PACE Force Field for Protein Simulations. 1. Full Parameterization of Version 1 and Verification

Wei Han,[†] Cheuk-Kin Wan,[†] Fan Jiang,^{†,‡} and Yun-Dong Wu^{*,†,‡,§}

Department of Chemistry, The Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong, China, School of Chemical Biology and Biotechnology, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, China, and College of Chemistry, Peking University, Beijing, China

Received June 8, 2010

Abstract: A further parametrization of a united-atom protein model coupled with coarse-grained water has been carried out to cover all amino acids (AAs). The local conformational features of each AA have been fitted on the basis of restricted coil-library statistics of high-resolution X-ray crystal structures of proteins. Potential functions were developed on the basis of combined backbone and side chain rotamer conformational preferences, or rotamer Ramachandran plots (ϕ , Ψ , χ_1). Side chain–side chain and side chain–backbone interaction potentials were parametrized to fit the potential mean forces of corresponding all-atom simulations. The force field has been applied in molecular dynamics simulations of several proteins of 56–108 AA residues whose X-ray crystal and/or NMR structures are available. Starting from the crystal structures, each protein was simulated for about 100 ns. The C α RMSDs of the calculated structures are 2.4–4.2 Å with respect to the crystal and/or NMR structures, which are still larger than but close to those of all-atom simulations (1.1–3.6 Å). Starting from the PDB structure of malate synthase G of 723 AA residues, the wall-clock time of a 30 ns simulation is about three days on a 2.65 GHz dual-core CPU. The RMSD to the experimental structure is about 4.3 Å. These results implicate the applicability of the force field in the study of protein structures.

Introduction

Protein modeling with molecular mechanics (MM) force fields plays an important role in computational biology.^{1–3} However, handling real systems with the current popular all-atom force fields is limited in both size and time scale due to the computationally demanding nature of these force fields.^{4–6} In recent years, there have been increasing efforts in developing coarse-grained (CG) force fields. Multiple interaction sites of biomolecules are simplified to one site, leading to a great increase in simulation capability. These force fields describe proteins or DNA/RNAs at various resolutions, ranging from the residue-based level,^{7–22} to the

intermediate level,^{23–34} and to the united-atom-based level.^{35–41} They have extended greatly the temporal and spatial scale of simulations.^{42–45} Some of them already have predictive capabilities.^{24,26,31,33,38,39}

Coping with environments such as explicit water requires a great deal of calculation time. Implicit solvent models are one way of reducing the computational cost. The generalized Born (GB) solvent model has been carefully optimized so that it fits well with the current all-atom force fields.^{1,46} Most CG force fields also do not consider solvent explicitly but treat solvent-induced effects such as hydrophobic interactions as pairwise additive interactions.^{7–13,20,23–34,38,39} The implicit solvent model may not be that efficient when the system becomes more complex, such as in the case of a heterogeneous protein/water/membrane system, in which the details of the membrane used are often of great interest. An alternative approach is to treat the environment explicitly in

* Corresponding author e-mail: chydwu@ust.hk.

[†] The Hong Kong University of Science & Technology.

[‡] Peking University Shenzhen Graduate School.

[§] Peking University.

a CG manner.^{14–19,21,22,35,37,40,41} A good example is the MARTINI force field derived by Marrink and co-workers,^{15–17} in which roughly four heavy atoms are reduced into one site for both the protein and environment. This force field has been successfully applied to protein/membrane interactions and membrane dynamics with a tremendous increase in efficacy.^{18,19} However, due to the loss of atomistic information of proteins in their model, the restraints on native contacts are needed to maintain the native structures of proteins.

Recently, we have been pursuing a protein force field^{40,41} that can be coupled with Marrink's CG environment but possesses adequate atomistic details for protein folding. By simulating the helix–coil transition of polyalanine-based peptides, we have initially demonstrated that such a force field coupled with the CG solvent is indeed possible.⁴⁰ Encouraged by the results, we are concentrating our effort on the extension of this force field to the simulation of real proteins.

To properly represent a protein, it is necessary to correctly describe the backbone (ϕ , Ψ) conformational preference of all 20 amino acids.^{47–52} For the amino acids with rotating side chains, the degree of freedom (χ) of the side chains needs to be considered.⁵³ Fitting gas phase Ramachandran plots of alanine and glycine dipeptides with high-level quantum mechanics (QM) calculations was a conventional way of optimizing backbone potentials (ϕ , Ψ).^{54–60} But during the development of the CHARMM–CMAP force field,⁴⁹ Brooks and co-workers found that even if the force field can reproduce exactly the high-level QM backbone potential in a vacuum, the systematic deviation in the (ϕ , Ψ) of the α helix and β sheet from experimental data is evident. They pointed out that the MM force field for the gas phase may not capture well the solvent dependence of the (ϕ , Ψ) potential in reality. Duan and co-workers⁴⁷ reoptimized the backbone potentials of the AMBER force field according to the high-level QM potential surface in a medium of $\epsilon = 4.0$. The improved force field can reproduce well the experimental structural properties of both dipeptides and short polypeptides.

More recently, Liu and co-workers⁵⁰ remodified the backbone potential of the GROMOS force field with a QM potential surface in water ($\epsilon = 78.0$) as a reference. Moreover, they empirically adjusted the parameters by fitting the free energy surface, instead of the potential surface, of their model to the QM data. The optimized force field is significantly improved in its ability to reproduce the native structures of proteins in time-intensive simulations. We here chose a similar way to optimize the backbone parameters of our force field through the calculation of free energy surfaces. The statistical Ramachandran surfaces from PDB were used as references.

In addition to the backbone parameters, we also need to optimize the parameters of the nonbonded interactions among atoms in 20 amino acids. Following a thermodynamics-based approach,^{14–17} a large part of these parameters were obtained by reproducing the experimental thermodynamic properties of organic molecules such as self-solvation free energies and hydration free energies.⁴¹ However, the parameters among

polar/charged groups remain absent due to a severe lack of direct experimental data on the pairwise interactions between polar/charged groups in solvent environments. As such, we plan to resort to explicit water simulations, from which the potential mean force (PMF) between polar/charged groups can be derived for the fitting. There are a number of successful precedent applications of explicit water PMF to the developments of force fields such as MARTINI⁶¹ and UNRES⁶² force fields. A similar strategy was used by Chen et al. for a GB implicit solvent model.⁶³

In this paper, we report our recent work toward the full parametrization of our force field, namely, the Protein in Atomistic details coupled with Coarse-grained Environment (PACE). Specifically, the backbone (ϕ , Ψ) potential was optimized to reproduce the Ramachandran plots from a protein data bank (PDB) coil library through aqueous simulations of dipeptides of 20 amino acids. The conformational (χ) distributions of side chains and the backbone preferences in different side chain conformations were also used in the fitting. The interactions between polar and charged sites in proteins were optimized by fitting the PMFs of simulations with the OPLS-AA/L force field in explicit water. Together with the previous parameters, our force field is already able to tackle proteins with real sequences. As a preliminary test, tens to hundreds of nanoseconds of aqueous simulations of several proteins (~ 50 AA to ~ 700 AA) were performed. Native structures were well preserved in all of the simulations (RMSD < 0.43 nm).

Models and Methods

The Coarse-Grained Protein Model. Our protein model is united-atom-based (UA). As shown in Figure 1, normally, each heavy atom together with the attached hydrogen atoms is represented by one site. But the hydrogen atoms in backbone amide groups and in the side chains of Asn, Gln, Trp, and His are also explicitly represented to better account for their hydrogen-bonding property. The total energy of our model is expressed in eq 1:

$$E = E_{\text{angle}} + E_{\text{torsion}} + E_{14\text{pair}} + E_{\text{improper}} + E_{\text{nonbond}} + E_{\text{HB}} \quad (1)$$

The detailed description of the above terms and their parametrization can be found in our previous papers.^{40,41} The first four terms describe the bonded interactions between the sites that are connected through not more than three covalent bonds. All bond lengths are constrained at their equilibrium values. All bond angles are restrained with a harmonic potential E_{angle} at their equilibrium values with a force constant of $K = 300$ kJ/mol/rad². The equilibrium values of bond lengths and angles are obtained through the QM calculations of 15 small molecules. All the values for bond length and bend angle are illustrated in Figure S1 in the Supporting Information (SI). E_{torsion} and $E_{14\text{pair}}$ describe the potential energy of a dihedral angle about a rotating bond, which has forms in eqs 2 and 3. $E_{14\text{pair}}$ stands for the interaction between sites separated by three bonds. For an amino acid, its side chain torsion potential is optimized by fitting the QM torsional potential

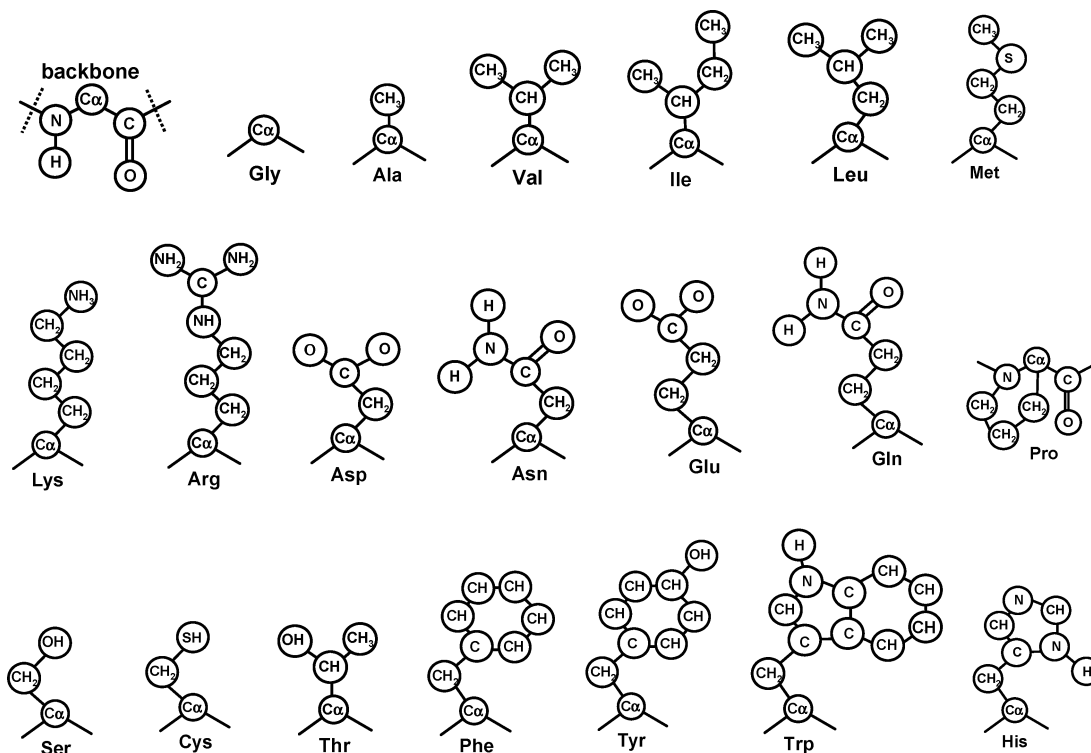


Figure 1. Schematic representations of amino acids in our model. Each circle indicates a site that is explicitly represented.

of 24 minima and 22 rotation barriers of simple molecules. But for the backbone part (ϕ , Ψ), the parameters will be optimized by fitting the experimental (ϕ , Ψ) maps in this work. E_{improper} is used to keep planar geometries and chiral centers in molecules. It imposes a harmonic potential on a dihedral of four sites with $K = 300$ kJ/mol/rad². All of the parameters for bonded terms can be found in Tables S1–S2 in the SI.

$$E_{\text{torsion}} = \sum_i K_{\text{torsion},i} [1 + \cos(n_i \zeta_i - \zeta_{0,i})] \quad (2)$$

$$E_{14\text{pair}} = \sum_{1-4\text{relationship}} 4\epsilon_{14-ij} \left(\frac{\delta_{14-ij}^{12}}{r^{12}} - \frac{\delta_{14-ij}^6}{r^6} \right) \quad (3)$$

E_{nonbond} and E_{HB} describe the interactions between sites beyond three-bond connections. E_{nonbond} is used for isotropic nonbonded interactions. These interactions consist of both van der Waals interactions and electrostatic interactions that are normally treated separately in all-atom force fields.^{58–60} But for simplicity, in our model, only a Lennard-Jones potential (eq 4) is used to represent the overall effect. Thus, all of the atomistic charges are set to zero in our current force field.

$$E_{\text{nonbond}} = \sum_{i \neq j} 4\epsilon_{ij} \left(\frac{\delta_{ij}^{12}}{r^{12}} - \frac{\delta_{ij}^6}{r^6} \right) \quad (4)$$

The CG water developed by Marrink et al. is a vdW sphere, representing a cluster of four water molecules. The LJ parameters ϵ and δ of the CG water site are 5.0 kJ/mol and 0.47 nm, respectively, which are optimized through reproducing the density and compressibility of

pure water. We have treated the interaction between CG water and protein sites with eq 4, with δ_{ij} as the summation of vdW radii of CG water and protein sites while ϵ_{ij} 's were optimized by fitting hydration free energies of 35 organic compounds.⁴¹ The average deviation from experimental data is about 1.1 kJ/mol. In our previous study, self-solvation free energies (or free energies of evaporation) of eight organic compounds were used to optimize E_{nonbond} parameters (ϵ_{ij} and δ_{ij}) for the interactions between protein sites.⁴¹ These compounds include cyclohexane, *n*-pentane, isopentane, benzene, diethyl ether, 2,3-dimethyl-2-butene, triethylamine, and dimethylsulfide. The average errors for density and self-solvation free energy are about 3.2% and 0.7 kJ/mol, respectively. These parameters of nonbonded parts are listed in Tables S3–S4 in the SI. It should be noted that to reproduce the solvation free energies and densities of these eight compounds only requires that E_{nonbond} parameters for the interactions between nonpolar sites and between nonpolar and polar sites are optimized. Thus, the interactions between polar sites such as HBs were left unparametrized.⁴¹ Therefore, one of our main focuses in this work is to obtain the parameters for the interactions between polar sites.

E_{HB} describes hydrogen bond (HB) interactions between two groups of sites such as amide groups. Because HB interactions have directionality, we have devised a set of attractive and repulsive potentials⁴⁰ which are simultaneously employed to maintain the directionality of HB between backbone amide groups (Figure 2a and eq 5), which has been adopted in the study by Dokholyan and co-workers.³⁹

$$E_{\text{HB}} = \sum_{|i-j|>2} \left[4\epsilon_{\text{attr}} \left(\frac{\delta_{\text{O}i-\text{NH}j}^{12}}{r_{\text{O}i-\text{NH}j}^{12}} - \frac{\delta_{\text{O}i-\text{NH}j}^6}{r_{\text{O}i-\text{NH}j}^6} \right) + 4\epsilon_{\text{rep}} \frac{\delta_{\text{O}i-\text{C}\alpha j}^{12}}{r_{\text{O}i-\text{C}\alpha j}^{12}} \right. \\ \left. + 4\epsilon_{\text{rep}} \frac{\delta_{\text{O}i-\text{C}j-1}^{12}}{r_{\text{O}i-\text{C}j-1}^{12}} + 4\epsilon_{\text{rep}} \frac{\delta_{\text{C}i-\text{NH}j}^{12}}{r_{\text{C}i-\text{NH}j}^{12}} \right] \quad (5)$$

Here, only the HB between two amides that are separated by at least two residues is considered. All of the parameters have been optimized by fitting the change of free energy, entropy, and enthalpy of the helix–coil simulations of polyalanine-based peptides in our previous work.⁴⁰ All of the parameters are kept unchanged in this work except for ϵ_{attr} between N and O atoms, which will be further optimized in this work. The parameters for E_{HB} are shown in Table S5 in the SI.

The backbone HB potential is not applicable to the HB involving side chain amide groups of Asn and Gln. This is because in these amide groups an amide nitrogen atom can have two hydrogen atoms as donors, but the original HB potential can only handle one hydrogen donor on each nitrogen atom. Therefore, these hydrogen atoms need to be explicitly represented, and another type of HB potential should be used (Figure 2b and eq 6).

$$E'_{\text{HB}} = \sum_{|i-j|>2} \left[4\epsilon'_{\text{attr}} \left(\frac{\delta'_{\text{O}-\text{H}}^{12}}{r_{\text{O}-\text{H}}^{12}} - \frac{\delta'_{\text{O}-\text{H}}^6}{r_{\text{O}-\text{H}}^6} \right) + 4\epsilon'_{\text{rep}} \frac{\delta'_{\text{C}-\text{H}}^{12}}{r_{\text{C}-\text{H}}^{12}} + 4\epsilon'_{\text{rep}} \frac{\delta'_{\text{O}-\text{N}}^{12}}{r_{\text{O}-\text{N}}^{12}} \right] \quad (6)$$

We have applied this approach to the HB interactions between the side chains of Asp, Asn, Glu, Gln, His, and Trp (Figure 1) and between these side chains and the backbone amide groups. The optimization of parameters for backbone–backbone interactions is based on the contents of secondary structures of peptides. The parameters for side chain–side chain and backbone–side chain HB interactions are optimized by fitting the all-atom PMF. The details of the optimization will be discussed in the Results and Discussions.

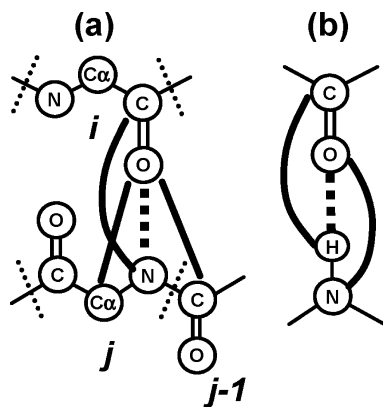


Figure 2. Schematic representation of hydrogen bond potentials (a) between backbone amide groups and (b) between donors and acceptors in side chains. The solid lines denote repulsive potentials, and the dotted lines denote attractive potentials.

The complete force field package is available upon readers' requests. This package includes the force field files that are compatible with the GROMACS software (v3.3) and a written code that can automatically modify topology files so that our force field can be used with GROMACS.

Dipeptide Simulations. All of the simulations were performed with the GROMACS software packages (version 3.3.1).⁶⁴ A dipeptide molecule (Ace–Xxx–NMe) that is capped at the two ends was placed in a dodecahedron box so that the minimum distance between box edges and the molecule was about 1.4 nm. The box was filled with 400–500 CG water particles. All bonds are constrained with the LINCS algorithm.⁶⁵ All nonbonded interactions were shifted to zero between distances of 0.9 and 1.2 nm. The system was optimized with the steepest descent methods. After a 5000-step optimization, the system was pre-equilibrated at 300 K and 1.0 atm for 50 ps. The pressure and the temperature were controlled by a thermostat and a pressure bath with coupling constants of 0.1 and 0.5 ps, respectively.⁶⁶ After the pre-equilibrium, the system was heated at 700 K with a constant volume for 10 ns. The last 5 ns of the heating simulation were used to generate the starting conformations for production simulations. Dummy atoms were used to remove the fastest motion of the system involved with hydrogen.⁶⁷ These dummy atoms are virtual interaction sites whose positions are determined by three nearby heavy atoms. Thus, in all of the normal molecular dynamics (MD) simulations including protein simulations, a large time step of 6 fs can be used, which has been shown to satisfy energy conservation during simulations.⁶⁷ As in our force field, all of the parameters related to hydrogen are optimized with the treatment of dummy sites; we suggest that for consistency, dummy sites should be used for explicit hydrogen atoms in simulations with our force field.

Replica exchange molecular dynamics (REMD) simulations provide an efficient method to perform equilibrium simulations.^{68,69} After combining with our fast UA model, the REMD method becomes beneficial to parametrization. Our REMD simulations contained 16 replicas with temperatures ranging from 300 to 431 K at 1 atm. Each replica started with a different conformation generated from the heating simulation at 700 K. Exchanges were attempted every 2 ps. The successful exchange rate was about ~15%. For each REMD simulation, each replica lasted for 50 ns. The last 40 ns of the simulation at 300 K were used for analysis.

Finally, in all of the REMD simulations of dipeptides, the Newton equation was integrated every 10 fs. Since most sites in our model represent only a single atom and are connected directly by strong covalent bonds, the system is prone to crashes at large time steps. This is particularly obvious in the REMD simulations, as simulation at a high temperature (up to 431 K) is involved. To avoid crashes, we treated all explicit hydrogen atoms as dummy sites and tripled the mass of all proteins. Our previous studies and the peptide simulations in an accompanying paper show that the mass scaling plus the use of dummy atoms should not affect the thermodynamic properties of the system, which is our main interest in the current study.⁴⁰

PMF Calculations. PMF calculations were performed in a rectangular box with a size of $7.0 \times 2.9 \times 2.9$ nm. A pair of small molecules was put in the center of the box solvated by about 500 CG water particles. To obtain the PMF of solutes at a particular orientation, solutes were constrained to move along a straight line. The PMF was calculated using the free energy perturbation method in this work. Free energy perturbation calculates the free energy change of a solute pair separated at different distances. The positions of the solute pair are restrained so that the pair is either in a close contact (~ 0.26 nm) or widely separated (~ 2.2 nm). The positions can be made a function of a coupling parameter (λ). As λ slowly varies from zero to unity (10 000 000 perturbation steps with a time step of 6 fs), the solutes are moved away from each other. The free energy difference between the pair in any intermediate state and the close contact state is determined as the accumulated energy difference between the conformation corresponding to λ and that corresponding to $\lambda + \delta\lambda$ at each perturbation step. The free energy at the longest distance is set to zero as the reference point. In each perturbation simulation, the free energies of 100 intermediate states were chosen for PMF plots. For each PMF calculation, six to eight perturbation simulations were performed to generate an averaged PMF curve. All simulations were kept at 300 K and 1 atm throughout the whole simulation.

PMFs were also calculated in all-atom simulations for comparison when necessary. The calculation protocol is similar to that of the CG simulation. The OPLS-AA force field⁷⁰ and TIP3P water model were used. As long-range electrostatics might be important in PMF calculations, PME summation⁷¹ was used in all atomistic simulations. In the CG PMF calculations, the shift potential was used for the nonbonded interactions like our normal MD and REMD simulations. About 2000 water molecules were placed in a box. Each perturbation simulation was performed for 5 000 000 perturbation steps with a time step of 2 fs. A total of 100 intermediate states were chosen in each simulation, and six to eight perturbation simulations were carried out for each PMF calculation.

Statistical Analysis. Protein X-ray crystal structures of a resolution < 0.2 nm and with an R factor < 0.2 were selected and downloaded from the Protein Data Bank (PDB),⁷² with a 50% sequence identity cutoff. There are a total of 4220 protein structures and 2.0×10^6 residues. When there are n identical chains in a protein structure, each chain is counted with a $1/n$ statistical weight. To obtain the local conformational preferences, we only chose the coil residues that are not adjacent to any secondary structures,⁷³ as previously suggested.⁷⁴ The residues that have backbone atoms with temperature factor $B > 36$ or preceding prolines⁷⁵ are also excluded from our new restricted coil library (Coil-R for short). Furthermore, we exclude the residues preceding a turn-like α conformation ($-60^\circ < \phi < +60^\circ$, both α_L and α_R). The reason is because the side chains of some amino acids such as Ser and Asx can form hydrogen bonds with the backbones of the next residues if they are in the α conformation.

Secondary Structure Analysis. As suggested by Garcia et al.,⁷⁶ a residue is considered to be helical only if this residue and both of its neighboring residues have their backbone dihedral (ϕ, Ψ) within ($-60^\circ \pm 30^\circ, -47^\circ \pm 30^\circ$). If the backbone dihedral (ϕ, Ψ) of two residues are within ($-135^\circ \pm 45^\circ, +135^\circ \pm 45^\circ$) and there is at least a HB between the adjacent backbone amide groups to the two residues, the two residues are considered as having β sheets. A HB is considered as formed when the donor–acceptor distance is shorter than 0.35 nm and the donor–hydrogen–acceptor angle is larger than 120.0° .

Results and Discussions

Parameterization of Backbone Potentials. As a first step, we carried out parametrization of backbone potentials using the dipeptides of glycine and alanine. We used our recently reported statistical potentials derived from a coil library of the high-resolution X-ray crystal structure database.⁷³ Wang and co-workers have shown that the high-level QM potential surfaces of alanine and glycine dipeptides in an implicit water solvent are remarkably similar to the statistical results.⁴⁸ A similar observation was reported by Hu et al.⁵² in their QM/MM simulations of alanine and glycine dipeptides. Table 1 shows a comparison between the surfaces from our coil library and the surfaces from the high-level QM calculation in a water solvent. The two types of surfaces agree well with each other in not only the depths of minima but also the heights of transition barriers. This is expected for the coil library, which reflects the conformational features of amino acids in solvent environments.

Our coil library⁷³ purposely avoids the nonlocal effects of other amino acids such as vdW and HB interactions. Therefore, the derived statistical surfaces correspond to the intrinsic conformational preference of amino acids, which is supported by the comparison with the QM results (Table 1). In folded proteins, besides the intrinsic preference of amino acids, nonlocal vdW and HB interactions are also important. These nonlocal effects (E_{nonbond} and E_{HB}) are separately treated and parametrized in our force field. This strategy of combining the intrinsic preference and the nonlocal effects appears to work well, as supported by our protein simulations in this work (see later) and the folding simulations of peptides in our accompanying paper.

The optimization of the backbone dihedral (ϕ, Ψ) potential is empirical; that is, each further adjustment of parameters is based on the free energy surface of the last simulation. All of the parameters of ϕ and Ψ are optimized, including the dihedral potentials of ϕ and Ψ and all of the related 1–4 pair interactions. The optimized values from our previous study⁴⁰ are used as the starting input. We also found that several nonbonded interactions (Figure 3) are important to the (ϕ, Ψ) free energy surface, such as $N_i \cdots H_{i+1}$, $H_i \cdots H_{i+1}$, $C_{\beta i} \cdots H_{i+1}$, $O_{i-1} \cdots O_i$, $O_{i-1} \cdots C_{\beta i}$, $O_{i-1} \cdots N_{i+1}$, and $O_{i-1} \cdots C_i$. Some of these interactions, such as $O_{i-1} \cdots O_i$, $O_{i-1} \cdots N_{i+1}$, and $O_{i-1} \cdots C_{\beta i}$, have been suggested to be the most effective in shaping the (ϕ, Ψ) surface in a geometric analysis of the PDB structures by Ho et al.⁷⁷ Therefore, the interaction parameters of these atom pairs are also optimized to reproduce the target surface.

Table 1. Comparison of Positions (ϕ , Ψ in deg); Free Energy Difference (kJ/mol) of β , PPII, PPII', α_R , and α_L Conformations; and Barrier Heights (kJ/mol) between the Conformations for Different Methods Including the Coil Library, Quantum Mechanics Calculations by Wang et al.,⁴⁸ and our UA Model^{a,b}

	β	PPII	PPII'	α_R	α_L	$\beta \rightarrow \alpha_R$	$\beta \rightarrow \text{PPII}$	PPII $\rightarrow \alpha_L$
Ala								
coil	(-155,155)	(-65,145)	(55, -135)	(-65, -35)	(55,45)	8	4	18
	3.0	0.0	12.4	2.1	6.9			
QM	(-156,144)	(-64,142)	N/A	(-70, -32)	(59,41)	7	1	24
	0.0	0.0		1.0	4.0			
UA	(-165,145)	(-75,145)	(65, -145)	(-75, -25)	(65,35)	8	4	20
	2.4 \pm 0.1	0.0	10.4 \pm 2.2	1.4 \pm 0.2	7.7 \pm 1.1			
Gly								
coil	(-175,175)	(-75,165)	(75, -165)	(-85, -15)	(75,15)	9	4	20
	3.3	1.4	1.2	3.0	0.0			
QM	(-166,173)	(-66,153)	(66, -153)	(-66, -26)	(66,26)	7		
	3.3	1.2	1.2	0.0	0.0			
UA	(-175,175)	(-75,145)	(75, -145)	(75,25)	(75, 25)	8	6	16
	3.7 \pm 0.2	1.5 \pm 0.4	1.7 \pm 0.4	0.1 \pm 0.5	0.0			

^a The calculation is made at the MP2/cc-pVTZ/HF/6-31G** level in $\epsilon = 78.0$ medium by Wang et al.⁴⁸ ^b The REMD simulations of dipeptide in CG water are performed at 300 K. Standard deviations were estimated from block averages with a block size of 10 ns.

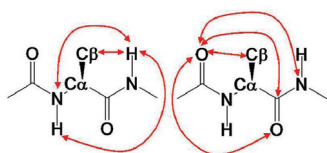


Figure 3. Schematic representation of remodified intraresidue nonbonded interactions, as denoted by red arrows.

Table 2. Summary of the Parameters of n , ζ_0 (deg), and K_{torsion} (kJ/mol) in eq 2 for ϕ ($\angle \text{C}-\text{N}-\text{C}_\alpha-\text{C}$) and Ψ ($\angle \text{N}-\text{C}_\alpha-\text{C}-\text{N}$); the Parameters of ϵ_{14} (kJ/mol) and δ_{14} (nm) for Related 1-4 Pairs; and the Parameters of ϵ (kJ/mol) and δ (nm) for Remodified Intrabackbone Lennard-Jones Interactions

AA	dihedral	K_{torsion}	n	ζ_0
Ala	ϕ	2.5	3	0
		1	1	-180
	Ψ	4	2	-180
		1	1	0
Gly	ϕ	2.5	3	0
		3.5	1	-180
	Ψ	4	2	-180
		1	1	-180

1-4 pair	ϵ_{14}	δ_{14}	1-4 pair	ϵ_{14}	δ_{14}
C-C	0.9	0.23	C-C β	0.9	0.28
H-C β	0	0	H-C	0	0
N-O	0.9	0.275	N-N	0.9	0.24
C β -O	0.9	0.28	C β -N	0.9	0.31

nonbond	ϵ	δ	nonbond	ϵ	δ
O $_{i-1}$ -C $_i^a$	0.6	0.27	O $_{i-1}$ -N $_{i+1}$	0.894	0.26
N $_i$ -H $_{i+1}$	2.98	0.183	O $_{i-1}$ -C $_{\beta i}$	0.894	0.32
O $_{i-1}$ -O $_i$	0.8	0.33	H $_i$ -H $_{i+1}^b$	1.95	0.22
C $_{\beta i}$ -H $_{i+1}$	0.447	0.315			

^a The subscripts indicate the identities of atoms in amino acids (refer to Figure 3). ^b For these interactions, only the repulsive part of the Lennard-Jones terms, $4\epsilon\delta^{12}/r^{12}$, is used.

Results for Alanine Dipeptide. The optimized parameters for alanine dipeptides are listed in Table 2. The simulated (ϕ , Ψ) map with the optimized parameters is shown in Figure 4c. Compared to the (ϕ , Ψ) map from the coil library (Figure

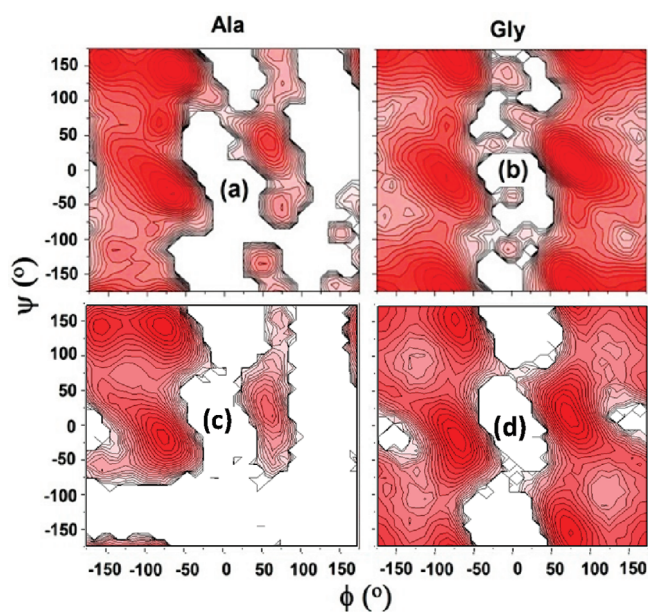


Figure 4. Statistical (ϕ , Ψ) PMFs of (a) Ala and (b) Gly dipeptides from the coil library. The calculated (ϕ , Ψ) PMFs with our UA model are shown in c for Ala and d for Gly dipeptides. The gap between the contour lines indicates a free energy difference of 1 kJ/mol.

4a), all of the major minima including α_R , α_L , β , and PPII are reproduced by our UA model. Table 1 contains a detailed comparison of the positions and relative free energies of these minima and the transition barrier heights between the minima from the coil library, the high-level QM calculations, and our simulations. The global minimum of the (ϕ , Ψ) surface in our simulations is the PPII conformer for Ala, in agreement with both the statistical analysis and QM calculations. The relative free energies of β , α_R , and α_L to PPII are 2.4, 1.4, and 7.7 kJ/mol, respectively. Our result of the stability of the β -conformer is closer to the statistical value (3.0 kJ/mol) than the QM value (0.0 kJ/mol). The relative stability (1.4 kJ/mol) of α_R conformers in our simulations is however closer to the QM results (1.0 kJ/mol) than to the statistical ones (2.1 kJ/mol). As for the α_L conformers, their stability estimated with our model is the same as that of the statistical

analysis. Considering the previously examined force fields with the relative stability of α_L at more than 12.5 kJ/mol,⁵² our model has considerably raised the stability of α_L . Besides the local minima, the barriers between the minima are also well reproduced. The heights of the barriers between β and α_R , between β and PPII, and between PPII and α_L are 8, 4, and 20 kJ/mol, respectively, which are very close to the statistical values (8, 4, and 18 kJ/mol, respectively).

Results for Glycine Dipeptide. As shown in Figure 4, Ala and Gly have quite different Ramachandran plots, due to the replacement of C_β in Ala by H_α in Gly. As our force field implicitly represents most H atoms, including H_α , the difference between Ala and Gly has to be reflected in another way, such as using different backbone potentials for Ala and Gly. We found that most of the backbone parameters remain unchanged, except for a minor change in the torsional potential of ϕ (Table 2), which is already good enough to reproduce the statistical result of Gly.

Figure 4d shows our (ϕ, Ψ) free energy surface of Gly. Since Gly has no chiral center, its (ϕ, Ψ) map should have C_2 symmetry about $(\phi \sim 0, \Psi \sim 0)$. In the plot of Gly from the coil library (Figure 4b), an asymmetry appears. That is, α_L is more populated than α_R . This indicates a trace of the coupling between Gly and its adjacent residues in proteins despite our effort to remove the coupling (see Models and Methods). Except for this, the two plots (Figure 4b and d) are quite similar. Five major minima are well reproduced, including α_R/α_L ($\phi \sim \pm\pi/2, \Psi \sim 0$), PPII/PPII' ($\phi \sim \pm\pi/2, \Psi \sim \pi$) and β ($\phi \sim \pi, \Psi \sim \pi$). The global minima are α_L and α_R , and the relative free energies of PPII, PPII', and β are 1.5, 1.7, and 3.7 kJ/mol, respectively, compared quite favorably with the statistical results (1.4, 1.2, and 3.3 kJ/mol, respectively) as well as the QM-calculated results (1.2, 1.2, and 3.3 kJ/mol).

Parameterization of Side Chain Potentials (χ). Except for Ala, Gly, and Pro, other amino acids have rotating side chains. In particular, statistical analyses^{73,78} have revealed that the degree of freedom of a side chain dihedral angle χ ($\angle N-C\alpha-C\beta-C\gamma$) plays a critical role in determining the conformational features of amino acids. Following the convention, according to the range of the χ angle, side chains are defined as being in the g+ ($-120^\circ < \chi < 0^\circ$), g- ($0^\circ < \chi < 120^\circ$), and t ($120^\circ < \chi < 240^\circ$) states (IUPAC-IUB Commission on Biochemical Nomenclature, 1970).

Just like the coupling between ϕ and Ψ , the coupling between χ and (ϕ, Ψ) is also controlled by a set of nonbonded interactions between special atom pairs.⁷⁸ Figure 5 illustrates these interactions. They can be divided into two kinds. The first kind is the steric repulsion between side chains (atoms at γ and δ positions) and their neighboring backbone amides (C, N, O, and H). In our testing simulations, normal nonbonded parameters appeared too repulsive for these interactions. The modified parameters should have reduced repulsive forces. One exception is the repulsion involving backbone H sites. Our previous QM calculations have shown that the repulsion between C_γ and backbone H sites is important in determining the relative energies of α_R and β conformations for amino acids such as Val.⁴⁰ Therefore, this repulsion was taken into account. The second

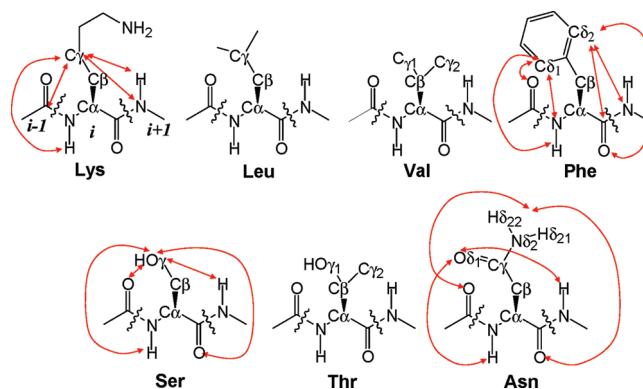


Figure 5. Schematic representation of local side chain–backbone interactions that are optimized.

kind of remodified interactions include those polar interactions between side chain polar groups and backbone amides. The normal nonbonded parameters can be too strong for these interactions. As revealed by our testing simulations, these parameters may distort the conformational distributions of backbones and side chains from the statistical results. This is especially severe for amino acids such as Asp, Asn, Ser, Thr, and His, all of which have polar groups close to their backbones. The local side chain–backbone interactions are major parameters to be optimized.

In an ideal case, all amino acids would share exactly the same set of parameters, which requires a global optimization. Although such global optimization will be expensive with our empirical fitting procedures, we are heading toward this end. The current approach is a compromise. That is, all backbone parameters including most of the ϕ , Ψ , and χ torsional terms and the intraresidue interaction terms are the same for all amino acids. However, the torsional terms of the ϕ , Ψ , and χ dihedral angles, the multiplicities (n) of which are one, were fine-tuned for individual amino acids in order to account for the distinct α_R , β , or even α_L propensities of different amino acids. Although this leads to a loss of transferability of the force field, it significantly simplifies the parametrization task and does not appear to lose applicability to real peptide systems. In the accompanying paper, we demonstrate that starting from random coils our force field can fold peptides with different sequences into their native structures.

Our parametrizations of the local interactions and torsional potentials are guided by two principals: (1) to match the populations of side chain rotamers (g+, g-, and t) and (2) to fit the Ramachandran plots of the three rotamers of each amino acid. In each side chain rotamer, there are four important minima including α_L , α_R , PPII, and β . For each amino acid, there are up to 12 minima (or basin). Their positions and basin depths can be fitted. Thus, the reference data should be more than enough to optimize all of the parameters. It should be noted that for most of the amino acids, their most favorable side chain conformations make up more than 50% of the population. Therefore, during the optimization, the backbone conformational distribution in the most favorable side chain conformations has a higher priority to be fitted than those in the other side chain conformations. All of the optimized parameters are summarized in Table 3.

Table 3. Summary of the Parameters of n , ζ_0 (deg), and K_{torsion} (kJ/mol) in eq 2 for ϕ ($\angle\text{C}-\text{N}-\text{C}_\alpha-\text{C}$), Ψ ($\angle\text{N}-\text{C}_\alpha-\text{C}-\text{N}$), and c ($\angle\text{N}-\text{C}_\alpha-\text{C}_\beta-\gamma$); the 1–4 Pair Parameters of ε_{14} (kJ/mol) and δ_{14} (nm) about χ ; and the Parameters of ε (kJ/mol) and δ (nm) for Remodified Local Backbone–Side Chain Lennard-Jones Interactions

	dihedral	K_{torsion}	n	ζ_0	1–4 pair	ε_{14}	δ_{14}	non-bond	ε	δ
All	ϕ	2.5	3	0						
	Ψ	2	4	–180						
	χ	4.9	3	0						
Lys/Arg	χ	0.5	1	180	$\text{C}_{\gamma i}-\text{N}_i^a$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}$	0.447	0.315
Gln/Glu	Ψ (Leu)	0.75	1	0	$\text{C}_{\gamma 1}-\text{C}_i$	0.1	0.33	$\text{C}_{\gamma i}-\text{N}_{i+1}$	0.447	0.35
Met/Leu	χ (Leu)	0.75	1	120				$\text{C}_{\gamma i}-\text{C}_{i-1}$	0.224	0.33
Val	ϕ	2	1	0	$\text{C}_{\gamma 1,i}/\text{C}_{\gamma 2,i}-\text{N}_i$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}$	0.447	0.29
	Ile	Ψ	2	1	0	$\text{C}_{\gamma 1,i}/\text{C}_{\gamma 2,i}-\text{C}_i$	0.1	0.33	$\text{C}_{\gamma i}-\text{N}_{i+1}$ $\text{C}_{\gamma i}-\text{C}_{i-1}$	0.447 0.224
Phe	ϕ	1.75	1	0	$\text{C}_{\gamma i}-\text{N}_i$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}$	0.224	0.3
	Tyr	Ψ	2.5	1	0	$\text{C}_{\gamma 1}-\text{C}_i$	0.1	0.3	$\text{C}_{\gamma i}-\text{C}_{i-1}$	0.224
Trp	χ	0.4	1	120				$\text{C}_{\gamma i}-\text{N}_{i+1}$	0.447	0.32
								$\text{C}_{\delta 1,i}/\text{C}_{\delta 2,i}-\text{C}_{i-1}/\text{C}_i$ $\text{C}_{\delta 1,i}/\text{C}_{\delta 2,i}-\text{N}_i$ $\text{C}_{\delta 1,i}/\text{C}_{\delta 2,i}-\text{N}_{i+1}$ $\text{C}_{\delta 1,i}/\text{C}_{\delta 2,i}-\text{H}_i$ $\text{C}_{\delta 1,i}/\text{C}_{\delta 2,i}-\text{O}_{i-1}$	0.224 0.447 0.447 0.224 0.447	0.3 0.3 0.34 0.3 0.29
Asn	ϕ	1	1	0	$\text{C}_{\gamma i}-\text{N}_i$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}^b$	7.5	0.235
	χ	2	1	–150	$\text{C}_{\gamma i}-\text{C}_i$	0.1	0.3	$\text{C}_{\gamma i}-\text{C}_{i-1}$ $\text{C}_{\gamma i}-\text{N}_{i+1}$ $\text{N}_{\delta 2,i}-\text{N}_i$ $\text{N}_{\delta 2,i}-\text{C}_{i-1}/\text{C}_i$ $\text{N}_{\delta 2,i}-\text{N}_{i+1}$ $\text{H}_{\delta 21,i}-\text{O}_i^c$ $\text{O}_{\delta 1,i}-\text{H}_i/\text{H}_{i+1}^c$ $\text{O}_{\delta 1,i}-\text{O}_{i-1}/\text{O}_i^b$ $\text{O}_{\delta 1,i}-\text{C}_{i-1}/\text{C}_i$	0.224 0.447 0.447 0.224 0.447 22 25 15 0.447	0.3 0.32 0.3 0.3 0.34 0.16 0.16 0.235 0.29
Asp	ϕ	1.5	1	0	$\text{C}_{\gamma i}-\text{N}_i$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}^b$	7.5	0.235
	χ	2.75	1	165	$\text{C}_{\gamma i}-\text{C}_i$	0.1	0.3	$\text{C}_{\gamma i}-\text{N}_{i+1}$ $\text{C}_{\gamma i}-\text{C}_{i-1}$ $\text{O}_{\delta 1,i}/\text{O}_{\delta 2,i}-\text{O}_{i-1}/\text{O}_i^b$ $\text{O}_{\delta 1,i}/\text{O}_{\delta 2,i}-\text{C}_{i-1}/\text{C}_i$ $\text{O}_{\delta 1,i}/\text{O}_{\delta 2,i}-\text{H}_i^c$ $\text{O}_{\delta 1,i}/\text{O}_{\delta 2,i}-\text{H}_{i+1}^c$ $\text{C}_{\gamma i}-\text{O}_i$	0.447 0.25 15 0.447 25 19.2 6	0.32 0.36 0.235 0.29 0.16 0.162 0.26
Cys	Ψ	0.5	1	0	$\text{S}_{\gamma i}-\text{N}_i$	0.1	0.31	$\text{S}_{\gamma i}-\text{N}_{i+1}$	0.447	0.35
	χ	1	1	–120	$\text{S}_{\gamma i}-\text{C}_i$	0.1	0.35	$\text{S}_{\gamma i}-\text{C}_{i-1}$	0.224	0.33
Ser	ϕ	1.5	1	–120	$\text{O}_{\gamma i}-\text{N}_i$	0.1	0.3	$\text{O}_{\gamma i}-\text{H}_i/\text{H}_{i+1}$	2.5	0.22
	χ	1.5	1	–120	$\text{O}_{\gamma i}-\text{C}_i$	0.1	0.3	$\text{O}_{\gamma i}-\text{N}_{i+1}$ $\text{O}_{\gamma i}-\text{O}_{i-1}$ $\text{O}_{\gamma i}-\text{O}_i$	0.894 2.5 0.1	0.32 0.27 0.353
Thr	ϕ	2	1	0	$\text{O}_{\gamma 1,i}-\text{N}_i$	0.1	0.3	$\text{C}_{\gamma 2,i}-\text{H}_i/\text{H}_{i+1}$	0.447	0.29
	χ	1	1	–120	$\text{O}_{\gamma 1,i}-\text{C}_i$ $\text{C}_{\gamma 2,i}-\text{N}_i$ $\text{C}_{\gamma 2,i}-\text{C}_i$	0.1 0.1 0.1	0.3 0.3 0.29 0.33	$\text{C}_{\gamma 2,i}-\text{N}_{i+1}$ $\text{C}_{\gamma 2,i}-\text{C}_{i-1}$ $\text{O}_{\gamma 1,i}-\text{H}_{i+1}$ $\text{O}_{\gamma 1,i}-\text{H}_i$ $\text{O}_{\gamma 1,i}-\text{N}_{i+1}$ $\text{O}_{\gamma 1,i}-\text{O}_{i-1}$ $\text{O}_{\gamma 1,i}-\text{O}_i$	0.447 0.224 2.5 0.1 0.894 2.5 0.1	0.35 0.33 0.22 0.288 0.32 0.27 0.353
Pro	Ψ	2.5	1	0	$\text{C}_{\gamma i}-\text{N}_i$	0.1	0.29	$\text{C}_{\gamma i}-\text{H}_i/\text{H}_{i+1}$	0.224	0.3
His	ϕ	2.25	1	0	$\text{C}_{\gamma i}-\text{C}_i$	0.1	0.3	$\text{C}_{\gamma i}-\text{C}_{i-1}$	0.224	0.3
	χ	0.3	1	120				$\text{C}_{\gamma 1}-\text{N}_{i+1}$ $\text{C}_{\delta 2,i}-\text{C}_{i-1}/\text{C}_i$ $\text{N}_{\delta 1,i}-\text{C}_{i-1}/\text{C}_i$ $\text{C}_{\delta 2,i}/\text{N}_{\delta 1,i}-\text{H}_i$ $\text{C}_{\delta 2,i}/\text{N}_{\delta 1,i}-\text{N}_i$ $\text{C}_{\delta 2,i}/\text{N}_{\delta 1,i}-\text{N}_{i+1}$ $\text{C}_{\delta 2,i}-\text{O}_{i-1}$	0.447 0.224 0.224 0.224 0.447 0.447 0.447	0.32 0.3 0.3 0.3 0.3 0.34 0.29

^a For these interactions, only a part of the Lennard-Jones terms, $4\varepsilon\delta^{12}/r^{12}$, is used to represent the repulsive forces between the pair of particles. ^b The subscripts indicate the identities of atoms in amino acids (refer to Figure 5). ^c The large values of ε are used for hydrogen bonding interactions. ^d A harmonic constraint at -63° with $K = 41$ kJ/mol rad² is applied to ϕ of Pro.

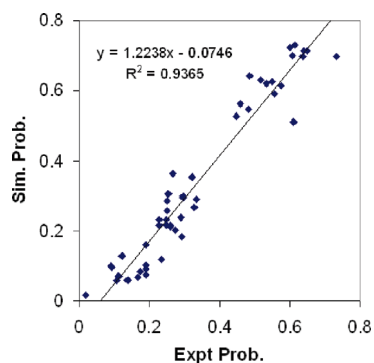
Results for Side Chain Conformers. Table 4 shows the preference of side chain conformers for all amino acids except for Ala, Gly, and Pro. Our results are obtained through REMD

simulations of dipeptides in CG water with the optimized parameters that are listed in Tables 2 and 3. As shown in Figure 6, the populations of three side chain rotamers of all amino

Table 4. Population of Side Chain Rotamers (g⁻, g⁺, and t) of Various Amino Acids from Our UA Model^a and the Coil Library

	g ⁺		g ⁻		t	
	CG	coil	CG	coil	CG	coil
Lys	0.712	0.649	0.057	0.106	0.231	0.246
Gln	0.712	0.640	0.070	0.114	0.218	0.246
Glu	0.723	0.601	0.061	0.139	0.216	0.260
Arg	0.701	0.606	0.069	0.166	0.230	0.228
Met	0.728	0.613	0.058	0.139	0.214	0.248
Leu	0.697	0.731	0.016	0.018	0.287	0.251
Val	0.613	0.574	0.290	0.333	0.097	0.093
Ile	0.511	0.612	0.362	0.267	0.127	0.121
Phe	0.629	0.516	0.074	0.190	0.296	0.295
Trp	0.526	0.447	0.119	0.233	0.355	0.320
Tyr	0.643	0.484	0.091	0.188	0.266	0.328
His	0.592	0.556	0.102	0.189	0.306	0.255
Cys	0.561	0.458	0.183	0.292	0.256	0.250
Ser	0.298	0.294	0.620	0.534	0.083	0.172
Thr	0.203	0.273	0.698	0.637	0.099	0.090
Asn	0.627	0.551	0.161	0.189	0.212	0.260
Asp	0.547	0.482	0.216	0.228	0.237	0.289
standard deviation	<0.020		<0.030		<0.020	

^a All of the calculated populations are obtained through dipeptide simulations in solution at 300 K. Standard deviations are estimated from block averages with a block size of 10 ns.

**Figure 6.** Plot of probabilities of side-chain rotamers (g⁺, g⁻, and t) of all amino acids except for Ala, Gly, and Pro in the coil library against the ones in our simulations.

acids expect for Ala, Gly, and Pro derived from our model match well with those from the coil library. The correlation coefficient R^2 is about 0.94, and the slope of the fitting line is 1.22. Interestingly, we found that simulations with the OPLS-AA/L and AMBERff03 force fields give no apparent correlation with the coil library in side chain rotamers.⁷³

Table 5 shows the relative free energies of α_L , α_R , PPII, and β conformers when the side chains adopt g⁺, g⁻, and t rotamers, from both our calculation and the statistical results. Not counting the conformers that are set as reference states with zero free energy or have too low a probability, the relative free energies calculated with our model are on average deviated from the statistical values by 1.98 kJ/mol. For the backbone conformers in the most favorable side chain rotamers, the average deviation is reduced to 1.12 kJ/mol.

Figure 7 shows the comparison of χ -dependent (ϕ , Ψ) Ramachandran plots between our model and the coil library for six representative amino acids. It is known that the shape and positions of the important minima on the (ϕ , Ψ) maps are different at the three χ conformers (g⁺, g⁻, and t).^{73,78} For example, in the g⁺ conformers for all amino acids, ϕ in the

left region spans the range of -150° to -60° , but the lower limit of ϕ extends to -180° in the g⁻ conformers, expect for β -branched amino acids such as Val and Thr. Unlike the g⁺ and g⁻ conformers, the upper limit of Ψ in the t conformers is lowered from 180° to 150° . In addition, in the g⁺ conformers, the α_L basins are well sampled, but they are normally less well sampled in the g⁻ and t conformers. Our model is able to capture most of these features for various amino acids.

The χ -dependent Ramachandran plots of other amino acids are given in the SI (Figure S2). The results of our force field are in good agreement with the coil library results in each case.

Revision of Hydration Parameters of Charged Side Chains and Amide Groups.

Globular proteins are soluble because there are polar side chains on the protein surface exposed to the hydrophilic environment. Although a protein core is normally composed of nonpolar groups, there are a significant number of polar or ionizable groups buried or partially buried inside a protein. Adolfsen et al. analyzed 124 PDB structures ranging from 100 amino acids to more than 600 amino acids.⁷⁹ About 37–61% of the surface area of ionizable groups is buried for 100-amino-acid-long proteins. Gunner et al. studied 490 proteins with a size of 36 to 1357 residues.⁸⁰ Among them, about 17% of the ionizable groups are fully buried in protein cores. Ionizable groups may not be charged if they are not exposed to water. For instance, Val66 of staphylococcal nuclease is buried in the core of the protein. When it is mutated to Lys, Asp, or Glu, the pK_a value of Lys is reduced to 5.7 and those of Asp and Glu are increased to 8.7 and 8.8, respectively.^{81,82} A similar effect is also observed when Leu38 is mutated to Asp or Glu.⁸³ Another example is Asp79 in ribonuclease S α , which has a pK_a value of 7.4.⁸⁴ The Asp79Phe mutant has a greater stability than the wide-type enzyme by 3.7 kcal/mol. This shows that ionizable amino acids may prefer a neutral state when they are placed in a hydrophobic environment. This can be physically understood since charged groups inside the protein induce a large desolvation penalty. During the folding process of a protein, ionizable residues would be neutralized before they move into the protein core. It may be particularly important to include this neutralization process in the force field in order to study the folding of large proteins that contain more buried ionizable groups.

In our previous work, we optimized the hydration parameters of charged groups by fitting the experimental hydration free energy of ions (ΔG_{hyd}).⁴¹ In that parametrization scheme, we assumed that all ionizable side chains are always charged in both water and a low dielectric medium. In fact, ionizable groups may become neutral when they are in protein cores which normally have low dielectric constants. To account for this effect, in our model, we designed a thermodynamic cycle, as shown in Figure 8. Our goal is to attain an effective ΔG_{hyd} for an ionizable group so that regardless of the charge state of this group, its partition between water and cyclohexane agrees with experimental data. As such, hydration parameters of ionizable groups are reoptimized to fit the effective ΔG_{hyd} , which is calculated according to eq 7.

$$\Delta G_{\text{hyd}} = \Delta G_{\text{chx}} + \Delta G_{(\text{hyd}-\text{chx})} + \Delta G_{pK_a} \quad (7)$$

Effective solvation free energy in water (ΔG_{hyd}) is the sum of the experimental solvation free energy in cyclohexane (ΔG_{chx}),

Table 5. Relative Free Energies (kJ/mol) of α_L , α_R , PPII, and β Conformers of All Amino Acids except for Ala and Gly When Side Chains Adopt g+, g-, and t Conformations^a

side chain conformer	coil library				PACE				
	α_L	β	PPII	α_R	α_L	β	PPII	α_R	
Lys	g+	3.0	1.0	0.0	0.0	5.0	3.2	0.2	0.0
	g-	11.0	0.0	2.0	2.0	8.4	0.0	4.3	2.9
	t	12.0	3.0	0.0	2.0	11.4	0.7	0.0	1.1
Gln	g+	4.0	1.0	0.0	0.0	5.7	3.2	0.1	0.0
	g-	10.0	0.0	2.0	3.0	n/a ^b	0.0	3.9	2.3
	t	10.0	3.0	0.0	1.0	11.4	0.2	0.0	1.2
Glu	g+	5.0	1.0	0.0	0.5	5.2	3.2	0.0	0.0
	g-	n/a	0.0	1.0	1.0	n/a	0.0	3.1	1.9
	t	11.0	4.0	0.0	1.0	10.4	0.4	0.0	1.6
Arg	g+	4.0	0.0	0.0	1.0	4.4	2.0	0.0	-0.3
	g-	12.0	0.0	2.0	4.0	10.9	0.0	4.3	2.2
	t	12.0	2.5	0.0	1.0	11.2	0.2	0.0	0.9
Met	g+	4.0	2.0	0.0	2.0	4.5	3.2	0.0	-0.1
	t	12.0	2.0	0.0	2.0	10.4	0.8	0.0	1.0
Leu	g+	7.0	2.0	0.0	2.0	6.7	4.2	0.0	1.6
	t	10.0	2.0	0.0	2.0	13.1	1.7	0.0	2.9
Val	g+	10.0	0.0	0.0	3.0	13.7	0.2	0.0	7.6
	g-	14.0	0.0	2.0	3.0	14.2	0.0	-0.3	3.7
	t	n/a	0.0	1.0	3.0	n/a	0.0	3.4	8.3
Ile	g+	11.0	0.0	0.0	4.0	14.7	0.0	0.1	6.7
	g-	13.0	0.0	3.0	3.0	n/a	0.0	0.2	3.7
	t	n/a	1.0	0.0	2.0	n/a	-3.5	0.0	4.6
Phe	g+	4.0	0.0	0.0	3.0	7.3	0.0	-1.3	3.2
	g-	n/a	0.0	4.0	5.0	n/a	0.0	5.6	8.4
	t	7.0	3.0	0.0	2.0	13.1	-0.5	0.0	6.9
Trp	g+	6.0	1.0	0.0	2.0	8.1	1.3	0.0	4.6
	g-	n/a	0.0	4.0	5.0	n/a	0.0	5.7	7.4
Tyr	g+	5.0	0.0	0.0	2.0	9.8	0.0	-0.7	3.8
	g-	n/a	0.0	4.0	5.0	13.8	0.0	5.9	12.1
	t	8.0	3.0	0.0	3.0	n/a	-0.7	0.0	6.1
His	g+	2.0	0.0	0.0	2.0	4.1	1.9	0.0	1.8
	g-	8.0	0.0	2.0	3.0	10.6	0.0	4.6	4.0
	t	4.0	1.0	0.0	1.0	9.4	-1.7	0.0	3.6
Cys	g+	4.0	2.0	0.0	3.0	6.6	2.5	0.0	0.4
	g-	n/a	0.0	3.0	1.0	n/a	0.0	2.3	2.2
Ser	g+	4.0	2.0	0.0	1.0	5.2	4.2	0.0	-0.2
	g-	13.0	1.0	1.0	0.0	10.4	4.0	2.0	0.0
	t	8.0	0.0	0.0	6.0	8.1	0.5	0.0	2.1
Thr	g+	8.0	2.0	0.0	2.0	11.9	0.0	0.0	6.8
	g-	11.0	0.0	2.0	0.0	10.1	0.0	-0.7	1.3
	t	17.0	0.0	3.0	7.5	n/a	0.0	3.9	6.2
Asn	g+	0.0	3.0	0.0	1.0	1.7	4.6	0.0	-0.8
	g-	11.0	0.0	3.0	0.0	9.1	3.9	2.5	0.0
	t	0.0	2.0	0.0	3.0	5.9	0.1	0.0	1.6
Asp	g+	3.0	4.5	0.0	1.0	2.5	7.4	0.0	-0.7
	g-	15.0	3.0	5.0	0.0	n/a	4.7	1.4	0.0
	t	2.0	2.0	0.0	3.0	5.1	0.0	0.0	0.2
Pro		n/a	n/a	0.0	4.0	n/a	n/a	0.0	3.3

^a Both the results from the coil library and our force field are listed. ^b The probability of the conformer is too low for free energy to be calculated. Standard deviations were estimated from block averages with a block size of 10 ns. The standard deviations for all conformers are less than 1.0 kJ/mol except for α_L conformers (<2.5 kJ/mol).

the experimental transfer free energy from cyclohexane to water ($\Delta G_{(\text{hyd}-\text{chx})}$), and the free energy change of ionization in water ($\Delta G_{\text{p}K_a}$). Experimental values of ΔG_{chx} and $\Delta G_{\text{hyd}-\text{chx}}$ are taken from the works of Radzicka and Wolfenden⁸⁵ and Wolfenden et al.⁸⁶ The $\Delta G_{\text{p}K_a}$ value can be obtained from eq 4.

$$\begin{aligned} \Delta G_{\text{p}K_a} &= -RT \ln K_{\text{eq}} \\ \ln K_{\text{eq}} &= 2.30(\text{pH} - \text{p}K_a) \end{aligned} \quad (8)$$

R is the Boltzmann constant; T is 300 K. The pH is assumed to be 7. Experimental $\text{p}K_a$ values of ionizable groups are taken from the works of Thurlkill et al.⁸⁷ and Nozaki and Tanford.⁸⁸

Solvation free energy values and revised parameters of ionizable groups are shown in Tables 6 and 7, respectively.

It should be noted that the hydration parameters of our ionizable groups are optimized in such a way that these groups can partition in water and a low dielectric medium as if they have distinct charge states in different environments. However, the parameters of nonbond interactions between these groups and other protein sites actually cannot vary in different environments.

As explicit hydrogen sites in side chain amide groups are included in the current model, the hydration scheme of this group is different from the previous model, in which the

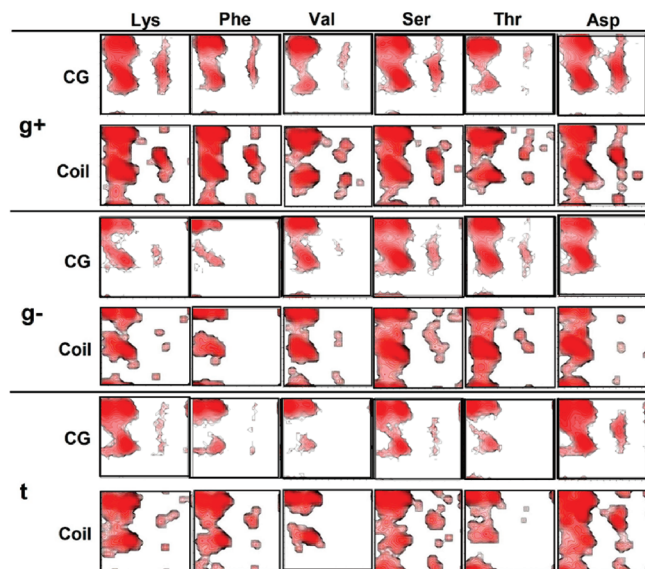


Figure 7. Backbone (ϕ , Ψ) distributions with side chains adopting different conformers (g^+ , g^- , and t) for Lys, Phe, Val, Ser, Thr, and Asp from our UA model in CG solvent (CG) and the statistical results (coil). The gap between the contour lines indicates the free energy difference of 1 kJ/mol.

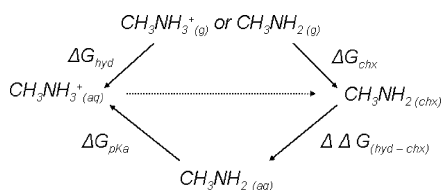


Figure 8. Revised hydration parametrization scheme for ionizable groups: Lys, Arg, Asp, and Glu. To account for the effect of the charged groups becoming neutral when they are in the protein core, the solvation free energy in water (ΔG_{hyd}) of ionizable groups is considered the sum of the solvation free energy in cyclohexane (ΔG_{chx}), the transfer free energy from cyclohexane to water ($\Delta G_{(\text{hyd}-\text{chx})}$), and the free energy change of charging in water (ΔG_{pKa}). Hydration parameters are optimized to fit ΔG_{hyd} .

Table 6. Calculated Solvation Free Energy of Ionizable Groups in Water (ΔG_{hyd}) from the Experimental Solvation Free Energy in Cyclohexane (ΔG_{chx}),^{85,86} the Transfer Free Energy from Cyclohexane to Water ($\Delta G_{(\text{hyd}-\text{chx})}$),^{85,86} and the Free Energy Change When Charged (ΔG_{pKa})^a

	ΔG_{hyd}	ΔG_{chx}	$\Delta G_{\text{hyd}-\text{chx}}$	ΔG_{pKa}	$\text{p}K_{\text{a}}$
Lys	-37.3	-16.4	-1.5	-19.4	10.40 ^a
Arg	-73.3	-20.6	-24.2	-28.5	12.0 ^b
Asp	-47.0	-9.2	-18.8	-19.0	3.67 ^a
Glu	-42.8	-13.9	-13.2	-15.7	4.25 ^a

^a All free energy values are in kJ/mol. $\text{p}K_{\text{a}}$ values are taken from the work of Thurlkill et al.⁸⁷ and Nozaki and Tanford.⁸⁸

hydrogen sites are implicitly represented. Thus, we also reoptimized the hydration parameters for interactions between CG water and the amide group. For the NH moiety of aromatic rings in His and Trp, its hydration parameters were transferred from the side chain amide group. As explicit amide hydrogen sites are also considered for backbone amide, the hydration parameters for backbone amide need a reoptimization. We find that most of the parameters for side chain

amide can be transferred to backbone amide except for the hydration parameters of hydrogen sites, which were separately optimized.

The hydration parameters for charged side chains and amide groups are obtained by fitting the experimental hydration free energies of organic compounds like we did before.⁴¹ All of the reoptimized parameters and calculated hydration free energies are listed in Table 7.

Potential of Mean Force of Polar Side Chains. It is important to accurately describe polar/charged interactions in order to study protein folding and protein–protein interactions. PMF calculation is a fundamental measure of these interactions in a water solvent. To accurately reproduce salt bridge and hydrogen bond interactions, interaction parameters of polar groups are optimized to fit PMFs obtained from all-atom simulations. As all nonbonded interactions in our model involve only two types of parameters, ϵ_{ij} representing interaction strength and δ_{ij} representing interaction distance, ϵ_{ij} and δ_{ij} parameters are optimized simultaneously in the PMF calculations. Optimized parameters for polar–polar interactions are shown in Table 8.

Figure 9 shows the PMFs of salt bridge or hydrogen bond interactions of eight polar side-chain pairs in water. As multidimensional PMF is still computationally challenging, 1D PMFs are obtained with the position restraint of polar groups on a straight line. The constrained orientations of polar groups are also shown in Figure 9. In our model, hydrogen(s) is implicitly incorporated into ammonium nitrogen, guanidinium nitrogen, and hydroxyl oxygen. Figure 9a shows the PMF of a typical salt bridge normally found in proteins between ammonium and carboxylate groups. This pair of groups was constrained by the collinear approach in that the $-\text{CH}_2\text{NH}_3^+$ moiety of ammonium and the $-\text{CH}_2-\text{C}<$ moiety of carboxylate were on the same line. It has an interaction energy of about -10.4 kJ/mol in our model, which is comparable to the all-atom simulation (-10.0 kJ/mol). The generalized Born (GB) model developed by Brooks et al. in 1999 underestimated the interaction energy by 3.3 kJ/mol, whereas EEF1 overestimated it by 6.3 kJ/mol.⁸⁹ Figure 9b shows the PMF of the guanidine–carboxylate interaction. This salt bridge pair was fixed on the same plane, and the $-\text{NH}-\text{C}<$ moiety of guanidine and the $-\text{CH}_2-\text{C}<$ moiety of carboxylate were constrained on a line. Our model is able to reproduce the interaction energy, which is about -18.4 kJ/mol in our force field and -18.8 kJ/mol in the all-atom model. The GB model is 2.5 kJ/mol less attractive, and EEF1 is about 17 kJ/mol too attractive.⁸⁹ We assume that HB acceptor groups such as amide carbonyl groups and acceptor nitrogen in aromatic rings share the same parameters with carboxylate groups when they interact with donor groups such as ammonium, amide, and hydroxyl. In addition, the parameters for like-charge groups are optimized in a similar manner, the PMFs of which are shown in Figure 9c–d.

Solvent-separated interactions can be observed in explicit solvent models only because this requires a single water molecule to be placed between two polar groups. Although this feature is apparent in our force field, the location of the solvent-separated interaction is shifted further away by about

Table 7. Optimized Parameters for Interactions between CG Water and Charged Side Chains or Amide Groups and the Calculated and Experimental ΔG_{hydr} of the Compounds That Are Used to Fit the Parameters^a

particle type ^b	ϵ_{ij} (kJ/mol)	δ_{ij} (nm)	compounds	ΔG_{hydr} (kJ/mol)	$\Delta G_{\text{hydr(expt)}}$ (kJ/mol)
-NH ₃ ⁺	12.00	0.340	butylammonium	-37.4	-37.3
-NH-C-(NH ₂) ₂	4.90	0.340	N-propylguanidinium	-73.1	-73.3
-NH-C-NH ₂	4.90	0.340			
-COO ⁻	6.00	0.340	acetate ion	-45.8	-47.0
-COO ⁻	2.00	0.415			
-CO-NH ₂	0.86	0.400	acetamide	-40.3	-40.6
-CO-NH ₂	9.00	0.340			
-CO-NH ₂	0.8	0.415			
-CO-NH ₂	3.40	0.280			
-CO-NH-	7.00	0.280	N-methylacetamide	-42.4	-42.1

^a The CG simulations are at 300 K. The calculated ΔG_{hydr} is an average over 6–8 simulations. The standard deviations are less than 1.5 kJ/mol. ^b -NH₃⁺ for ammonium; -COO⁻ for carboxylate; -CO-NH₂ for side chain amide; -CO-NH- for backbone amide; -NH-C-(NH₂)₂ for guanidine.

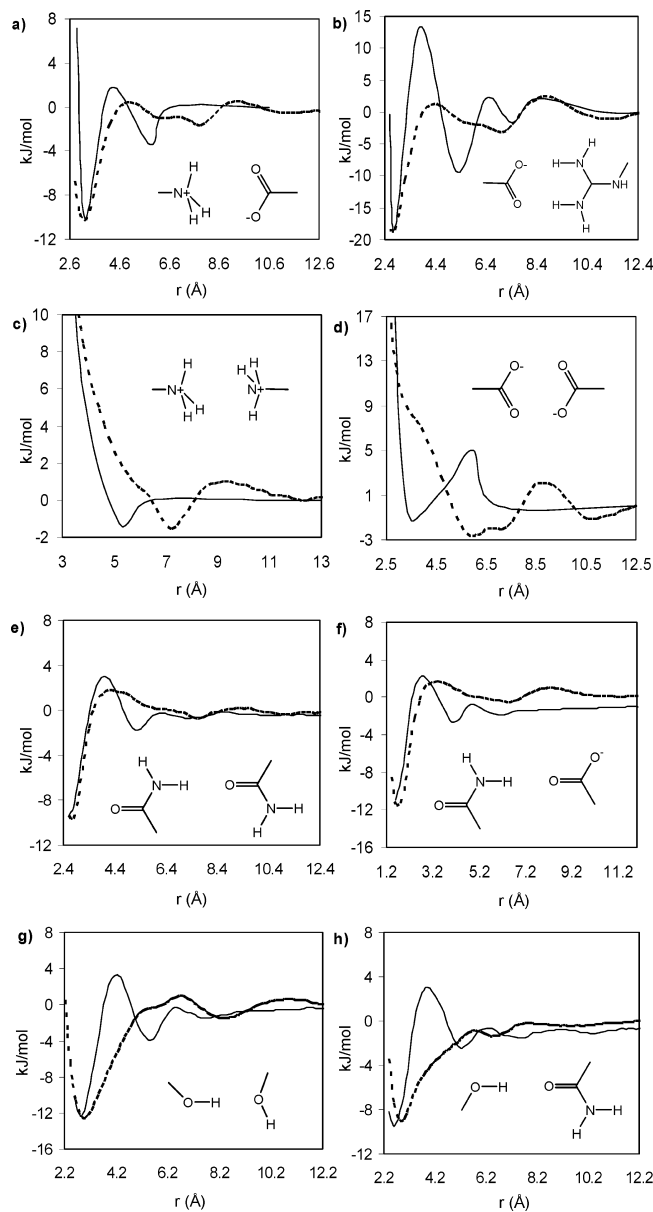
Table 8. Optimized Parameters for Polar–Polar Interactions of Charged/Polar Groups

interacting particles ^a	ϵ_{ij}	δ_{ij} (nm)
-NH ₃ ⁺ ...-COO ⁻	10.00	0.260
-NH ₃ ⁺ ...-CO-NH ₂	10.00	0.260
-NH ₃ ⁺ ...=N-	10.00	0.260
-NH-C-(NH ₂) ₂ ...-COO ⁻	10.00	0.260
-NH-C-(NH ₂) ₂ ...-COO ⁻ ^b	1.00	0.360
-NH-C-(NH ₂) ₂ ...-CO-NH ₂	3.00	0.300
-CO-NH ₂ ...-COO ⁻	23.00	0.160
-CO-NH ₂ ...-COO ⁻ ^c	15.00	0.235
-CO-NH ₂ ...-COO ⁻ ^c	15.00	0.235
-CO-NH ₂ ...-CO-NH ₂	23.00	0.160
-CO-NH ₂ ...-CO-NH ₂ ^c	15.00	0.235
-CO-NH ₂ ...-CO-NH ₂ ^c	15.00	0.235
-CO-NH ₂ ...=N-	23.00	0.160
-CO-NH ₂ ...=N ⁻ ^c	15.00	0.235
-OH...-COO ⁻	7.20	0.260
-OH...-CO-NH ₂	7.20	0.260
-OH...-OH	6.20	0.260
-OH...=N-	7.20	0.260
-CO-NH...-COO ⁻	27.00	0.16
-CO-NH...-CO-NH ₂	27.00	0.16
-NH ₃ ⁺ ...-NH ₃ ⁺ ^b	4.00	0.33
-COO ⁻ ...-COO ⁻ ^c	2.00	0.28
-NHC(NH ₂) ₂ ...-NHC(NH ₂) ₂ ^b	4.00	0.33

^a -NH₃⁺ for ammonium; -COO⁻ for carboxylate; -CO-NH₂ for side chain amide; -CO-NH- for backbone amide; -NH-C-(NH₂)₂ for guanidine; =N- for acceptor nitrogen in heterocycle; -OH for hydroxyl group. ^b The following potential is used for this interaction: $E_{\text{nonbond}} = \sum_{i \neq j} (4\epsilon_{ij}\delta_{ij}^6)/r^6$. ^c The following potential is used for this interaction: $E_{\text{nonbond}} = \sum_{i \neq j} (4\epsilon_{ij}\delta_{ij}^{12})/r^{12}$.

2 Å (Figure 9). This is because each of our CG solvent particles represents four water molecules. In addition, the solvent-mediated interactions are examined only for pairs of small molecules. Further studies are needed to investigate these interactions involved with much larger molecules such as protein–protein interactions. Thus, we should be cautious in studying proteins that involve significant solvent-separated interactions. Like implicit solvent models, the height of the first PMF peak is underestimated in our force field. However, it is not clear whether there is any significant consequence in capturing such fine details of the interactions. While the folding kinetics might be altered, the lower height of the first PMF peak might in fact accelerate the conformation sampling without introducing much thermodynamic bias.

In our model, polar hydrogen is explicitly represented in amide (Asn, Gln), methylimidazole (His), and methylindole (Trp) groups. With explicit hydrogens, hydrogen bonds

**Figure 9.** PMFs of six polar group pairs in the UA model (dashed lines) and all-atom model (solid lines). r is the distance between atoms of two groups. The all-atom PMFs in a–d come from ref 89. In g and h, the value of 10.0 kJ/mol is used for ϵ_{attr} .

involving these polar groups could have better orientation and directionality. Here, we use the same parameters for the NH moiety in these different donor groups. As an example, amide/amide and amide/carboxylate PMFs are shown in Figure 9e–f. Two polar groups are constrained on the same plane in the PMF calculation. The hydrogen bond interaction energy of amide/amide is -9.4 and -9.5 kJ/mol in our model and the all-atom model, respectively. The amide/carboxylate pair also shows comparable results to the all-atom simulation. The contact energy in the all-atom model is -9.2 kJ/mol and is -9.4 kJ/mol in our model.⁶³ As the backbone NH moiety has a larger hydration parameter than the NH moiety in side chain amide (Table 7), we also optimized its interaction parameters with the acceptors like carbonyl and carboxylate groups.

As a hydroxyl group is treated as one site currently, the inclusion of both its donor and acceptor ability may cause trouble when it simultaneously interacts with multiple donor and acceptor groups. We here are only focused on its HB donor property, except for the case of hydroxyl/hydroxyl interactions. A statistical analysis of the PDB structures showed that a hydroxyl group is 3 times more likely to be a donor than an acceptor.⁹⁰ PMFs of hydroxyl/hydroxyl and hydroxyl/amide hydrogen bond interaction are shown in Figure 9g and h, respectively. For all hydrogen bond cases, polar group pairs are constrained so that the hydrogen bond donor, the hydrogen atom, and the hydrogen bond acceptor are on the same line. The hydroxyl/hydroxyl pair has about -12.6 and -12.3 kJ/mol of contact energy in our force field and the all-atom model, respectively. The hydrogen bond interaction energy of the hydroxyl/amide pair is -9.0 kJ/mol in our model and about -9.5 kJ/mol with the all-atom model. As the hydrogen site is implicit in our force field, the potential derived in the above corresponds to a situation in which the OH bond is always in the best orientation to form a HB when it is close to an acceptor group. However, because the OH bond can rotate freely, placing it in the HB direction can cause a loss of entropy. We considered this effect by adding a correction of $RT \ln(1/3)$ onto the attraction parameters between hydroxyl and acceptor groups due to the alignment of the OH bond in the HB direction from three rotamer states. Similarly, we added $RT \ln(1/3 \times 2/3)$ for the hydroxyl/hydroxyl interaction. Therefore, the original attraction parameter (ϵ) of 10.0 kJ/mol for hydroxyl/acceptor and hydroxyl/hydroxyl now becomes 7.2 and 6.2 kJ/mol, respectively.

Optimization of Backbone/Backbone HB. Backbone/backbone HB interactions are critical for defining the geometrical feature of proteins, especially for secondary structures. They are the most abundant interaction type in proteins. As expected, any small variation in the parameters of backbone/backbone interactions can significantly influence the simulation results of a force field. Unlike the parametrization of side chain amide groups, we tried to optimize the backbone/backbone HB parameters through peptide simulations. Such a strategy has been used before. For instance, Brooks and co-workers refined backbone potentials and atomistic input radii for continuum electrostatics in the development of their CHARMM-CAMP force field.⁶³ Irback

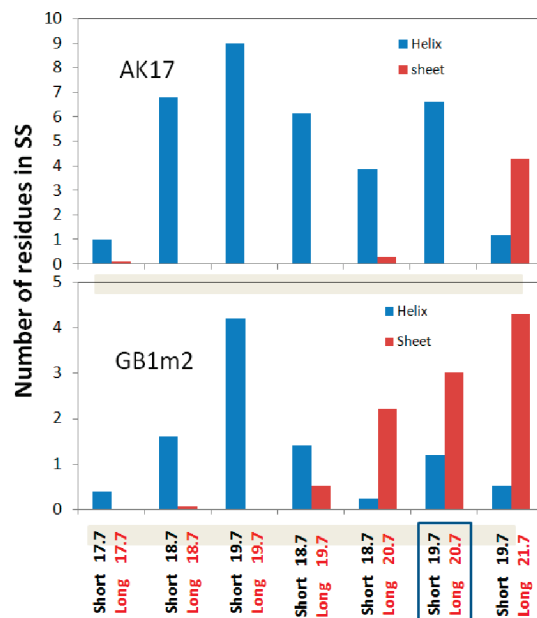


Figure 10. Number of helical (blue) and sheet (brown) residues of the AK17 and GB1m2 peptides with different ϵ_{attr} parameters (kJ/mol). “Short” indicates the values of ϵ_{attr} for the short-range O_i/N_{i+3} and O_i/N_{i+4} HBs. “Long” indicates the values of ϵ_{attr} for the other HBs.

and Mohanty used a series of folding simulations of peptides and mini-proteins to optimize nonbonded interactions of their all-atom force field with an implicit solvent model.⁹¹ These studies raise the importance of balancing between an α helix and β sheet. We here chose two peptides, the AK17 [Ac-(AAKAA)₃GY-NMe]⁹² peptide that is ~ 30 – 35% helical and the GB1m2 (Ac-GEWTYNPATGKFTVTE-NMe)⁹³ peptide that is a mutant of the N-terminal β -hairpin of the protein G B1 domain. The extremely fast kinetics of these peptides enable us to perform numerous folding simulations to optimize the backbone/backbone HB parameters. Each folding simulation was conducted in 16 REMD replicas with a temperature ranging from 300 to 430 K. Each simulation lasts for 100–200 ns. The average helical and sheet contents were calculated over the last 50 ns of the simulation.

Equation 5 and Figure 2a show all of the potential terms of the backbone/backbone HBs. Only the attraction parameter ϵ_{attr} that controls HB strength was further optimized. It should be noted that no attempts have been made to reproduce the native structures of the peptides. Only the total contents of the secondary structures of the peptides are fitted in the optimization. Figure 10 shows the secondary structure contents of AK17 and the GB1m2 with different ϵ_{attr} values. With ϵ_{attr} increasing from 17.7 to 19.7 kJ/mol, the helical content of AK17 increases from $\sim 7\%$ to $\sim 61\%$. However, instead of a β sheet, a significant amount of α helix is also developed in the GB1m2, indicating that a single ϵ_{attr} value is not enough. We therefore adopt two ϵ_{attr} parameters. One is for short-range hydrogen bonds between O_i and N_{i+3} and between O_i and N_{i+4} that are crucial to turns and helices. The other is for the remaining long-range HB bonds. After a nontrivial trial-and-error optimization, we find that ϵ_{attr} is about 19.2 kJ/mol for the short-range HBs and ϵ_{attr} is about 20.7 kJ/mol for the long-range HBs.

Table 9. Comparison between the Averaged C_{α} RMSD (nm) from the Respective Experimental Structures for the Nine Proteins Simulated with Our Force Field and Previous All-Atom Models^a

PDB ID	no. of AAs	UA		All-atom	
		simulation length (ns)	C_{α} RMSD (nm)	simulation length (ns)	C_{α} RMSD (nm) ^b
3gb1	NMR	56	100	50	0.106 (0.020)
1ctf	X-ray	68	100	100	0.16 ^c
1d3z	NMR	76	100	22	0.141 (0.020)
1bta	NMR	89	100	142.9	0.134 (0.016)
1fks	NMR	107	100	143.5	0.358 (0.074)
1fw7	NMR	108	100	148.0	0.171 (0.015)

^a The values in parentheses are standard deviations. Average RMSD values were calculated from the beginnings of simulations. ^b These are from ref 50. ^c This value is from ref 95 and is an average over the last 90 ns.

Aqueous Simulations of Proteins. Like all-atom force fields, our force field is built upon the parametrization of small molecules such as side chain analogues or dipeptides that contain constituent functionality. There remains a question of whether it is effective and accurate to combine all of the optimized parameters for the study of real proteins. A stringent test should be the folding of peptides and mini-proteins into their native structures by first principle means. Such a test is presented in our accompanying paper. Here, we show a preliminary test on protein simulations. Aqueous simulations of proteins have become a consensual way of examining the protein portions of force fields.⁹⁴ Using this approach and starting with NMR/X-ray structures, multi-nanosecond MD simulations of proteins in water have been conducted. During the simulations, the maintenance of experimental structures as well as other thermodynamic properties is deemed an indication of the feasibility of force fields for protein simulations.

Verification of Our Force Field. To examine our force field, we selected seven proteins of small to medium size (56–108 AA) that have a cross-section of various secondary structures. Most of them have recently been used to examine the performance of a modified version of the CHARMM force fields.⁹⁵ Another (PDB code: 1ctf) was used to validate the modification of the GROMOS96 force fields.⁵⁰ The RMSD values of C_{α} carbon atoms from native structures were employed to monitor whether the native structures are kept during the simulations. Table 9 lists the average C_{α} RMSD values in 100 ns simulations of the six proteins with our force field. The full trajectories of the RMSD can be found in Figure S3 in the SI. All of the calculated RMSDs are around or below 0.3 nm except for 1bta. The average RMSD for all of the proteins is about 0.295 nm. Thus, the native structures can be considered maintained. Our results are generally larger than but comparable to the RMSD of the same proteins derived from all-atom simulations. In one case (1fks), our RMSD is even better than that of all-atom models (0.248 nm vs 0.358 nm). In terms of the ability to reproduce experimental native structures, our force field is roughly comparable to the all-atom models.

Finally, we examined whether our force field can be extended into more complex systems such as large proteins. We chose to study malate synthase G (PDB code: 2jqx), which contains 723 amino acids.⁹⁶ Starting with the NMR-derived structure, a 30 ns MD simulation was carried out. Throughout the simulations, the RMSD is kept around 0.4

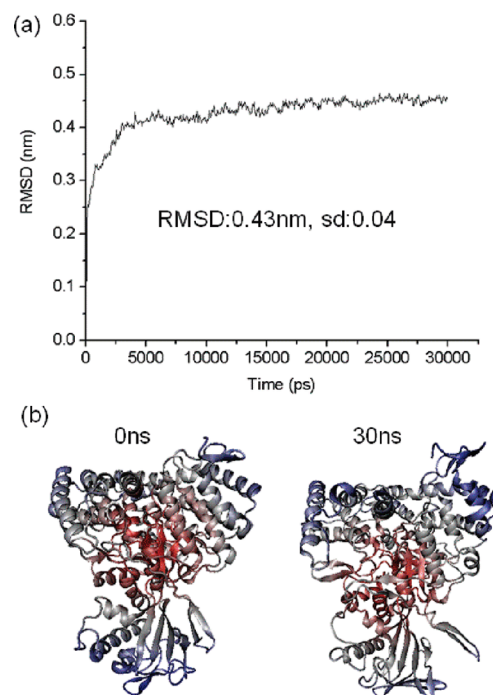


Figure 11. (a) Change of the root-mean-square deviation (nm) of C_{α} carbons during the simulation. (b) The PDB structure of malate synthase G and its simulated structure at the end of the simulation.

nm (Figure 11a). The average RMSD is 0.43 nm. Considering the large size of the protein, this RMSD is at least small enough as an indication of the maintenance of the native topology (Figure 11b). It should be noted that the simulation was conducted with a 2.66 GHz dual-core CPU at a speed of 10 ns/day. Combining the solvents, our force field system is equivalent to an all-atom system of more than 100 000 atoms. Simulating tens or hundreds of nanoseconds of such systems is still a formidable task for all-atom force fields.

We do notice that several proteins including 1bta and 1fw7 partially unfold during the simulations (Table 9 and Figure S3, SI). Visual inspection reveals that the unfolding is mainly caused by a considerable loss of native polar interactions such as HBs. We suspect that our force field may underestimate polar interactions in a low dielectric medium such as a large protein. This is because all of the parameters for polar interactions are optimized to reproduce all-atom results in a solvent-exposed environment. In a low dielectric medium,

the polar interactions are supposed to be strengthened due to the removal of the screening effect of water. Nevertheless, the screening effect is missing in the current CG model that is basically LJ fluid. Thus, our force field may be more suitable for proteins with large proportions of exposed parts.

Conclusion

In this paper, we describe our recent effort in the improvement of our protein force field model, which includes two aspects. First, the backbone (ϕ , Ψ) potential is reoptimized to reproduce shapes, positions, and depths of β , PPII, α_R , and α_L basins on the statistical (ϕ , Ψ) surfaces for Ala and Gly. Local side chain–backbone interactions are optimized to reproduce correct preferences of side chain conformers (g+, g–, and t) of various amino acids as well as the dependence of backbone conformations (ϕ , Ψ) on side chain conformations (χ). Second, we also parametrize the interactions between polar groups by fitting the PMFs of polar group pairs that are derived from all-atom simulations. Together with previously derived parameters, our force field is able to reproduce the native structures of series of proteins of small to medium size (<150 AA). Moreover, the stability of the native structures can be maintained for even larger proteins (>700 AA). Simulating tens of nanoseconds of such large proteins is easy (several days of calculation with a single computer) for our force field, and simulations of microseconds are within reach. These results imply the applicability of our force field to the structural study of real proteins.

Acknowledgment. We are grateful to the Research Grants Council of Hong Kong (CA06/07.SC05, 663509) for financial support of this research.

Supporting Information Available: Tables S1–S5, the parameters of our force field; Figure S1, the equilibrium values of bond lengths and bond angles of our force field; Figure S2, the χ -dependent Ramachandran plots of Gln, Glu, Arg, Met, Leu, Ile, Trp, Tyr, His, Cys, and Asn; and Figure S3, the plots of RMSDs in all six protein simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140.
- van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. *Curr. Opin. Struct. Biol.* **2008**, *18*, 149.
- MacKerell, A. D.; Nilsson, L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 194.
- Schaeffer, R. D.; Fersht, A.; Daggett, V. *Curr. Opin. Struct. Biol.* **2008**, *18*, 4.
- Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470.
- Villa, E.; Balaeff, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6783.
- Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694.
- Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945.
- Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534.
- Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859.
- Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.
- Sippl, M. J. *J. Comput.-Aided Mol. Des* **1993**, *7*, 473.
- Matysiak, S.; Clementi, C. *J. Mol. Biol.* **2006**, *363*, 297.
- Khurana, E.; DeVane, R.; Kohlmeyer, A.; Klein, M. L. *Nano Lett.* **2008**, *8*, 3626.
- Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.
- Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.
- Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819.
- Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593.
- Treptow, W.; Marrink, S. J.; Tarek, M. *J. Phys. Chem. B* **2008**, *112*, 3277.
- Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123.
- Aksimentiev, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4337.
- Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806.
- Miyazawa, S.; Jernigan, R. L. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 357.
- Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646.
- Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697.
- Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.
- Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874.
- Yap, E. H.; Fawzi, N. L.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 626.
- Khatun, J.; Khare, S. D.; Dokholyan, N. V. *J. Mol. Biol.* **2004**, *336*, 1223.
- Dokholyan, N. V. *Curr. Opin. Struct. Biol.* **2006**, *16*, 79.
- Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. *J. Phys. Chem. B* **2009**, *113*, 267.
- Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 394.
- Fujitsuka, Y.; Chikenji, G.; Takada, S. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 381.
- Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616.
- Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.
- Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116.
- Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289.

- (38) Ding, F.; Tsao, D.; Nie, H. F.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010.
- (39) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147.
- (40) Han, W.; Wu, Y.-D. *J. Chem. Theory Comput.* **2007**, *3*, 2146.
- (41) Han, W.; Wan, C.-K.; Wu, Y.-D. *J. Chem. Theory Comput.* **2008**, *4*, 1891.
- (42) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511.
- (43) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10.
- (44) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.
- (45) Klein, M. L.; Shinoda, W. *Science* **2008**, *321*, 798.
- (46) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.
- (47) 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.
- (48) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781.
- (49) MacKerell, A. D.; Fieg, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400.
- (50) Cao, Z.; Lin, Z.; Wang, J.; Liu, H. Y. *J. Comput. Chem.* **2009**, *30*, 645.
- (51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712.
- (52) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451.
- (53) Kaminski, G.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
- (54) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908.
- (55) 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.
- (56) Head-Gordon, T.; Head-Gordon, M.; Frisch, M. J.; Brooks, C. L., III; Pople, J. A. *J. Am. Chem. Soc.* **1991**, *113*, 5989.
- (57) Brooks, C. L., III; Case, D. A. *Chem. Rev.* **1993**, *93*, 2487.
- (58) Cornell, W. D.; Cieplak, P.; Bayly, C.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (59) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Feig, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-MacCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (60) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
- (61) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4* (5), 819.
- (62) Makowski, M.; Sobolewski, E.; Czaplowski, C.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2008**, *112*, 11385.
- (63) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728.
- (64) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (65) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.
- (66) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (67) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.
- (68) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (69) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435.
- (70) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
- (71) Bogusz, S.; Cheatham, T. E.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070.
- (72) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
- (73) Jiang, F.; Han, W.; Wu, Y.-D. *J. Phys. Chem. B* **2010**, *114*, 5840.
- (74) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691.
- (75) Lovell, S. C.; Davis, I. W.; Arendall, W., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437.
- (76) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 2782.
- (77) Ho, B. K.; Thomas, A.; Brasseur, R. *Protein Sci.* **2003**, *12*, 2508.
- (78) Chakrabarti, P.; Pal, D. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1.
- (79) Kajander, T.; Kahn, P. C.; Passila, S. H.; Cohen, D. C.; Lehtio, L.; Adolfsen, W.; Warwicker, J.; Schell, U.; Goldman, A. *Structure* **2000**, *8*, 1203.
- (80) Kim, J.; Mao, J.; Gunner, M. R. *J. Mol. Biol.* **2005**, *348*, 1283.
- (81) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B. *Biophys. J.* **2002**, *82*, 3289.
- (82) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; García-Moreno, E. B. *Biophys. J.* **2007**, *92*, 2041.
- (83) Harms, M. J.; Castañeda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; García-Moreno, E. B. *J. Mol. Biol.* **2009**, *389*, 34.
- (84) Trevino, S. R.; Gokulan, K.; Newsom, S.; Thurlkill, R. L.; Shaw, K. L.; Mitkevich, V. A.; Makarov, A. A.; Sacchettini, J. C.; Scholtz, J. M.; Pace, C. N. *J. Mol. Biol.* **2005**, *354*, 967.
- (85) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664.
- (86) Wolfenden, R.; Anderson, L.; Cullis, P.; Southgate, C. *Biochemistry* **1981**, *20*, 849.
- (87) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2006**, *15*, 1214.
- (88) Nozaki, Y.; Tanford, C. *Methods Enzymol.* **1967**, *11*, 715.

- (89) Masunov, A.; Lazridis, I. *J. Am. Chem. Soc.* **2003**, *125*, 1722.
- (90) Eswar, N.; Ramaakrishnan, C. *Protein Eng.* **2000**, *13*, 227.
- (91) Irback, A.; Mohanty, S. *Biophys. J.* **2005**, *88*, 1560.
- (92) Luo, P.; Baldwin, R. L. *Biochemistry* **1997**, *36*, 8413.
- (93) Feinmeyer, R. M.; Hudson, F. M.; Anderson, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238.
- (94) Price, D. J.; Brooks, C. L., III. *J. Comput. Chem.* **2002**, *23*, 1045.
- (95) Feig, M. *J. Chem. Theory Comput.* **2008**, *4*, 1555.
- (96) Grishaev, A.; Tugarinov, V.; Kay, L. E.; Trewella, J.; Bax, A. *J. Biomol. NMR* **2008**, *40*, 95.

CT1003127

PACE Force Field for Protein Simulations. 2. Folding Simulations of Peptides

Wei Han,[†] Cheuk-Kin Wan,[†] and Yun-Dong Wu^{*,†,‡,§}

Department of Chemistry, The Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong, China, School of Chemical Biology and Biotechnology, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, China, and College of Chemistry, Peking University, Beijing, China

Received June 8, 2010

Abstract: We present the application of our recently developed PACE force field to the folding of peptides. These peptides include α -helical (AK17 and Fs), β -sheet (GB1m2 and Trpzip2), and mixed helical/coil (Trp-cage) peptides. With replica exchange molecular dynamics (REMD), our force field can fold the five peptides into their native structures while maintaining their stabilities reasonably well. Our force field is also able to capture important thermodynamic features of the five peptides that have been observed in previous experimental and computational studies, such as different preferences for a helix–turn–helix topology for AK17 and Fs, the relative contribution of four hydrophobic side chains of GB1p to the stability of β -hairpin, and the distinct role of a hydrogen bond involving Trp-H_ε and a D9/R16 salt bridge in stabilizing the Trp-cage native structure. Furthermore, multiple folding and unfolding events are observed in our microsecond-long normal MD simulations of AK17, Trpzip2, and Trp-cage. These simulations provide mechanistic information such as a “zip-out” pathway of the folding mechanism of Trpzip2 and the folding times of AK17 and Trp-cage, which are estimated to be about 51 ± 43 ns and 270 ± 110 ns, respectively. A 600 ns simulation of the peptides can be completed within one day. These features of our force field are potentially applicable to the study of thermodynamics and kinetics of real protein systems.

Introduction

The successful application of molecular dynamics (MD) simulations to structural and dynamic studies of proteins relies on the efficiency of the sampling protocol and the quality of the underlying force fields.^{1,2} With the rapid growth in computing power, advanced computing techniques³ and accelerating sampling methods such as multicanonical simulations,⁴ replica exchange molecular dynamics (REMD),⁵ and metadynamics,⁶ computer simulations have become far more powerful than those of even a few years ago.^{7–9} Despite this progress, all-atom force fields with

explicit treatment of solvents, which are arguably the most accurate, are only capable of attaining converged sampling for small peptides with not more than 20 amino acids at considerable computational expense.^{9–11}

To further enhance sampling, it is necessary to reduce the number of degrees of freedom. The key challenge here is how to simplify models with a minimal loss of accuracy. Several approaches have been developed. For example, an explicit solvent model can be replaced by an implicit solvent model such as the generalized Born model with surface area (GB/SA).^{12–14} In conjunction with all-atom protein models, GB/SA has been accepted and used widely for its improved efficacy.¹ Very recently, the combination of the AMBERff96 all-atom force field and the GB model by Onufriev et al.¹⁴ has made a breakthrough in folding a millisecond slow folder NTL9(1–39),⁷ with the help of large-scale distributed

* Corresponding author e-mail: chydwu@ust.hk.

[†] The Hong Kong University of Science & Technology.

[‡] Peking University Shenzhen Graduate School.

[§] Peking University.

Table 1. Sequences and Experimental Stabilities of the Five Peptides Studied in This Work and Previous Representative Computational Studies of These Peptides

peptide name	sequence ^a	stability at 300 K	refs	used to optimize parameters? ^b
AK17	Ac-(<u>AAKAA</u>) ₃ GY-Nme	~30–35% ^c	42, 46	overall helical content
Fs	Ac-A ₅ (<u>AAAAA</u>) ₃ -Nme	~50–55%; $T_m = 303$ K ^c	43–45, 62, 71, 84	no
GB1m2	GEWTYNPATGKFTVTE	~70–75%; $T_m = 320$ K ^d	68, 71	overall β -sheet content
Trpzip2	SWTWENKQWTWK	>90%; $T_m = 345$ K ^d	62, 68, 69, 98, 99	no
Trp-cage	<u>NLYIQWLKDG</u> GGPSSGRPPPS	~70%; $T_m = 317$ K ^d	9, 52–58, 68, 70, 71	no

^a The underlined residues are in helical states; the bold residues are in sheet states; the italic residues are in loops, turns, or coils. ^b If yes, this column indicates which propensities of the peptides are used for parametrization. ^c Based on CD measurement.^{79,80} ^d Based on NMR measurement.^{33,35,36}

computations and GPU coding.³ Besides the implicit model, another approach is to use a coarse-grained (CG) protein representation,^{15,16} which is coupled with either an implicit^{17–19} or a CG solvent model.^{20–22}

Combined with the fast sampling method REMD, the simplified model has reached the point where converged sampling for small peptides and mini-proteins only costs moderate computation time. Thus, a comparison with experimental observations can be readily made. This comparison is a critical test of the quality of force fields even for all-atom models with explicit solvents, as most of the models are derived from small molecules or dipeptide models.^{23–26} On the experimental side, there are a number of designed peptides that fold at the microsecond time scale.²⁷ These peptides possess α -helices,^{28–30} β -hairpins,^{31–35} or mixed topologies,^{36–41} and tertiary structures can be found in some of them.^{36–41} As a result, they are perfect targets in extensive computational studies.^{9,19,40,42–59}

The comparisons reveal that it is not trivial for force fields to achieve a balance between α -helical and β -sheet structures. Okamoto and co-workers have shown that different all-atom force fields with explicit solvent models have quite different preferences for helical and extended structures.^{60,61} Shell et al. performed extensive REMD simulations of short helices and hairpins with four versions of AMBER force fields and three sets of parameters for the GB/SA implicit solvent, which were proposed previously.⁶² They reported that only the AMBERff96 coupled with the GB/SA parameters proposed by Onufriev et al.¹⁴ could attain the balance. A CG model with implicit solvent has also been found to be problematic in transferability when used to predict α , β , and mixed structures.⁶³ The peptide simulations clearly demonstrate that current force fields need further optimization.^{64–67} A good balance for a number of small peptides has been achieved in several encouraging cases such as a modified version of the CHARMM22/CMAP with the GB solvent by Chen et al.,⁶⁸ the OPLS-AA with the GB model by Ulmschneider and Jorgensen,⁶⁹ the CG OPEP force fields by Chebaro et al.,⁷⁰ and an all-atom model with implicit solvent by Irback and Mohanty.⁷¹

In our preceding paper,⁷² we presented the full parametrization of a united-atom protein model, namely, the PACE (Protein in Atomistic details coupled with Coarse-grained Environment) force field, for all 20 amino acids, which is coupled with a CG solvent model.^{73–76} We have shown that the force field is capable of maintaining the native structures of several proteins in MD simulations. It is thus critical to

evaluate the applicability of the force field in folding simulations of peptides, and in particular to examine whether the force field attains a balance between the α -helix and β -sheet. In this paper, we report the results of the folding simulations of five small peptides using REMD techniques. As listed in Table 1, these peptides include (1) designed α -helical peptides such as AK17²⁹ and Fs⁷⁷ peptides; (2) β -hairpin peptides such as GB1m2,³³ a mutant of the N-terminal hairpin in the protein G B1 domain that has much faster folding kinetics than the wide type, and Trpzip2;³⁵ and (3) the α -helical/coil Trp-cage³⁶ that possesses a tertiary topology. These peptides have been extensively studied in MD simulations and used to validate parameters of force fields (Table 1). As normal MD can also provide kinetic information about folding, we have also carried out normal MD folding simulations for AK17, Trpzip2, and Trp-cage. In addition, possible further improvements on this model will also be discussed.

Model and Methods

All of the simulations were performed with the GROMACS software package (version 3.3.1).⁷⁸ The PACE force field, presented in the preceding paper, was used for peptides.^{21,22,72} Water was modeled by the CG solvent model of Marrink et al.⁷³ All of the peptides were capped with an acetyl group at N-terminals and N-methylamide at C-terminals. All of the starting conformations were full helices. The peptide was placed into a dodecahedron box. The minimum distance between the peptide and the edges of the box was about 1.3 nm. About 1100–1300 CG solvent particles, depending on the size of the peptide, were filled in the box. The whole system was first optimized with the steepest descent method by 5000 steps. A 120 ps pre-equilibrium simulation was carried out on the optimized system at $T = 300$ K and $P = 1$ atm with a time step of 6 fs. A 6 ns NVT denaturing simulation was then performed at $T = 700$ K. Snapshots after 2 ns of the denaturing simulation were randomly chosen, and 120 ps simulations of these snapshots were performed at $T = 300$ K and $P = 1$ atm to prepare the starting conformations for the production simulation. The time step of the denaturing simulation is 4 fs due to the high temperature.

Our REMD simulations were composed of 16 replicas. Their temperatures ranged from 300 to 430 K. The pressure was kept at 1 atm. A time step of 10 fs was used. The mass of peptides was tripled to avoid the instability of the

simulations at high temperatures. It will be shown later that this has little effect on the thermodynamic properties of the simulated systems. An attempt to exchange configurations between replicas was made every 2 ps. The average ratio of a successful exchange was about 15%. With a cluster of eight two-way 2.66 GHz CPUs, the wall time for 300 ns ($\times 16$ replicas) of REMD simulation was about 30 h.

Following Garcia and Sanbonmatsu,⁴⁴ a residue was considered as helical only if this residue and its two neighboring residues had their (ϕ, Ψ) within $(-60^\circ \pm 30^\circ, -47^\circ \pm 30^\circ)$. The helical probability (or fraction) is calculated as the ratio of the number of helical residues to the total residue number minus two, which accounts for two terminal residues that can never be helical. A hydrogen bond was considered formed only if the donor/acceptor distance was shorter than 0.35 nm and the donor/hydrogen/acceptor angle was larger than 120° .

Results and Discussions

Simulations of Helical Peptides. The total helical content of the AK17 peptide at 300 K is $\sim 41\%$, which is averaged over the last 70 ns of 100 ns REMD simulations. Since this quantity has been used to fit our force field parameter,⁷² as expected, it is consistent with $\sim 30\%$ of the helical content of this peptide by CD measurements.⁷⁹ It is natural to immediately wonder if the optimized parameters can be transferred to other helical peptides. We therefore also carried out the REMD simulations of a longer helical peptide [Ac-A₅(AAARA)₃A-NMe], known as the Fs peptide,^{30,77} which has been explored experimentally.^{80,81} The CD-based measurement gave a melting temperature T_m of 303 K for this peptide.⁸⁰ The average helical content in our simulation (the last 50 ns of 80 ns REMD simulation) is $\sim 43\%$, in agreement with experimental results.

In addition to the total helical content, the residual helical contents of the AK17 and Fs peptides at 300 K were investigated (Figure 1). Their residual helical contents share two common features: (1) The residues in the N-terminus have a higher helical content than the C-terminal residues. In AK17, residues 2 and 3 at the N-terminus have helical contents of $\sim 40\%$ and $\sim 60\%$, respectively, while residues 14–16 at the C-terminus have helical contents of $\sim 0\text{--}20\%$. In Fs, residues 2 and 13 are $\sim 40\%$ and $\sim 65\%$ helical, respectively, but residues 19 and 20 are $\sim 25\%$ and $\sim 10\%$ helical, respectively. The asymmetry of the helical content distribution has been well characterized by previous NMR studies of polyaniline-based peptides.^{82,83} (2) The residual contents in the middle of the helices are slightly lower than the residual contents closer to the terminals. In AK17, the middle residues 7–9 are $\sim 40\%$ helical, but the highest residual helical content exceeds 60%. In Fs, the helical contents of residues 10 and 11 are also $\sim 40\%$, lower than those of residues 3–6 and 15–17. This feature has also been observed in the previous MD simulation studies of the A21 and Fs peptides.^{43,45}

The Fs peptide has been the subject of numerous MD studies.^{43–45,62,71} Using the AMBER force field with the GB solvent, Zhang et al. performed multiple 100 ns MD

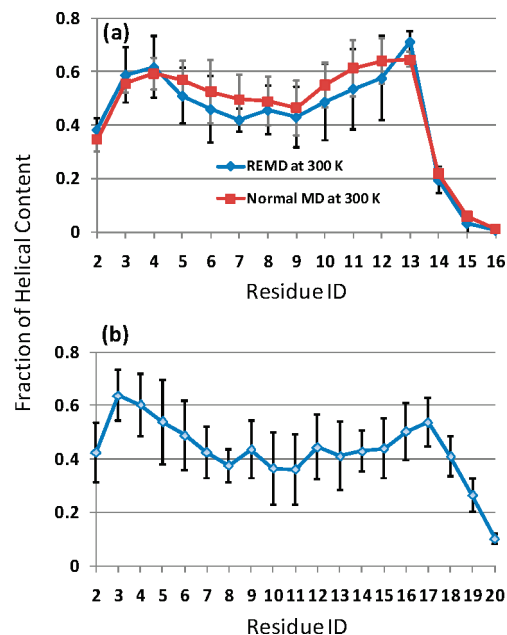


Figure 1. Residual helical contents of the AK17 (a) and Fs 17 (b) peptides at 300 K. The error bars were estimated from block averages, with block sizes of 10 and 500 ns for REMD and normal MD simulations, respectively.

simulations and found that a helix–turn–helix structure is predominant for Fs at 300 K.⁴³ In a seminal work, Garcia and Sanbonmatsu⁴⁴ applied the REMD techniques so that equilibrium sampling could be achieved. They found that the AMBER-ff94 force field with explicit solvent significantly overestimates the T_m of Fs ($T_m \sim 400$ K). This can be corrected by ignoring all backbone torsion terms in this force field. Using the modified force field, they suggested that Fs' preference for the helix–turn–helix structure in the GB simulations is likely an artifact of the implicit solvent.⁸⁴ Sorin and Pande⁴⁵ proposed another modified version of the AMBER force field (parm99 ϕ) with explicit solvent that not only renders correct thermodynamics of this peptide but also reproduces the non- α conformations and helix–coil kinetics better than Garcia's modified version. Interestingly, their results showed that the population of the helix–turn–helix is considerable.

To reveal the structural features of our Fs model in more detail, we clustered the structures sampled from the 30th to 100th nanosecond of the simulation at 300 K according to the root-mean-square deviation (RMSD) of backbone atoms with a cutoff of 0.25 nm. A total of 125 clusters were identified, and the top 10 clusters which account for $>50\%$ of the total population are shown in Figure 2a. Clearly, various helical topologies can be sampled with our force fields. Following Zhang et al.,⁴³ we defined seven helical topologies, tagged them on each cluster, and then counted the overall populations of each topology. The topologies include (1) full helix (F–H), (2) one-helix (1-H), which has only one helical fragment, (3) helix–turn–helix (HTH), which has two helical fragments and an intervening loop/turn in the middle of the peptide, (4) N-helix–turn–helix (N-HTH), which has a helix–turn–helix topology with the loop/turn closer to the N-terminus, (5) C-helix–turn–helix (C-HTH) with the loop/turn closer to the C-terminus, (6)

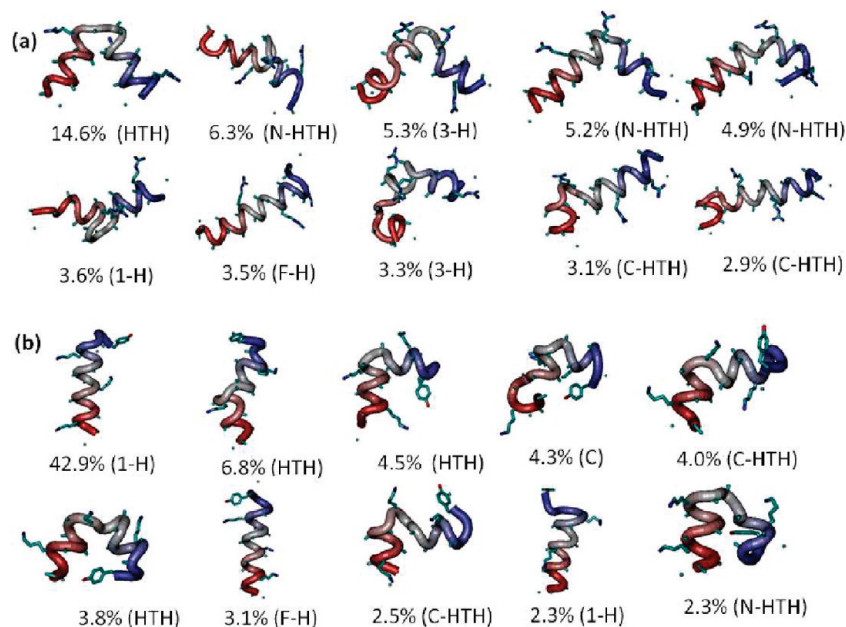


Figure 2. Representative structures of the most populated clusters of the Fs (a) and AK17 (b) peptides.

Table 2. Probabilities of Different Helical Topologies Sampled in the Simulations of the Fs and AK17 Peptides

	Fs	AK17
full helix	3.5%	3.1%
one helix ^a	17.1%	55.0%
helix–turn–helix ^b	19.3%	15.7%
N-helix–turn–helix ^c	20.0%	6.8%
C-helix–turn–helix ^d	15.7%	7.4%
three-turn–helix ^e	19.9%	7.0%
full coil	4.5%	5.0%

^a A conformation with only one helical fragment. ^b Two helical fragments intervened by a loop/turn in the middle of the peptide. ^c A helix–turn–helix with the loop/turn closer to the N-terminus. ^d A helix–turn–helix with the loop/turn closer to the C-terminus. ^e A conformation with three helical fragments.

three-turn–helix (3-H), which has three helical fragments, and (7) full coil (C).

It can be seen from Table 2 that the structures with a two-helix topology are the most abundant in our Fs simulations. The overall population of HTH, N-HTH, and C-HTH is ~55%, comparable to the ~53% in implicit solvent simulations⁴³ and slightly larger than the ~42% in explicit solvent simulations.⁴⁵ The probability of a full helical structure is ~3%, smaller than the ~20% and the ~17% estimated from the implicit solvent⁴⁵ and explicit solvent models,⁴⁵ respectively. This is because, in our Fs model, the fraying ends prevent the formation of a perfect full helix. If the one-helix topology is also considered, the combined population reaches ~21%, which is much closer to the previous studies. The topological distribution here can also be cross-checked using the radius of gyration (R_g) of the conformational ensemble. Small-angle X-ray scattering (SAXS) measurements reported an R_g value of ~0.9 nm for the Fs.⁴⁵ Our calculated $\langle R_g \rangle$ is $\sim 0.85 \pm 0.07$ nm, slightly smaller than but still in agreement with experimental results.

One may wonder if the preference to the HTH topologies is inherent to our model. We therefore conducted the same analysis for the AK17 peptide. The top 10 clusters and the

populations of each helical topology are shown in Figure 2b and Table 2, respectively. The most stable cluster forms almost all possible helical HBs (on average ~12 helical HBs) except for the C-terminal Gly. The backbone of the Gly residue, known as a helix breaker,⁸⁵ adopts an α_L conformation in the most stable cluster. Even though the AK17 peptide has only four fewer residues than the Fs peptide, the structures with only one helical fragment make up ~58% of the total population. On the other hand, the population of two-helix topologies decreases to ~30%. This is in line with an implicit solvent MD study by Chowdhury et al.⁴² where a similar peptide [Ac–YG(AAKAA)₂AAKA–NH₂] was found to be ~60% full helix. Moreover, the calculated R_g of the AK17 is 0.78 ± 0.06 nm, close to the experimental value of 0.82 nm.⁴⁶ The topological analyses of the Fs and AK17 peptides demonstrate that our force field is able to capture the common topological features observed in previous implicit and explicit solvent models.

Besides the REMD simulations, a 3 μ s normal MD simulation was performed for the AK17 at 300 K. Our purpose was 2-fold. In the REMD simulation, a 10 fs time step was used, and the mass of peptides is tripled to maintain replicas at high temperatures. In the normal MD, a 6 fs time step was used without scaling of the mass of the peptide. This time step is well below the upper limit of a 7.5 fs time step for the conservation of the energy of a system when using dummy atoms in simulations.⁸⁶ Thus, any effect of the REMD setup on the thermodynamics properties can be seen from the comparison with normal MD results. Our second purpose was to examine how fast our model can sample the conformational space with a normal MD protocol. Unlike the REMD, a normal MD is able to provide information about the transition between states, which can be valuable for an assumption-free analysis of the kinetics of protein folding.^{7,8}

The helical content and the average R_g during the normal MD simulation are ~43% and 0.79 ± 0.06 nm, respectively,

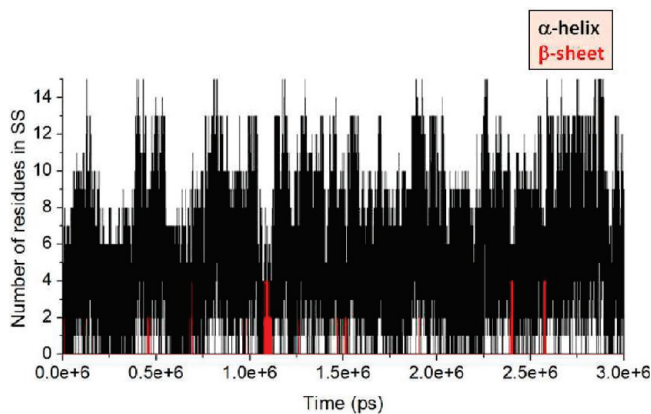


Figure 3. Change in the numbers of helical and sheet residues during a normal MD simulation of the AK17 peptide at 300 K.

which agree well with our REMD results (41% and 0.78 ± 0.06 nm, respectively). In addition, the residual helical contents of each amino acid of the REMD simulation also match well with those of the normal MD simulations (Figure 1a). These suggest that our REMD protocol has little effect on the thermodynamics.

Figure 3 shows the change in the numbers of residues in helices and sheets during the normal MD simulations. Multiple helix–coil and sheet–coil transitions are observed. Although a detailed kinetic analysis of the folding mechanism is out of the scope of this paper, we wanted to roughly estimate how fast the model travels in the conformational space. Following what we have done before,²¹ we tracked the dynamic change in the number of helical residues. After a coil is reached (zero helical residues), a folding event is thought to occur once a nearly full helix is formed (≥ 12 helical residues). An unfolding event is assumed as a transition from a nearly full helix to a coil. In the $3 \mu\text{s}$ simulations, there are a total of 32 folding/unfolding events. The averaged times for folding (τ_f) and unfolding (τ_u) events are $\sim 51 \pm 43$ and $\sim 44 \pm 33$ ns, respectively, which are similar to our previous calculations.²¹ The relaxation time (τ) of helix–coil transition in our MD simulation is estimated to be ~ 24 ns according to $\tau = \tau_u \tau_f / (\tau_u + \tau_f)$. The kinetics of helix–coil transition here are approximately an order of magnitude faster than the helix–coil kinetics from experimental measurements ($\tau \sim 100\text{--}500$ ns).^{81,80,87,88} Two reasons may account for the acceleration. One is the CG solvent model that we used. Marrink et al. have shown that coarse-graining on water can lead to an overall increase in the diffusion rate of solutes by 4–5 times.⁷³ The other factor may be our simplification to the protein model, leading to a smoother free energy surface for the helix–coil transition.

Simulations of β -Hairpin Peptides. *Folding of the GB1m2 Peptide.* With the same force field, we now turn to β -sheet peptides. The peptide GB1p (GEWYDDATKFTVTE), the N-terminal fragment of the protein G B1 domain, has been known to possess a native-like β -hairpin structure in aqueous solution.³² Although a Trp fluorescence study reported that this peptide is 80% folded at 297 K and has a folding time of $\sim 6 \mu\text{s}$,⁸⁹ recent NMR data suggested that its folded population is only $\sim 30\%$ at 298 K³³ and its

folding time is $17\text{--}20 \mu\text{s}$,³⁴ slower than previously estimated. Several mutants of GB1p have been derived to increase the stability and folding speed of the β -hairpin.³³ For instance, a loop mutant, GB1m2 (GEWYTNPATGKFTVTE), can have $\sim 74\%$ folded structures at 298 K, and its folding time decreases to $\sim 5 \mu\text{s}$.^{33,34}

Because of their well-characterized kinetics and thermodynamics, the GB1p series have been investigated in numerous computational studies.^{47–51,62,68,70,71,90} They have been used to test the quality of force fields. For example, the AMBER ff99SB and ff99 sets of parameters with the GB solvent fold the GB1p into helical structures instead of β -hairpins.⁶² On the other hand, several force fields such as OPLS-AA and the AMBER ff03 and ff99ci sets overestimate the stability of GB1p considerably.⁹⁰ The calculated T_m 's are 60–80 K above the experimental T_m values (293 K). Interestingly, using their OPEP force fields with the REMD, Chebaro et al. reproduced the thermal stability of GB1p.⁷⁰ The folding of GB1p series has been used in the optimization of force fields.^{68,71} The specially optimized force fields by Chen et al.⁶⁸ and by Irback and Mohanty⁷¹ are even able to discriminate the stability difference among the GB1p and its mutants.

We chose the GB1m2 mutant for the optimization and examination of our force field⁷² due to its stability and fast kinetics. It should be noted that in the optimization we only tried to reproduce the total β -sheet content by changing backbone–backbone HB strengths. Other factors such as native topologies were not considered at all. Since the β -hairpin folds slower than the α -helix, we found that the REMD simulations of the GB1m2 converge at a slower rate than the REMD simulations of the AK17 and Fs (Figure S1 in the SI), which is basically consistent with previous computational studies.⁶⁸ We therefore carried out three REMD simulations of the GB1m2 starting from three different sets of denatured conformations, two lasting for 300 ns and one lasting for 600 ns. The last 100 ns of each simulation were analyzed.

The generated conformations are clustered according to backbone and C_β atoms excluding two terminal residues with a RMSD cutoff of 0.2 nm, following the clustering scheme proposed by Daura et al.⁹¹ Figure 4a shows the superposition between the GB1p PDB structures (1pgb) and the representative structures of the largest clusters of the three REMD simulations at 300 K. The RMSDs of the three representative structures with respect to the experimental structure are 0.09, 0.09, and 0.14 nm. The slightly larger RMSD for the latter one is due to the fraying ends. The populations of the largest clusters for the three REMD simulations are 43.0%, 45.6%, and 45.5%, indicating a convergence for the sampling of native structures.

Another way to estimate native conformations is to count native backbone HBs.^{51,68,71} The number of native HBs in a given conformation is denoted by N_{nHB} . Figure 5a shows the probability distribution of N_{nHB} at 300 K. Assuming that a folded conformation has $N_{\text{nHB}} \geq 3$, the probabilities of folded conformations in the three REMD simulations are 43.2%, 45.4%, and 45.4%, respectively, consistent with the estimation by RMSD. The fractions of folded structures were

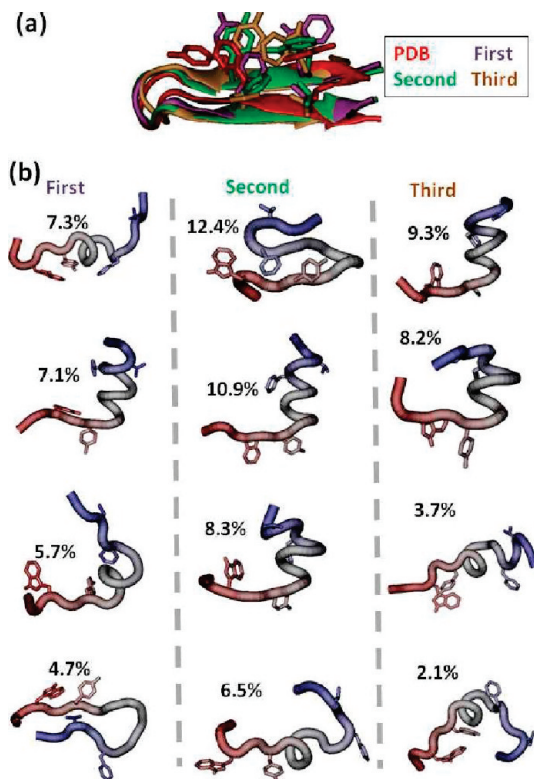


Figure 4. (a) Superimpositions of the native structure of the N-terminal hairpin of the protein G B1 domain (red) and the representative structures of the largest clusters of the three REMD simulations of the GB1m2 peptide at 300 K (purple, green, and brown). (b) The representative structures of the second to fifth most stable clusters of the three REMD simulations at 300 K.

computed for all replicas. The change in the folded fractions with temperature is shown in Figure 5b. Because the CG solvent developed by Marrink et al. has a frozen point of as high as ~ 290 K,⁷³ we could not obtain the melting curve at too low a temperature. However, since the folded fraction at 300 K is already close to one-half, we roughly estimated the T_m to be ~ 295 – 300 K by interpolating the melting curve to 50% of folded fraction. Our estimated T_m is ~ 20 – 25 K lower than experimental data (320 K).³³

There is an interesting puzzle about the hydrophobic interactions in the GB1p hairpin.⁹² There are four nonpolar side chains, Trp3, Tyr5, Phe12, and Val14. In both the crystal and NMR structures of the whole domain, all four residues contribute significantly to interstrand hydrophobic interactions to stabilize the hairpin.^{93,94} This can be seen clearly from our counting of the numbers of atomistic contacts between the side chains of the two strands (cross-strand atomistic contacts). As shown in Table 3, all four residues have at least 25 cross-strand atomistic contacts. Furthermore, the number of contacts between Trp3 and Val14 is 25, and the number of contacts between Tyr5 and Phe12 is 35. There are few contacts between Trp-Val14 and Tyr5-Phe12. Thus, Trp3-Val14 and Tyr5-Phe12 form two separated hydrophobic clusters. Nevertheless, a mutagenesis study revealed that the influence of mutation of Val14 on the thermal stability of the isolated GB1p hairpin is negligible compared with the others.⁹² Blanco et al. have also shown several interstrand

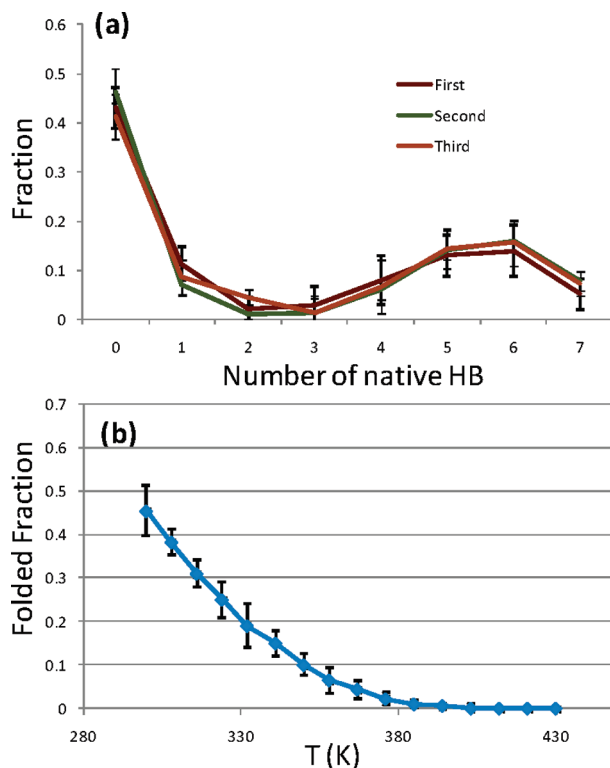


Figure 5. (a) Probability distribution of the number of native HBs in the three REMD simulations of the GB1m2 peptide. The native HBs are Pro7(O)–Gly10(NH), Asn6(O)–Lys11(NH), Asn6(NH)–Lys11(O), Thr4(O)–Thr13(NH), Thr4(NH)–Thr13(O), Glu2(O)–Thr15(NH), and Glu2(NH)–Thr15(O). (b) The melting curve of the REMD simulations of the GB1m2 peptide, which is the average of the three simulations. The error bars were estimated from block averages with a block size of 20 ns.

Table 3. Numbers of Cross-Strand Contacts Involved with Trp3, Tyr5, Phe12, and Val14 for the Folded and Unfolded Ensembles in the REMD Simulation of the GB1m2 Peptide at 300 K and the Native β -Hairpin from the Crystal Structure (PDB ID: 1pgb)

	Trp3	Tyr5	Phe12	Val14
folded ^a	22.9 ^b \pm 5.2	20.1 \pm 4.1	34.5 \pm 6.9	8.5 \pm 2.7
unfolded ^a	5.8 \pm 1.5	3.5 \pm 1.0	8.8 \pm 2.2	0.6 \pm 0.2
PDB	31	35	41	25

^a The folded ensemble includes all of the conformations that have a RMSD of backbone and C_{β} atoms less than 0.2 nm; the other conformations are considered the unfolded structures. ^b For example, “22.9” here means that the folded ensemble has on average 22.9 atomistic contacts between the side chain of Trp3 and the side chains of hydrophobic amino acids on the opposing strand. A cutoff distance of 0.54 nm is used to count contacts.

NOEs in the NMR study of the GB1p hairpin, but no interstrand NOE between the Val14 side chain and other residues.³² In our folded ensemble (Table 3), the numbers of interstrand contacts involving Trp3, Tyr5, and Phe12 are 22.9 ± 5.2 , 20.1 ± 4.1 , and 34.5 ± 6.9 , respectively, which are still comparable to those in the crystal structures (31, 35, and 41, respectively). However the interstrand contact of Val14 is significantly reduced from 25 to 8.5 ± 2.7 , indicating that Val plays a less significant role in the cross-strand hydrophobic cluster. This agrees with experimental

observations. A visual investigation reveals that the loss of contacts of Val14 is because the side chain of Trp3 prefers to rotate to interact more with Tyr5 and Phe12, and consequently, the Trp3-Val14 cluster in the crystal structure is broken. Our results are also in line with a simulation study on this GB1p, in which only Val14 is detached but the other three aromatic side chains remain in the hydrophobic cluster at an early stage of an unfolding process.⁹⁵

Although the GB1p series are β -hairpin peptides, GB1p has been thought to possess a minor α -helical conformer. The simulation study of GB1p by Levy and co-workers⁴⁸ was able to estimate the free energy difference between the major β -hairpin state and the minor α -helix state at room temperature by applying the temperature-weighted histogram analysis method. According to their calculation, the α -helical state makes up $\sim 8\%$ of the population. The α -helical probabilities of our REMD simulations of GB1m2 are calculated (see Model and Methods) to be $\sim 6.7\text{--}10.7\%$.^{48,96} The microstates accounting for this helical probability could be found from the most stable clusters in the REMD simulations. Figure 4b shows the second to fourth most stable clusters of the three REMD simulations. In all of the REMD simulations, there are persistent clusters with helical fragments spanning from Asn6 to Lys11 or even extending to the C-terminus. Using OPEP-REMD simulations, Chebaro et al. also found that GB1p has an α -helix spanning Asp6 to Thr11.⁷⁰ We suspect that the minor α -helical propensities of the GB1p series may be a consequence of their loop structures, since their six-residue loop requires the formation of an α -helical turn.

Folding of the Trpzip2 Peptide. Trpzip2 (SWTWENGK-WTWK), a 12-residue tryptophan zipper, is the smallest peptide adopting a unique β -fold.³⁵ With the high propensity of a type I' turn in the ENGK region and a characteristic structural motif of Trp-Trp cross-strand pairs, it has exceptional stability ($T_m \sim 345$ K) and fast folding kinetics ($\tau_f \sim 1.8 \mu\text{s}$).⁹⁷ The folding of Trpzip2 has been examined by numerous force fields.^{62,68,98,99} In the CHARMM-CMAP force field simulations with the GB solvent model, Chen et al. found that their REMD simulations are more difficult to converge for Trpzip2 than for the GB1p series despite the experimental fact that Trpzip2 folds much faster than the GB1p series.⁶⁸ In the study by Irback and Mohanty,⁷¹ their force field can fold the GB1p series correctly but cannot fold Trpzip2. This indicates that the folding of Trpzip2 is an interesting test of the ability of a force field to fold a β -sheet structure.

We performed two REMD simulations of the Trpzip2 peptide with 16 replicas from denatured conformations, one lasting for 120 ns and the other for 150 ns. It appears that convergence can be achieved after 50 ns of simulation (Figure S2 in SI), faster than in the simulations of the GB1m2 peptide (Figure S1 in SI). Thus, the last 70 ns of each REMD simulation of Trpzip2 were used for analysis. The sampled conformations were clustered according to the RMSD of backbone and C_β atoms of residues 2–11 with a cutoff of 0.2 nm. The representative structures of the largest clusters at 300 K in the two simulations are shown in Figure 6a, which are superimposed with the NMR structure of Trpzip2

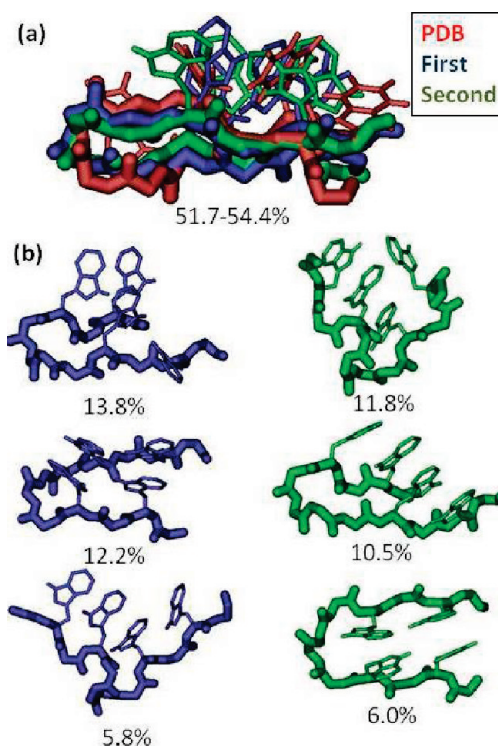


Figure 6. (a) Superposition of the native structure of Trpzip2 (red) and the representative structures of the largest clusters of the REMD simulations of this peptide at 300 K (blue and green, respectively). (b) The representative structures of the second to fourth most stable clusters of the two REMD simulations at 300 K.

(PDB ID: 1LE1).³⁵ Their RMSDs to the NMR structure are 0.14 and 0.16 nm. The populations of the largest cluster in the two simulations are 51.7% and 54.5%. Inspection of other top clusters (Figure 6b) of the two REMD simulations shows that the stabilities of the other minor non-native states are also consistent in the two simulations, suggesting that convergence has been achieved in the simulations.

The N_{nHB} analysis is also done for the REMD simulations of Trpzip2. The probability distributions of the two REMD simulations are shown in Figure 7a. Assuming the native structure with $N_{\text{nHB}} \geq 3$, the estimated populations of folded structures are 50.3% and 52.0% for the two REMD simulations, respectively, which are close to our estimation by clustering conformations. The melting curve of Trpzip2 is plotted by calculating the fraction of folded structures at different temperatures (Figure 7b), which gives a T_m of $\sim 300\text{--}305$ K. The calculated T_m is about 40–45 K lower than the experimental T_m for the Trpzip2 (345 K).³⁵

Our model underestimates the stability of the Trpzip2 native structure. In the folded ensemble of our simulations, the key stabilizing structural elements such as native HBs and type I' β -turns are well preserved. However, the packing of side chains of four Trp residues differs a lot from the NMR structure (Figure 6a). In the NMR structure, Trp11/Trp2 and Trp4/Trp9 form cross-strand pairs through an edge-to-face packing. To satisfy this packing motif, the hairpin is highly twisted.³⁵ In our folded structures, there is no edge-to-face packing among the Trp side chains. Instead, Trp11/Trp2 and Trp4/Trp9 form cross-strand pairs through ring

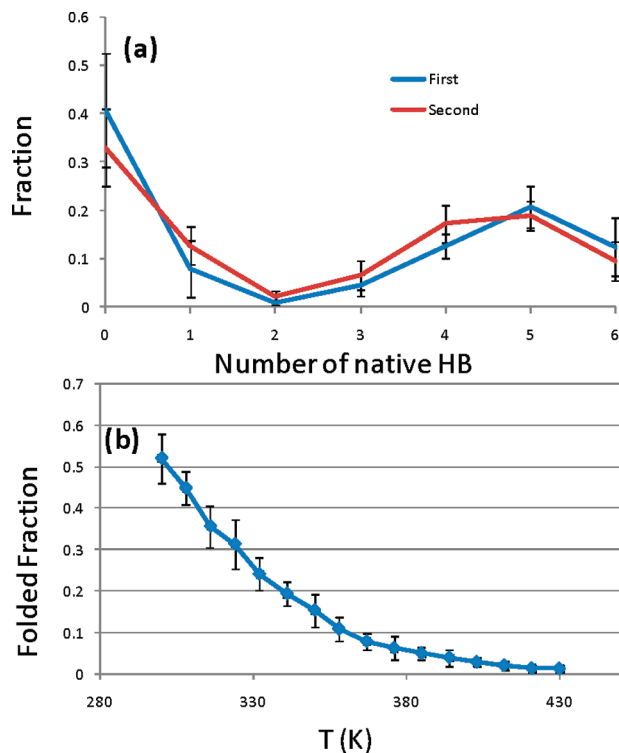


Figure 7. (a) Probability distribution of the number of native HBs in the two REMD simulations of the Trpzip2 peptide. The native HBs are Glu5(O)–Lys8(NH), Glu5(NH)–Lys8(O), Thr3(O)–Thr10(NH), Thr3(NH)–Thr10(O), Ser1(O)–Lys12(NH), and Ser1(NH)–Lys12(O). (b) The melting curve of the REMD simulations of the GB1m2 peptide, which is the average of the two simulations. The error bars were estimated from block averages with a block size of 20 ns.

stacking. As a result, the hairpin is less twisted. The loss of the experimental packing motif arises because the edge-to-face packing between aromatic rings is driven by electrostatic multipole interactions,¹⁰⁰ which is missing in our model. The aromatic groups in our model are only composed of van der Waal (vdW) particles. The free energy calculation in explicit solvent by Guvench and Brooks⁵⁰ revealed that for the Trpzip2 peptide, the edge-to-face packing between Trp side chains with the side chain multipole interactions provides ~ 4 kcal more stabilization than the stacked packing with only the side chain vdW interactions. This may account for the loss of stability in our Trpzip2 model. This may also be the reason why the force field by Irback and Mohanty,⁷¹ having a model of aromatic groups similar to ours, cannot fold the native β -hairpin topology of the Trpzip2 since it lacks the multipole interactions. Thus, inclusion of the multipole features of aromatic groups in explicit or implicit ways should be one of our next steps in the improvement of the force field.

Slow sampling for β -sheet peptides has been observed in previous computational studies. In a Monte Carlo (MC) study, Ulmschneider and Jorgensen⁶⁹ carried out MC simulations of two hairpin peptides, U(1–17)T9D¹⁰¹ and Trpzip2. They used a concerted-rotation MC move which is supposed to enhance sampling speed. For U(1–17)T9D, in eight simulations lasting for 20–80 million MC steps, only one simulation reached the native state. The other simulations

were trapped in non-native hairpins. For Trpzip2, despite much longer simulations (350–550 million steps), only two out of eight simulations reached the native state. Even with the REMD sampling method, Chen et al.⁶⁸ found that the native state of Trpzip2 was formed in one REMD simulation but was not formed in another. To examine the sampling speed for the β -sheet, we performed five 4.8–5 μ s normal MD simulations of Trpzip2 at 300 K starting from different denatured conformations. The folding of Trpzip2 is indeed slower than the folding of AK17. As shown in Figure 8a, four of the five simulations reached the native structures.

The folding time ranges from ~ 0.5 to ~ 3.1 μ s. The secondary structure analysis of the five trajectories reveals (Figure 8b) that before the native state is reached, the peptide experiences non-native β -hairpins and transient helical structures. In the fastest folding simulation (MD-1), no significant non-native secondary structures are observed. In the simulation (MD-5) that did not reach the native states, the system is trapped in several exchanging non-native hairpins. The non-native hairpins have fewer β -sheet structures than the native state, and their dwell time ranges from ~ 0.1 to about 1 μ s. However, once the native state is achieved, it lasts until the end of the simulation, indicating that the native state is much more stable than the non-native hairpins.

The four folding trajectories (MD 1–4) reveal two features of the folding pathways of Trpzip2. One is that all of the non-native secondary structures need to be completely unfolded before the folding of the native state starts, which is reflected by the secondary structure analysis (Figure 8b). The other feature is revealed by monitoring the structural change during the folding, which is shown in Figure 9. In all four simulations, the folding initiates at the type I' turn of Asn6–Gly7 and propagates toward the tails by forming cross-strand HBs sequentially. This folding mechanism is consistent with the “zip-out” model proposed by Munoz et al.¹⁰² in their experiments with the GB1p peptide, which was supported by later experimental and computational studies.^{103,104}

Simulations of the Trp-cage Peptide. The Trp-cage mini-protein (NLYIQ₅WLKDG₁₀GPSSG₁₅RPPPS₂₀) is a designed peptide that behaves like large globular proteins (Figure 10a).³⁶ It contains an α -helix, a 3_{10} helix, and a polyproline II segment. Its tertiary structure is stabilized by a compact hydrophobic core centered at Trp6. The side chain of Trp6 is caged by the side chains of Tyr3, Leu7, Pro12, and Pro18. In addition, a distant HB between the side chain of Trp6 and the backbone carbonyl group of Arg16 and a salt bridge between the side chains of Asp9 and Arg16 are suggested to contribute to the stability of the Trp-cage. The temperature jump experiments show that the folding of the Trp-cage is extremely fast ($\tau_f \sim 4.1$ μ s),¹⁰⁵ which makes the Trp-cage an ideal model for computational study. Many simulations with explicit or implicit solvent have reproduced the folded structure of Trp-cage with excellent accuracy.^{9,52–58,68,70,71}

Because there are no attempts to fit the parameters of our force field by simulation of the Trp-cage, we think that the Trp-cage is a good case to test the transferability of our force field. We performed both REMD and normal MD simulations

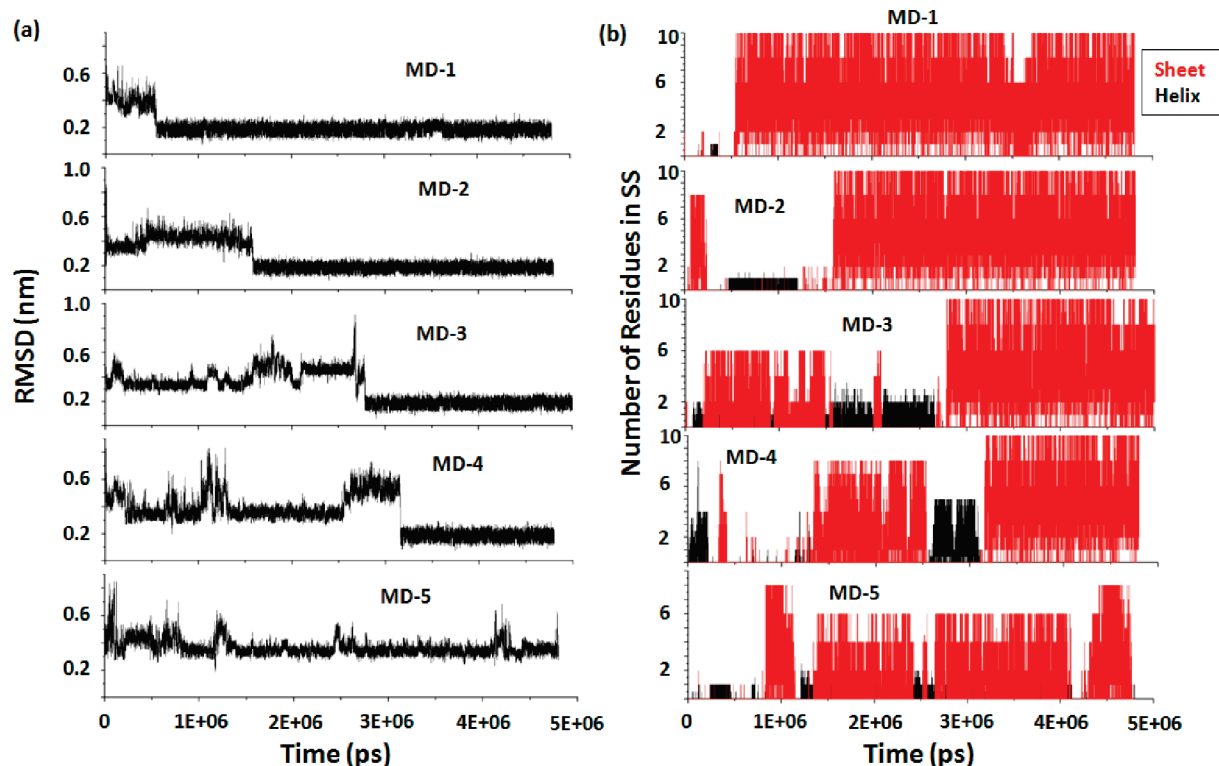


Figure 8. (a) RMSD with respect to the native structure (1LE1) of the normal MD simulations of Trpzip2. (b) The numbers of residues in helical (black) and sheet (red) structures.

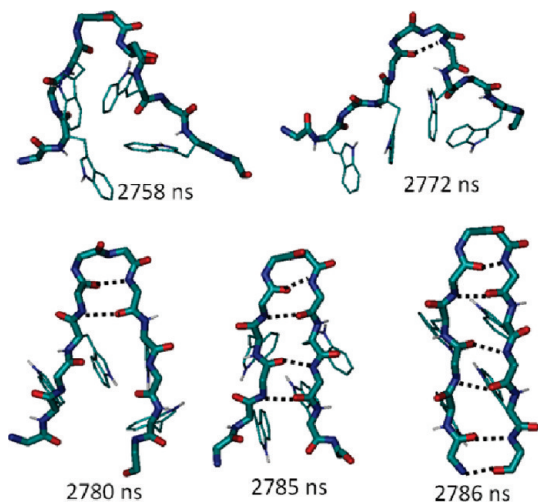


Figure 9. Representative pathway of the folding of Trpzip2 observed in our simulations.

for the Trp-cage peptide. The REMD simulation of the Trp-cage lasted for 180 ns, and the last 80 ns of simulations were used for structural analysis. The sampled conformations at 300 K were clustered according to the RMSD of backbone atoms of residues 2–19, discarding the fraying ends. The representative structures of the five most stable clusters are shown in Figure 10. The representative structure of the largest cluster has a RMSD of 0.20 nm with respect to the NMR native structure (PDB ID: 1L2Y).³⁶ In addition to a good fitting of the backbone structures, the native RMSD of all heavy atoms except for two terminal residues is ~ 0.27 nm for the representative structure of the largest cluster. In particular, the caged arrangement of the hydrophobic side chains is well reproduced. Assuming that the largest cluster

is folded, the fraction of folded conformations is 33.3%. We also applied the method of Garcia and co-workers to the calculation of a folded population, in which the native RMSD is calculated for each conformation and the conformation having a RMSD greater than a certain threshold is considered a folded state. We chose a threshold of 0.24 nm, slightly larger than Garcia et al.'s (0.22 nm).⁹ The folded population is estimated to be about 35.2%, consistent with our clustering analysis. Since the experimental probability of the folded structure is $\sim 70\%$ at 300 K,³⁶ our force field currently underestimates the stability of the native structure of Trp-cage.

We also carried out a 3 μ s normal MD simulation of Trp-cage at 300 K starting from a denatured conformation. The RMSD from the NMR structure is monitored during the simulations, which is plotted in Figure 11. Clearly, multiple folding and unfolding events occurred during the simulation. Discarding the first 500 ns of simulation, about 36.7% of sampled conformations have a RMSD from the NMR structure of less than 0.24 nm. This agrees very well with our REMD result. While the detailed folding mechanism of the normal MD trajectory is outside the scope of this paper, an estimation of the time scale of conformational change would be valuable. We estimated a folding time τ_f as the time the system takes to go deep into the folded basin (RMSD < 0.14 nm) after it enters the unfolded basin (RMSD > 0.70 nm). By this definition, six folding events are identified in our normal MD trajectory. The average folding time τ_f is approximately 270 ± 110 ns. Compared to the experimental τ_f of 4.1 μ s,¹⁰⁵ our model again has kinetics 1 order of magnitude faster than experimental data, similar to our observation in the normal MD of AK17.

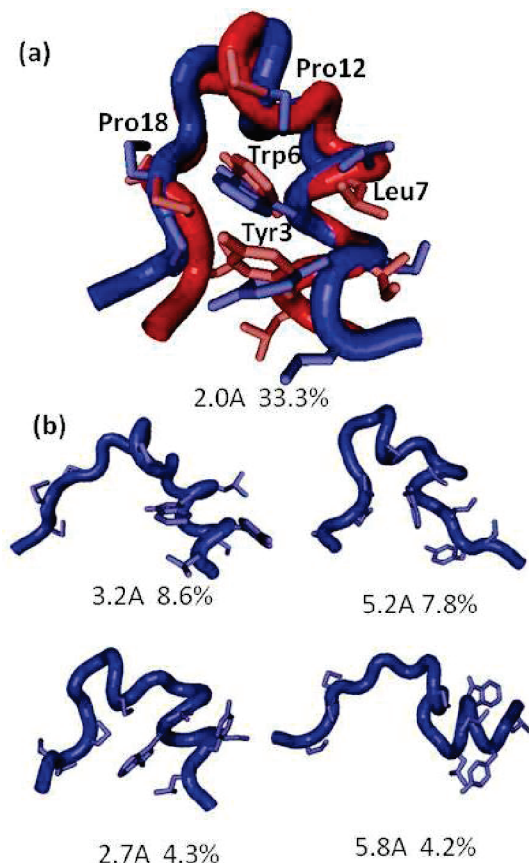


Figure 10. Representative structures, the probabilities, and the RMSD of the representative structures to the NMR structure (1L2Y) of the five most stable clusters (blue) of the REMD simulation of the Trp-cage at 300 K. (a) The representative structure of the most stable cluster is superimposed onto the NMR structure (red).

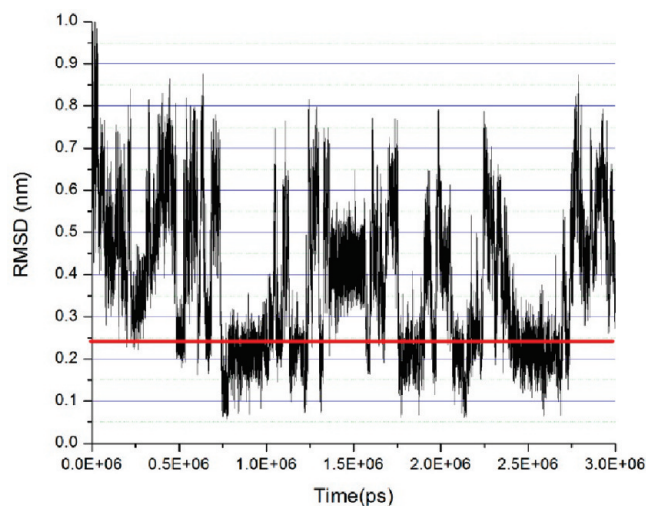


Figure 11. Change in the backbone RMSD (nm) to the NMR structure (1L2Y) during the normal MD simulation of the Trp-cage. The red line indicates the cutoff used for the calculation of the folded probability.

The HB (W6H_ε/R16O) between the H_ε of the Trp6 side chain and the backbone carbonyl of Arg16 and the salt bridge (D9/R16) between the side chains of Asp9 and Arg16 have been widely studied.^{36,37} The W6H_ε/R16O interaction is well supported by NOE signals.³⁷ The mutation of Trp6 by other

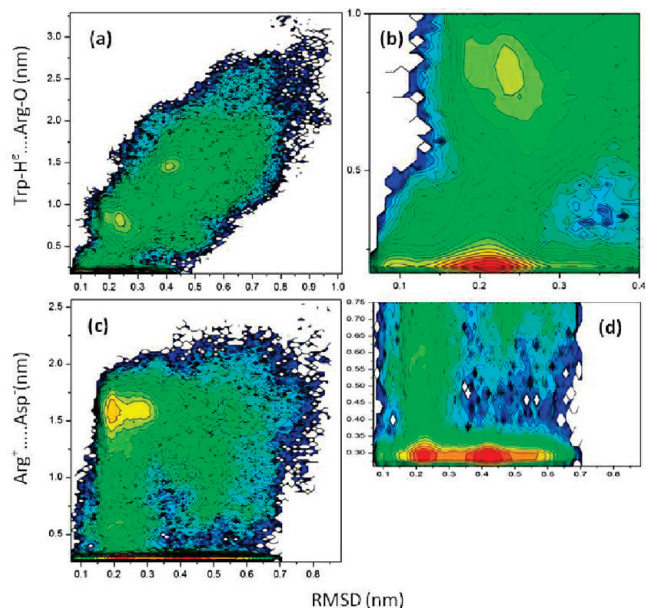


Figure 12. (a, b) The 2D statistical map of the backbone RMSD (nm) from the NMR structure against the W6H_ε/R16O distance (nm). (c, d) The 2D statistical map of the backbone RMSD (nm) from the NMR structure against the distance between the guanidine of Arg16 and the carboxylate of Asp9 (nm). Each contour line denotes 1.0 kJ/mol of free energy difference.

aromatic groups, including a naphthalene group, destabilizes the native structure by about 9–12.5 kJ/mol.¹⁰⁶ On the other hand, the D9/R16 salt bridge is less certain because the side chain/side chain NOEs across the bridging sites are absent.³⁶ The mutation that breaks the salt bridge only destabilizes the native structure by 3.4–6 kJ/mol.³⁷ In many MD simulations of the Trp-cage, the D9/R16 salt bridge is observed either as an interaction facilitating the folding^{52,107} or as a kinetic trap.^{56,108}

It is desirable to examine the roles of the W6H_ε/R16O HB and D9/R16 salt bridge in our folding simulations. To do so, we projected the normal MD trajectory onto a 2D map of the RMSD and the distance between W6H_ε and R16O (Figure 12a,b) and a 2D map of the RMSD and the distance between the guanidinium of Arg16 and the carboxylate of Asp6 (Figure 12c,d). It is clear that on the RMSD–W6H_ε/R16O map (Figure 12a) the W6H_ε/R16O distance correlates well with the RMSD. The W6H_ε/R16O HB can only be formed in the native basin (Figure 12b). On the contrary, there is no apparent correlation between the D9/R16 distance and the RMSD (Figure 12c). The salt bridge can form in both the native basin (RMSD ~ 0.21 nm) and the non-native basin (RMSD ~ 0.42 nm) (Figure 12d). Our observation is in line with the recent astonishing 100 μs all-atom simulations with explicit solvent by Garcia and co-workers.⁹ They found that the probability of W6H_ε/R16O HB matches well with the folded fraction at all temperatures. However, at high temperatures, where the folded population is scarce, there is still 20%–30% D9/R16 salt bridge. A simple explanation for our observation is that since HBs have directionality, to form the W6H_ε/R16O HB, the side chain of Trp6 and the backbone of Arg16 should be arranged in a certain orientation, which can only be satisfied in the native conformation.

However, salt bridge interactions have little directionality, and the long side chain of Arg16 is quite flexible. Therefore, the D9/R16 salt bridge can be accommodated in various conformations and does not exclusively stabilize the native structure.

Conclusions

It is important to perform converged sampling of peptide folding to evaluate the quality of a force field.^{62,68,70,71} In this paper, we report our effort to fold peptides with our recently developed PACE force field in tandem with a CG solvent model. The peptides include AK, Fs, GB1m2, Trpzip2, and Trp-cage, which have been shown to fold in various topologies such as α -helix, β -sheet, and mixed helix/coil structures. Despite our force field not being optimized for the folding of these peptides, their native structures are all identified as the dominant conformations in our REMD simulations. Normal MD simulations have also been performed for AK17, Trpzip2, and Trp-cage. Convergence can be reached in microsecond simulations which give very similar results to the REMD simulations. It is noted that the simulated kinetics of AK17 and Trp-cage by the force field are about 1 order of magnitude faster than the experimental ones. We attribute this to the coarse-graining of the water solvent and the simplification of protein potentials. The force field appears to underestimate the stability of native structures of Trpzip2 and Trp-cage, indicating the need for further improvement of the force field, such as considering multipole interactions in aromatic rings. As pointed out in our preceding paper,⁷² our force field was parametrized to reproduce the results of all-atom force fields in a solvent-exposed environment. However, as our CG water model is simply LJ fluid, the dielectric screening effect of water may be implicitly incorporated into our parameters for interactions between polar sites. Thus, the current force field may only be suitable for the simulations of small peptides which have reasonably good exposure to a solvent. Our ongoing parametrization is underway to account for environment-dependent electrostatic interactions.

Acknowledgment. We are grateful to the Research Grants Council of Hong Kong (CA06/07.SC05, 663509) for financial support of this research.

Supporting Information Available: Figure S1, the change in the helical content and the folded fraction of the GB1m2 and AK17 peptides during the REMD simulations; Figure S2, the change in the helical content and folded fraction of the Trpzip2 peptide during the REMD simulations. This information is available free of charge via the Internet at <http://pubs.acs.org>

References

- (1) Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140.
- (2) Van Gunsteren, W. F.; Dolene, J.; Mark, A. E. *Curr. Opin. Struct. Biol.* **2008**, *18*, 149.
- (3) Freiderichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. J. *J. Comput. Chem.* **2009**, *30*, 864.
- (4) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9.
- (5) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (6) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435.
- (7) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526.
- (8) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702.
- (9) Day, R.; Paschk, D.; Garcia, A. E. *Proteins* **2010**, *78*, 1889.
- (10) Khandogin, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880.
- (11) Liang, C.; Derreumaux, P.; Wei, G. *Biophys. J.* **2007**, *93*, 3353.
- (12) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, I. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
- (13) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122.
- (14) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.
- (15) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10.
- (16) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.
- (17) Fujitsuka, Y.; Chikenji, G.; Takada, S. *Proteins* **2006**, *62*, 381.
- (18) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394.
- (19) Ding, F.; Tsao, D.; Nie, H. F.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010.
- (20) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116.
- (21) Han, W.; Wu, Y.-D. *J. Chem. Theory Comput.* **2007**, *3*, 2146.
- (22) Han, W.; Wan, C.-K.; Wu, Y.-D. *J. Chem. Theory Comput.* **2008**, *4*, 1891.
- (23) Brooks, B. R.; Brucoleri, Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (24) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1.
- (25) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.
- (26) Jorgenson, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
- (27) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76.
- (28) Marqusee, S.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 8898.
- (29) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5286.
- (30) Lockhart, D. J.; Kim, P. S. *Science* **1993**, *260*, 198.
- (31) Hughes, R. M.; Waters, M. L. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514.
- (32) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584.

- (33) Fesinmeyer, R. M.; Hudson, F. M.; Anderson, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238.
- (34) Olsen, K. A.; Fesinmeyer, R. M.; Stewart, J. M.; Anderson, N. H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *42*, 15483.
- (35) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578.
- (36) Neidigh, J. W.; Fesinmeyer, R. M.; Anderson, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425.
- (37) Barua, B.; Lin, J. C.; Williams, V. D.; Kummeler, P.; Neidigh, J. W.; Anerson, N. H. *Protein Eng. Des. Select.* **2008**, *21*, 171.
- (38) Struthers, M.; Ottesen, J. J.; Imperiali, B. *Folding Des.* **1998**, *3*, 95.
- (39) Struthers, M. D.; Cheng, R. C.; Imperiali, B. *Science* **1996**, *271*, 342.
- (40) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102.
- (41) Fung, A.; Li, P.; Godoy-Ruiz, R.; Sanchez-Ruiz, J. M.; Munoz, V. *J. Am. Chem. Soc.* **2008**, *130*, 7489.
- (42) Chowdhury, S.; Zhang, W.; Wu, C.; Xiong, G.; Duan, Y. *Biopolymers* **2003**, *68*, 63.
- (43) Zhang, W.; Lei, H.; Chowdhury, S.; Duan, Y. *J. Phys. Chem. B* **2004**, *108*, 7479.
- (44) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 2782.
- (45) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472.
- (46) Zagrovic, B.; Jayachandran, G.; Millett, I. S.; Doniach, S.; Pande, V. S. *J. Mol. Biol.* **2005**, *353*, 232.
- (47) Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50.
- (48) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722.
- (49) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795.
- (50) Guvench, O.; Brooks, C. L. *J. Am. Chem. Soc.* **2005**, *127*, 4668.
- (51) Weinstock, D. S.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R. M.; Wu, K.-P.; Baum, J. *J. Am. Chem. Soc.* **2007**, *129*, 4858.
- (52) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548.
- (53) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258.
- (54) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587.
- (55) Schug, A.; Herges, T.; Wenzel, W. *Phys. Rev. Lett.* **2003**, *91*, 158102.
- (56) Zhou, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280.
- (57) Zhan, L.; Chen, J. Z. Y.; Liu, W.-K. *Proteins* **2007**, *66*, 436.
- (58) Paschek, D.; Nymeyer, H.; Garcia, A. *J. Struct. Biol.* **2007**, *157*, 524.
- (59) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6465.
- (60) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2004**, *386*, 460.
- (61) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys.* **2004**, *307*, 269.
- (62) Shell, M. S.; Ritterson, R.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 6878.
- (63) Liwo, A.; Khalili, M.; Czaplowski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H. A. *J. Phys. Chem. B* **2007**, *111*, 260.
- (64) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008.
- (65) Ferrara, P.; Apostolakis, J.; Caffisch, A. *J. Phys. Chem. B* **2000**, *104*, 5000.
- (66) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56*, 310.
- (67) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362.
- (68) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728.
- (69) Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1849.
- (70) Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. *J. Phys. Chem. B* **2009**, *113*, 267.
- (71) Irbach, A.; Mohanty, S. *Biophys. J.* **2005**, *88*, 1560.
- (72) Han, W.; Wan, C.-K.; Jiang, F.; Wu, Y.-D. *J. Chem. Theory Comput.* **2010**, DOI: 10.1021/ct1003127.
- (73) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.
- (74) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Mariink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819.
- (75) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 592.
- (76) Kasson, P. M.; Kelley, N. W.; Singhai, N.; Vrljic, M.; Brunger, A. T.; Pande, V. S.; Biophys, J. *Proc Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11916.
- (77) Lockhart, D. J.; Kim, P. S. *Science* **1992**, *257*, 947.
- (78) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (79) Luo, P.; Baldwin, R. L. *Biochemistry* **1997**, *36*, 8413.
- (80) Thompson, P. A.; Eaton, W. A.; Hofrichter, J. *Biochemistry* **1997**, *36*, 9200.
- (81) Lednev, I. K.; Karnoup, A. S.; Sparrow, M. C.; Asher, S. A. *J. Am. Chem. Soc.* **2001**, *123*, 2388.
- (82) Lyu, P. C.; Liff, M. I.; Marky, L. A.; Kallenbach, N. R. *Science* **1990**, *250*, 669.
- (83) Miick, S. M.; Casteel, K. M.; Milhauser, G. L. *Biochemistry* **1993**, *32*, 8014.
- (84) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934.
- (85) Chakrabarty, A.; Kortemme, T.; Baldwin, R. L. *Protein Sci.* **1994**, *3*, 843.
- (86) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.
- (87) Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B. *Biochemistry* **1996**, *35*, 691.
- (88) Huang, C.-Y.; Getahun, Z.; Zhu, Y.; Klemke, J. W.; DeGrado, W. F.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2788.

- (89) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196.
- (90) Lwin, T. Z.; Luo, R. *Protein Sci.* **2006**, *15*, 2642.
- (91) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236.
- (92) Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564.
- (93) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657.
- (94) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721.
- (95) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062.
- (96) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801.
- (97) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Acad. Natl. Sci. U.S.A.* **2004**, *101*, 4077.
- (98) Okur, a.; Strockbine, B.; Hornak, V.; Simmerling, C. *J. Comput. Chem.* **2003**, *24*, 21.
- (99) Chen, C.; Xiao, Y. *Bioinformatics* **2008**, *24*, 659.
- (100) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104.
- (101) Zerella, R.; Chen, P. Y.; Evans, P. A.; Raine, A.; Williams, D. H. *Protein Sci.* **2000**, *9*, 2142.
- (102) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196.
- (103) Du, D.; Zhu, Y.; Huang, C. Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915.
- (104) Zhang, J.; Qin, M.; Wang, W. *Proteins* **2006**, *62*, 672.
- (105) Qiu, L.; Pabit, S. A.; Roiberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952.
- (106) Barua, B.; Andersen, N. H. *Lett. Pept. Sci.* **2002**, *8*, 221.
- (107) Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. *J. Mol. Biol.* **2003**, *327*, 711.
- (108) Ding, F.; Buldyrev, s. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147.

CT100313A

JCTC

Journal of Chemical Theory and Computation

Spectroscopic Properties of Formaldehyde in Aqueous Solution: Insights from Car–Parrinello and TDDFT/CASPT2 Calculations

Paola Lupieri,^{†,‡,§} Emiliano Ippoliti,^{†,‡} Piero Altoè,^{||} Marco Garavelli,^{||} M. Mwalaba,^{⊥,∇} and Paolo Carloni^{*,†,‡,§,#}

German Research School for Simulation Sciences GmbH, 52425 Jülich and RWTH Aachen, Germany, SISSA, via Bonomea 265, 34136 Trieste, Italy, Department of Chemistry “G. Ciamician”, University of Bologna, via Selmi 2, I-40126 Bologna, Italy, International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34151 Trieste, Italy, and Democritos Modeling Center for Research in Atomistic Simulation, via Bonomea 265, 34136 Trieste, Italy

Received July 9, 2010

Abstract: We present Car–Parrinello and Car–Parrinello/molecular mechanics simulations of the structural, vibrational, and electronic properties of formaldehyde in water. The calculated properties of the molecule reproduce experimental values and previous calculations. The $n \rightarrow \pi^*$ excitation energy, calculated with TDDFT and CASPT2, agrees with experimental data. In particular, it shows a blue shift on going from the gas phase to aqueous solution. Temperature and wave function polarization contributions have been disentangled.

Introduction

The carbonyl group is a key component of (bio)organic molecules, with interesting optical properties. As we know from textbooks, the oxygen’s electron lone pairs form nonbonding orbitals (n), whose electrons can be promoted to an antibonding π orbital localized over the C=O bond. In the ground states of many carbonyl compounds, the n state is the highest energy occupied molecular orbital (HOMO).¹ In that case, the lowest transition energy is the singlet–singlet transition $n \rightarrow \pi^*$.

Formaldehyde, the simplest system containing such functionality, has been a test case for quantitative prediction of this transition (and recently also of the corresponding

emission^{2,3}) in the isolated molecule and aqueous solution. The approaches used range from semiempirical AM1,⁴ first-principle DFT,^{5–7} ab initio HF,^{8,9} and to post-HF techniques such as CISD and CASSCF^{6,10–18} for formaldehyde using either molecular mechanics in a QM/MM scheme,^{4,6,9–13,17,18} an implicit solvent,^{5,19,20} or full quantum representation^{7,15,16} for the solvent (see the Supporting Information for details).

In the gas phase (at 0 K), the calculated values range between 3.3 and 4.5 eV. This may be compared with the experimental values measured at 330 K, ranging from 3.3 to 5 eV, with a maximum around 3.8 eV.^{21,22}

In solution, at room temperature (298 or 300 K), these computational approaches predict values in the range 3.5–5.7 eV, indicating a blue shift due to the solvent as well as the temperature between 0.07 and 0.43 eV. Experimentally, formaldehyde undertakes a reaction in water, forming methyleneglycol²³



Since the equilibrium constant of the reaction is 10^4 , the absorption cannot be measured in submolar solutions. However, it can be detected in highly concentrated solutions (10 M or more) at about 4.2 eV,²³ where molecular clusters

* Corresponding author e-mail: p.carloni@grs-sim.de.

[†] German Research School for Simulation Sciences.

[‡] RWTH Aachen.

[§] SISSA.

^{||} University of Bologna.

[⊥] ICTP.

[#] Democritos Modeling Center for Research in Atomistic Simulation.

[∇] Permanent address: Department of Physics, University of Zambia, School of Natural Sciences, Lusaka 10101, Zambia.

of formaldehyde are likely to exist. The spectra are likely to retain the properties of a single formaldehyde solvated molecule.^{7,9,11–16} Thus, these experimental findings are consistent with a solvent blue-shift effect.

Ab initio Car–Parrinello (CP) molecular dynamics (MD) and hybrid CPMD/MM calculations in conjunction with techniques for excited states such as time-dependent density functional theory (TDDFT)^{24,25} are mature techniques to study electronic absorption spectra of solutes in water at room temperature.^{26–29} Ab initio MD allows realistic simulations to be performed without adjustable parameters, which is a necessary drawback of empirical approaches and continuum methods. In addition, it allows for an explicit description of H bonding, whose polarization effects might be difficult to be captured in continuum models.²⁵

Here we use such approaches to investigate formaldehyde in water. The effects of the environment on the structural and electronic properties of formaldehyde are discussed in detail. Excitation energies are computed using not only TDDFT³⁰ but also complete active space with second-order perturbation theory (CASPT2).³¹ CASPT2 has been already widely used as a reference for high-level calculations, and it has been applied in the past to study solvent effects on carbonyl-bearing molecules.^{2,32–34} The calculations are run over a set of snapshots extracted from the simulations.

Computational Methods

The quantum problem was solved within the framework of DFT,^{35,36} using the BLYP gradient-corrected functional.^{37,38} This provides good geometrical features and has been shown to treat H-bonded systems with fairly good accuracy.^{39–42} The absorption spectrum of formaldehyde in the gas phase was computed also with the PBE functional,⁴³ which is known from the literature to reproduce absorption energies with good accuracy.⁴⁴

The wave functions were expanded on a plane wave basis set, up to a cutoff of 70 Ry. The atomic all-electrons potentials were substituted with norm-conserving Troullier–Martins pseudopotentials,⁴⁵ and the Kleinman–Bylander approach was used to treat their nonlocal part.⁴⁶ The simulations were run with dispersion-corrected pseudopotentials,⁴⁷ which are shown to reproduce quite well London forces.

The in vacuo geometry of formaldehyde (\mathbf{FA}_{opt} hereafter) was optimized using the GDIIS algorithm⁴⁸ up to a variation of the gradient on the nuclei positions of 10^{-6} au and on the wave function of 10^{-8} au.

DFT CPMD simulations for formaldehyde in vacuo ($\mathbf{FA}_{\text{CPMD}}$ hereafter) and in water ($\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$) as well as hybrid CPMD/MM simulations²⁴ for formaldehyde in water ($\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$) were carried out. In the latter scheme, the QM part was given by the formaldehyde molecule and the MM part by water. The second was described with the TIP3P model.⁴⁹

Formaldehyde was inserted in a box of edges $9.01 \text{ \AA} \times 7.39 \text{ \AA} \times 8.65 \text{ \AA}$ (\mathbf{FA}_{opt}), in a water box of edges $11.33 \text{ \AA} \times 11.22 \text{ \AA} \times 9.63 \text{ \AA}$ ($\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$), and in a box of $8.8 \text{ \AA} \times 9.3 \text{ \AA} \times 8.8 \text{ \AA}$ ($\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$). In $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$, the box contained 43 water molecules (density = 1.09 g/cm^3). In $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$, the

QM simulation box was further inserted in a $28.4 \text{ \AA} \times 26.1 \text{ \AA} \times 28.9 \text{ \AA}$ orthorhombic box of 697 water molecules (density = 1.03 g/cm^3). Both $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ were first equilibrated by classical MD (1 ns classical MD was performed in the NPT ensemble ($T = 300 \text{ K}$, $P = 1 \text{ atm}$), obtained using a Langevin-like thermostat and barostat,⁵⁰ using a time step of 1 fs and periodic boundary conditions. Interactions were described with the parm99 force field⁵¹ and the TIP3P model of water molecules.⁴⁹ The lengths of bonds involving a hydrogen were constrained with the Shake algorithm.⁵² The PME algorithm was adopted to treat electrostatic interactions.⁵³ A cutoff of 14 \AA was set for van der Waals interactions and the real part of the Coulomb interactions. The MD was run with the software NAMD.⁵⁴)

In \mathbf{FA}_{opt} , $\mathbf{FA}_{\text{CPMD}}$, and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$, isolated system conditions within the plane-wave formalism were achieved with the scheme of Martyna and Tuckerman.⁵⁵

$\mathbf{FA}_{\text{CPMD}}$, $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$, and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ underwent 3, 3, and 12 ps of MD, respectively. These simulations were performed in the NVT ensemble at 300 K. Constant temperature conditions were applied through a Nosé–Hoover chain of thermostats.⁵⁶ $\mathbf{FA}_{\text{CPMD}}$ also underwent geometry optimization prior to the dynamics (same procedure as for \mathbf{FA}_{opt}).

Calculated Properties. (1) The dipole moment of \mathbf{FA}_{opt} was computed from real-space integration in the isolated cell.⁵⁷ The vibrational normal modes of the molecule were calculated from the diagonalization of the Hessian matrix. This was calculated using the finite difference technique with a step size of 0.01 bohr and a convergence criterion of 10^{-8} au over the wave function gradient.

The dipole moment of formaldehyde in $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ was computed from the positions of the nuclei and those of the Wannier centers of the molecule,^{58,59} calculated every 50 steps of the dynamics. The vibrational spectrum of formaldehyde was obtained from the Fourier transform of the autocorrelation of the velocities of its atoms,⁶⁰ also in this case calculated every 50 steps of the dynamics. Normal mode vibrations were derived projecting this spectrum on the normal modes of the isolated molecule.⁶¹

(2) The radial distribution functions (rdf) of water oxygens and hydrogens around formaldehyde's oxygen in $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ were calculated up to one-half the smallest dimension of the simulation box, discretized in steps of 0.05 \AA , and taking periodic boundary conditions into account.

(3) The first electronic excitations were obtained by TDDFT⁶² and CASPT2³¹ calculations. They were calculated on \mathbf{FA}_{opt} and on 100 equispaced snapshots from $\mathbf{FA}_{\text{CPMD}}$, $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$, and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ simulations. Casida's formulation of the linear-response TDDFT⁶³ was used to perform the direct calculation of the electronic excitations on the molecule. The Tamm–Dancoff approximation was applied in this case.³⁰

CASPT2 calculations exploited a zeroth-order CASSCF wave function with no IPEA shift. Here, the complete active space included 4 electrons and 3 orbitals (n , π , and π^*). This active space was chosen after careful tests and turned out to be the best compromise between accuracy and reliability. The wave function was expanded over the ANO-S basis set⁶⁴ with contraction 7s6p3d for carbon and oxygen and 4s3p for hydrogen.

Table 1

	\mathbf{FA}_{opt} (BLYP)	\mathbf{FA}_{opt} (PBE)	exp.	$\mathbf{FA}_{\text{CPMD}}$	$\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$	$\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$
$d(\text{C}=\text{O})$ (Å)	1.22	1.22	1.20 ^a	1.23 ± 0.01	1.23 ± 0.01	1.22 ± 0.01
$d(\text{C}-\text{H})$ (Å)	1.12	1.12	1.10 ^a	1.12 ± 0.01	1.11 ± 0.01	1.11 ± 0.01
$\angle(\text{HCH})$ (deg)	116	116	116 ^a	116 ± 1	117 ± 2	117 ± 2
μ (D)	2.30	2.25	2.34 ^b	2.86 ± 0.16	3.72 ± 0.29	3.45 ± 0.25

^a Value from ref 70. ^b Value from ref 72.

CASPT2 and TDDFT calculations of the solute in the presence of the solvent field were calculated for $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ (and also $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ for the CASPT2 calculations) following refs 24, 30, and 65.

All DFT and TDDFT calculations were performed with the BLYP functional.^{37,38} The calculations at 0 K were also performed at the PBE level.⁴³ All DFT and TDDFT calculations were performed with the CPMD code.⁵⁷ The code was interfaced with the Gromos classical MD program⁶⁶ in $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$. The CASSCF/CASPT2 calculations were performed with the Molcas program.⁶⁷ Here, the solvent was introduced as a perturbation of the CASPT2 wave function by the point charges of the water molecules solvating formaldehyde. No charge screening was applied. The classical van der Waals contributions were not included.

Results and Discussion

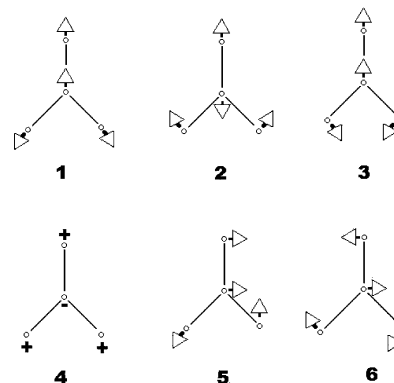
Ground State. The geometrical features of formaldehyde in vacuo calculated using the BLYP^{37,38} and the PBE⁴³ functionals are the same (Table 1). They are similar to those obtained by DFT-B3LYP,^{68,69} MRCI,¹⁸ CASSCF, and CCSD calculations^{10,14} (Table S4 in the Supporting Information). The bond lengths are slightly longer than those observed based on microwave spectra,^{70,71} whereas the bond angle is the same (Table 1). Formaldehyde is a planar molecule. Its symmetry elements are those of the C_{2v} point symmetry group: one C_2 axis (passing through the carbon and oxygen atoms and bisecting the HCH angle) and two σ_v planes (the molecular plane and the plane perpendicular to it, bisecting the HCH angle). Here, we report formaldehyde's structural determinants and dipole moment from our simulations and from the literature. Our calculations are at the DFT/BLYP level, unless otherwise specified.

The calculated dipole moment $\bar{\mu}$ (Table 1) is aligned with the O=C bond and has a modulus $\mu = 2.25\text{--}2.3$ D, in accordance with DFT-B3LYP^{6,7} and PBE0,⁵ CASSCF,^{10,12} MRCI,¹⁸ as well as CCSD calculations¹¹ (Table S5 in the Supporting Information). It is also in agreement with the value obtained by microwave spectroscopy.⁷²

The frequencies of the calculated normal modes (Table 2) agree within 50–100 cm^{-1} with B3LYP calculations,^{68,69} whereas they are rather far from the CASSCF results of ref 10 (shifts of 50–500 cm^{-1}). However, it is known from the literature⁷³ that CAS-calculated vibrational frequencies are overestimated, due to the lack of electronic correlation, and that a scaling factor of 0.9 should be applied. Our calculations agree within 70 cm^{-1} with experimental data⁷¹ (Table 2).

We next analyze solvation and temperature effects on formaldehyde by performing DFT CPMD ($\mathbf{FA}_{\text{CPMD}}$) and hybrid CPMD/MM simulations²⁴ of formaldehyde in water ($\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$) at 300 K.

Table 2. Frequencies of the Six Normal Modes of Formaldehyde^{68,74} (cm^{-1}) Calculated at the BLYP Level of Theory^a



C_{2v} -symmetric species	A ₁		B ₁		B ₂	
	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6
exp.	2766	1746	1500	1167	2843	1251
\mathbf{FA}_{opt}	2803	1736	1472	1144	2839	1174
$\Delta\nu \mathbf{FA}_{\text{CPMD}}$	-309	-104	-150	-75	-368	-117
$\Delta\nu \mathbf{FA}_{\text{CPMD}}^{\text{aq}}$	-100	-149	-98	-64	-85	-79
$\Delta\nu \mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$	-180	-108	-61	-14	-188	12

^a Comparison is made with experimental data for formaldehyde in the gas phase.⁷⁵ Mode 1: symmetric C–H stretching. Mode 2: C=O stretching. Mode 3: CH₂ bending. Mode 4: CH₂ out of plane bending. Mode 5: antisymmetric C–H stretching. Mode 6: CH₂ rocking.

Both temperature and solvation slightly affect the structural parameters. In both DFT CPMD and hybrid CPMD/MM simulations of formaldehyde in water ($\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$, respectively), the average C=O bond is elongated and the C–H bonds are shortened by about 0.01 Å, which falls within our statistical uncertainty (see Table 1), and is therefore insignificant. The average value of μ is instead remarkably larger than that of \mathbf{FA}_{opt} (more than 50%, see Table 1). Since the extent of the temperature and solvation effects on the structure is negligible, we conclude that in our approach the solvent affects the dipole moment intensity of formaldehyde by polarizing its electronic wave function, rather than by modifying its structure. This finding is in line with B3LYP/MM calculations on large clusters of water molecules around formaldehyde, which suggest that μ increases by 42% on passing from the gas phase to water solution⁷ (Table S5 in the Supporting Information). They are also consistent with CPMD/MM calculations of acetone in water, which predict the value of μ to increase by about 60% upon going from the gas phase to aqueous solution.²⁴

In both $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$, the radial distribution function (rdf) between the carbonyl oxygen atom and water hydrogens exhibits a first peak around 1.9–2.0 Å and a minimum around 2.5 Å (Figure S2 and Table S6 in the

Table 3. Absorption Energies (in eV) of Formaldehyde in Vacuo^a

	\mathbf{FA}_{opt}	$\mathbf{FA}_{\text{CPMD}}$
TDDFT (this work)	3.85, 3.82, ^b 3.89 ^c	3.77 ± 0.04
CASPT2 (this work)	3.84, 3.86 ^c	3.76 ± 0.05
PBE0(6-311G(d,p))// TD-PBE0(6-311++G(d,p)) ⁵	3.84	
B3LYP(cc-pVTZ)//LR-B3LYP, LR-CCSD(6-31++G) ⁶	3.81, 3.84	
exp.//CASSCF(ANO) ¹²	4.04	
MP2(cc-pvtz)// MRCI(cc-pvtz) ¹⁸	4.04	
experimental values	3.79, ²¹ 3.86 ²²	

^a Values are reported for the TDDFT and CASPT2 calculations performed in this work for formaldehyde at 0 and 300 K. The reported TDDFT values are referred to calculations at the BLYP level, unless otherwise specified. Experimental as well as some previous theoretical results are reported in the last 5 lines of the tables. Some details of the levels of theory of the cited calculations are indicated (ground-state geometry//absorption energy). See the Supporting Information for further details. ^b Calculation at the PBE level. ^c Calculation performed over the CASSCF-optimized structure of formaldehyde.

Table 4. Absorption Energies (in eV) of Formaldehyde in Water Solution^a

	$\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$		$\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$	
	FF	0	FF	0
TDDFT		3.77 ± 0.04	4.0 ± 0.1	3.77 ± 0.07
CASPT2	4.08 ± 0.12	3.77 ± 0.05	4.07 ± 0.11	3.77 ± 0.09

^a Columns “FF” (columns “0”) indicate single-point calculations with water molecules taken explicitly into account (with water molecules removed). The reported TDDFT values are referred to calculations at the BLYP level.

Supporting Information). Integration of this peak up to the minimum yields a coordination number of about 1.9. Rather similar results are obtained from other QM and QM/MM techniques:^{4,9,10,15,16} ref 4 predicts the peak of the oxygen–hydrogen rdf at 1.8 Å (integrating to 1.4 hydrogens), ref 15 at 1.7 Å (integrating to 1.9 hydrogens), and ref 16 at 1.9 Å. The oxygen–oxygen rdf (Table S6 in the Supporting Information) shows a first peak around 2.9–3.0 Å; integration up to the first minimum of the distribution gives a value of 2.5 and 2.9 water molecules, for the QM and QM/MM simulation. Reference 16 reports an oxygen–oxygen rdf with a first peak at 2.8 Å. Its integration yields a value of 2.2 water molecules, fairly similar to our result. Notably, ref 18 obtains the rdf peaks at a slightly smaller distance (1.75 Å for hydrogen and 2.75 Å for oxygen), resulting in about one water molecule H bonded to formaldehyde; ref 17, instead, reports the peaks at 2.11 and 2.9 Å, respectively, again with one water molecule in the first solvation shell.

The specific influence of temperature effects on formaldehyde’s structure and dipole is here investigated by comparing structural and electronic properties of \mathbf{FA}_{opt} with those of a CPMD simulation of formaldehyde in vacuo at 300 K ($\mathbf{FA}_{\text{CPMD}}$). The change in the CPMD-averaged C=O and C–H bond lengths with respect to those of \mathbf{FA}_{opt} is null or negligible. The average dipole moment is larger by about 24% (Table 1), showing a much smaller change than in $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ or $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$.

Temperature effects therefore do not cause a significant distortion to the structure of the molecule. The small number

of degrees of freedom of the system in $\mathbf{FA}_{\text{CPMD}}$ does not allow for a proper dissipation of the thermal energy provided by the coupling to the thermostat. This provokes rather large fluctuations in the kinetic energy of formaldehyde (Figure S4 in the Supporting Information). As a consequence, the structure of the latter is particularly flexible.

We finally turn our attention to the vibrational spectrum. All vibrational frequencies, with the exception of the CH₂ rocking in $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$, are red shifted on going from \mathbf{FA}_{opt} to $\mathbf{FA}_{\text{CPMD}}$, $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$, and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ (Table 2, Figure S3 in the Supporting Information). These comparisons should be taken with caution, since the vibrational frequencies for \mathbf{FA}_{opt} are calculated in a different way from those of $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$ (see the Computational Methods section). In solution, the shift is more significant for the stretching modes (85–188 cm⁻¹) than for the bending ones (14–98 cm⁻¹). This contrasts with RIS-(CAS)SCF calculations,¹⁰ which predict that the frequencies are shifted toward the blue, that is, the molecule becomes more rigid in the solvent. Early ab initio calculations of geometry-optimized complexes of H₂CO···HO⁷⁶ suggest that solvation effects are qualitatively similar to our results, though the red shifts in this case are much smaller (within 40 cm⁻¹). The effects of temperature can be clearly assessed also in the vibrational spectra: the shifts in $\mathbf{FA}_{\text{CPMD}}$ are even larger than those of $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$. However, this might be due at least in part to scarce thermal energy dissipation.

In conclusion, the most important contribution of the solvent effects to structural, vibrational, and electronic properties is the polarization of the wave function, rather than the geometric distortion of formaldehyde. This is reflected by the remarkable change in the intensity of formaldehyde’s dipole moment occurring between the gas phase and the solution case. As for temperature effects, these partially affect the vibrational spectrum of formaldehyde. The structure of the latter is more flexible than in the gas phase.

First Electronic Excitation Energy. TDDFT and CASPT2 approaches have been used to compute the electronic transition energy between the ground (S0) and the first excited state (S1). The latter is characterized by a single occupation of two orbitals. The *n* molecular orbital (HOMO) is localized on the oxygen and the π^* molecular orbital (LUMO) along the C=O double bond (Figure S1 in the Supporting Information).

The absorption energy values for the *n* → π^* transition (ΔE) of \mathbf{FA}_{opt} , calculated with TDDFT and CASPT2 methods, are similar (ranging from 3.82 to 3.85 eV). They do not depend on the functional used (BLYP or PBE). Thus, only BLYP calculations are reported here. The results are collected in Table 3, where we also report the values of the excitation energies obtained on the formaldehyde geometry optimized at the CASSCF level, showing a good agreement (within 0.04 eV) with the values obtained for \mathbf{FA}_{opt} . Our calculated values of ΔE at 0 K agree with other DFT,^{5,6} coupled cluster,⁶ CISD, and CASSCF¹³ as well as with the experimental values of refs 21 and 22 (Table 3). The latter are measured from setups which keep formaldehyde at rather high temperatures (around 330 K). However, this is expected

Table 5. Literature Values of the Temperature- and Water-Induced Blue Shift of ΔE (in eV)^a

full QM approaches	QM/MM approaches	ΔE
(IVC)-PCM//PCM opt PBE0(6-311++G(d,p))// TD-PBE0(6-311++G(d,p)) ⁵		0.07–0.12
IEF-PCM//gpo CASSCF, MR-CI(6-31G**)// CASSCF, MR-CI(6-31G**) ¹⁹	QM/MMpol//gpo AM1//CI ⁴	0.14
FMO//FMO-HF(6-31G)//CIS(D)(6-31G*) ¹⁶	RISM-SCF//gpo RHF(DZP)//CASSCF(DZP) ⁹	0.15–0.2
	ASEP//gas phase exp//CASSCF(ANO) ¹²	
	QM/classical//gpo MP2(ANO)// CASSI(ANO) ¹⁷	
	ASEP//gpo MP2(cc-pvtz)//MRCI(cc-pvtz) ¹⁸ effective potential//cluster opt	
	HF(DZP)//MRD-CI(DZP) ⁸¹	
	RISM-SCF//RISM-SCF opt CASSCF(DZP)// CASSCF(DZP) ¹⁰	0.21–0.25
	QM/MM//gpo B3LYP(cc-pVTZ)// LR-CCSD(aug-cc-pVTZ) ¹¹	
	ASEP//gpo RHF(6-31G+d)//CASPT2(6-311+G/G**) ¹³	
QM//gpo MP2(aug-cc-pVDZ)//TDDFT(6-311++G(d,p)) ⁷	QM/MM//CCSD/MM(6-31++G), B3LYP/MM(6-31++G)// LR-CCSD(6-31++G), LR-B3LYP(6-31++G) ⁶	0.25–0.35
QM//INDO//CIS ¹⁵	QM/MM//gpo RHF(6-31G+d)//RHF(6-31G+d) ⁸	
	QM/MM-pol-vib/CAV//gpo CCSD(TZ+2P)// CASSCF(cc-pVTZ) ¹⁴	
	QM/MM//gpo RHF(6-31G+d)//CIS(6-31G**) and dipolar coupling//gas phase opt RHF(6-31G+d)//CIS(6-31G**) ²⁰	
dielectric continuum PBF//gas phase opt RHF(6-31G+d)// CIS(6-31G**) and dipolar coupling//gas phase opt RHF(6-31G+d)//CIS(6-31G**) ²⁰		0.43

^a Details of the cited schemes used to incorporate the water environment in the calculations//theoretical methodologies and basis sets used for the calculation of solute geometries//and of the absorption energies. opt, geometry optimization of the system; gpo, gas-phase optimization: formaldehyde geometry was constrained to the gas-phase-optimized one.

not to have a significant effect on the absorption of formaldehyde in the UV range.

It was not possible to obtain the value of ΔE for $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ with TDDFT, because of charge transfer between the solute and the solvent.^{77,78} Indeed, all excitations we could compute with TDDFT on the $\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ setup were localized on water molecules and not on formaldehyde.

The ΔE values in solution as calculated with TDDFT, and, more, with CASPT2, are larger than those of \mathbf{FA}_{opt} (Table 4). The H-bond interactions between water and the solute oxygen in fact decrease the energy of the n state relative to \mathbf{FA}_{opt} , increasing ΔE .⁷⁹ The predicted values are close to the experimental data²³ and a variety of calculations (Table 5). These are in the 3.5–5 eV range and predict a blue shift within 0.07–0.43 eV.

To disentangle temperature and wave function polarization contributions of the ground and excited state, we calculated $n \rightarrow \pi^*$ excitation energies in vacuo also on the geometries of formaldehyde extracted from the MD simulation in water solution ($\mathbf{FA}_{\text{CPMD}}^{\text{aq}}$ and $\mathbf{FA}_{\text{QM/MM}}^{\text{aq}}$). Notably, the results are identical to those calculated from $\mathbf{FA}_{\text{CPMD}}$ (compare the results reported in Tables 3 and 4). Hence, we attribute the whole blue shift to polarization of the wave functions of the two states.

An analogous consideration was raised in ref 80 for a similar molecule (acrolein) in water solution. However, in that case the distortion of geometry was nevertheless pointed out as a major factor of changes in the absorption spectra.

We do not find significant differences in the excitation energies obtained with CASPT2 over configurations obtained with full QM water and those from the mixed QM/MM

approach. This is justified by the small differences that we could detect in the structural features (geometry and dipole moment of formaldehyde, rdf of water around the carbonyl) of the systems described at the full QM and at the hybrid one.

Since geometry distortion plays no significant role in modifying the absorption energy of formaldehyde, the use of an explicit or implicit solvent model should bring the same result in terms of ΔE . It is worth noticing that implicit solvent approaches, such as those of refs 5 and 19, predict rather small solvent blue shifts, with respect to those obtained from QM/MM techniques such as the LR-CCSD/MM of ref 11, the RISM-SCF approach of ref 10, and the CASPT2-ASEP of ref 13 as well as with respect to full QM calculations (the CIS of ref 15, and TDDFT calculations of ref 7). Adding polarizability to such schemes may result in enlarging the solvent shift values, as shown in ref 20. Nevertheless, molecular solvent schemes tend to produce higher shifts.²⁰

In conclusion, we presented a Car–Parrinello QM and QM/MM MD study with excited-state calculations.

The geometry of formaldehyde, its normal modes, and its dipole moment in the gas phase agree with experiments^{70,71,75} as well as with the past literature. The effects of solvent on these structural features, as known from previous calculations,^{10,14,18,68,69} are also recovered with the simulations performed here.

The excitation energies calculated with both the TDDFT and the CASPT2 approaches are in good agreement with experimental and other theoretical results.^{9–13,17,18,81} In particular, they predict a blue shift in the $n \rightarrow \pi^*$ absorption energy that is here assigned to wave function polarization effects by the solvent.

Supporting Information Available: Brief summary of other studies about the absorption spectra of formaldehyde, along with further details about the analysis performed on the calculations presented in the main text of this article. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Guilbault, G. G. *Practical Fluorescence*, 2nd ed.; Marcel Dekker, Inc.: New York, 1990.
- Öhrn, A.; Karlström, G. *J. Phys. Chem. A* **2006**, *110*, 1934–1942.
- Improta, R.; Scalmani, G.; Frisch, M. J.; Barone, V. *J. Chem. Phys.* **2007**, *127*, 074504–074513.
- Thompson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492–14507.
- Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708–4718.
- Nielsen, C. B.; Christiansen, O.; Mikkelsen, K. V.; Kongsted, J. *J. Chem. Phys.* **2007**, *126*, 154112–154130.
- Malaspina, T.; Coutinho, K.; Canuto, S. *J. Braz. Chem. Soc.* **2008**, *19*, 305–311.
- Blair, J. T.; Krogh-Jespersen, K.; Levy, R. M. *J. Am. Chem. Soc.* **1989**, *111*, 6948–6956.
- Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443–7454.
- Naka, K.; Morita, A.; Kato, S. *J. Chem. Phys.* **1999**, *110*, 3484–3493.
- Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Astrand, P.-O.; Christiansen, O. *J. Chem. Phys.* **2004**, *121*, 8435–8446.
- Martín, M. E.; Sánchez, M. L.; del Valle, F. J. O.; Aguilar, M. A. *J. Chem. Phys.* **2000**, *113*, 6308–6315.
- Sánchez, M. L.; Martín, M. E.; Aguilar, M. A.; Olivares del Valle, F. J. *Chem. Phys. Lett.* **1999**, *310*, 195–200.
- Kawashima, Y.; Dupuis, M.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 248–258.
- Coutinho, K.; Saavedra, N.; Serrano, A.; Canuto, S. *J. Mol. Struct.: THEOCHEM* **2001**, *539*, 171–179.
- Mochizuki, Y.; Komeiji, Y.; Ishikawa, T.; Nakano, T.; Yamataka, H. *Chem. Phys. Lett.* **2007**, *437*, 66–72.
- Ohrn, A.; Karlstrom, G. *Mol. Phys.* **2006**, *104*, 3087–3099.
- Xu, Z.; Matsika, S. *J. Phys. Chem. A* **2006**, *110*, 12035–12043.
- Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1998**, *109*, 2798–2808.
- Bader, J. S.; Cortis, C. M.; Berne, B. J. *J. Chem. Phys.* **1997**, *106*, 2372–2388.
- Walzl, K. N.; Koerting, C. F.; Kuppermann, A. *J. Chem. Phys.* **1987**, *87*, 3796–3803.
- Taylor, S.; Wilden, D. G.; Comer, J. *J. Chem. Phys.* **1982**, *70*, 291–298.
- Bercovici, T.; King, J.; Becker, R. *J. Chem. Phys.* **1972**, *56*, 3956–3963.
- Röhrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Röthlisberger, U. *ChemPhysChem* **2003**, *4*, 1177–1182.
- Conte, A. M.; Ippoliti, E.; Del Sole, R.; Carloni, P.; Pulci, O. *J. Chem. Theory Comput.* **2009**, *5*, 1822–1828.
- Sulpizi, M.; Carloni, P.; Hutter, J.; Röthlisberger, U. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4798–4805.
- Murugan, N. A.; Rinkevicius, Z.; Ågren, H. *J. Phys. Chem. A* **2009**, *113*, 4833–4839.
- Spezia, R.; Duvail, M.; Vitorge, P.; Cartailier, T.; Tortajada, J.; Chillemi, G.; D'Angelo, P.; Gaigeot, M. P. *J. Phys. Chem. A* **2006**, *110*, 13081–13088.
- Tilocca, A.; Fois, E. *J. Phys. Chem. C* **2009**, *113*, 8683–8687.
- Hutter, J. *J. Chem. Phys.* **2003**, *118*, 3928–3935.
- Andersson, K.; Malmqvist, P. Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- Liao, D.-W.; Mebel, A. M.; Chen, Y.-T.; Lin, S.-H. *J. Phys. Chem. A* **1997**, *101*, 9925–9934.
- Muñoz Losa, A.; Fdez.-Galván, I.; Aguilar, M. A.; Martín, M. E. *J. Phys. Chem. B* **2007**, *111*, 9864–9870.
- Serrano-Andrés, L.; Fülischer, M. P.; Karlström, G. *Int. J. Quantum Chem.* **1997**, *65*, 167–181.
- Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.
- Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- Becke, A. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B* **1988**, *37*, 785–789.
- Piana, S.; Carloni, P. *Proteins* **2000**, *39*, 26–36.
- Bucher, D.; Guidoni, L.; Carloni, P.; Röthlisberger, U. *Biophys. J.* **2010**, *98*, L47–L49.
- Alves, C. N.; Martí, S.; Castillo, R.; Andrés, J.; Moliner, V.; Tuóúin, I.; Silla, E. *Bioorg. Med. Chem.* **2007**, *15*, 3818–3824.
- Bucher, D.; Raugei, S.; Guidoni, L.; Dal Peraro, M.; Röthlisberger, U.; Carloni, P.; Klein, M. L. *Biophys. Chem.* **2006**, *124*, 292–301.
- Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5037.
- Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425–1428.
- von Lilienfeld, O.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004–153008.
- Császár, P.; Pulay, P. *J. Mol. Struct.* **1984**, *114*, 31–34.
- Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–936.
- Quigley, D.; Probert, M. *J. Chem. Phys.* **2004**, *120*, 11432–11442.
- Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327–341.
- Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577–8594.
- Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

- (55) Martyna, G.; Tuckerman, M. *J. Chem. Phys.* **1999**, *110*, 2810–2922.
- (56) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (57) CPMD, version 3.13.2; IBM Corp. 1990–2008, MPI für Festkörperforschung Stuttgart 1997–2001, 2009; <http://www.cpmid.org/>.
- (58) Silvestrelli, P. L.; Marzari, N.; Vanderbilt, D.; Parrinello, M. *Solid State Commun.* **1998**, *107*, 7–11.
- (59) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12865.
- (60) Wilson, E.; Decius, J.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover Publications Inc.: Mineola, NY, 1980.
- (61) Carloni, P.; Sprik, M.; Andreoni, W. *J. Phys. Chem. B* **2000**, *104*, 823–835.
- (62) Runge, E.; Gross, E. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.
- (63) Casida, M. E. Time-Dependent Density Functional Response Theory for Molecules. In *Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, Chapter 5, pp 155–192.
- (64) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.
- (65) Altoé, P.; Stenta, M.; Bottoni, A.; Garavelli, M. *Theor. Chim. Acta* **2007**, *118*, 219–240.
- (66) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kroger, 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 Zurich: Zurich, 1996.
- (67) Karlstrom, G.; Lindh, R.; Malmqvist, P.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.
- (68) Jalbouta, A. F.; El-Nahas, A. M. *J. Mol. Struct.: THEOCHEM* **2004**, *671*, 125–132.
- (69) George, W.; Jones, B.; Lewisa, R.; Pricea, J. *J. Mol. Struct.* **2000**, *550–551*, 281–296.
- (70) Takagi, K.; Oka, T. *J. Phys. Soc. Jpn.* **1963**, *18*, 1174–1180.
- (71) Moule, D. C.; Walsh, A. D. *Chem. Rev.* **1975**, *75*, 67–84.
- (72) Schoolery, J. N.; Sharbaugh, A. H. *Phys. Rev.* **1951**, *82*, 95–95.
- (73) Garavelli, M.; Negri, F.; Olivucci, M. *J. Am. Chem. Soc.* **1999**, *121*, 1023–1029.
- (74) Bernal, R.; Lemus, R. *J. Mol. Spectrosc.* **2006**, *235*, 218–234.
- (75) Job, V. A.; Sethuraman, V.; Innes, K. K. *J. Mol. Spectrosc.* **1969**, *30*, 365–426.
- (76) Dimitrova, Y. *J. Mol. Struct.: THEOCHEM* **1997**, *391*, 251–257.
- (77) Röthlisberger, U.; Carloni, P. Drug-Target Binding Investigated by Quantum Mechanical/Molecular Mechanical (QM/MM) Methods. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*; Ferrario, M., Ciccotti, G., Binder, K., Eds.; Springer Berlin/Heidelberg: Berlin, Heidelberg, 2006; Vols. 2 and 704, Chapter 17, pp 449–479.
- (78) Dal Peraro, M.; Raugei, S.; Carloni, P.; Klein, M. L. *ChemPhysChem* **2005**, *6*, 1715–1718.
- (79) Bayliss, N. S.; McRae, E. G. *J. Phys. Chem.* **1954**, *58*, 1002–1006.
- (80) Martín, M. E.; Losa, A. M.; Fdez.-Galván, I.; Aguilar, M. A. *J. Chem. Phys.* **2004**, *121*, 3710–3717.
- (81) Frank, I.; Grimme, S.; von Arnim, M.; Peyerimhoff, S. D. *Chem. Phys.* **1995**, *199*, 145–153.

CT100384F

Position Isomerism on One and Two Photon Absorption in Multibranching Chromophores: A TDDFT Investigation

Claudine Katan,^{*,†,‡} Mireille Blanchard-Desce,[†] and Sergei Tretiak^{*,§}

Université Européenne de Bretagne, CNRS—Chimie et Photonique Moléculaires (CPM), Université de Rennes 1, 35042 Rennes, France, and Université Européenne de Bretagne, CNRS—Fonctions Optiques pour les Technologies de l'Information (FOTON), INSA de Rennes, CS70839, 35708 Rennes, France, and Center for NonLinear Studies (CNLS) and Center for Integrated NanoTechnologies (CINT), Los Alamos, New Mexico 87545, United States

Received August 7, 2010

Abstract: Recently, branching and click chemistry strategies have been combined to design a series of optically active chromophores built from triazole moieties. These triazole-based multipolar chromophores have been shown to be promising candidates for two-photon absorption (TPA) transparency optimization in perspective of optical limiting in the visible region. In this work, the nature of one- and two-photon absorption properties in a family of triazole-based chromophores has been investigated using hybrid time-dependent density functional theory (TD-DFT). We use recent extensions of TD-DFT to determine nonlinear optical responses and natural transition orbitals to analyze the underlying electronic processes. Our results are also interpreted in the framework of the Frenkel exciton model. In agreement with experimental data, we found that introducing a triazole moiety into multibranching chromophores substantially modifies their optical behavior due to changes in electronic delocalization and charge-transfer properties between donating end groups and the branching center that can be controlled by the triazole ring. Structural conformations via modulation of the torsion between phenyl and triazole rings significantly alter the excited state electronic structure. Moreover, isomer positioning also greatly influences both linear and nonlinear optical responses such as TPA. Our theoretical findings allow elucidation of these differences and contribute to the general understanding of structure–property relations. Consequently, the interplay of donor/acceptor strength, triazole regioisomerism, and branching are shown to provide flexible means allowing for precise tuning of both linear and nonlinear optical responses, thus opening new perspectives toward synergic TPA architectures.

I. Introduction

1,2,3-Triazole chemistry has recently gained renewed interest thanks to the click methodology offering far more ease of synthesis.^{1,2} This has prompted synthetic effort to design novel molecular architectures in which the triazole moiety is assumed to play quite different roles, such as electron

donor,³ electron acceptor,^{2,4,5} active π -linker,^{6,7} metal ligating entity,^{8,9} or site-isolation group.^{10–12} In particular, 1,2,3-triazole click chemistry has been explored in the field of small push–pull^{3,6,9,13} and larger multipolar^{4,13} chromophores up to dendrimers¹⁰ and polymers with linear⁵ or hyperbranched^{7,11,12} polytriazole structures. In addition, 1,2,3-triazole has four points of attachment that allow for the design of various regioisomers which may show quite different behaviors.

This recent synthetic effort was partly driven by interest in their photophysical properties, which have long been overlooked and are still underdeveloped.⁶ Recently, on the

* Corresponding authors. E-mail: claudine.katan@insa-rennes.fr; serg@lanl.gov.

[†] CNRS Université de Rennes 1.

[‡] CNRS INSA de Rennes.

[§] CNLS and CINT.

basis of the assumption of electron-donating properties of triazole, ligation of this ring has been shown to effectively modulate the emission properties of coumarin, leading to increased fluorescence quantum yield by more than an order of magnitude.³ Investigation of the photophysical properties of isomeric donor–acceptor-substituted 1,2,3-triazoles has also been reported.⁶ It has been demonstrated both experimentally and theoretically that positioning of the donor and acceptor on the triazole ring results in important modulations of both transition energies and absorption intensities.⁶ In the case of 1,4- and 1,5-linked hyperbranched polytriazoles, the different emission colors of the regiostructures were related to the differences in aggregation in the solid state.⁷ On the basis of the site-isolation concept,¹⁴ different hyperbranched polytriazoles have shown enhancement of their second-order nonlinear optical (NLO) responses overcoming the asymptotic dependence of electro-optic activity on chromophore density.^{10–12,15} Finally, triazole-based multipolar⁴ and functional poly(aryltriazole)s⁵ have also been promising for their nonlinear optical properties, in particular, two photon absorption (TPA) capacity.

Indeed, among the numerous NLO processes, simultaneous absorption of two photons has gained increased attention over recent years^{16–21} as it opens the way for improved and novel technological capabilities. This, in turn, led to the search for adequate functional materials with enhanced TPA responses. Thanks to their inherent modularity and potentially high TPA responses, conjugated organic molecular systems are of particular interest. From that perspective, the branching strategy has received much attention over recent years.^{16–18} It relies on multibranch systems obtained by gathering several molecular units which are either of the dipolar^{21–31} (also known as push–pull chromophores) or quadrupolar^{32–38} type. In such nanoarchitectures, the interbranch coupling plays a central role. Depending on the nature and the strength of the interaction between the molecular units, optical properties such as TPA can show cooperative enhancement, additive behavior, or weakening. The interchromophore interactions are related to quite a few structural parameters: nature and length of the monomeric building blocks, number of chromophoric units, as well as nature of the branching center. All of these parameters will directly influence intermonomer distances and orientation and control the overall molecular symmetry.²⁷ Symmetry will, in turn, result in optically allowed and forbidden electronic transitions and drive oscillator strength redistribution among excited states (ESs) when going from the monomeric unit to the multi-chromophoric architecture. Besides the nature of the excitation, these parameters also influence the subsequent excitation energy transfer dynamics.³⁹

Recently, branching and click chemistry strategies have been combined to design a series of TPA-active chromophores built from triazole moieties.^{4,13,40,41} These triazole-based multipolar chromophores have been shown to be promising candidates for TPA transparency optimization from the perspective of optical limiting in the visible region. Obtained quadrupolar and three-branched chromophores have been shown to combine efficient synthesis, high solubility, intense absorption in the near UV region, full transparency,

high excited state lifetimes, and good TPA optical responses.^{4,13} Moreover, emission wavelengths can be tuned, and excited state lifetimes can be significantly increased by increasing the polarity of the solvent. Related derivatives such as four-branched systems or position isomers have also been designed and experimentally investigated.^{40,41} Experimental data reveal quite different behaviors as compared to the related quadrupolar, octupolar, or other multibranch compounds for both linear and NLO responses. As an example, when compared to similar three-branched systems also built by gathering branches through a triphenylamine central core, the triazole derivative deviates significantly. Indeed, no marked solvatochromic behavior of emission has been observed, while it is very effective for the quadrupolar analog^{13,40} for which symmetry breaking in the relaxed excited state is thus clearly evidenced.⁴² In addition, TPA cross sections of the three-branched chromophore still increase at 577 nm.⁴ Thus, no maximum is observed near twice the one-photon absorption (OPA) wavelength (694 nm), as is usually the case.

To get a better understanding of triazole-based chromophores and to provide a deep physical insight into ongoing phenomena, theoretical calculations can be quite useful. Among the many different approaches ranging from simple few-state models up to high-level quantum methods (QM), the Frenkel exciton model has been shown to provide a valuable qualitative tool^{25,26,38,43–47} with a few exceptions. Indeed, to be applicable, not only should the strength of the coupling (solely due to Coulomb interactions) between monomeric units be weak but the inherent nature (e.g., symmetry) of the building blocks should not be affected by the branching center.³⁷ Among the different QM methods available nowadays, adiabatic time-dependent density functional theory (TD-DFT) is currently the method of choice for calculating the ES of sizable molecular systems.^{48–55} Arbitrary frequency-dependent NLO polarizabilities can be subsequently derived on the basis of the quasi-particle formalism of the TD-Kohn–Sham equations.⁵⁶ This method has been applied to calculate OPA and TPA responses of various substituted chromophores.^{17,25,26,31,35,37,57–63} These studies have shown the good performances of the TD-DFT-based hybrid functionals for molecular NLO properties.

In this paper, we present a theoretical investigation, mainly based on the above mentioned density matrix formalism of TD-DFT for NLO responses,¹⁷ of two 1,2,3-triazole-based series compiling a total of 19 compounds (Figure 1). The case of **dai**, **dfi**, **2dai**, and **3dfi** has already been analyzed in detail elsewhere,³¹ and we recall here the main findings as well as further interpretation for the sake of comparison to other **nd** and **ndi** chromophores. Our purpose is to address two main points: (i) the branching effect, which appears to show a quite different behavior than that reported for analogous multibranch chromophores, and (ii) the influence of position isomerism on both OPA and TPA responses. To our knowledge, it is the first time that TPA properties have been investigated on four-member branched chromophores^{64,65} as well as in relation to regioisomerism. Our findings will also be analyzed in the framework of the Frenkel exciton model. After a brief description of computational details,

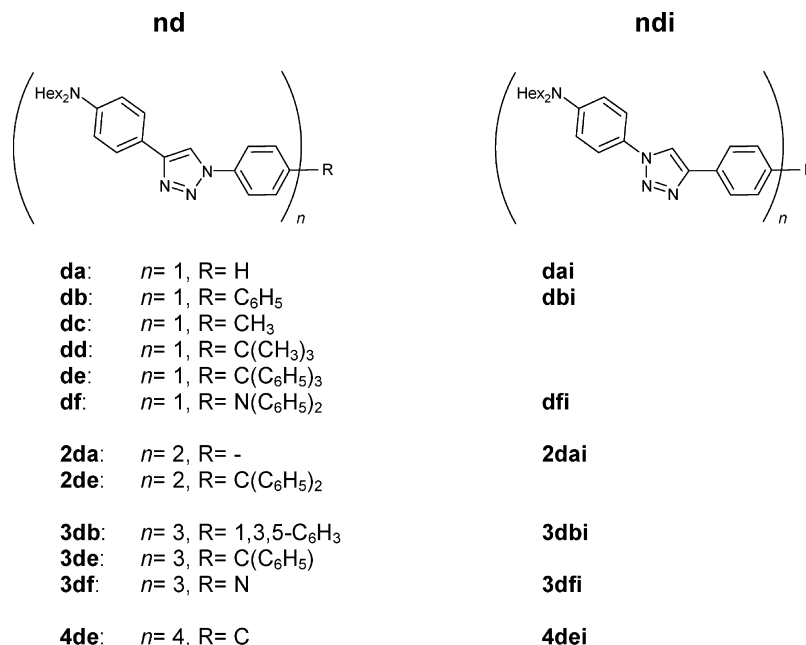


Figure 1. Molecular structures of 1,2,3-triazole-based **nd** chromophores (left) and corresponding **ndi** position isomers (right). Hexyl solubilizing chains have been replaced by methyl groups.

already well documented, we will first validate our computational approach for both series of chromophores. This is thought to allow inference with what extent a comparison to experiments is reasonable, i.e., underline all sources of possible discrepancies. After analysis of the main geometrical features, we will sequentially investigate OPA and TPA responses. For both properties, we first analyze trends within each of the two series and subsequently discuss effects of position isomerism. Where appropriate, a comparison to related branched structures will be made. The main findings will be summarized in the concluding section. We believe that this detailed theoretical study performed on a large set of compounds contributes to the overall understanding of structure–properties relationships of NLO chromophores and opens new perspectives for further improvements within the branching strategy.

II. Computational Methodology

A set of 19 chromophores, sketched in Figure 1, have been theoretically investigated. They are classified in two different series, series **nd** and corresponding 1,2,3-triazole position isomers **ndi**. For the sake of simplicity and reduction of the numerical cost, the hexyl solubilizing chains of original chromophores^{4,13,40,41} have been replaced by methyl groups. These two series contain nine dipoles, three quadrupoles, five three-branched chromophores, and two four-branched compounds. All QM calculations have been performed using either Gaussian 98⁶⁶ or Gaussian 03 packages⁶⁷ in vacuum. For ground state geometries (Figures 2 and S1 and S2, Supporting Information), it was previously found that the Hartree–Fock (HF) method is superior to the DFT-based approaches by reproducing bond length alternation in similar conjugated molecules.⁵⁷ It was also shown to be a good compromise for subsequent calculation of optical spectra.^{17,25,26,37,57–60,63} Thus, all ground state geometries have

been optimized at the HF/6-31G level of theory, starting from a few different conformations, but no exhaustive sampling has been conducted.

Linear and nonlinear optical responses have been calculated using the density matrix formalism, as described in ref 56. Subsequent computation of TPA cross-sections is detailed in ref 17. We retained the TD-B3LYP/6-31G//HF/6-31G level of theory, in conventional quantum chemical notation “single point//optimization level”, for the calculation of optical spectra, as it was shown to be a good compromise between numerical resources and accuracy and led to quite good agreement with experimental results for a large set of related chromophores. The forthcoming section will be devoted to discussing the validity of such an approach for the present triazole-based chromophores and include a comparison to results obtained with a larger amount of orbital exchange, namely, the hybrid BHandH functional instead of B3LYP. The calculated ES structures include 20, 30, 40, and 60 ESs respectively for dipolar, quadrupolar, three-branched, and four-branched compounds. The number of ESs was checked for at least one dipole, quadrupole, and octupole. It was found to be sufficient as asymptotic values are reached for the absorption spectra, and only small effects (the largest deviation observed is 12% for the first band of **3dbi** when using 24 instead of 40 ESs) are discerned for the amplitude of the TPA spectra.

Besides the number of ESs, a few other parameters influence the calculated response amplitudes. First, the damping factor introduced to simulate finite line width in the resonant spectra has been fixed to 0.20 eV for all chromophores, on the basis of typical broadening seen in available experimental absorption spectra (Table S1, Supporting Information). Next, no solvation effects have been accounted for in the present work. In fact, due to different solute and solvent polarizabilities, the commonly used

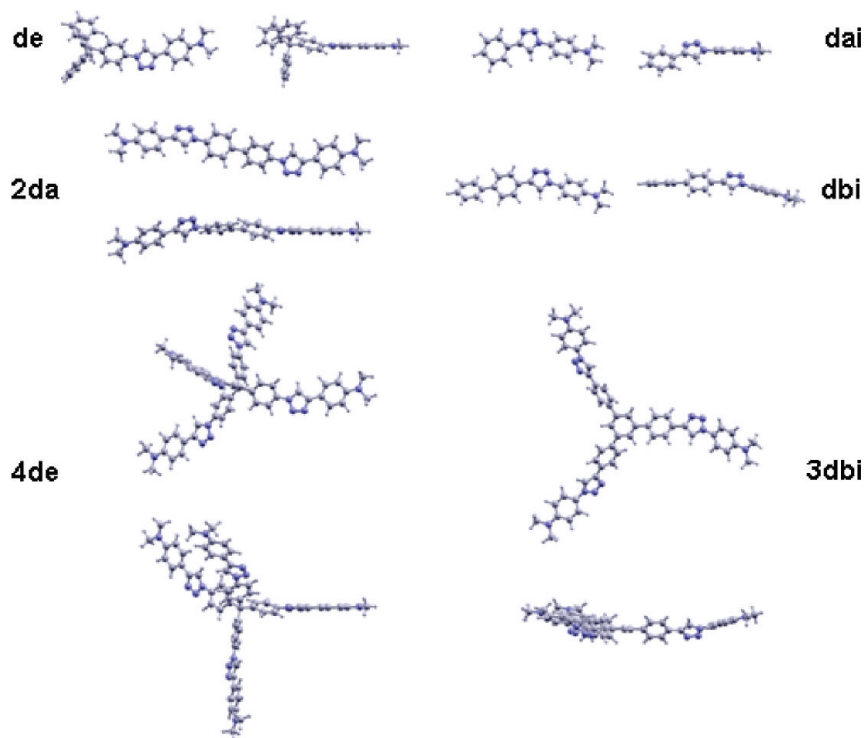


Figure 2. Optimized ground state molecular geometries at the HF/6-31G level of theory for selected **nd** (left) and **ndi** (right) chromophores.

Table 1. Calculated One-Photon Vertical Absorption Maxima at the TD-B3LYP/6-31G//HF/6-31G Level in Vacuum and Experimental One-Photon Absorption Maxima in Toluene^{4,13,40 a}

	2da		3db		dai	dbi	2dai	3dbi	3dfi	4dei
λ^{calcd} nm	388	280	381	274	286	299 ^b	313 ^b	303	348 ^b	292
(eV)	(3.19)	(4.42)	(3.25)	(4.52)	(4.34)	(4.15)	(3.96)	(4.09)	(3.56)	(4.25)
λ^{exptl} nm	350 ^c	296 ^c	347 ^c	293 ^c	303 ^c	310 ^d	325 ^{d,e}	317 ^c	347 ^e	311 ^c
(eV)	(3.54)	(4.19)	(3.57)	(4.23)	(4.09)	(4.00)	(3.82)	(3.91)	(3.57)	(3.99)

^a In parentheses, corresponding transition energies in eV. ^b From ref 31. ^c From ref 40. ^d From ref 13. ^e From ref 4.

Lorentz local field factor correction is clearly inappropriate. In addition, relevant treatment of solvent effects is more time-consuming and may lead to further charge transfer overestimation.¹⁷ Last, no vibrational contributions have been considered. All of these effects will influence computed OPA and TPA spectra, and these approximations should always been kept in mind when comparison to experimental findings is sought.

Natural transition orbital (NTO) analysis of the ES⁶⁸ has been used to analyze the nature of the ES involved in the photophysical processes of interest. They offer a compact representation of transition densities in terms of their expansion into single-particle transitions. Figures showing molecular geometries and NTOs were obtained with XcrySDen.⁶⁹ Use of the Frenkel exciton model within the multibranch strategy is already well documented, and the reader may refer to ref 17 for quadrupoles and three-branched chromophores and to ref 43 for three- and four-branched compounds.

III. Results and Discussion

Validation of TD-DFT Approach. First, let us validate the TD-DFT methodology used to determine the optical responses of the chromophores investigated in this work. A

comparison of TD-B3LYP/6-31G//HF/6-31G results with available experimental photophysical data in toluene^{4,13} showed that the employed level of theory fits nicely for members of **ndi** series.³¹ Further comparison to experimental data⁴⁰ for **dai**, **3dbi**, and **4dei** leads to a comparable agreement with a calculated first absorption band slightly blue-shifted with respect to the experimental band position (Table 1). Such a good agreement is not observed for members of **nd** series for which the calculated transition energies of the first absorption band are red-shifted by more than 0.3 eV with respect to experimental ones (Table 1). In addition, experimental absorption spectra of **2da** and **3db** show two bands in the 285–400 nm spectral region.⁴⁰ For both compounds, the first one shows up near 3.5 eV and has a molar extinction coefficient 2 to 3 times smaller than the strongest absorption band visible near 4.2 eV.⁴⁰ Within TD-B3LYP/6-31G//HF/6-31G, both the ratio between oscillator strengths and the band splitting are significantly overestimated (Table 1, Figures 3 and S7 and S8, Supporting Information). Similar deviations have already been reported for a series of related 1,2,3-triazole-based CT regioisomers⁶ at TD-B3LYP/6-31G(d)//B3LYP/6-31G(d), for which the agreement of experimental and calculated band positions and amplitude much depends on the attachment points of

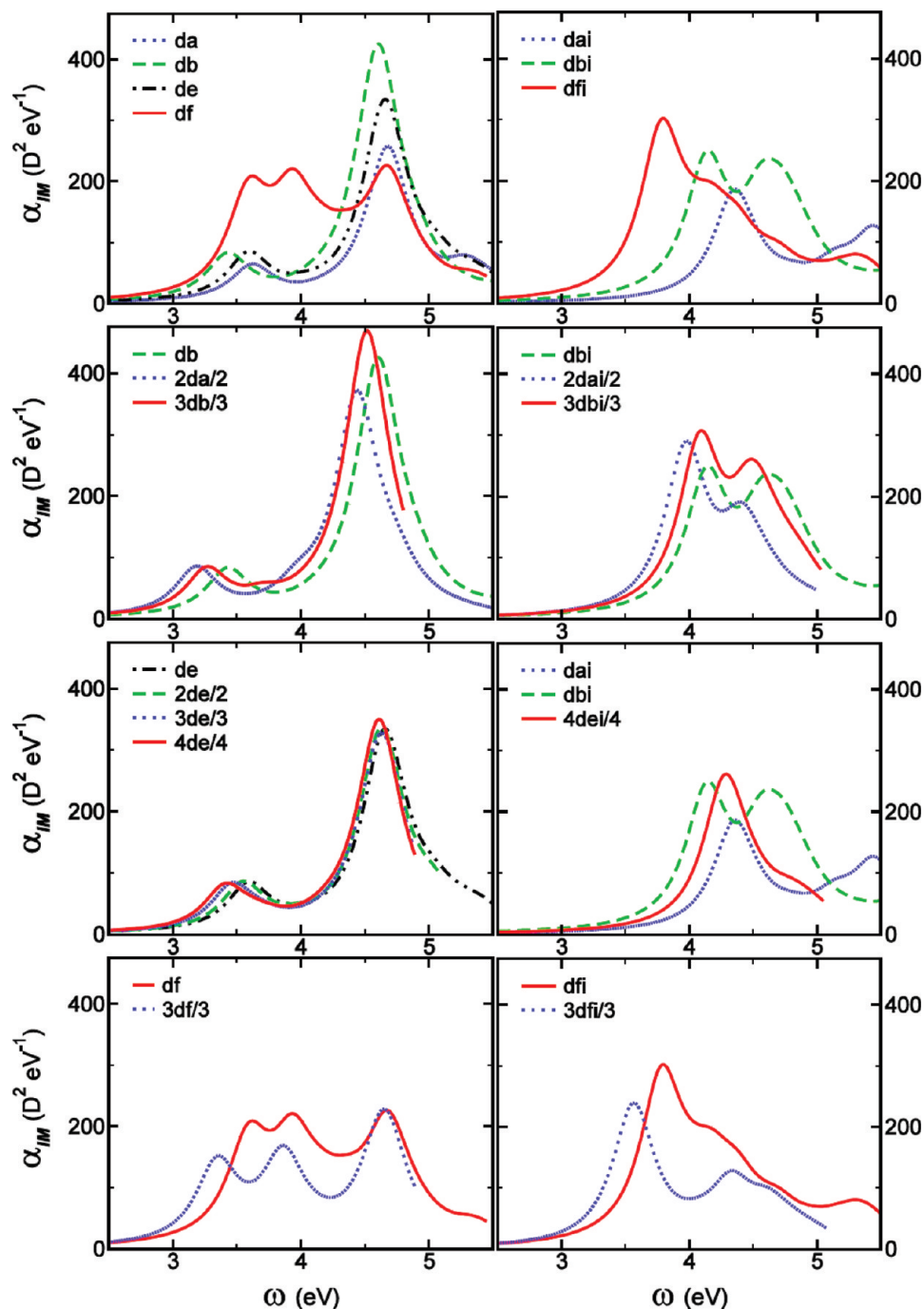


Figure 3. Calculated one-photon absorption (OPA) spectra, normalized for the number of branches: **nd** series (left panels), **ndi** series (right panels).

substituents. Here, we may encounter the well-known TD-DFT problem related to an improper description of the excited state with pronounced charge transfer (CT) character. Even hybrid functionals with a moderate amount of orbital exchange such as the B3LYP model are affected by this problem by placing the CT states energetically too low and by miscalculating their optical transition dipoles compared to experimental results.⁵⁴ Functionals that contain a larger fraction of Fock-like orbital exchange have been shown to overcome this problem, but at the expense of accuracy for the absolute band positions.⁵⁴ Indeed, using the early hybrid BHandH functional (50% of orbital exchange; TD-BHandH/6-31G//HF/6-31G results are available as Supporting Infor-

mation) leads to an overall blue shift of the first absorption band and to a large deviation from the experimental band position (Figures S3, S7, and S8, Supporting Information). At the same time, the splitting between the first two absorption bands becomes comparable to experimental findings, while the ratio of corresponding oscillator strengths becomes close to unity and thus becomes too small (Figure S3, Supporting Information). As a consequence, the correct empirical amount of orbital exchange for given compounds lies somewhere in between B3LYP and BHandH hybrid functionals. In addition, NTOs obtained within these two hybrid schemes, for a larger set of **nd** compounds (Figures S7–S10, Supporting Information), show an overall good

Table 2. Average Torsional Angles in Degrees between Phenyl and Triazole Rings

attachment point	da	db	de	df	2da	3db	3df	4de	dai	dbi	dfi	2dai	3dbi	3dfi	4dei
N	31	31	30	34	30	30	33	28	42	42	42	42	42	42	42
C	2	2	3	3	2	2	2	3	1	0	2	3	5	9	7

correspondence, between B3LYP and BHandH results, for the ESs relevant for the main OPA and TPA bands, with a few exceptions. As expected, one observes a somewhat larger delocalization within BHandH, particularly visible on corresponding holes. However, the few exceptions confirm that the optimal fraction of exchange is a system-dependent parameter.⁵⁴ Even so, B3LYP results suffer from a certain amount of inaccuracy; we believe that the main trends to be discussed in the present work are trustworthy. For the sake of consistency, we will stick to B3LYP throughout the manuscript, while BHandH results will be given as Supporting Information. We emphasize that a comparison between **nd** and **ndi** series should not concern absolute amplitudes, and all other approximations, namely fixed line width, limited number of ESs, and absence of solvent and vibrational contributions, should always be kept in mind, especially for comparison to experimental data.

Molecular Geometry. When experiments are performed in solution at room temperature, various molecular conformations coexist, leading to inhomogeneous broadening of optical spectra. Even if the expected symmetry is not strictly retained, reminiscence of symmetry elements, such as centrosymmetry for quadrupoles⁷⁰ or C_3 ⁴⁷ or T^{43} symmetry for three- or four-branched chromophores, is still observed in experimental findings and is often very useful in supporting an interpretation. In addition, the amount of possible conformations increases with the number of monomeric units building up the multibranching molecule of interest. For these reasons, we did not perform an exhaustive review of attainable molecular arrangements. For the chromophores studied in the present work, the main conformational degree of freedom concerns torsional angles and the sequence of corresponding signs. Explored geometries showed small energy differences compared to kT at room temperature. End groups may be either approximately coplanar or perpendicular to the molecular mean plane (Figures 2 and S1 and S2, Supporting Information). This may modify the state dipole moment, both in amplitude and vector components. But, it remains small enough to keep all chromophores in the so-called *neutral* class of chromophores experiencing no or little solvatochromic behavior of their absorption spectra.^{42,71–73}

For **2d** compounds, we retained the conformation closest to inversion symmetry (Figure 2). For **3d** compounds, the central triphenylamine (TPAmine) or triphenylbenzene (TPB) cores allow for either left- or right-handed skewness of the propeller core. Successive twist angles have been selected to ensure C_3 symmetry and approximate planar geometries (Figure 2). The case of **4de** and **nde** chromophores is somewhat more involved. First, chiral tetrahedral symmetry cannot be satisfied due to the geometry of the molecular branches. Among the carbon-cored tetraphenyl analogs reported in the Cambridge Structural Data Base,⁷⁴ some have almost D_2 symmetry with phenyls related two-by-two like two butterflies;⁷⁵ other have two-by-two almost perpendicular

branches.⁷⁶ For the present study, we retained the latter type of conformation (Figure 2), which also approximates tetrahedral symmetry well enough to lead to quasi-degeneracy of the three first ESs.⁴³

The ground state optimized geometries of the branched structures show geometrical parameters for each branch quasi-identical to those of their dipolar counterparts. Twist angles around BP, TPAmine, or TPB cores are all of about 45° , in line with those reported for related dipolar,^{25,26,29} quadrupolar,³⁷ and three-branched^{25,26,29} chromophores. Despite a reduction of conjugation, as a result of the large twist angles, it has been demonstrated that significant electron communications can still be in action and can lead to enhanced NLO responses.¹⁷ The triazole moiety introduces additional twist angles along the molecular backbone which are significantly larger within **ndi** series than for members of **nd** series, with a larger torsion for the phenyl ring connected to the nitrogen atom of the triazole ring (Table 2). The latter is about 42° for members of **ndi** series, while it is about 30° for compounds of **nd** series. The latter is located in the middle of **nd**'s molecular branches and thus favors electronic decoupling of both molecular ends. This may accentuate the CT problem discussed in the previous section. The twist with the phenyl ring connected to the carbon atom is much smaller, ranging between 1 and 12° (Table 2). These findings are consistent with the structural parameters of related triazole-based structures reported in the Cambridge Structural Data Base.⁷⁴

One-Photon Absorption (OPA). nd Series. Given the 0.3 eV red shift observed between available experimental data and calculated transition energies (Table 1), all members of **nd** series are predicted to show good transparency in the visible region (Figure 3). The first transition has strong CT character, somewhat overestimated within B3LYP; thus the overlap between electron and hole wave functions is small and results in a small oscillator strength (Figure 3). As visible from the respective NTOs (Figures 4–6 and S7–S12, Supporting Information), the nature of the partial CT is preserved all along the series and corresponds to CT from the triazole–phenyl–NMe₂ end group to the opposite end (core in multibranching systems). As the chemical composition of the latter changes along the series, the transition energy of the first band changes as well. This is not the case for the main contribution to absorption, which is dominated by a (few for multibranching systems) higher lying ES near 4.6 eV for all members of the **nd** series. NTOs illustrate the nice correspondence observed all along the series and the large wave function overlap facilitating the large oscillator strength (Figures 6 and S11, Supporting Information). It should be noticed that for multibranching compounds, the oscillator strength is shared between several ESs, which are not all reported in Figure 4 (**4de**) for the sake of conciseness (Figures S11 and S12, Supporting Information). As both hole

	hole	electron		hole	electron
1> w=0.84 3.39eV 366nm f ₀₁ =0.19			1> w=0.83 4.25 eV 292 nm f ₀₁ =0.85		
2> w=0.82 3.39eV 365.5nm f ₀₂ =0.20			2> w=0.83 4.27 eV 290.5 nm f ₀₂ =0.97		
3> w=0.77 3.41eV 364nm f ₀₃ =0.19			3> w=0.45 w=0.35 4.30 eV 288 nm f ₀₃ =1.19		
4> w=0.87 3.44eV 361nm f ₀₄ =0.08			4> w=0.53 w=0.30 4.38 eV 283 nm f ₀₄ =0.11		
39> w=0.28 4.63eV 268nm f ₀₃₉ =2.4			27> w=0.49 w=0.32 4.81eV 258 nm f ₀₂₇ =0.0		

Figure 4. Natural transition orbitals⁶⁸ of **4de** (left panels) and **4dei** (right panels). Text quote in sequence excited state number, associated eigenvalues, transition energies, transition wavelengths, and oscillator strengths.

and electron are located on the peripheral triazole–phenyl–NMe₂ moiety, it is little influenced by chemical substitution on the other end, by the nature of the core or by interbranch interactions in multimers, and stays around 270 nm over all **nd** series. Thus, transition energies are hardly affected by the other molecular extreme or branching center.

It should be noticed that, in the case of quadrupolar series, a conclusion is less straightforward regarding the nature of the relevant ES bearing the main OPA oscillator strength. Within B3LYP, ES |9> bears the largest oscillator strength but has both the hole and electron located at the molecular center (Figure S7, Supporting Information), while BHandH

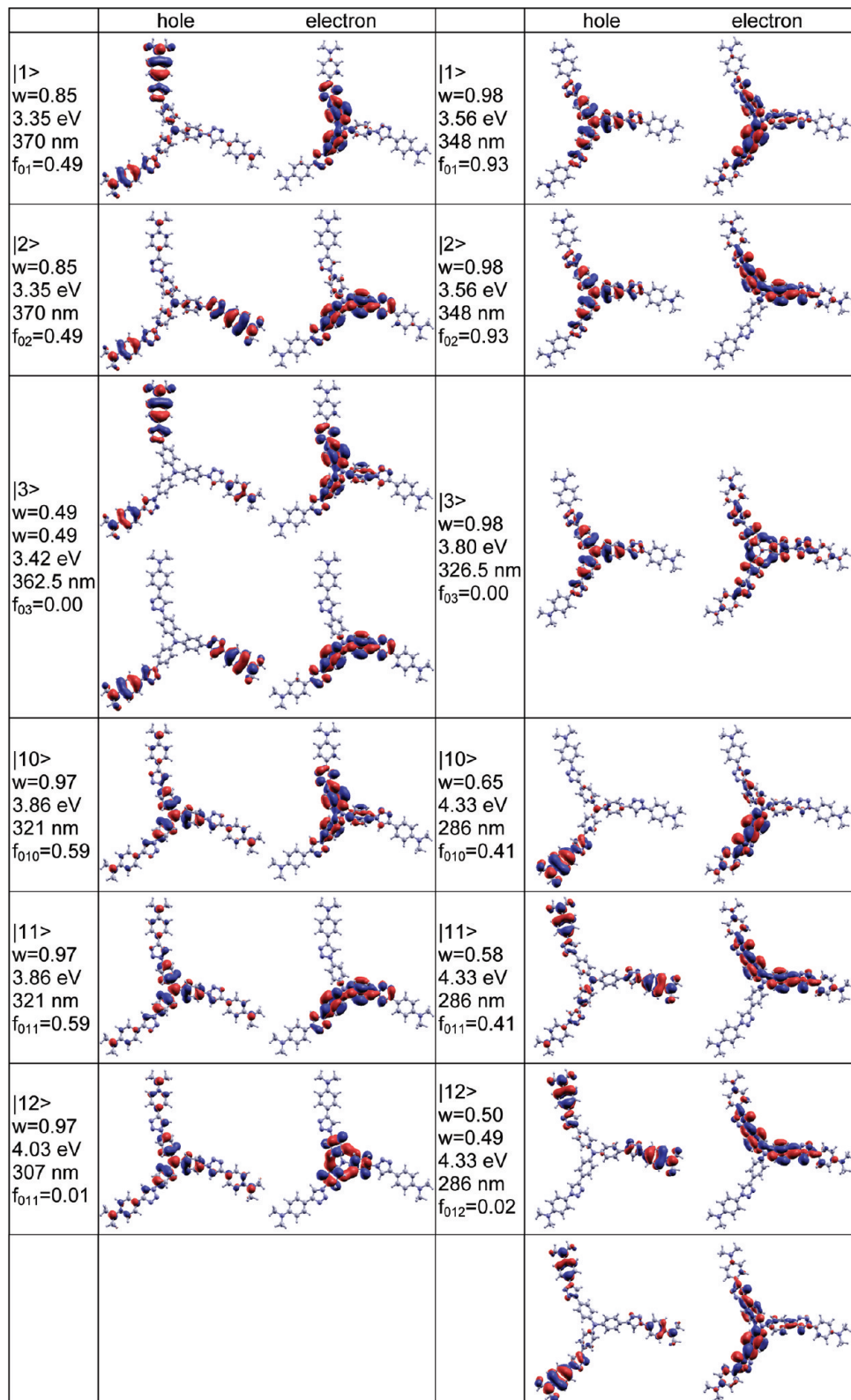


Figure 5. Natural transition orbitals⁶⁸ of **3df** (left panels) and **3dfi** (right panels). Text quote in sequence excited state number, associated eigenvalues, transition energies, transition wavelengths, and oscillator strengths.

leads to bright state |5> with major weight on the molecular ends (Figure S7, similar to ESs |10> and |11> of **3db** shown Figure S8, Supporting Information), consistent with the above-mentioned picture. Within B3LYP, ES |12> shows up at 4.68 eV with comparable shape but is a dark state for symmetry reasons, while ES |13> is a mixture with only a

small amount of oscillator strength (Figure S7, Supporting Information).

In addition, comparison of related dipolar chromophores (**da** with **db** (Figure 3) and **dc** up to **de** (Figure S4, Supporting Information)) shows a slight but systematic red shift and amplitude increase with increasing molecular size. This shows

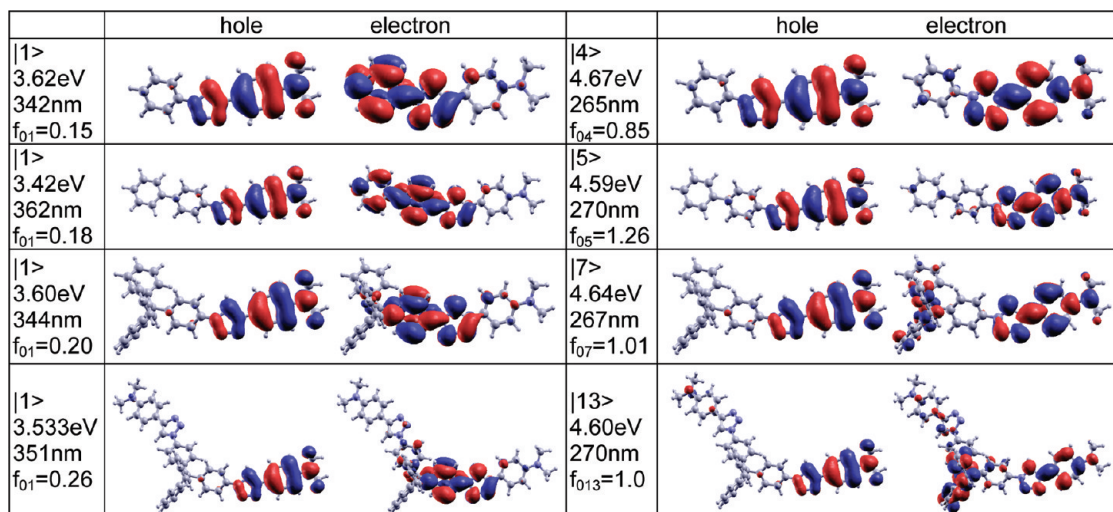


Figure 6. Natural transition orbitals⁶⁸ relevant for the first (left panels) and main (right panels) one-photon absorption (OPA) band of **da** (top), **db** (middle top), **de** (middle bottom), and **2de** (bottom). Text quote in sequence excited state number, associated transition energies, transition wavelengths, and oscillator strengths.

that the choice for the dipolar unit used to investigate efficiency of the branching strategy may be important. Normalized absorption spectra show an overall additive behavior upon branching with no major enhancement or weakening (Figures 3 and S3, Supporting Information).

ndi Series. For all members of this series, the main contribution to the first absorption band arises from the first (first few for multibranching chromophores) ES, which bears significant oscillator strength (Figure 3). Despite the expected red shift upon branching, all chromophores maintain excellent transparency in the visible region. Comparison to the related chromophores³¹ shows a systematic blue shift of the first ES but only a slight reduction of the corresponding oscillator strength. This decrease has to be related to the large twist angle introduced by the triazole moiety (Table 2). Close-lying higher ESs bearing significant oscillator strength lead to broad absorption spectra. NTOs illustrate the nature of corresponding electron redistribution upon excitation. Interestingly, while the holes and electrons associated with the first ES of **dai**, **dbi**,³¹ **2dai**,³¹ **3dbi**, and **4dei** are very similar (Figures 4 and 7 and S13, Supporting Information), with a periphery to core electronic redistribution, those of **dfi**³¹ (Figure S9, Supporting Information) and **3dfi**³¹ (Figure 5) are much different, as they both involve the TPamine core up to the triazole moiety. The latter are quite similar to those reported for a three-branched structure built by gathering three quadrupolar chromophores connected via common nitrogen.³⁷ In both cases, there is a competition between the donating TPamine core and donating NMe₂ end groups, resulting in a larger donating character for the former. Normalized absorption spectra shown Figure 3 illustrate the different behavior upon branching. Again, to discuss OPA amplitude enhancement, the choice of monomeric analog is crucial, as can be seen from the comparison of **dai** and **dbi** spectra (Figure 3). Using **dbi** as a reference, both **2dai** and **3dbi** show significant enhancement of the first absorption band (Figure 3), due to significant delocalization across the branching cores (Figures 7 and S13, Supporting Information), while that of **4dei** is almost additive. On the contrary, the normalized amplitude of **3dfi** decreases (Figure 3) as a result

of slightly reduced delocalization of both the hole and the electron of its first ES³¹ (Figure 5 and S9, Supporting Information).

Position Isomerism. A comparison between members of **nd** and **ndi** series shows quite a few dramatic changes directly related to position isomerism of the 1,2,3-triazole ring. Due to the inverted position of the triazole ring, compounds of **nd** series have a large twist angle in the middle of the molecular backbone which potentially may disrupt their electronic communication between both molecular ends (Figures 2 and S1 and S2, Supporting Information; Table 2). As a consequence, NTOs associated with the first excited state are less delocalized over the molecular backbone and corresponding oscillator strength are smaller than what is observed within **ndi** series (Figures 4 and 5). While for all **ndi** members the first ES contributes predominantly to the main absorption band, **nd** compounds have a first absorption band with reduced amplitude followed by a stronger band showing up at approximately the same position for all members of the series (Figure 3).

Further comparison requires splitting both series into two subassemblies. Let us first concentrate on chromophores that do not contain the TPamine moiety. For all of them, the terminal NMe₂ group acts as the donating unit while the opposite core or end group, which may extend up to the triazole moiety, has electron-withdrawing character. For these compounds, one observes a systematic red shift for **nd** members as compared to the **ndi** analogs, but all maintain transparency in the visible region (Figure 3). The situation is quite different for the TPamine-based chromophores as the donating characters of the different amino groups compete. First, despite the two donating end groups, neither **df** nor **dfi** can be assimilated to a quadrupole (Figure S9, Supporting Information). In fact, the branches deviate significantly from quadrupoles, as both the chemical difference between TPamine and NMe₂ and the dissymmetry of the 1,2,3-triazole moiety significantly break molecular symmetry. A comparison of NTOs of **df** and **3df** with those of **dfi** (Figure S9, Supporting Information) and **3dfi** (Figure 5), respectively, shows a nice correspondence and nicely

	hole	electron		hole	electron
1> w=0.99 4.34 eV 286 nm f ₀₁ =0.56			1> w=0.65 4.09 eV 303 nm f ₀₁ =1.22		
1> w=0.99 4.15 eV 299 nm f ₀₁ =0.67			2> w=0.65 4.09 eV 303 nm f ₀₂ =1.22		
5> w=0.66 5.14 eV 241 nm f ₀₅ =0.15			3> w=0.84 4.58 eV 270 nm f ₀₃ =0.39		
3> w=0.84 4.58 eV 270 nm f ₀₃ =0.39			9> w=0.94 5.45 eV 227 nm f ₀₉ =0.38		
9> w=0.94 5.45 eV 227 nm f ₀₉ =0.38			4> w=0.82 4.73 eV 262 nm f ₀₄ =0.33		
4> w=0.82 4.73 eV 262 nm f ₀₄ =0.33			11> w=0.83 5.77 eV 215 nm f ₀₁₁ =0.12		
11> w=0.83 5.77 eV 215 nm f ₀₁₁ =0.12			30> w=0.76 4.92 eV 252 nm f ₀₃₀ =0.00		
6> w=0.75 5.03 eV 246 nm f ₀₆ =0.04					

Figure 7. Natural transition orbitals⁶⁸ of **dai** (left panels, first line), **dbi** (left panels, second line), and **3dbi** (right panels). Text quote in sequence excited state number, associated eigenvalues, transition energies, transition wavelengths, and oscillator strengths.

illustrates the competition between the donating moieties. Indeed, one observes a clear correspondence between the ES |1> (|3>) of **df** with the ES |6> (|11>) of **dfi**. The same holds when ESs |1> and |2> (|10> and |11>) of **3df** are compared to the |10> and |11> (|1> and |2>) of **3dfi**. The slight differences especially visible on the triazole moiety are directly related to the amplitude of the torsional angles with neighboring phenyl rings. Given the CT problem encountered for **ndf** members, the small residual differences with **ndfi** transition energies after a systematic blue shift of **ndf** ones does not allow for a conclusion about the sign of the shift induced by position isomerism.

We underline that none of these NTOs evidence a strong participation of the triazole ring as electron donating or withdrawing species, when compared to all other moieties building the chromophores. Its main effect is to allow for a

significant blue shift as compared to related dipolar,^{25,26,29} quadrupolar,⁷⁰ and three-branched^{26,27,34,37,77} chromophores and triggers the competition between similar end groups while maintaining good OPA responses. These findings corroborate with what has been previously inferred for isomeric triazole-based CT chromophores⁶ and for hyper-branched poly(1,2,3-triazole)s.⁷ In fact, Tang et al.⁷ conclude that “there exist ground-state interaction between TPamine and triazole moieties in the polymers and that the electronic communications are dictated by the regioisomeric structures”, and Diederich et al.⁶ state that “1,2,3-triazole makes a particularly intriguing linker where the CT path is interrupted... CT can occur and may be even quite efficient depending on the substitution pattern... responsible for about

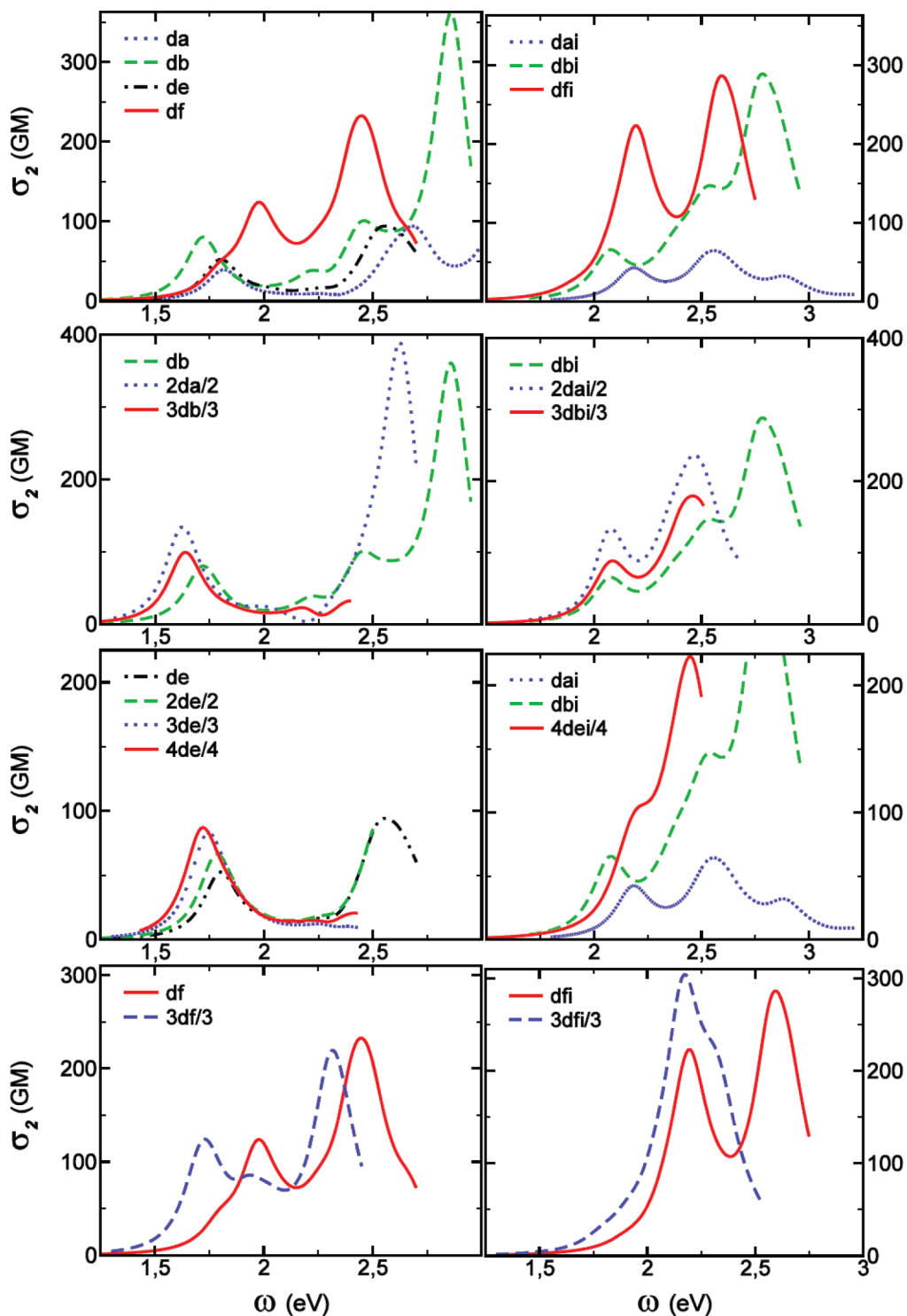


Figure 8. Calculated two-photon absorption (TPA) spectra, normalized for the number of branches: **nd** series (Left panels), **ndi** series (right panels).

50 nm range of transition energies and also affects strongly the absorption intensities”.

Finally, despite the expected red shift upon branching, all chromophores investigated in the present work maintain excellent transparency in the visible region. These findings corroborate experimental data^{4,13,40,41} and confirm that these triazole-based chromophores are good candidates for optical limiting in the visible region.⁴

Two-Photon Absorption (TPA). *nd* Series. Given that our results for this series may be affected by the CT problem

discussed above, let us first determine to what extent an analysis of TPA properties is reasonable at the B3LYP level of theory. Comparison to BHandH calculations shows that the ES at the origin of the first TPA band keeps the same number, and corresponding NTOs have similar shapes (Figures S7–S10, Supporting Information) except for **df**. In both types of calculations, higher lying ESs are responsible for the appearance of additional TPA bands having larger TPA amplitudes (Figures 8 and S5 and S6, Supporting Information). The behavior upon branching regarding am-

plitudes is also consistent within the two levels of theories (Figures 8 and S5, Supporting Information). The main difference concerns transition energies and especially state splitting, which prevents any inference about the strength of the interbranch coupling within the Frenkel excitonic scheme. Indeed, B3LYP would lead to coupling energies less than 0.02 eV, while BHandH gives up to 0.09 eV for **3df** (Figures S7–S10, Supporting Information).

For all but **df**, the first TPA maximum is found where expected, i.e., at the energy of the first, second, third, and fourth ES respectively for all other dipoles, two-branched, three-branched, and four-branched compounds. This is consistent with the Frenkel excitonic picture where dipolar species are approximated by a single ES. The corresponding electronic redistribution involves a periphery NMe₂ to core electronic redistribution (Figures 4, 5, and 7 and S7 and S8, Supporting Information). Interestingly, **df** has only a weak shoulder near the first ES, while the main contribution to the first TPA band arises from the third ES (Figures 8 and S5 and S9, Supporting Information). This may be a consequence of a pseudoquadrupolar behavior of **df**, which bears nitrogen-based end groups that compete. Further inference about the underlying electronic redistribution of relevant ESs is hindered by significant differences between NTOs obtained at the B3LYP and BHandH levels of theory (Figure S9, Supporting Information). The case of **3df** is less tricky (Figure 5) thanks to the additional dissymmetry introduced on the TPAmine core, which is trifunctionalized in the multimer, in a similar way to what has been obtained for an analogous branched chromophore built by gathering three quadrupolar units through a TPAmine core.³⁷ Even larger TPA amplitudes are predicted at higher energies (Figures 8 and S5) for all but **3db**, **3de**, and **4de**, for which analysis is limited by the number of ESs accounted for in our calculations. Unfortunately, a systematic interrelationship with given ESs is difficult, and no general feature emerges from an inspection of NTOs that have been identified (Figures S9 and S10, Supporting Information).

In the investigation of branching strategy, it is worthwhile to emphasize the crucial role played by the choice of the benchmark monomer. Whereas only slight variations of both amplitudes and band position occur for various substitution of the sp³ carbon branching center (going from **dc** to **de**, Figure S6, Supporting Information), the differences are large when replacing the phenyl end group of **da** with a BP moiety to build **db** (Figure 8). The BP moiety not only stabilizes the first ES but also leads to a 2-fold increase of the corresponding TPA amplitude thanks to an expansion of delocalization over the BP end group (Figure 6).

Whereas an overall additive behavior was observed for OPA, the branching strategy is shown to be efficient for TPA responses (Figure 8). Once more, the TPAmine-based chromophores are the exceptions among the series, as the amplitude of the main TPA bands reveals essentially an additive behavior. Enhancement and reduction in a specific spectral region are also observed. The overall behavior is quite similar to what has been reported for a branched quadrupole,³⁷ but careful inspection of the relevant NTOs shows that the nature of the ES in question is much less

modified by the branching center. There is a clear correspondence between ES |1⟩ of **df** (Figure S9, Supporting Information) and |3⟩ of **3df** (Figure 5). Within this picture, the amplitude of the first TPA band of **3df** has to be compared to the shoulder near 1.8 eV in the TPA spectra of **df**, and we recover the large enhancement expected for TPAmine branched dipoles.^{25,26} Despite similarities between the NTO of |3⟩ of **df** (Figure S9, Supporting Information) and |27⟩ of **3df** (Figure 5), the large blue shift of the latter prompts care in further interpretation. On the contrary, the picture is simple for all other members of this series, as the first TPA band of monomers can be safely attributed to ES |1⟩ and that of multimers to what predicts the exciton model. The quadrupole **2da** shows the largest enhancement (a factor of 2 with respect to **db**) thanks to through-BP-core communication (Figures 8 and S7, Supporting Information). **3db** evidences much less enhancement consistently with the other TPBcore¹⁷ analogs for which functionalization in meta-positions inhibits interbranch conjugation (Figures 8 and S8, Supporting Information). It should be noticed that comparison to the benchmark chromophore bearing a TPB end group instead of BP would most probably lead at best to additive and, at worst, to reduced TPA responses.²⁹ The case of C-centered branched chromophores is quite surprising, as little through-bond cross-talk is expected. Surprisingly, we observe a large enhancement that increases with an increasing number of branches from **2de** up to **4de** (Figure 8). Unfortunately, without any reasonable estimate for the interbranch coupling, it is hard to conclude about the underlying nature of the interbranch interactions that may be mainly electrostatic or involve coherent interbranch interaction.

ndi Series. Let us start with the monomeric benchmark chromophores. For the sake of conciseness, we did not consider any of the monomers substituted by sp³ carbons. Results obtained on **nd** series reveal only slight modifications between **da** and sp³ carbon analogs, the latter having properties in between those of **da** and **db** (Figure 8). Thus, **dai** and **dbi** will respectively be used as lower and upper limits for investigating the effect of branching in all branched chromophores missing TPAmine. Figure 8 reports calculated TPA spectra for the three monomers **dai**, **dbi**, and **dfi**. Replacement of the phenyl end group by either BP or TPAmine leads to a dramatic increase of the TPA response, with the same trends previously observed for the respective OPA (Figure 3). These substitutions induce respectively significant red and slight blue shifts of the first TPA band. The latter is quite surprising given the large red shift observed for the main OPA band (Figure 3). In fact, the first TPA band of both **dai** and **dbi** mainly involves ES |1⟩, which also dominates the OPA spectra. On the contrary, that of **dfi** originates from ES |6⟩ (Figure 8 and S9, Supporting Information), while ES |1⟩ appears as a very slight shoulder on the low-energy part of the spectrum. A comparison of corresponding NTOs is informative, as ES |6⟩ of **dfi** (Figure S9, Supporting Information) shows a similar periphery to core electronic redistribution to ES |1⟩ of **dai** and **dbi** (Figure 7). Thus, while for the first OPA band, the nature of the underlying electronic redistribution is completely different

for **dfi**, it is not for the final TPA state, giving rise to the first TPA band. Furthermore, the NTO of an electron for **dfi** is the same for ES |1⟩ and ES |6⟩, while the respective hole orbitals involve either TPamine or NMe₂ end groups, indicating the competition between the two donating nitrogen-based moieties (Figure S9, Supporting Information). The second TPA maxima near 2.5 eV (Figure 8) can be assigned to ESs |5⟩, |6⟩, and |13⟩ respectively for **dai**, **dbi** (Figure 7), and **dfi** (Figure S9, Supporting Information).

TPA spectra normalized for the number of branches, reported in Figure 8, reveal significant enhancement for all branched chromophores with coincident peak positions (shoulder for **4dei**) for the first TPA band. Higher-lying ESs lead to even larger TPA responses on the high-energy part of the spectra (Figures 4 and 7), except for **3dfi**, for which analysis is hindered by the limited number of ESs accounted for. For all branched chromophores but **3dfi**, the first TPA maximum is found consistently with the Frenkel exciton model predictions: the second, third, and fourth ES for **2dai**, **3dbi**, and **4dei**, respectively. The final ES involves a periphery (NMe₂) to core electronic redistribution (Figures 4 and 6 and S13, Supporting Information) that nicely compares with ES |1⟩ of **dai** and **dbi** (Figure 7), especially for **3dbi** and **4dei**. However, predicted state splitting with respect to benchmark monomers is not at all satisfied and manifests itself by correspondence in band position of the first TPA bands of monomers and multimers. From the state splitting calculated for each multimer, estimates of the interbranch coupling lead to 0.02, 0.03, and 0.08 eV for **3dbi**, **4dei**, and both **2dai** and **3dfi**, respectively. These findings correlate nicely with the shapes of corresponding NTOs as the hole and electrons have at least significant if not prevailing weight on the branching center for the last two chromophores. **3dbi** evidences the lowest enhancement among the series, as was the case for **3db**, and the same reasoning holds. The case of C-centered branched chromophores **4dei** is similar to that of **3dbi**, as the enhancement near 2.2 eV partly results from the tail of the band located near 2.4 eV. The latter can be assigned to a couple of higher-lying excited states such as ES |27⟩ (Figure 4) showing similitude to some degree with ESs |5⟩ and |6⟩ of **dai** and **dbi**, respectively (Figure 7). For **2dai**, the strong through-bond communication allowed by the BP core leads to a significant mixture of dipolar ESs (Figure S13, Supporting Information) and 2-fold enhancement of the normalized TPA amplitude at the first maximum as compared to **dbi**.³¹ Contrary to the three other branched **ndi** compounds, but consistent with what has been observed for the benchmark chromophore **dfi**, no TPA maximum is visible in the vicinity of ES |3⟩ of **3dfi** (Figures 5 and 8). The main TPA band has to be related to ES |12⟩, and the corresponding NTOs show a marked correspondence with those of ES |6⟩ of **dfi**³¹ (Figures 8 and S9, Supporting Information). Thus, for all **ndi** chromophores, the first TPA band corresponds to an ES with quasi-identical holes and electrons, i.e., underlying electronic redistribution from the periphery (NMe₂ side) to the core. Moreover, significant enhancement of the normalized TPA amplitude is also in action for **3dfi**. Therefore, the branching strategy is shown to be efficient for all four

investigated cores. The main change occurring for TPamine-based chromophores is the decoupling of OPA and TPA states, leading to a consequent blue shift of the first TPA maximum, which shows the largest values among both monomers and multimers. Among the reasons for large TPA amplitudes, we notice that higher-lying ESs open the possibility of manifold coupling to the intermediate lower-lying ESs which have sizable transition dipole moments.

Position Isomerism. The overall effects of position isomerism of the 1,2,3-triazole ring on TPA responses are less dramatic than those reported for OPA responses, except for TPamine-based chromophores. In fact, ESs at the origin of the main OPA band are quite different for the two series, but this is not the case for the first TPA band. Consistently with the Frenkel exciton picture, where monomers are reduced to a single ES (being the first ES), large TPA responses originate from the first, second, third, and fourth ESs, respectively, for monomers and two-, three-, and four-branched compounds, for both position isomers. In addition, NTOs of the final ES show comparable periphery (NMe₂ end groups) to core electron redistribution for all. Overall, their delocalization over the molecular branches is significantly larger for **ndi** than for **nd**; even so, comparison is limited due to the occurrence of CT problems for **nd** compounds. Accordingly, transition dipole moments are significantly enhanced for **ndi** chromophores with respect to their **nd** analogs and so are the related TPA responses.

The ranking of TPA amplitudes and relative peak positions among the dipoles stays the same for both position isomers. Substitution of the phenyl ring by a BP moiety leads to a noteworthy increase of TPA amplitudes and red shifts. Further substitution by the TPamine core leads to an additional TPA enhancement but with a significant red shift of the first TPA band.

Interestingly, none of the investigated branched chromophores showed any quantitative agreement with the Frenkel exciton model such as band splitting around the first ES of the benchmark monomer. TPA spectra of **nd** chromophores undergo significant red shifts with respect to their monomeric counterpart, while the first TPA maximum of each **ndi** chromophore shows up almost at the same energy as its dipolar benchmark. On the contrary, the behavior upon branching is quite similar in both series with moderate enhancement for the TPB core, sizable enhancement for the sp³ carbon core, and a 2-fold increase for the BP core. Thus, combining the branching strategy with inversion of the triazole moiety allows for a spectral tuning of TPA responses to a larger extent than OPA responses, while maintaining the efficiency of cooperative effects.

The different behavior upon triazole inversion in TPamine-based chromophores is directly related to the presence of two donating units instead of one. As already mentioned for investigation of their OPA responses, the triazole moiety controls competition between the two amino groups. This is nicely illustrated by NTOs of relevant ESs. Indeed, there is a clear correspondence between NTOs of ESs |3⟩ of **3df** and |12⟩ of **3dfi** (Figure 5). These final ESs contribute predominantly to the first TPA band. They involve periphery to core electronic redistribution comparable to what was found for

the compounds missing the TPAmine moiety. The same conclusions hold for ESs 11) of **df** and 16) of **dfi** (Figure S9, Supporting Information). Moreover, NTOs of ESs 12) of **3df** and 13) of **3dfi** (Figure 5) also have very similar shapes and are thus consistent with the picture of competing end groups. Because corresponding electronic redistributions are localized on the molecular center, they do not contribute strongly to the TPA responses. Thus, the overall competition between the donating end groups explains why both **dfi** and **3dfi** have their first TPA maximum far blue-shifted from the position expected from simple reasoning.

IV. Conclusion

In conclusion, our quantum-chemical study based on TD-DFT methodology reveals that the introduction of triazole leads to dramatic changes in optical properties of chromophores especially for multibranching systems. The triazole ring substantially modifies electronic communication and delocalization between the electron donating end groups and the branching center via torsion between phenyl and triazole rings (which modifies π conjugation) as well as through its position isomerism. This adds to the multibranching interactions allowing for stark and intuitively unexpected modifications of the excited state structure and, subsequently, relevant linear and nonlinear optical responses. In particular, when compared to the related analogs, introduction of the triazole moiety in the molecular backbone allows for an overall blue shift of both OPA and TPA bands with a moderate reduction of their response amplitudes. Their ability to be promising candidates for optical limiting in the visible region^{4,13} is thus confirmed for a diversified set of chromophores. In addition, larger TPA responses are predicted at higher energies due to contribution from high-lying ESs, as has been found for similar chromophores.³⁷

The changes observed between the two investigated 1,2,3-triazole regioisomers can be related to differences in the amplitude of the torsional angles with neighboring phenyl rings. The twist angle with the phenyl substituted at the nitrogen position is significantly larger than that of phenyl attached to the carbon atom. The former is located in the middle of **nd**'s molecular branch(es), thus favoring electronic decoupling of both molecular ends. Besides the overall blue shift when going from **nd** series to **ndi**, this electronic decoupling increases the CT character of the first absorption band and consequently decreases its OPA oscillator strength. The main absorption band shows up near 4.6 eV independently of the **nd** chromophore of interest. This is a direct consequence of hole and electron localization on the peripheral NMe₂ donating moiety that ensures little influence on the relevant ES(s) of chemical substitution on the other end, including the nature of the core or interbranch interactions in multimers. Position isomerism has less impact on TPA properties except for TPAmine-based chromophores. The latter bear nitrogen-based donating moieties on both molecular ends or in the periphery and core for monomers and three-branched derivatives, respectively. Thus, monomers could be thought of as pseudoquadrupoles, but a detailed inspection of NTOs reveals more subtle behavior. In fact, the respective role of TPAmine and NMe₂ moieties is

interchanged and triggered by the triazole moiety. Calculations show that the nature of the electronic redistribution (periphery to TPAmine core) associated with the first TPA band of the three-branched compounds is preserved. As a consequence, while the main contribution to this TPA band arises from the third ES for **3df**, it corresponds to the 12's ES for **3dfi**. In addition, none of the monomers has its first TPA band in the vicinity of the first ES. Thus, the apparent behavior upon branching is dramatically different for both isomers with a large red shift and almost additive behavior for **3df** as compared to **df** and concomitance and enhancement for **3dfi** as compared to **dfi**. This intricate finding renders the Frenkel exciton model useless for these TPAmine-based chromophores.

For all other chromophores, results are qualitatively consistent with the Frenkel exciton picture, as the main TPA band can be associated with the first, second, third, and fourth ESs respectively for dipoles, quadrupoles, and three- and four-branched derivatives. As found for related chromophores, the donating end groups confer an electron-withdrawing character to the BP,³⁷ TPB,²⁹ and sp³ carbon moieties, and monomers have significant dipolar character. The branching strategy is shown to be efficient, as a sizable enhancement of TPA responses are observed with moderate effect for the TPB core and an about 2-fold enhancement for both the BP core and carbon-cored tetramers. However, we underline that the amount of enhancement strongly depends on the choice of the benchmark monomer. In addition, for large chromophores, the increasing number of conformational degrees of freedom may lead to experimentally broadened and reduced TPA responses when performed in solutions at room temperature, as a result of oscillator strength redistribution (intramonomer) and inhomogeneous broadening.

These results and further inspection of NTOs clearly show that the 1,2,3-triazole moiety has neither electron-donating nor an electron-withdrawing character. This corroborates earlier findings where triazole has been described as an intriguing linker that introduces conjugation disruption while maintaining significant electronic communication and optical properties dictated by the regioisomeric structure. This subtle acting of the triazole moiety may be combined with complementary routes of molecular engineering. Replacement of donating end groups by electron-withdrawing ones should lead to further blue shifts that may allow keeping full transparency in the visible region for even larger branched systems.^{25,26,29} Substitution of the BP core by a fluorene core should allow for increased conjugation along the quadrupolar backbone, and both improve OPA and TPA responses.³⁷ Slight dissymmetrical functionalization could also offer interesting perspectives for spatial control of the fluorescence properties while maintaining sizable TPA responses.⁶³

From the theoretical perspective, these series are also a kind of tutorial example of isomeric compounds that are quite sensitive to the specific exchange-correlation DFT functional in use. In fact, while the description of one of the isomers shows very good agreement with experimental results, e.g., for transition energies and oscillator strength, the other is quite poorly described with the very same functional. In fact, besides comparison to available experimental data, in the

B3LYP model, the first ESs of **nd** members have small integral overlap between their electron and hole orbitals, which renders these states to be optically weak. This small overlap between NTO orbitals of ES indicates its CT nature (in a general sense) and points to a possible overestimated red shift of the transition energy in semilocal DFT models. Nevertheless, an increase of the amount of exact exchange (going from B3LYP to BHandH) allows for at least qualitative interpretation of the optical responses of the studied chromophores. The present example illustrates the need for further improvements in developing a better XC functional, as it shows that an increased amount of exact exchange improves respective oscillator strengths of different electronic transitions, while at the same time the respective transition energies become overestimated.

Acknowledgment. We wish to thank Dr. Olivier Mongin and Manuel Parent for stimulating discussions. This work was performed in part at the U.S. Department of Energy, Center for Integrated Nanotechnologies (CINT), at Los Alamos National Laboratory (LANL; contract DE-AC52-06NA25396). This work was granted access to the HPC resources of CINES under the allocation 2005-[c20050822414] and 2008-[x20080825087] made by GENCI (Grand Equipement National de Calcul Intensif).

Supporting Information Available: Table of spectral linewidths; additional calculated OPA and TPA spectra; and additional calculated NTOs and corresponding transition energies, transition wavelengths, and oscillator strengths. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 2004.
- Qin, A.; Lam, J. W. Y.; Tang, B. Z. Click polymerization. *Chem. Rev.* **2010**, *39*, 2522.
- Zhou, Z.; Fahmi, C. J. A Fluorogenic Probe for the Copper(I)-Catalyzed Azide-Alkyne Ligation Reaction: Modulation of the Fluorescence Emission via $3(n,\pi^*)-1(\pi,\pi^*)$ Inversion. *J. Am. Chem. Soc.* **2004**, *126*, 8862.
- Parent, M.; Mongin, O.; Kamada, K.; Katan, C.; Blanchard-Desce, M. New chromophores from click chemistry for two-photon absorption and tuneable photoluminescence. *Chem. Commun.* **2005**, 2029.
- Qin, A.; Jim, C. K. W.; Lu, W.; Lam, J. W. Y.; Häußler, M.; Dong, Y.; Sung, H. H. Y.; Williams, I. D.; Wong, G. K. L.; Tang, B. Z. Click Polymerization: Facile Synthesis of Functional Poly(aroyltriazole)s by Metal-Free, Regioselective 1,3-Dipolar Polycycloaddition. *Macromolecules* **2007**, *40*, 2308.
- Jarowski, P. D.; Wu, Y.-L.; Schweizer, W. B.; Diederich, F. 1,2,3-Triazoles as Conjugative π -Linkers in Push-Pull Chromophores: Importance of Substituent Positioning on Intramolecular Charge-Transfer. *Org. Lett.* **2008**, *10*, 3347.
- Qin, A.; Lam, J. W. Y.; Jim, C. K. W.; Zhang, L.; Yan, J.; Häußler, M.; Liu, J.; Dong, Y.; Liang, D.; Chen, E.; Jia, G.; Tang, B. Z. Hyperbranched polytriazoles: Click polymerization, regioisomeric structure, light emission, and fluorescent patterning. *Macromolecules* **2008**, *41*, 3808.
- Meudtner, Robert, M.; Ostermeier, M.; Goddard, R.; Limberg, C.; Hecht, S. Multifunctional IdoClickatesrdquo as Versatile Extended Heteroaromatic Building Blocks: Efficient Synthesis via Click Chemistry, Conformational Preferences, and Metal Coordination. *Chem.—Eur. J.* **2007**, *13*, 9834.
- Schweinfurth, D.; Hardcastle, K. I.; Bunz, U. H. F. 1,3-Dipolar cycloaddition of alkynes to azides. Construction of operationally functional metal responsive fluorophores. *Chem. Commun.* **2008**, 2203.
- Li, Z. a.; Yu, G.; Wu, W.; Liu, Y.; Ye, C.; Qin, J.; Li, Z. Nonlinear Optical Dendrimers from Click Chemistry: Convenient Synthesis, New Function of the Formed Triazole Rings, and Enhanced NLO Effects. *Macromolecules* **2009**, *42*, 3864.
- Li, Z. a.; Yu, G.; Hu, P.; Ye, C.; Liu, Y.; Qin, J.; Li, Z. New Azo-Chromophore-Containing Hyperbranched Polytriazoles Derived from AB₂Monomers via Click Chemistry under Copper(I) Catalysis. *Macromolecules* **2009**, *42*, 1589.
- Xie, J.; Hu, L.; Shi, W.; Deng, X.; Cao, Z.; Shen, Q. Synthesis and nonlinear optical properties of hyperbranched polytriazole containing second-order nonlinear optical chromophore. *J. Polym. Sci., Part B: Polym. Phys.* **2008**, *46*, 1140.
- Zoon, P. D.; Stokkum, I. H. M. v.; Parent, M.; Mongin, O.; Blanchard-Desce, M.; Brouwer, A. M. Fast photo-processes in triazole-based push-pull systems. *Phys. Chem. Chem. Phys.* **2010**, *12*, 2706.
- Fréchet, J. M. J. Dendrimers and supramolecular chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4782.
- Sullivan, P. A.; Rommel, H.; Liao, Y.; Olbricht, B. C.; Akelaitis, A. J. P.; Firestone, K. A.; Kang, J.-W.; Luo, J.; Davies, J. A.; Choi, D. H.; Eichinger, B. E.; Reid, P. J.; Chen, A.; Jen, A. K. Y.; Robinson, B. H.; Dalton, L. R. Theory-guided design and synthesis of multichromophore dendrimers: An analysis of the electro-optic effect. *J. Am. Chem. Soc.* **2007**, *129*, 7523.
- He, G. S.; Tan, L. S.; Zheng, Q.; Prasad, P. N. Multiphoton absorbing materials: Molecular designs, characterizations, and applications. *Chem. Rev.* **2008**, *108*, 1245.
- Terenziani, F.; Katan, C.; Badaeva, E.; Tretiak, S.; Blanchard-Desce, M. Enhanced Two-Photon Absorption of Organic Chromophores: Theoretical and Experimental Assessments. *Adv. Mater.* **2008**, *20*, 4641.
- Rumi, M.; Barlow, S.; Wang, J.; Perry, J. W.; Marder, S. R. Two-Photon Absorbing Materials and Two-Photon-Induced Chemistry. In *Photoresponsive Polymers I; Advances in Polymer Science*; Springer-Verlag: Berlin, 2008; Vol. 213, pp 1–95.
- Pawlicki, M.; Collins, H. A.; Denning, R. G.; Anderson, H. L. Two-Photon Absorption and the Design of Two-Photon Dyes. *Angew. Chem., Int. Ed. Engl.* **2009**, *48*, 3244.
- Specht, A.; Bolze, F.; Omran, Z.; Nicoud, J.-F.; Goeldner, M. Photochemical tools to study dynamic biological processes. *HFSP J.* **2009**, *3*, 255.
- Kim, H. M.; Cho, B. R. Two-photon materials with large two-photon cross sections. Structure-property relationship. *Chem. Commun.* **2009**, 153.
- Joshi, M. P.; Swiatkiewicz, J.; Xu, F. M.; Prasad, P. N. Energy transfer coupling of two-photon absorption and reverse

- saturable absorption for enhanced optical power limiting. *Opt. Lett.* **1998**, *23*, 1742.
- (23) Chung, S. J.; Kim, K. S.; Lin, T. H.; He, G. S.; Swiatkiewicz, J.; Prasad, P. N. Cooperative enhancement of two-photon absorption in multi-branched structures. *J. Phys. Chem. B* **1999**, *103*, 10741.
- (24) Cho, B. R.; Son, K. H.; Lee, S. H.; Song, Y. S.; Lee, Y. K.; Jeon, S. J.; Choi, J. H.; Lee, H.; Cho, M. H. Two Photon Absorption Properties of 1,3,5-Tricyano-2,4,6-tris(styryl)benzene Derivatives. *J. Am. Chem. Soc.* **2001**, *123*, 10039.
- (25) Katan, C.; Terenziani, F.; Mongin, O.; Werts, M. H. V.; Porres, L.; Pons, T.; Mertz, J.; Tretiak, S.; Blanchard-Desce, M. Effects of (multi)branching of dipolar chromophores on photophysical properties and two-photon absorption. *J. Phys. Chem. A* **2005**, *109*, 3024.
- (26) Katan, C.; Terenziani, F.; Droumaguet, C. L.; Mongin, O.; Werts, M. H. V.; Tretiak, S.; Blanchard-Desce, M. Branching of dipolar chromophores: effects on linear and nonlinear optical properties. *Proc. SPIE, Int. Soc. Opt. Eng.* **2005**, 5935, 593503.
- (27) Terenziani, F.; Morone, M.; Gmouh, S.; Blanchard-Desce, M. Linear and Two-Photon Absorption Properties of Interacting Polar Chromophores: Standard and Unconventional Effects. *ChemPhysChem* **2006**, *7*, 685.
- (28) Allain, C.; Schmidt, F.; Lartia, R.; Bordeau, G.; Fiorini-Debuisschert, C.; Charra, F.; Tauc, P.; Teulade-Fichou, M.-P. Vinyl-Pyridinium Triphenylamines: Novel Far-Red Emitters with High Photostability and Two-Photon Absorption Properties for Staining DNA. *ChemBioChem* **2007**, *8*, 424.
- (29) Terenziani, F.; Le Droumaguet, C.; Katan, C.; Mongin, O.; Blanchard-Desce, M. Effect of branching on two-photon absorption in triphenylbenzene derivatives. *ChemPhysChem* **2007**, *8*, 723.
- (30) Kim, H. M.; Seo, M. S.; Jeon, S.-J.; Cho, B. R. Two-photon absorption properties of hexa-substituted benzene derivatives. Comparison between dipolar and octupolar molecules. *Chem. Commun.* **2009**, 7422.
- (31) Katan, C.; Tretiak, S.; Even, J. Two-photon transitions in triazole based quadrupolar and octupolar chromophores: a TD-DFT investigation. *Proc. SPIE, Int. Soc. Opt. Eng.* **2010**, *7712*, 77123D.
- (32) Drobizhev, M.; Karotki, A.; Dzenis, Y.; Rebane, A.; Suo, Z. Y.; Spangler, C. W. Strong cooperative enhancement of two-photon absorption in dendrimers. *J. Phys. Chem. B* **2003**, *107*, 7540.
- (33) Yoo, J.; Yang, S. K.; Jeong, M. Y.; Ahn, H. C.; Jeon, S. J.; Cho, B. R. Bis-1,4(p-diarylaminostyryl)-2,5-dicyanobenzene derivatives with large two-photon absorption cross-sections. *Org. Lett.* **2003**, *5*, 645.
- (34) Mongin, O.; Porres, L.; Katan, C.; Pons, T.; Mertz, J.; Blanchard-Desce, M. Synthesis and two-photon absorption of highly soluble three-branched fluorenylene-vinylene derivatives. *Tetrahedron Lett.* **2003**, *44*, 8121.
- (35) Bartholomew, G. P.; Rumi, M.; Pond, S. J. K.; Perry, J. W.; Tretiak, S.; Bazan, G. C. Two-photon absorption in three-dimensional chromophores based on 2.2 -paracyclophane. *J. Am. Chem. Soc.* **2004**, *126*, 11529.
- (36) Varnavski, O.; Yan, X. Z.; Mongin, O.; Blanchard-Desce, M.; Goodson, T.; Goodson, T. Strongly interacting organic conjugated dendrimers with enhanced two-photon absorption. *J. Phys. Chem. C* **2007**, *111*, 149.
- (37) Katan, C.; Tretiak, S.; Werts, M. H. V.; Bain, A. J.; Marsh, R. J.; Leonczek, N.; Nicolaou, N.; Badaeva, E.; Mongin, O.; Blanchard-Desce, M. Two-photon transitions in quadrupolar and branched chromophores: Experiment and theory. *J. Phys. Chem. B* **2007**, *111*, 9468.
- (38) Rumi, M.; Pond, S. J. K.; Meyer-Friedrichsen, T.; Zhang, Q.; Bishop, M.; Zhang, Y.; Barlow, S.; Marder, S. R.; Perry, J. W. Tetrastyrilarene Derivatives: Comparison of One- and Two-Photon Spectroscopic Properties with Distyrylarene Analogues. *J. Phys. Chem. C* **2008**, *112*, 8061.
- (39) Varnavski, O. P.; Ostrowski, J. C.; Sukhomlinova, L.; Twieg, R. J.; Bazan, G. C.; Goodson, T. Coherent Effects in Energy Transport in Model Dendritic Structures Investigated by Ultrafast Fluorescence Anisotropy Spectroscopy. *J. Am. Chem. Soc.* **2002**, *124*, 1736.
- (40) Parent, M. Ph.D. Thesis, University of Rennes, Rennes, France, 2008.
- (41) Blanchard-Desce and Coll. Unpublished.
- (42) Terenziani, F.; Painelli, A.; Katan, C.; Charlot, M.; Blanchard-Desce, M. Charge instability in quadrupolar chromophores: Symmetry breaking and solvatochromism. *J. Am. Chem. Soc.* **2006**, *128*, 15742.
- (43) Lambert, C.; Schmalzlin, E.; Meerholz, K.; Brauchle, C. Synthesis and nonlinear optical properties of three-dimensional phosphonium ion chromophores. *Chem.—Eur. J.* **1998**, *4*, 512.
- (44) Tretiak, S.; Chernyak, V.; Mukamel, S. Localized Electronic Excitations in Phenylacetylene Dendrimers. *J. Phys. Chem. B* **1998**, *102*, 3310.
- (45) Poliakov, E. Y.; Chernyak, V.; Tretiak, S.; Mukamel, S. Exciton-scaling and optical excitations of self-similar phenylacetylene dendrimers. *J. Chem. Phys.* **1999**, *110*, 8161.
- (46) Chernyak, V.; Poliakov, E. Y.; Tretiak, S.; Mukamel, S. Two-exciton states and spectroscopy of phenylacetylene dendrimers. *J. Chem. Phys.* **1999**, *111*, 4158.
- (47) Beljonne, D.; Wenseleers, W.; Zojer, E.; Shuai, Z. G.; Vogel, H.; Pond, S. J. K.; Perry, J. W.; Marder, S. R.; Bredas, J. L. Role of dimensionality on the two-photon absorption response of conjugated molecules: The case of octupolar compounds. *Adv. Funct. Mater.* **2002**, *12*, 631.
- (48) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. Molecular excitation energies to high-lying bound states from time-dependent density-functional response theory: Characterization and correction of the time-dependent local density approximation ionization threshold. *J. Chem. Phys.* **1998**, *108*, 4439.
- (49) Furche, F.; Ahlrichs, R. Adiabatic time-dependent density functional methods for excited state properties. *J. Chem. Phys.* **2002**, *117*, 7433.
- (50) Grimme, S. Calculation of the electronic spectra of large molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 2004; Vol. 20, pp 153–218.
- (51) Dreuw, A.; Head-Gordon, M. Single-reference ab initio methods for the calculation of excited states of large molecules. *Chem. Rev.* **2005**, *105*, 4009.
- (52) Champagne, B.; Guillaume, M.; Zutterman, F. TDDFT investigation of the optical properties of cyanine dyes. *Chem. Phys. Lett.* **2006**, *425*, 105.

- (53) Miura, M.; Aoki, Y.; Champagne, B. Assessment of time-dependent density functional schemes for computing the oscillator strengths of benzene, phenol, aniline, and fluorobenzene. *J. Chem. Phys.* **2007**, *127*, 084103.
- (54) Magyar, R. J.; Tretiak, S. Dependence of spurious charge-transfer excited states on orbital exchange in TDDFT: Large molecules and clusters. *J. Chem. Theory Comput.* **2007**, *3*, 976.
- (55) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. Extensive TD-DFT Benchmark: Singlet-Excited States of Organic Molecules. *J. Chem. Theory Comput.* **2009**, *5*, 2420.
- (56) Tretiak, S.; Chernyak, V. Resonant nonlinear polarizabilities in the time-dependent density functional theory. *J. Chem. Phys.* **2003**, *119*, 8809.
- (57) Masunov, A.; Tretiak, S. Prediction of two-photon absorption properties for organic chromophores using time-dependent density-functional theory. *J. Phys. Chem. B.* **2003**, *108*, 899.
- (58) Kobko, N.; Masunov, A.; Tretiak, S. Calculations of the third-order nonlinear optical responses in push-pull chromophores with a time-dependent density functional theory. *Chem. Phys. Lett.* **2004**, *392*, 444.
- (59) Kauffman, J. F.; Turner, J. M.; Alabugin, I. V.; Breiner, B.; Kovalenko, S. V.; Badaeva, E. A.; Masunov, A.; Tretiak, S. Two-photon excitation of substituted enediynes. *J. Phys. Chem. A.* **2005**, *110*, 241.
- (60) Badaeva, E. A.; Timofeeva, T. V.; Masunov, A.; Tretiak, S. Role of donor-acceptor strengths and separation on the two-photon absorption response of cytotoxic dyes: A TD-DFT study. *J. Phys. Chem. A.* **2005**, *109*, 7276.
- (61) Clark, A. E. Time-Dependent Density Functional Theory Studies of the Photoswitching of the Two-Photon Absorption Spectra in Stilbene, Metacyclophenadiene, and Diarylethene Chromophores. *J. Phys. Chem. A* **2006**, *110*, 3790.
- (62) Wu, C.; Tretiak, S.; Chernyak, V. Y. Excited states and optical response of a donor-acceptor substituted polyene: A TD-DFT study. *Chem. Phys. Lett.* **2007**, *433*, 305.
- (63) Katan, C.; Charlot, M.; Mongin, O.; Le Droumaguet, C. I.; Jouikov, V.; Terenziani, F.; Badaeva, E.; Tretiak, S.; Blanchard-Desce, M. Simultaneous Control of Emission Localization and Two-Photon Absorption Efficiency in Dissymmetrical Chromophores. *J. Phys. Chem. B* **2010**, *114*, 3152.
- (64) Cross-shaped chromophores consisting of four donor substituted branches linked to an aromatic core have already been investigated but show significant differences, e.g., in OPA spectra, and can well be described as a dimer of corresponding quadrupolar benchmarks.³⁸
- (65) Ultrafast fluorescence spectroscopy has already been performed on a related carbon-cored tetramer.³⁹
- (66) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Pople, J. A. *Gaussian 98*, revision A.11; Gaussian, Inc.: Pittsburgh, PA, 2001.
- (67) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (68) Martin, R. L. Natural transition orbitals. *J. Chem. Phys.* **2003**, *118*, 4775.
- (69) Kokalj, A. Computer graphics and graphical user interfaces as tools in simulations of matter at the atomic scale. *Comput. Mater. Sci.* **2003**, *28*, 155.
- (70) Mongin, O.; Porres, L.; Charlot, M.; Katan, C.; Blanchard-Desce, M. Synthesis, fluorescence, and two-photon absorption of a series of elongated rodlike and banana-shaped quadrupolar fluorophores: A comprehensive study of structure-property relationships. *Chem.—Eur. J.* **2007**, *13*, 1481.
- (71) Lu, D.; Chen, G.; Perry, J. W.; Goddard, W. A. Valence-Bond Charge-Transfer Model For Nonlinear-Optical Properties Of Charge-Transfer Organic-Molecules. *J. Am. Chem. Soc.* **1994**, *116*, 10679.
- (72) Barzoukas, M.; Blanchard-Desce, M. Molecular engineering of push-pull dipolar and quadrupolar molecules for two-photon absorption: A multivalence-bond states approach. *J. Chem. Phys.* **2000**, *113*, 3951.
- (73) Terenziani, F.; Sissa, C.; Painelli, A. Symmetry breaking in octupolar chromophores: Solvatochromism and electroabsorption. *J. Phys. Chem. B* **2008**, *112*, 5079.
- (74) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380.
- (75) Basavoju, S.; Aitipamula, S.; Desiraju, G. R. Host-guest and network structures of some tetraphenylmethane derivatives. *CrystEngComm.* **2004**, *6*, 120.
- (76) Oldham, W. J.; Lachicotte, R. J.; Bazan, G. C. Synthesis, spectroscopy, and morphology of tetrastilbenoidmethanes. *J. Am. Chem. Soc.* **1998**, *120*, 2987.
- (77) Porres, L.; Mongin, O.; Katan, C.; Charlot, M.; Pons, T.; Mertz, J.; Blanchard-Desce, M. Enhanced two-photon absorption with novel octupolar propeller-shaped fluorophores derived from triphenylamine. *Org. Lett.* **2004**, *6*, 47.

Computing Relative Free Energies of Solvation Using Single Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange

Ilya V. Khavrutskii* and Anders Wallqvist

Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland 21702, United States

Received June 15, 2010

Abstract: This paper introduces an efficient single-topology variant of Thermodynamic Integration (TI) for computing relative transformation free energies in a series of molecules with respect to a single reference state. The presented TI variant that we refer to as Single-Reference TI (SR-TI) combines well-established molecular simulation methodologies into a practical computational tool. Augmented with Hamiltonian Replica Exchange (HREX), the SR-TI variant can deliver enhanced sampling in select degrees of freedom. The utility of the SR-TI variant is demonstrated in calculations of relative solvation free energies for a series of benzene derivatives with increasing complexity. Of note, the SR-TI variant with the HREX option provides converged results in a challenging case of an amide molecule with a high (13–15 kcal/mol) barrier for internal cis/trans interconversion using simulation times of only 1 to 4 ns.

A. Introduction

The ability to predict solvation free energies for a series of small molecules is important in drug design.^{1,2} It allows for an assessment of solvation and, in general, partition properties of novel molecules before their syntheses. Furthermore, solvation free energy is an essential ingredient for predicting proton affinities of small molecules and their binding affinities to biomolecular drug targets in water.^{2–6} Therefore, computational tools that provide accurate solvation free energies are indispensable in drug design, particularly in the lead optimization stages where molecules of interest are not immediately available for experimental evaluation.

Sustainable high quality predictions of solvation free energies require rigorous computational methods that can properly account for explicit ligand–solvent interactions and ligand flexibility.^{4,7–12} Although various empirical methods exist that have been tuned to reproduce solvation free energies of available compounds, the quality of their predictions decays quickly for novel, flexible molecules.^{10,11} Therefore, more rigorous methods are required in such

cases.^{4,12–20} Currently, the most rigorous methods are Thermodynamic Integration (TI) and the closely related Free Energy Perturbation (FEP).^{2,5,6,21–30} Because of the computational equivalence of TI and FEP methods, we limit ourselves to TI in this paper. While the TI method incorporates ligand flexibility via Molecular Dynamics (MD) or Monte Carlo (MC) simulations, it inherits sampling issues associated with these simulation methods.^{23,31–33} These sampling issues become particularly severe in systems with hindered conformational transitions, often rendering the results of TI calculations unreliable.^{20,23,31–35}

Proper accounting for ligand flexibility requires an enhanced sampling technique. Previous attempts to achieve enhanced sampling have produced a multitude of sophisticated computational methods, such as conformational flooding,^{36,37} hyper dynamics,^{38–40} accelerated molecular dynamics,^{41–45} simulated scaling,^{46–48} and generalized ensemble methods such as temperature and Hamiltonian replica exchange methods.^{20,49–60} In principle, all of these methods could be coupled with TI to compute solvation free energies. However, it is not clear which combination would give the most accurate results most efficiently.

* To whom correspondence should be addressed. E-mail: ikhavrutskii@bioanalysis.org.

A combination of TI with Hamiltonian Replica Exchange (HREX) has been demonstrated to enhance sampling and improve convergence in solvation free energy calculations.^{58,59} This combination improves upon standard TI at the smallest expense and is, therefore, the most promising first step. The improvements are attributed to the so-called “Hamiltonian tunnel” that enriches otherwise isolated MD simulations with configurations from all of the other TI windows. Recently, the HREX-TI method has been further enhanced by introducing the double-tunneling scheme.³⁵ This latter work highlights the importance of the simulation setup and provides a way to accelerate specific hindered degrees of freedom. In addition, it demonstrates that the cost of the simulations could be further reduced via an optimal selection of TI windows. Enhanced sampling of the hindered degrees of freedom is achieved through the “Hamiltonian tunnel” that connects the molecules on the original, hindered potentials to their unhindered counterparts.^{20,35,56,57}

Computing relative as opposed to absolute solvation free energies could be more advantageous for enhanced sampling. To rank a series of molecules according to their solvation free energies, one can use either an absolute or a relative scale. For a ranking based on relative solvation free energies, one needs to relate all of the molecules from the series to a single reference state.^{61,62} This seeming disadvantage has prompted many researchers to turn to the absolute scale. We note, however, that relative TI simulations provide opportunities for incorporating procedures to select specific degrees of freedom for enhanced sampling.³⁵ For example, hindered internal rotations can greatly benefit from enhanced sampling.^{20,35,56,57} Such opportunities appear to be absent in absolute TI calculations. Therefore, while TI can yield both absolute and relative solvation free energies, the latter can provide additional sampling benefits.

Finally, the choice of topology appears critical to achieving the most efficient sampling with HREX-TI. Thus, Yang et al. demonstrated that dual topology TI simulations benefited greatly from HREX, whereas in the single topology case, the effect of replica exchange was small.³⁵ Hence, the authors concluded that dual topology TI would be the method of choice to combine with HREX for enhanced sampling.

In the present paper, we developed an approach to enhance sampling of hindered degrees of freedom in the single topology TI framework that was inspired by the single reference state extrapolation method.^{15,19} Thus, we remove the roadblock identified by Yang et al.³⁵ This new strategy allows for enhanced sampling of hindered degrees of freedom in a controlled way by choosing a reference state in which the desired degrees of freedom are sampled freely. The added flexibility of this approach comes from the fact that the reference state does not have to correspond to a real molecule as in standard TI.

This paper is organized as follows. We first briefly recap the conventional TI methodology and point out its drawbacks. We then present the Single Reference-TI (SR-TI) approach and highlight its features that help overcome the referred drawbacks. Following the computational details, we describe results of SR-TI simulations for a series of benzene derivatives. Specifically, we focused on *para*-substituted phenols,

hydroxylated benzenes, and aryl alcohols. Finally, we take on a real challenge—an amide molecule with a hindered internal rotation that separates its *cis* and *trans* isomers.⁶³ This difficult test case presents a sampling problem that can only be solved using SR-TI with the HREX option. This test case also validates the SR-TI approach by comparing the results of the regular and HREX SR-TI simulations against each other and against the amide bond torsional Potential of Mean Force (PMF). To the best of our knowledge, this is the first direct calculation of the amide hydration free energy that automatically accounts for *cis/trans* interconversion.

B. Methodology

B.1. General TI. TI methods are simulation methods concerned with computing free energy or reversible work for an alchemical transformation of a molecule A to a different molecule B using rigorous statistical mechanical principles.^{21,23,28,64} Depending on the topology of the alchemical system, TI methods can be divided into single and dual topology methods.^{65,66} Dual topology TI methods simultaneously propagate Hamiltonians (H_A and H_B) of both molecules coupled through their shared environment, whereas single topology methods create a hybrid molecule, a union of the molecules A and B, and propagate the resulting single Hamiltonian (H_{AB}). While both approaches have their advantages and disadvantages, in this paper, we will focus on single topology TI methods.

To better understand the chemistry behind TI simulations, it is useful to consider the potential energy function V that along with kinetic energy K comprises the Hamiltonian of the system ($H = K + V$). For a system of N atoms described by a configuration R with $3N$ coordinates, its potential energy depends on the identity of the atoms and the way they are connected by covalent bonds. Typically, TI methods use Molecular Mechanical (MM) potentials that have assigned atom types and do not break covalent bonds. Therefore, an MM potential can be expressed as a simple sum of bonded and nonbonded terms. The general form of all of these latter terms is well-known,^{67–69} whereas the fine details of each term depend on the types of atoms involved.^{70,71}

In brief, bonded terms include harmonic terms for stretching, V^{bond} (two atoms with one bond), and bending, V^{angle} (three atoms with two bonds), as well as an anharmonic, torsional term for twisting, V^{torsion} (four atoms with three bonds in a chain). In some cases additional harmonic, improper torsional terms, V^{improper} (four atoms with three bonds sharing one of the atoms), are added. These latter terms enforce a particular configuration of the shared atom with respect to the plane formed by the other atoms. Hence, improper torsions can be used to enforce planarity or a particular chirality of the four-atom group. On the other hand, the anharmonic, torsional terms are essential in representing different conformations of the molecule separated by corresponding barriers. Because the bonded terms cannot act through space, the MM potential is completed by adding nonbonded terms. The essential nonbonded terms include pairwise, long-range electrostatic, V^{Coulomb} , and short-range

Lennard-Jones (LJ), V^{LJ} , potentials. The LJ terms among other things ensure that atoms of the opposite charge do not collapse on top of each other. This level of description is more than sufficient to understand the alchemical transformations behind the TI method and their important consequences.

In single topology TI, most common alchemical transformations can be represented by two simple operations, namely, “atom substitution” and “atom creation/annihilation.” Both transformations modify the MM potential terms, but to a different extent. Atom substitution transforms one atom type into another, by changing all bonded and nonbonded parameters correspondingly. This transformation proportionally modifies all of the associated bonded terms, including the anharmonic torsional terms. In contrast, atom annihilation completely voids the anharmonic torsional terms associated with the annihilated atom but does not modify the harmonic terms for stretching and bending.^{65,66} The improper torsional terms that involve the annihilated atoms can be handled either way depending on the situation. In addition, atom annihilation “switches off” all of the corresponding nonbonded terms. Therefore, the atoms that are “switched off” remain attached to the molecule via their original harmonic interactions but no longer interact with the other atoms through space. Consequently, atom creation requires that the atom be present in the hybrid molecule in the “switched off” state. The “switched off” atoms are often referred to as *dummy* atoms, as opposed to *real* atoms. Needless to say, when dummy atoms are involved with the hybrid molecule, standard rules of valency do not have to apply.

The reversible work for the alchemical transformation is derived by following the changes in the MM potential of the hybrid system. During the alchemical transformation from molecule A to molecule B, the total number of atoms in the system, including the dummy atoms, does not change. However, their contributions to the MM potential energy do change depending on whether the atoms are being “switched on/off” or substituted. Throughout the alchemical transformation, the hybrid potential V_{AB} remains well-defined. Therefore, the associated reversible work can be computed. This reversible work formally corresponds to changing the potential of the system V_{AB} from the state that represents the original molecule A to the state that corresponds to molecule B.

The efficiency of computing the reversible work associated with the alchemical transformation depends on the way the hybrid Hamiltonian is constructed. Typically, the hybrid Hamiltonian of the system (H_{AB}) that includes the potential V_{AB} is linearly interpolated between the end points, molecules A and B, using a coupling parameter λ :

$$H_{\text{AB}}(\lambda) = (1 - \lambda)H_{\text{A}} + \lambda H_{\text{B}} \quad (1)$$

The coupling parameter in TI spans an interval from $\lambda = 0.0$ (molecule A) to $\lambda = 1.0$ (molecule B) and is an analogue of the reaction coordinate. However, linear interpolation of Coulomb electrostatic and LJ nonbonded interactions presents a sampling problem in cases with dummy atoms, known as the end-point catastrophe.^{72,73} To alleviate this problem, several approaches have been proposed.^{16,72–75} Among those, we find the use of soft-core potentials to be the most suited

for our purposes. Instead of going to infinity at zero interaction distances (when atoms collapse on top of each other), soft-core potentials take on finite values due to built-in distance offsets. It is worth mentioning that although the approach introduced by Yang et al. employed the soft-core potentials with the HREX option, it could also work without them.³⁵ Hence, this approach could be regarded as an alternative solution to the end-point catastrophe problem.

In the present work, we employ a free energy integration procedure that requires a finite potential and its derivatives and hence necessitates the use of soft-core potentials. The available soft-core potentials also differ in styles. In the present work, we employ the original GROMOS style soft-core (SC) potentials as implemented in GROMACS.^{37,73,76}

$$V_{\text{AB}}^{\text{sc}}(r) = (1 - \lambda)V_{\text{A}}(R_{\text{A}}(r, \lambda)) + \lambda V_{\text{B}}(R_{\text{B}}(r, \lambda)) \quad (2)$$

$$R_{\text{A}}(r, \lambda) = (\alpha\sigma_{\text{A}}^6\lambda^p + r^6)^{1/6} \quad (3)$$

$$R_{\text{B}}(r, \lambda) = (\alpha\sigma_{\text{B}}^6(1 - \lambda)^p + r^6)^{1/6} \quad (4)$$

where $p = 2$, r is the distance between a given pair of atoms, α is the soft core parameter, and σ is the radius of interaction computed from LJ parameters. Although different strategies combine electrostatic and LJ soft-core transformations differently, sometimes completely decoupling the two,^{14,75,77–80} a simultaneous change of both potentials should reduce the computational burden.

Given all of the Hamiltonian interpolation rules underlying a particular alchemical transformation, we can begin collecting necessary statistics for computing the corresponding reversible work. One of the most efficient procedures to do that is the multiconfigurational TI.^{23,64} In this method, the Hamiltonian derivatives with respect to λ are collected in M independent MD or MC simulation windows. The λ values of these windows span the interval $[0; 1]$. Thus, each i th window corresponds to the hybrid Hamiltonian with a fixed value of λ_i , and its simulation provides the ensemble averaged Hamiltonian derivatives, $\langle(\partial H_{\text{AB}})/(\partial \lambda)|_{\lambda_i}\rangle$. Integration of the mean Hamiltonian derivatives from all the windows yields the reversible work.

B.2. Fourier Bead Integration. The reversible work can be computed from the multiconfigurational TI data in many different ways. Methods such as the Weighted Histogram Analysis Method (WHAM),^{81–84} Bennett’s Acceptance Ratio (BAR),^{16,85–87} or simple line integration have been used most frequently. Recently, overlap histogramming³⁵ was used to integrate the results of TI simulations where the distributions of the Hamiltonian derivatives in two adjacent windows overlap. Here, we employ a Fourier beads^{88–90} variant of the line integration procedure suggested by Bennett.⁸⁵ Single topology TI simulations have been noted to provide faster convergence for the mean Hamiltonian derivatives than dual topology ones,³⁵ thus justifying the use of the line integration approach. In this method, we do not need to have the histograms of the adjacent windows overlap. However, we require that changes in the mean Hamiltonian derivatives are sufficiently smooth that they could be interpolated between all of the windows. We caution that in the absence

of the histogram overlap, TI simulations with the HREX option will become inefficient.^{35,57–59,91}

The adaptation of the Fourier bead integration procedure to TI is straightforward. The Fourier interpolated values of the Hamiltonian derivatives and of the corresponding λ values are represented as continuous functions of a parameter ξ defined on the interval [0; 1]:

$$\tilde{X}(\xi) = \langle X|_{\lambda=0} \rangle + (\langle X|_{\lambda=1} \rangle - \langle X|_{\lambda=0} \rangle)\xi + \sum_{k=1}^M a_k \sin(k\pi\xi) \quad (5)$$

Here, X refers to either the derivative of the Hamiltonian or the corresponding λ value and a_k 's are the Fourier amplitudes. The brackets indicate ensemble averaging, and “ \sim ” indicates the continuous function representing the interpolated ensemble-averaged values. Note that we only need to interpolate the values of the coupling parameter λ if the TI windows are not distributed evenly on the interval [0; 1]. The corresponding Fourier interpolation amplitudes are derived by the discrete Fourier transform procedure using the simulation data from all M windows:

$$a_k = 2 \sum_{l=1}^M [\langle X|_{\lambda=\lambda_l} \rangle - \langle X|_{\lambda=0} \rangle - (\langle X|_{\lambda=1} \rangle - \langle X|_{\lambda=0} \rangle)\xi_l] \sin(l\pi\xi_l)\Delta\xi \quad (6)$$

where the set $\{\xi_l | l = \overline{1, M}\}$ forms a uniform grid by construction:

$$\xi_l = \frac{l-1}{M-1} \quad (7)$$

$$\Delta\xi = \frac{1}{M-1} \quad (8)$$

Using the interpolated values, we can take all of the required derivatives analytically and then compute the work via the line integral, also, in principle, analytically. In practice, we use a fine grid to evaluate the integral using a simple trapezoidal rule.

$$W(\xi) = \int_0^\xi \left\langle \frac{\partial \tilde{H}_{AB}}{\partial \lambda} \right\rangle_{\lambda(\xi')} \frac{d\lambda(\xi')}{d\xi'} d\xi' \quad (9)$$

We stress that all of the Fourier beads methodology can be applied to any multiconfigurational TI calculations that collect the mean derivatives of the hybrid Hamiltonian with respect to the coupling parameter and can be trivially generalized to multiple coupling parameters.^{88–90}

B.3. Drawbacks of Conventional TI. Multiconfigurational TI with MD simulations, like other methods employing independent windows, suffers an inherent drawback that can deteriorate the quality of the computed free energy estimates. Because of the random nature of the thermal fluctuations, certain rare transitions can occur in some windows but not others, creating incoherence across the windows over time. With different windows exploring nonoverlapping regions of configurational space, the final free energy estimates might get skewed.^{23,31–34,64,77,92–95} Another related issue that affects the results of the multiconfigurational TI is the bias from

the initial configuration that causes the simulation to explore only a limited region of the configurational space near the starting configuration. Both of these issues are due to sampling limitations that stem from the inability of MD simulations to consistently overcome large barriers ($\gg k_B T$) that correspond to rare events. Because their effect on the quality of the final free energy estimates can be significant, they need to be addressed. Although different solutions have been proposed to tackle some of these issues,^{31,41,42,55–60,94–96} we believe that HREX-TI is one of the most cost-effective improvements of TI methodologies in general.

B.4. Single Reference-TI Approach. The SR-TI approach presented here computes relative solvation free energies for a series of related molecules. It has been developed to aid lead optimization efforts in drug design. The SR-TI approach is a variant of single topology TI methods and has been inspired by the Single Reference State Extrapolation (SRSE) method.^{15,18,19} The SRSE method attempts to compute free energies for an alchemical transformation of a series of related molecules A to a common reference state B by using only a single simulation of the reference state B . The reference state B , by construction, includes all of the atoms necessary to form any molecule A in the series. Hence, the alchemical transformation free energies are derived by evaluating the differences in the MM potential of the real and reference states over the configurations sampled by the reference state only. Computationally, this is roughly equivalent to running a single TI window for the series of molecules and is, therefore, very economical. Unfortunately, the accuracy of the SRSE method depends on the overlap between the configurations of the reference simulation and those of the real state. This makes the choice of the reference state complicated and the results approximate. The SR-TI approach naturally avoids these issues albeit at a higher cost.

In a nutshell, SR-TI computes free energies or reversible works for the alchemical transformations of each molecule A in a series down to a common reference molecule B . This is similar to previous TI calculations using a common reference state.^{18,61,62} The fact that the reference molecule B is a single substructure of all of the molecules A in the series greatly simplifies the chemistry in SR-TI. Thus, SR-TI needs to annihilate atoms of molecule A that extend beyond the common reference structure B and substitute their shared atoms to match atom types to those within the reference. This is illustrated in Figure 1 with the example of pyridine and toluene as a series of two molecules. Thus, the SR-TI approach never needs to create new atoms. The SR-TI approach makes comparing different molecules A from a given series simple. Importantly, unlike in earlier TI simulations using a common reference,^{18,61,62} in SR-TI, reference molecule B does not need to represent a real molecule. For example, in Figure 1, reference molecule B comprises only the heavy atoms of the benzene. Once the common core structure for the series has been determined, the hybrid topology and parameter files for each molecule A can be set up independently. Nevertheless, because the real atoms (but not necessarily the dummy atoms) in the common reference B are the same across the series, any

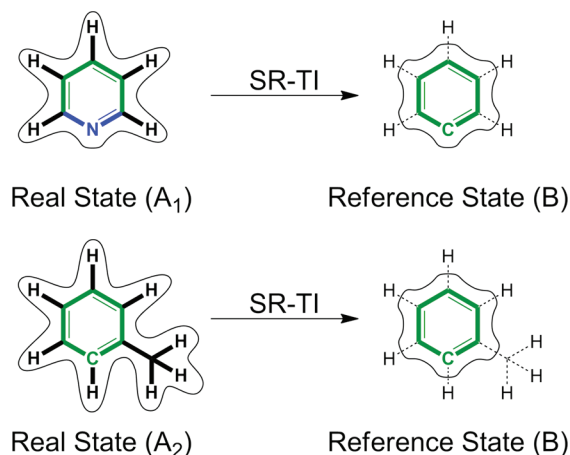


Figure 1. Two kinds of alchemical operations and a single reference state in SR-TI. The top pane represents the transformation of pyridine (A_1) to a benzene core reference state (B), by “atom substitution” of the nitrogen atom and “atom annihilation” of the hydrogen atoms. The bottom pane shows another transformation of toluene (A_2) to the same reference state (B) with only “atom annihilations.” In this latter case, the internal rotation of the methyl group becomes enhanced in the reference state. The common reference substructure is highlighted in bold green. The atoms and bonds that are shown in bold correspond to the real atoms, whereas regular labels with thin dashed bonds represent the dummy atoms. The contours surrounding molecules represent molecular volume, which is always reduced in the reference state. We emphasize that the reference states on the right are considered identical, despite the differences in the dummy atoms.

molecule A can be conveniently compared to all of the other molecules from the series. Therefore, the SR-TI approach can be used to study differences between stereoisomers, including *cis/trans* isomers and enantiomers.

Besides the computational convenience, SR-TI can overcome the bias of the initial configuration inherent in conventional TI methods. In SR-TI, the volume of molecule A always reduces down to that of the common reference structure B (see Figure 1). At the same time, it is possible to also reduce the complexity of the reference state and its overall polarity. As a result, the reference end state can sample confined spaces, such as solvent cages, more efficiently than the real end state. In addition, if reference molecule B were less polar than molecule A, it could escape traps due to specific interactions with the solvent. Last but not least, the TI windows that are closer to reference state B enjoy enhanced sampling of the torsional degrees of freedom, which involve the dummy atoms (recall that these terms become void in the reference molecule). Therefore, by appropriately choosing the reference substructure, one can enhance sampling not only of the orientational, but also of the specific torsional degrees of freedom. Although, these sampling benefits can remove the initial configuration bias and ease the overlap issues at or near the reference end state, they do not apply to the real end state or molecule A automatically.

B.5. Single Reference–TI with Hamiltonian Replica Exchange. With an appropriate technology, SR-TI can achieve enhanced sampling across all TI windows. We have

noted that SR-TI naturally achieves enhanced sampling in the windows at and near the reference state. To enhance sampling in all TI windows, we can invoke exchanges of configurations between the SR-TI windows by employing Hamiltonian replica exchange (HREX).^{20,53,55–60,96} Note that unphysical states have been used to enhance sampling of hindered degrees of freedom with the help of HREX before.^{56,57} Although HREX has been proposed⁵³ and applied^{35,57–59,96} in combination with conventional TI previously, the benefits are most pronounced in dual topology HREX-TI.³⁵

In the case of the NPT ensemble, for example, where all of the TI windows are run at the same temperature T and pressure P , the HREX option can be implemented as follows. Consider a vertical excitation for an i th window with configuration R_i from the Hamiltonian H_{AB} on the alchemical energy surface V_{AB} at λ_i to that at λ_j :

$$\Delta_{ij}(R_i) = \beta(V_{AB}(R_i, \lambda_j) - V_{AB}(R_i, \lambda_i)) \quad (10)$$

where $\beta = 1/k_B T$ is the inverse temperature and k_B is the Boltzmann constant. Now consider the energy change upon exchange of the two adjacent windows i and j . The total change in the energy of the generalized ensemble upon the Hamiltonian exchange between two configurations R_i and R_j is as follows:

$$\Delta\Delta = \Delta_{ij}(R_i) + \Delta_{ji}(R_j) \quad (11)$$

and the final Metropolis acceptance criterion is

$$W(R_i \leftrightarrow R_j) = \begin{cases} 1, & \text{for } \Delta\Delta \leq 0 \\ \exp(-\Delta\Delta), & \text{for } \Delta\Delta > 0 \end{cases} \quad (12)$$

Here, W is the probability of the exchange. Although different exchange protocols exist that vary in efficiency,^{53,54,91,97–100} we follow the standard exchange protocol developed previously for temperature replica exchange (TREX).^{53,54,97} Specifically, we perform HREX by alternating exchange attempts for all odd and all even pairs of replicas. Odd pairs are replicas $i = 2n - 1$ and $j = 2n$, and even pairs are replicas $i = 2n$ and $j = 2n + 1$, where n starts at 1 with the largest index not to exceed the total number of TI windows M . The exchanges are attempted at regular, predetermined time intervals. Best of all, because HREX maintains the original ensembles within each window, the methodology for computing the reversible work remains unchanged. All we need to do is follow λ values through all of the exchanges. Thus, the only additional cost associated with evaluation of the Metropolis exchange criteria is well compensated for by the benefits of the enhanced sampling.

The combination of HREX with single-topology SR-TI addresses both the initial bias and the overlap drawbacks of conventional TI methods in much the same way as the previously developed dual topology HREX-TI variant.³⁵ Conveniently, SR-TI provides flexibility in partitioning a given molecule between reference and dummy atoms that grants control over hindered torsional degrees of freedom, molecular volume, and polarity of the system. The HREX option creates a flow of replicas across the Hamiltonian space of the whole alchemical transformation. Such a flow of

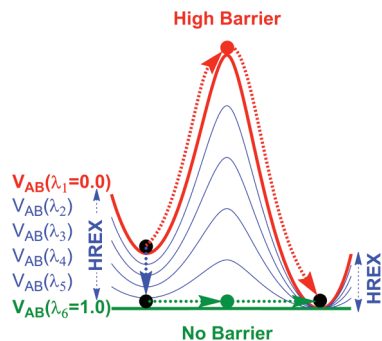


Figure 2. An illustration of the “Hamiltonian tunnel” opened by the HREX option during SR-TI calculations. The potential energy surface at $\lambda = 0$ has an insurmountable barrier, which disappears completely in the potential at $\lambda = 1.0$. The red and green dashed arrows represent regular transitions on the same surface through corresponding transition states. The blue arrows represent the Hamiltonian replica exchanges that allow the system to hop between the surfaces. Instead of going over a high barrier on the $\lambda = 0$ surface, the system arrives at the surface with the lower barrier, say at $\lambda = 1$ through HREX, crosses over to the other minimum, and is then brought back to the $\lambda = 0$ surface on the other side of the barrier.

replicas has two main benefits. First, it opens a “Hamiltonian tunnel” (see Figure 2) that allows consistent crossing of high energy barriers over a short period of time that would have been impossible otherwise.^{46,48,56} Second, the flow of replicas mixes configurations from different TI windows and thus provides superb overlap in configurational space. Ultimately, HREX SR-TI is an improvement over regular SR-TI and permits calculations of high quality solvation free energies. In tandem, regular and HREX SR-TI calculations can be used as a tool to identify sampling issues.

The HREX option should be applied with caution to SR-TI calculations involving chiral atoms outside the common reference structure, as the chiral atoms can invert their configurations in the dummy state. However, this issue may be avoided if the chirality is maintained through harmonic, improper torsional terms.

C. Computational Details

C.1. Small Molecule Parameters. We obtained initial coordinates of the small molecules from their SMILES^{101,102} using a program called CORINA.¹⁰³ An exhaustive conformational search was attempted using a companion program ROTATE,¹⁰⁴ followed by structural refinement with the semiempirical AM1 method,¹⁰⁵ as implemented in open source MOPAC7, version 1.11.¹⁰⁶ Where multiple conformations existed, the AM1 partial charges for each conformation were symmetrized across the equivalent atoms and combined into a conformation-independent set of charges using Boltzmann weighting by the AM1 energies at the target temperature of 300 K with appropriate degeneracies. Finally, the AM1 charges were augmented through the BCC procedure^{107,108} as implemented in the Antechamber program^{109,110} from Amber Tools, version 1.2. The resulting conformation-independent, properly symmetrized set of AM1BCC charges is expected to reproduce HF/6-31G*

RESP charges to a good approximation. The remainder of the parameters, including vdW and bonded terms, were generated in an automated fashion by the Antechamber program^{109,110} to comply with the GAFF force field.⁷¹

C.2. Single Reference State. The choice of the reference state for SR-TI simulations is flexible. Recall that the single reference state does not have to correspond to a real molecule. The size of the reference state cannot exceed that of the largest common substructure and cannot be less than one atom from that substructure. For computational convenience, the total charge of the reference state should be an integer. To derive parameters for the reference state, we can employ parameters of its nearest real molecule. We do that by turning the atoms used to complete the reference structure to the nearest molecule (preferably hydrogens) into dummy atoms and adding their partial charges to those of the nearest real atoms of the common substructure. This procedure is similar to the one used in AutoDock4.0 to derive united-atom parameters from all-atom ones.^{111–113} For the molecules studied in the present work, we chose the benzene substructure, comprising only six heavy atoms, as the reference state (see Figure 1).

C.3. SR-TI Setup. The alchemical transformation turns all the atoms of the original molecules that are outside the benzene core reference state into dummy atoms by altering their partial charges and vdW parameters simultaneously. To avoid the end-point catastrophe, we employ GROMOS style soft-core electrostatic and LJ potentials⁷³ as implemented in GROMACS.^{76,114–116} Specifically, we employ $p = 2$, the soft-core parameter $\alpha = 1.5$, and the radius of interaction $\sigma = 0.3$ (see eqs 2–4 and the GROMACS 4.0 manual for a description¹¹⁷). Thus, the SR-TI simulations in this paper assess the reversible works needed to alchemically change each molecule to the benzene core reference state.

To automate the alchemical transformation setup using molecular mechanical Amber 99SB⁷⁰ and GAFF force fields⁷¹ in GROMACS, we developed an in-house PERL script by augmenting an existing script used to convert Amber parameter and topology files to GROMACS format.⁹⁵

C.4. MD Simulations. All of the simulations were run using a single precision version of GROMACS. Water solution was modeled with an explicit TIP3P water model using a cubic simulation box that extended at least 10 Å beyond the solute molecule. The box was prepared using the LEAP module from Amber Tools, version 1.2.¹¹⁸ For water simulations, prior to production runs in the NPT ensemble at $T = 300$ K and $P = 1$ atm, each system was first minimized and then equilibrated in a series of runs with gradually vanishing harmonic restraints on all of the atoms of the solute. The equilibration protocol involved 500 steps of steepest descent minimization with a force constant, $fc = 100$ kJ/mol/nm², followed by 5000 steps of NVT run with $fc = 100$ kJ/mol/nm², followed by a series of NPT runs of 10 000 steps with fc progressing as 100, 10, 1, 0.1, 0.01, and 0.0 kJ/mol/nm². The whole equilibration procedure totaled 65 000 steps or 130 ps. The production run employed a Langevin thermostat and Berendsen barostat,^{114–117} with identical collision frequencies of 2 ps⁻¹. For gas phase

simulations, the equilibration procedure in the canonical ensemble at $T = 300$ K was performed in a way similar to that in water, but without any harmonic restraints.

Throughout the simulations, all of the bonds containing hydrogen atoms were constrained using LINCS,¹¹⁹ and the integration time step was set to 2 fs. To compute nonbonded interactions in water, we employed periodic boundary conditions (PBC) with particle mesh Ewald for electrostatics^{114–117} using a 1 nm real space cutoff, while switching van der Waals interactions off over the range between 0.8 and 0.9 nm. In the gas phase, no PBCs were used, and all of the interactions were computed explicitly without any cutoffs. Unless stated otherwise, all of the simulations have been repeated two times with different random seeds to generate initial velocities. Typically, production runs were 1-ns-long, but in some cases the runs were extended to 4 ns per window. The coordinates of the system were recorded every 1000 steps.

C.5. Regular SR-TI Simulations. To obtain the alchemical free energies or reversible works, the TI procedure split the interval from the real state of the molecule at $\lambda = 0$ to the reference state at $\lambda = 1$ into M windows, separated by equal λ intervals. The majority of work was done with $M = 12$ windows, but in some cases additional simulations were performed with $M = 23$ windows. For each molecule, all TI windows had identical starting configurations. For each window, we recorded $(\partial V)/(\partial \lambda)$ values at every time step. The mean values $\langle (\partial V)/(\partial \lambda) \rangle$ for all of the windows were assembled into the final work using the Fourier beads integration procedure described in the Methodology section. The standard deviations were calculated from two independent runs. The differences between the work values in gas and water phases yielded the relative hydration free energies with respect to the reference state.

C.6. HREX SR-TI Simulations. To use the HREX option, we have developed a PERL script interfaced with GROMACS. Unless otherwise stated, replica exchanges were attempted every 1000 MD steps or 2 ps. For the majority of simulations, we employed a total of 500 exchange attempts over 1 ns simulation time. In special cases, the number of exchanges was increased to 2000, extending the simulation time to 4 ns. The exchanges were attempted using a standard procedure described in the Methodology section. Upon acceptance, only λ values were exchanged between windows, while keeping the coordinates and velocities in place to expedite restarting of the simulations. Following the exchange attempts, simulations in all of the windows were restarted with different random seeds provided by the PERL random number generator to reinitialize the Langevin dynamics and avoid possible random seed artifacts.¹²⁰ All of the other simulation details were the same as with regular SR-TI simulations.

C.7. Torsional PMF. To get the PMF for the hindered amide bond rotation, we performed umbrella sampling simulations¹²¹ using 72 equally spaced windows to span the range from -180 to $+180^\circ$. We employed the harmonic biasing potential with a force constant of 2000 KJ/mol/rad². All of the equilibration and production protocols were identical to those in the regular SR-TI approach. The

coordinates were saved every 50 steps during the 4 ns simulation time. The results of the simulations were unbiased and reconstructed into the final PMF using two independent methods. Specifically, we applied WHAM^{83,84,122} and the HFB^{88–90} method to the same data set, without regard to periodicity. Both methods gave nearly identical PMFs.

C.8. *cis/trans* Ratio from the HREX SR-TI Simulations. To compute the *cis/trans* ratios for the amide molecule from the HREX SR-TI simulations, we developed an additional PERL script that assembled a full length simulation trajectory from the individual short pieces produced by HREX SR-TI between exchange attempts for the specified λ value. In the present paper, we only used the trajectory for the real state of the molecule ($\lambda = 0.0$). The values of the dihedral angle were then extracted using the TRJCONV and G_ANGLE tools from GROMACS.^{114–116} The configurations with the dihedral angles in the range between -100 and $+100^\circ$ were considered *cis*, and all the other configurations were considered *trans*.

D. Results and Discussion

D.1. *para*-Phenols. To test the SR-TI approach, we first computed hydration free energies for a series of *para*-phenols $p\text{-C}_6\text{H}_4(\text{OH})(\text{X})$, where $\text{X} = \text{H}, \text{F}, \text{Cl}, \text{Br}, \text{I}, \text{CH}_3, \text{C}_2\text{H}_5, \text{CN}$, and OCH_3 , along with benzene, for a total of 10 molecules. For this series, we did not employ the HREX option and used the benzene molecule without the hydrogen atoms as the common reference. Thus, for the *para*-phenols, the alchemical transformation annihilates all of the hydrogen atoms of the benzene ring, along with the OH and the X groups. The results are summarized in Table 1, along with experimental as well as computational data from other research groups for 7 out of the 10 molecules.

Table 1 demonstrates that SR-TI results are in good agreement with previous benchmark TI calculations.⁷⁹ The largest difference between the results of the two calculations is 0.37 kcal/mol. Our uncertainties (based on two independent simulations) are in general slightly higher than those from previous benchmark calculations. This is to be expected, as the bootstrap method used in the latter case is known to underestimate the uncertainties by a factor of 3.^{14,80} The exception is provided by *p*-ethylphenol ($\text{X} = \text{C}_2\text{H}_5$), which shows the largest uncertainty of 0.21 kcal/mol for the reversible work in water. Overall, SR-TI systematically overestimated hydration free energies compared to the previous TI benchmarks. Note that the referred benchmark calculations were done using NVT simulations, while annihilating each molecule as a whole (absolute scale). In contrast, the SR-TI simulations were performed in the NPT ensemble and annihilated only those parts of each molecule that extended beyond the common reference substructure. Given these differences, we find the agreement between our and previous benchmark simulations particularly satisfying.

Like previous benchmark calculations, SR-TI overestimates the experimental relative hydration free energies for *para*-phenols. The SR-TI predictions for the phenols with simple aliphatic substituents, namely, *p*-cresol ($\text{X} = \text{CH}_3$) and *p*-ethylphenol ($\text{X} = \text{CH}_2\text{CH}_3$), are the closest to the

Table 1. Alchemical and Relative Hydration Free Energies for a Series of *para*-Substituted Phenols^a

compound X,Y	SR-TI results				previous results	
	ΔG_G (SD)	ΔG_W (SD)	$\Delta\Delta G$ (SD)	$\Delta\Delta G^{\text{SR-TI}}$ (SD)	$\Delta\Delta G^E$	$\Delta\Delta G^{\text{TI}}$ (SD)
H,H	-8.04 (0.02)	-5.24 (0.01)	-2.80 (0.02)	0.00 (0.03)	0.00	0.00 (0.03)
H,OH	9.27 (0.04)	16.64 (0.08)	-7.37 (0.09)	-4.57 (0.09)	-5.75	-4.97 (0.04)
F,OH	7.52 (0.02)	14.29 (0.04)	-6.77 (0.05)	-3.97 (0.05)	-5.33	-4.29 (0.04)
Cl,OH	9.72 (0.02)	16.81 (0.01)	-7.09 (0.02)	-4.29 (0.03)	-6.17	-4.66 (0.03)
Br,OH	11.56 (0.01)	19.06 (0.01)	-7.51 (0.01)	-4.70 (0.03)	-6.27	-4.77 (0.03)
I,OH	11.46 (0.00)	18.50 (0.07)	-7.04 (0.07)	-4.24 (0.07)	N/A	N/A
CH ₃ ,OH	11.14 (0.01)	18.39 (0.06)	-7.25 (0.06)	-4.45 (0.07)	-5.27	-4.66 (0.03)
CH ₂ CH ₃ ,OH	9.90 (0.02)	17.06 (0.21)	-7.15 (0.21)	-4.35 (0.21)	-5.27	-4.38 (0.04)
CN,OH	7.99 (0.03)	17.33 (0.02)	-9.35 (0.04)	-6.55 (0.04)	-9.31	-6.91 (0.04)
OCH ₃ ,OH	8.44 (0.00)	17.41 (0.04)	-8.97 (0.04)	-6.17 (0.04)	N/A	N/A

^a Compound labels X,Y refer to the benzene substituent groups in the *para* position to each other. The free energy values are given in kcal/mol. The standard deviations (SDs) are computed on the basis of two independent simulations. The SR-TI simulations employed 12 windows run for 1 ns each in canonical (gas phase) and in NPT (water) ensembles at $T = 300$ K and $P = 1$ atm. ΔG_G and ΔG_W are the SR-TI work values for the alchemical transformation to the benzene core in the gas phase and water, respectively, and $\Delta\Delta G = \Delta G_G - \Delta G_W$ is the corresponding relative hydration free energy. The values $\Delta\Delta G^{\text{SR-TI}}$, $\Delta\Delta G^{\text{TI}}$, and $\Delta\Delta G^E$ are the hydration free energies from SR-TI, earlier TI calculations, and the experiment with respect to benzene. The corresponding absolute numbers for the reference benzene are -2.80, -0.70, and -0.86 kcal/mol.⁷⁹

Table 2. Alchemical and Relative Hydration Free Energies for Benzene and its Hydroxylated Derivatives^a

compound		M	EX_G [EX_W], %	ΔG_G (SD)	ΔG_W (SD)	$\Delta\Delta G$ (SD)	$\Delta\Delta G^{\text{SR-TI}}$ (SD)	$\Delta\Delta G^{\text{SRSE}}$
benzene		12		-8.04 (0.02)	-5.24 (0.01)	-2.80 (0.02)	0.00 (0.03)	0.00
	HREX	12	75 [34]	-8.04 (0.05)	-5.23 (0.01)	-2.82 (0.05)	-0.02 (0.05)	
		23		-8.07 (0.03)	-5.20 (0.03)	-2.87 (0.04)	-0.07 (0.05)	
phenol		12		9.27 (0.04)	16.64 (0.08)	-7.37 (0.09)	-4.57 (0.09)	-5.28
	HREX	12	74 [25]	9.25 (0.01)	16.68 (0.03)	-7.42 (0.03)	-4.62 (0.04)	
		23		9.26 (0.03)	16.90 (0.07)	-7.64 (0.08)	-4.84 (0.08)	
catechol		12		1.64 (0.07)	11.87 (0.11)	-10.23 (0.13)	-7.43 (0.13)	-7.00
	HREX	12	64 [21]	1.51 (0.14)	11.80 (0.07)	-10.28 (0.15)	-7.48 (0.15)	
		23		1.58 (0.15)	12.15 (0.08)	-10.57 (0.17)	-7.77 (0.17)	
pyrogallol		12		9.34 (0.05)	21.46 (0.09)	-12.12 (0.11)	-9.32 (0.11)	-9.61
	HREX	12	58 [19]	9.31 (0.07)	21.54 (0.03)	-12.23 (0.08)	-9.43 (0.08)	
		23		9.27 (0.06)	21.98 (0.03)	-12.71 (0.07)	-9.91 (0.07)	

^a The free energy values are given in kcal/mol. The standard deviations (SDs) are computed on the basis of two independent simulations. The SR-TI simulation employed M windows run for 1 ns each in canonical (gas phase) and in NPT (water) ensembles at $T = 300$ K and $P = 1$ atm. The HREX label indicates that the option was turned on during the simulations, in which case exchanges were attempted every 2 ps. EX_G and EX_W indicate the average acceptance ratio in the gas phase and water, respectively. ΔG_G and ΔG_W are the SR-TI work values for the alchemical transformation to the reference benzene core in the gas phase and water, respectively, and $\Delta\Delta G = \Delta G_G - \Delta G_W$ is the corresponding relative hydration free energy. $\Delta\Delta G^{\text{SR-TI}}$ and $\Delta\Delta G^{\text{SRSE}}$ are hydration free energies relative to benzene from the SR-TI approach and single reference state extrapolation (SRSE)¹⁸ method, respectively.

experimental values and overestimate the hydration free energy by 0.82 and 0.92 kcal/mol, respectively. The largest disagreement (of 2.76 kcal/mol) is found for *p*-cyanophenol.

D.2. Hydroxylated Benzenes. To validate the HREX SR-TI approach, we computed hydration free energies for a series of hydroxylated benzenes with and without the HREX option. We selected benzene, phenol, catechol (benzene-1,2-diol), and pyrogallol (benzene-1,2,3-triol) for this study following previously published work.¹⁸ For these four molecules, we used the same benzene core reference as above. We anticipated that for the benzene and phenol molecules, both of which have been studied in the previous series, the SR-TI approach with and without the HREX option should give identical results. The actual results are summarized in Table 2. For this series, we compare the results with the previous calculations employing the more approximate SRSE method that inspired the present work.¹⁸

Table 2 reveals that the regular SR-TI results correspond very well with the SRSE results. As expected, the discrepancies between SR-TI and the SRSE method are much larger than between the two TI approaches in the previous section. The largest discrepancy of 0.72 kcal/mol is found for the

phenol. Nevertheless, given the computational savings that the more approximate SRSE method provides over the SR-TI, this level of agreement can be considered satisfactory.

Table 2 further demonstrates good agreement of the results from the regular SR-TI approach with those from the HREX SR-TI approach. Thus, we find that the relative free energies computed with and without the HREX option are identical within the specified uncertainties for all four molecules.

Here, we assess the dependence of the computed work estimates on the number of windows used in SR-TI. For simplicity, we employ the regular SR-TI approach in this test. Up to this point, the results discussed were obtained using 12 TI windows. With the regular SR-TI, we can simply insert and simulate 11 new windows precisely between the 12 old windows. In this way, we obtain a TI simulation with a total of 23 equally spaced windows (see Table 2). As is seen from Table 2, while benzene has nearly identical hydration free energies, the other three molecules show significant differences with 12 and 23 windows. In particular, pyrogallol shows the largest difference of 0.59 kcal/mol. Generally, the hydration free energies computed with 23 windows are lower than those with 12 windows. Interest-

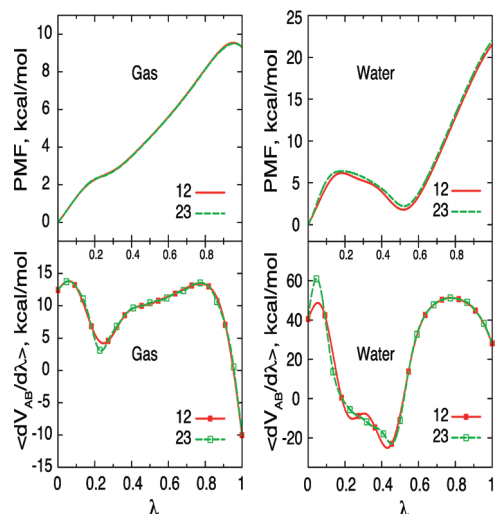


Figure 3. Integration of the TI data for pyrogallol using the Fourier beads method. The bottom panels show the actual simulation data as squares and the Fourier beads fit as lines for 12 and 23 windows for a regular SR-TI simulation in the gas phase and in water. The top panels show the corresponding potentials of mean force (PMFs) from the Fourier fitted curves with respect to the coupling parameter λ . The values of these PMF curves at $\lambda = 1.0$ give the corresponding reversible works for alchemical transformation of pyrogallol to the benzene core.

ingly, the gas phase work values are independent of the number of windows. Therefore, the water phase contributions create the observed disparity.

To understand the dependence of the computed work in water on the number of windows, we plotted the mean force $\langle \langle \partial V_{AB} / (\partial \lambda) \rangle \rangle_{\lambda_i}$ profiles, along with their integrals (see Supporting Information). For brevity, Figure 3 shows only the results for pyrogallol. As seen from Figure 3, the mean force peaks sharply near $\lambda = 0.045$. Going from pyrogallol down to catechol and then to phenol gradually lowers the peak, which vanishes completely at benzene (see Supporting Information). Therefore, this relatively sharp peak causes the 12-window interpolation to overestimate the relative hydration free energies.

Appearance of the sharp peak in the mean force profile near $\lambda = 0$ follows from the properties of the GROMOS soft-core potential used in this work.⁷³ This particular soft-core potential (see eqs 2–4) is known to create peaks in the $\langle \langle \partial V_{AB} / (\partial \lambda) \rangle \rangle_{\lambda_i}$ profile near $\lambda = 0.0$ due to the so-called “soft-core effect of hydrogens without the LJ interactions.”^{76,115} In the present case, the peak comes from the real hydrogen atoms of the hydroxyl groups. Clearly, this problem can be alleviated by introducing additional windows, as was done here. In addition, the peak size can be manipulated by reducing the value of the SC- σ parameter. Finally, the issue can be addressed by using a different form of the soft-core potential that peaks precisely at $\lambda = 0.0$.¹⁶ Although trivial to implement, this latter option requires further study and has not been pursued in this work.

With HREX SR-TI, as with any replica exchange method, it is useful to know the exchange rate. Table 2 provides the exchange acceptance ratios for both gas and water phase transformations. As expected, Table 2 demonstrates that the

acceptance ratio is significantly larger in the gas phase than in explicit water. However, we also notice a trend that the more atoms of molecule A need to be “switched off” to get to the reference state B, the lower the acceptance ratio becomes. Nevertheless, in the present case, the acceptance ratio does not affect the results significantly in either phase. The differences in free energies computed with the SR-TI approach with and without the HREX option suggest that regular MD achieves sufficient sampling for the molecules studied here.

D.3. Aryl-Alcohols. To further assess the HREX SR-TI approach for computing hydration free energies, we turned to a more complex set of molecules. Specifically, we looked at a series of aryl-alcohols that proved challenging for both SRSE and regular TI approaches.¹⁸ For completeness, we have considered terminal aryl-alcohols of the form $C_6H_5-(CH_2)_n-OH$ and their dimethylated analogues $C_6H_5-C(CH_3)_2-(CH_2)_{n-1}-OH$ where $n = 1, 2,$ and 3 . These molecules have additional torsional degrees of freedom that might benefit from the enhanced sampling of HREX SR-TI. To enhance sampling of these degrees of freedom, we once again used the benzene core as the reference. Armed with the results of the previous sections, we can use the differences between the values from SR-TI with and without the HREX option as a measure of nonergodicity stored in these torsional degrees of freedom. The results are summarized in Table 3 along with comparisons to the earlier calculations and experimental data where available.

We find that hydration free energies computed with SR-TI always fall within the error bars from the earlier SRSE extrapolated values. However, the error bars on the SRSE results are rather large in this case, reducing their predictive power compared to that of SR-TI. In addition, for the three linear molecules $C_6H_5-(CH_2)_n-OH$, our results compare favorably with previous TI calculations and experimental measurements.⁷⁹ In particular, we find that our calculations overestimate the experimental relative solvation free energies for the linear aryl-alcohols by at most 1.57 kcal/mol. Compared to previous TI benchmarks, SR-TI free energies are overestimated by at most 0.44 kcal/mol. In both cases, the discrepancy is systematic.

Similarly to hydroxylated benzenes, the hydration free energies for the aryl-alcohols computed with and without the HREX option are nearly identical. Analysis of the data for this series of molecules reinforces our previous observation that the exchange rate is inversely proportional to the number of atoms that is “switched off” within the series. The largest molecule with 22 out of 28 atoms “switched off” in the reference state exhibits the lowest acceptance ratio of 16% in water (see Table 3). Recall that we enhance sampling in all of the torsional degrees of freedom outside the benzene core of the molecules when using the HREX option. The lack of sizable differences between the results with and without the replica exchange suggests that regular sampling of the torsional degrees of freedom is sufficiently ergodic. Therefore, to clearly demonstrate the utility of the HREX option, we examined a molecule with a hindered torsional degree of freedom.

D.4. N-(2-Hydroxy-phenyl)formamide. To demonstrate the utility of the HREX option, we have investigated

Table 3. Alchemical and Relative Hydration Free Energies for a Series of Aryl Alcohols^a

compound	EX _G [EX _W], %	SR-TI results				previous results		
		ΔG _G (SD)	ΔG _W (SD)	ΔΔG (SD)	ΔΔG ^{SR-TI} (SD)	ΔΔG ^{SRSE} (SD)	ΔΔG ^E	ΔΔG ^{TI} (SD)
L1		-8.36 (0.01)	-1.29 (0.08)	-7.07 (0.08)	-4.27 (0.09)		-5.76	-4.71 (0.04)
HREX	71 [22]	-8.37 (0.06)	-1.26 (0.05)	-7.11 (0.08)	-4.31 (0.08)			
L2		-7.70 (0.04)	-0.46 (0.13)	-7.24 (0.14)	-4.44 (0.14)	-5.90 (1.84)	-5.93	-4.63 (0.04)
HREX	68 [21]	-7.70 (0.05)	-0.61 (0.11)	-7.09 (0.12)	-4.29 (0.12)			
L3		-6.52 (0.02)	0.77 (0.01)	-7.30 (0.02)	-4.49 (0.03)	-7.01 (3.09)	-6.06	-4.80 (0.04)
HREX	65 [19]	-6.47 (0.04)	0.83 (0.03)	-7.31 (0.05)	-4.50 (0.05)			
B1		5.35 (0.02)	12.23 (0.11)	-6.88 (0.11)	-4.08 (0.11)	-5.45 (5.40)		
HREX	63 [18]	5.35 (0.01)	12.17 (0.00)	-6.82 (0.01)	-4.02 (0.03)			
B2		-15.67 (0.12)	-8.86 (0.06)	-6.81 (0.14)	-4.01 (0.14)	-8.64 (6.04)		
HREX	57 [17]	-15.68 (0.03)	-8.97 (0.16)	-6.72 (0.16)	-3.91 (0.16)			
B3		-13.58 (0.11)	-6.39 (0.01)	-7.19 (0.11)	-4.39 (0.11)	-2.12 (5.61)		
HREX	52 [16]	-13.59 (0.01)	-6.33 (0.12)	-7.26 (0.12)	-4.46 (0.12)			

^a The free energy values are given in kcal/mol. In the compound column, Ln refers to linear C₆H₅-(CH₂)_n-OH, and Bn refers to their branched, dimethylated analogues C₆H₅-C(CH₃)₂(CH₂)_{n-1}-OH. The HREX label indicates that the option was turned on during the simulations, in which case exchanges were attempted every 2 ps. EX_G and EX_W indicate the average acceptance ratio in the gas phase and water, respectively. The standard deviations (SDs) are computed on the basis of two independent simulations. The SR-TI simulations employed 12 windows run for 1 ns each in canonical (gas phase) and in NPT (water) ensembles at *T* = 300 K and *P* = 1 atm. ΔG_G and ΔG_W are the SR-TI work values for the alchemical transformations to the reference benzene core in the gas phase and water, respectively, and ΔΔG = ΔG_G - ΔG_W is the corresponding relative hydration free energy. ΔΔG^{SR-TI}, ΔΔG^{SRSE}, ΔΔG^E, and ΔΔG^{TI} are relative hydration free energies with respect to benzene from the SR-TI, single reference state extrapolation (SRSE),¹⁸ experiment and earlier TI calculations, correspondingly.⁷⁹

molecules with an N-C amide bond. The rotation about the amide bond is so strongly hindered that amides are typically considered locked in one of the two conformations at *T* = 300 K, namely *cis* or *trans*.^{13,28,123-128} Of the two isomers, the *trans* isomer is considered the most favorable, and the *cis* isomer is often completely ignored.^{63,128-130} Earlier attempts to compute hydration free energies of some amides produced results conflicting with experimental data.^{13,123,125,127} Interestingly, for simple amides, the hydration free energies of the *cis* and *trans* isomers have been found experimentally to be nearly identical.¹²⁸ This argued that the preference for the *trans* isomer, which has important implications for peptide and protein structure in general, is not due to hydration. For the purposes of our study, we wanted to examine an amide with substantially different hydration free energies in the *cis* and *trans* conformations. Consequently, we placed a formamide group on the benzene ring and added a hydroxyl group in the *ortho* position allowing for intramolecular interactions. In what follows, we refer to the resulting N-(2-hydroxy-phenyl)formamide simply as “amide” for brevity.

An exhaustive conformational search for the amide identified five groups of isomers in the gas phase (see Supporting Information). Specifically, employing the AM1 semiempirical potential,¹⁰⁵ we found two groups for *trans* and three groups for *cis* conformations of the amide. All of the groups, except the first *trans* group, contained two iso-energetic conformations that were mirror reflections of each other. The *trans* isomer that did not have such a degeneracy was completely planar. Thus, we identified nine distinct conformations in total. Interestingly, the AM1 potential¹⁰⁵ ranked the first *cis* group to be the most favorable in the gas phase, while the second *trans* group that had a possibility for intramolecular hydrogen bonding was ranked the least favorable.

We used the information from all *cis* and *trans* conformations to derive a conformation-independent set of AM1BCC atomic charges^{107,108} as described in the Methodology section. Reoptimizing geometries of these conformations

using the conformation-independent charges with the GAFF molecular mechanical (MM) potential^{71,110} changed the ranking only slightly (see the Supporting Information). Most importantly, the MM potential strongly favored (by 3.0 kcal/mol) the second *trans* group of conformations with an intramolecular hydrogen bond. This group became the new ground state and was 1.3 kcal/mol lower than the lowest energy *cis* group. We did not attempt to make any empirical adjustments to correct for this behavior in the MM potential.

Using the MM potential derived above, we computed the reversible works or potentials of mean force (PMFs) for the amide bond torsion. The PMFs were computed from MD simulations with a relatively stiff harmonic biasing potential on the amide bond dihedral angle. To unbiased the results, we employed two independent approaches, namely Weighted Histogram Analysis Method (WHAM)^{83,84,122} and umbrella integration with Harmonic Fourier Beads (HFB).⁸⁸⁻⁹⁰ Figure 4 depicts the final PMFs in the gas and water phases for 360° rotation about the amide bond. In computing the PMFs, we ignored the periodicity and treated the data near -180 and near +180 independently. In what follows, we use the PMF extrema, i.e., the local minima and maxima (see the Supporting Information), instead of introducing *cis* and *trans* indicator functions,⁹⁴ to make quantitative comparisons. The resulting asymmetry in the PMFs provides a measure of uncertainty, which is between 0.3 and 0.5 kcal/mol. Importantly, the PMFs confirm that the rotation about the amide bond is strongly hindered with barriers ranging between 13.3 and 15.0 kcal/mol. Hence, *cis* and *trans* isomers will not interconvert during regular MD simulations on a nanosecond time scale. Furthermore, in the gas phase, the *cis* isomer is lower than the *trans* by 0.3 kcal/mol, despite the bias in the force field toward the *trans* isomer. In contrast, in water, the *trans* isomer is lower than *cis* by 1.5 kcal/mol. Such a significant change in the PMF upon hydration suggests that the selected molecule is well suited for the ultimate HREX SR-TI validation.

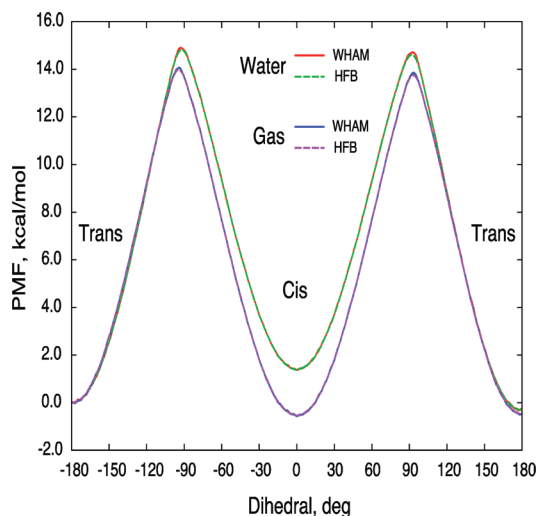


Figure 4. Potential of mean force (PMF) for rotation about the amide bond in the gas phase and in water. The PMFs were computed from umbrella sampling simulations with harmonic dihedral restraints using two independent methods, namely, the weighted histogram analysis method (WHAM) and umbrella integration with harmonic Fourier beads method.

To validate the SR-TI approach, its results must satisfy certain criteria. Specifically, in the regular SR-TI simulations, the amide molecule should not cross the barrier separating the *cis* and *trans* conformations and, therefore, should maintain the conformation of the initial configuration near $\lambda = 0$. Thus, the regular SR-TI simulations in either phase should provide corresponding relative free energies of *cis* and *trans* isomers directly. Furthermore, regardless of the initial configuration, each HREX SR-TI simulation should yield the *cis/trans* ratio matching the free energy differences computed with the regular SR-TI approach. Similarly, the averaged hydration free energy computed with HREX SR-TI, along with its gas and water phase components, should be bounded by the corresponding values for *cis* and *trans* conformations from regular SR-TI. Finally, the computed results should be consistent with the amide torsional PMFs.

Table 4 summarizes the results of multiple SR-TI simulations with and without the HREX option. To ensure that the amide dihedral is activated in the HREX simulations, we employed the benzene core as the reference structure. All nine conformations of the amide identified during the exhaustive search were used to initiate SR-TI simulations, resulting in nine independent SR-TI simulations per option. In each case, we ran two simulations per isomer, one with 12 and another with 23 windows. All of the results can be found in the Supporting Information. Table 4 shows only the averaged results from these simulations, for all isomers together, and for *cis* and *trans* isomers individually.

The regular SR-TI simulations demonstrate that *cis* and *trans* conformations have similar free energies in the gas phase, and that hydration significantly favors the *trans* conformation over the *cis*. These results are in good agreement with the torsional PMFs.

Comparing the results with 12 and 23 windows, we find a discrepancy of about 0.4 kcal/mol between relative free energies of *cis* and *trans* conformations both in the gas and

water phases. Similar discrepancy was observed in the torsional PMFs. Thus, the regular SR-TI simulations with 12 windows underestimate the free energy of the *trans* conformation, which is the opposite of the reversible work for the alchemical transformation reported in Table 4. This discrepancy is due to the appearance of a sharp peak in the mean force profiles near $\lambda = 0$ (see the Supporting Information) as was seen for the hydroxylated benzenes in water. However, because in the amide case, this discrepancy is present in both the gas and water phases, it fortuitously cancels out in the final free energy difference result. The fact that the *trans* and not the *cis* conformation is affected strongly argues that this peak is related to the intramolecular hydrogen bond formed between the hydroxyl group and the amide carbonyl.

Using the results of the regular SR-TI simulations, we can construct the expectation values for the *cis/trans* ratios (C/T) in the gas and water phases as follows:

$$C/T = \exp\left[\frac{\Delta G(cis) - \Delta G(trans)}{k_B T}\right] \quad (13)$$

where k_B is the Boltzmann constant and $\Delta G(cis)$ and $\Delta G(trans)$ are the reversible works to alchemically transform the amide in the *cis* and *trans* configurations to the reference state, respectively. Because these expectation values employ the relative free energies of the *cis* and *trans* conformations, they are particularly sensitive to small variations and hence depend on the number of windows. With 12 windows, the C/T ratios are 1.91 and 0.10, in the gas phase and water, respectively, whereas with 23 windows, these numbers change to 0.98 and 0.06, respectively.

Now we can examine the results of the HREX SR-TI simulations. Table 4 shows that with 12 windows the rate of exchanges in water is only 11%. Note that merely 11 atoms out of a total 17 are “switched off” here. This is the lowest exchange rate we have seen. However, even at such a low rate, the HREX option cuts the standard deviation of the computed hydration free energy in half (from 0.87 in regular SR-TI to 0.42 kcal/mol in HREX SR-TI). Nevertheless, over 4 ns of simulation time, the *cis/trans* ratios in the gas and water phases are far from the expected values (1.91 and 0.10, respectively). Attesting to the validity of HREX SR-TI, the computed hydration free energies stay well within the bounds set by regular SR-TI.

Increasing the number of windows can significantly improve the HREX SR-TI results. We have seen that the number of windows can affect the results of regular SR-TI simulations and their derived expectation values. Here, we demonstrate that increasing the number of windows about two times greatly affects the exchange rate and consequently improves the overall HREX SR-TI performance. In particular, going from 12 to 23 windows raises the acceptance ratio in water from 11 to 40%. This, in turn, reduces the standard deviation in the computed hydration free energy by about a factor of 4 (from 0.86 to 0.20 kcal/mol). Furthermore, the averaged hydration free energies for simulations initiated from *cis* and *trans* conformations get within 0.36 kcal/mol of each other. With 23 windows, the C/T ratios in the gas and water phases begin to approach the expected values (of

Table 4. Results of Regular and HREX SR-TI Simulations for N-(2-Hydroxy-phenyl)formamide^a

selection	regular SR-TI, 4 ns			EX _G [EX _W], %	HREX SR-TI, 4 ns				HREX SR-TI, last 3 ns	
	ΔG _G	ΔG _W	ΔΔG		ΔG _G	ΔG _W	ΔΔG	C/T _G [C/T _W]	ΔΔG	C/T _G [C/T _W]
12 windows										
all-avg	16.35	30.92	-14.57	56 [11]	16.32	31.06	-14.74	0.99 [0.88]		
all-SD	0.20	0.68	0.87	1 [0]	0.05	0.43	0.42	0.38 [0.95]		
cis-avg	16.48	30.47	-13.99	56 [11]	16.31	30.84	-14.53	0.99 [1.31]		
cis-SD	0.04	0.05	0.07	1 [0]	0.05	0.34	0.36	0.36 [0.89]		
trans-avg	16.09	31.82	-15.73	56 [11]	16.34	31.49	-15.15	1.00 [0.03]		
trans-SD	0.07	0.04	0.08	2 [0]	0.06	0.13	0.09	0.50 [0.06]		
23 windows										
all-avg	16.49	31.05	-14.56	77 [40]	16.51	31.58	-15.08	1.18 [0.24]	-15.24	1.20 [0.08]
all-SD	0.05	0.86	0.86	1 [0]	0.08	0.19	0.20	0.52 [0.16]	0.12	0.60 [0.03]
cis-avg	16.49	30.48	-13.99	77 [40]	16.51	31.46	-14.96	1.27 [0.33]	-15.20	1.12 [0.09]
cis-SD	0.02	0.05	0.06	1 [0]	0.10	0.07	0.10	0.61 [0.08]	0.10	0.60 [0.02]
trans-avg	16.50	32.20	-15.71	77 [40]	16.50	31.82	-15.32	0.99 [0.04]	-15.33	1.34 [0.05]
trans-SD	0.08	0.04	0.10	0 [0]	0.04	0.11	0.08	0.26 [0.01]	0.13	0.70 [0.00]

^a The free energy values are given in kcal/mol. In the selection column, all, cis, and trans labels refer to average (avg) and standard deviation (SD) values computed over all nine, just six *cis*, and just three *trans* conformations of the amide, respectively. The number of windows used in the simulations is indicated on a separate line. The HREX SR-TI simulations attempted exchanges every 2 ps. EX_G and EX_W indicate the average acceptance ratio in the gas phase and water, respectively. Similarly, C/T_G and C/T_W refer to cis/trans ratios in gas and water phases. All simulations employed the specified number of windows run for 4 ns each in canonical (gas phase) and in NPT (water) ensembles at $T = 300$ K and $P = 1$ atm. ΔG_G and ΔG_W are the SR-TI work values for the alchemical transformations to the reference benzene core in the gas phase and water, respectively, and ΔΔG = ΔG_G - ΔG_W is the corresponding relative hydration free energy. The expectation values for C/T in either phase were computed using alchemical free energies from regular SR-TI as follows: $C/T = \exp[(\Delta G(\text{cis}) - \Delta G(\text{trans})) / (k_B T)]$, where k_B is the Boltzmann constant. With 12 beads, these values were 1.91 and 0.10 for the gas and water phases, respectively, whereas with 23 windows they changed to 0.98 and 0.06.

0.98 and 0.06, respectively). Hence, in difficult cases like the amide, increasing the number of windows can greatly improve the results of the HREX simulations.

To assess the effect of exchange frequency on the results, we repeated SR-TI simulations with the HREX option attempting simulations every 500 steps (every 1 ps) as opposed to every 1000 steps. While doubled the number of exchange attempts over the 4 ns run, it did not significantly affect the final solvation free energy (see the Supporting Information). This confirms that our previous simulations have reached convergence. We note that increasing the frequency of exchanges could provide additional computational savings as long as the time it takes to evaluate the Metropolis acceptance criteria is much less than the time to run MD simulations between the exchanges. However, we did not attempt to identify the corresponding limit on exchange frequency. A previous work employing HREX with FEP suggested that attempting exchanges every 100 steps (0.2 ps) is close to the limit.⁹⁶

Equilibrating the generalized ensemble improves the HREX SR-TI predictions. Typically, before we begin the Hamiltonian exchanges, each window is equilibrated the same way as in regular SR-TI (see the Methodology section). However, this equilibration does not involve any exchanges, and hence at the beginning, the simulations are still biased by their starting configurations. Therefore, additional equilibration is desired for the generalized HREX ensemble itself. To demonstrate this, we use the simulation with 23 windows. Specifically, we divide the 4 ns of simulation time, which has 2000 exchange attempts, into two periods $t_e + t_p = 4$ ns, where t_e and t_p are the equilibration and production periods. We use only the t_p portion to recalculate the hydration free energies and the *cis/trans* ratios. In particular, we start with $t_p = 3$ and proceed in decrements of 1 ns or 500 exchange

attempts. Note that we have already discussed the results with $t_p = 4$ ns, which corresponds to using all of the HREX simulation time, in previous paragraphs. In Table 4, we only show additional results for $t_p = 3$ ns, which has the lowest hydration free energy of -15.24 kcal/mol with the lowest standard deviation of 0.12 kcal/mol. Clearly, the recomputed hydration free energy remains bounded by the corresponding regular SR-TI values for *cis* and *trans* conformations. Finally, for $t_p = 3$ ns, the C/T ratio in water is 0.08 and practically matches the expected value of 0.06. Thus, introducing an equilibration time for the generalized ensemble can greatly improve the quality of predictions.

This work has demonstrated that the HREX option is absolutely essential to get high quality results for solutes with multiple configurations that have distinct solvation properties and are separated by high energy barriers. In the case of the amide studied here, HREX SR-TI consistently circumvents barriers as high as 15 kcal/mol over the course of 4 ns. Importantly, HREX SR-TI achieves this outstanding result using a modest number of simulation windows. Such a dramatic enhancement in sampling is simply impossible with either regular SR-TI or conventional TI.

We feel that the difficult amide test case has helped us validate the SR-TI approach and establish best protocols for its use. The C/T ratios predicted with regular SR-TI and HREX SR-TI are in excellent agreement. Furthermore, the final hydration free energy from HREX SR-TI is independent of whether the simulation starts from the *cis* or *trans* isomer. The final value stays within the clear bounds defined by regular SR-TI for *cis* and *trans* isomers separately and, independently, by the dihedral PMF results. All of these results provide confidence in the approach necessary for future applications.

E. Conclusion

To conclude, we have presented a single-topology TI variant, called SR-TI augmented with HREX, that provides reliable estimates of relative solvation free energies for series of related molecules even in the presence of hindered conformational transitions. The key difference from conventional TI methods is that SR-TI transforms all of the molecules from a particular series down to a single reference state that does not have to correspond to a physical state. The choice of the reference state is flexible and allows rational selection of torsional degrees of freedom for enhanced sampling. Furthermore, a reduction in molecular volume in the reference state allows for better mobility in confined spaces. The benefits of the enhanced sampling and mobility in the reference state can be passed on to the real state using the HREX option. In addition, the HREX option improves overlap in configuration space between the TI simulation windows. Therefore, combining the SR-TI approach with Hamiltonian replica exchange brings considerable improvements over current single-topology TI methods. Thus, we feel that the SR-TI approach with and without the HREX option is a useful addition to the family of rigorous, high quality TI methods. Application of this methodology to more complex problems, including ligand binding, is currently in progress in our laboratory and will be described in a subsequent paper.

Acknowledgment. We would like to thank Dr. In-Chul Yeh, Dr. Michael S. Lee, and Dr. Hyung-June Woo for helpful discussions. Also, we acknowledge the National Cancer Institute (NCI) for allocation of computing time and staff support at the Advanced Biomedical Computing Center (ABCC) at NCI–Frederick. This work was sponsored by the U.S. Department of Defense High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) initiative. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or the U.S. Department of Defense.

Supporting Information Available: Figures 1S–4S show the results of integration of the simulation data for the series of hydroxybenzenes that demonstrate the appearance of a sharp peak near $\lambda = 0.0$ due to hydrogen atoms of the hydroxyl groups. Figures 5S and 6S show the same results for the *cis* and *trans* isomers of the amide. For the amide, all of the isomers identified during the exhaustive conformation search and their AM1 and AM1BCC/GAFF energies are presented. A summary of the torsional PMF extrema is also presented. Additional simulation and analysis results are provided in several tables. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Gunthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501.
- (2) Lybrand, T. P.; Ghosh, I.; McCammon, J. A. *J. Am. Chem. Soc.* **1985**, *107*, 7793.
- (3) Jorgensen, W. L.; Nguyen, T. B. *J. Comput. Chem.* **1992**, *14*, 195.
- (4) Ewing, T. J. A.; Lybrand, T. P. *J. Phys. Chem.* **1994**, *98*, 1748.
- (5) Lybrand, T. P.; McCammon, J. A.; Wipff, G. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 833.
- (6) Tembe, B. L.; McCammon, J. A. *Comp. Chem.* **1984**, *8*, 281.
- (7) Severance, D. L.; Essex, J. W.; Jorgensen, W. L. *J. Comput. Chem.* **1995**, *16*, 311.
- (8) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2002**, *106*, 11009.
- (9) Woolf, T. B.; Roux, B. *J. Am. Chem. Soc.* **1994**, *116*, 5916.
- (10) Richards, N. G. J.; Williams, P. B.; Tute, M. S. *Int. J. Quantum Chem.* **1992**, *44*, 219.
- (11) Mannhold, R.; Waterbeemd, H. v. d. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337.
- (12) Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. *J. Chem. Theory Comput.* **2009**, *5*, 2436.
- (13) Morgantini, P.-Y.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 6057.
- (14) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769.
- (15) Pitner, J. W.; Gunsteren, W. F. v. *J. Phys. Chem. B* **2001**, *105*, 11264.
- (16) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508(13).
- (17) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919.
- (18) Mordasini, T. Z.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 360.
- (19) Liu, H.; Mark, A. E.; Gunsteren, W. F. v. *J. Phys. Chem.* **1996**, *100*, 9485.
- (20) Hritz, J.; Oostenbrink, C. *J. Phys. Chem. B* **2009**, *113*, 12711.
- (21) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.
- (22) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.
- (23) Mitchel, M. J.; McCammon, J. A. *J. Comput. Chem.* **1991**, *12*, 271.
- (24) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050.
- (25) Okazaki, S.; Nakanishi, K.; Touhara, H. *J. Chem. Phys.* **1979**, *71*, 2421.
- (26) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. *Faraday Symp. Chem. Soc.* **1982**, *17*, 55.
- (27) Singh, U. C.; Brown, F. K.; Bash, P. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1987**, *109*, 1509.
- (28) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564.
- (29) Mezei, M.; Beveridge, D. L. *Ann. Acad. Sci. N. Y.* **1986**, *482*, 1.
- (30) Warshel, A. *J. Phys. Chem.* **1982**, *86*, 2218.
- (31) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *101*, 5032.
- (32) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1989**, *90*, 3300.

- (33) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1989**, *91*, 3631.
- (34) Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255.
- (35) Min, D.; Li, H.; Li, G.; Bitetti-Putzer, R.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 144109.
- (36) Grubmueller, H. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 2793.
- (37) Lange, O. F.; Schaefer, L. V.; Grubmueller, H. *J. Comput. Chem.* **2006**, *27*, 1693.
- (38) Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908.
- (39) Voter, A. F. *J. Chem. Phys.* **1997**, *106*, 4665.
- (40) Perez, D.; Uberuaga, B. P.; Shim, Y.; Amar, J. G.; Voter, A. F. *Annu. Rep. Comput. Chem.* **2009**, *5*, 79.
- (41) Fajer, M.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1565.
- (42) Fajer, M.; Swift, R. V.; McCammon, J. A. *J. Comput. Chem.* **2009**, *30*, 1719.
- (43) Hamelberg, D.; McCammon, J. A. *Annu. Rep. Comput. Chem.* **2006**, *2*, 221.
- (44) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.
- (45) Hamelberg, D.; Shen, T.; McCammon, J. A. *J. Am. Chem. Soc.* **2005**, *127*, 1969.
- (46) Li, H.; Fajer, M.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 024106/1.
- (47) Zheng, L.; Carbone, I. O.; Lugovskoy, A.; Berg, B. A.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 034105/1.
- (48) Zheng, L.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 124107/1.
- (49) Hansmann, U. H. E. *Phys. A (Amsterdam, Neth.)* **2010**, *389*, 1400.
- (50) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.
- (51) Hansmann, U. H. E.; Okamoto, Y. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1996**, *54*, 5863.
- (52) Lyubartsev, A. P.; Martsinovskii, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.
- (53) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042.
- (54) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (55) Itoh, S. G.; Okumura, H.; Okamoto, Y. *J. Chem. Phys.* **2010**, *132*, 134105/1.
- (56) Kwak, W.; Hansmann, U. H. E. *Phys. Rev. Lett.* **2005**, *95*, 138102(4).
- (57) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2008**, *128*, 144121(10).
- (58) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703.
- (59) Woods, C. J.; King, M. A.; Essex, J. W. *Lect. Notes Comput. Sci. Eng.* **2006**, *49*, 251.
- (60) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058.
- (61) Jorgensen, W. L.; Nguyen, T. B. *J. Comput. Chem.* **1993**, *14*, 195.
- (62) Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. *J. Comput. Chem.* **1993**, *14*, 1240.
- (63) Drakenberg, T.; Forsen, S. *J. Chem. Soc. D* **1971**, 1404.
- (64) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1991**, *95*, 1175.
- (65) Boresch, S.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 103.
- (66) Boresch, S.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 119.
- (67) Levitt, M.; Warshel, A. *Nature (London)* **1975**, *253*, 694.
- (68) Karplus, M.; McCammon, J. A. *Nature (London)* **1979**, *277*, 578.
- (69) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature (London)* **1977**, *267*, 585.
- (70) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.
- (71) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.
- (72) Simonson, T. *Mol. Phys.* **1993**, *80*, 441.
- (73) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529.
- (74) Cross, A. J. *Chem. Phys. Lett.* **1986**, *128*, 198.
- (75) Steinbrecher, T.; Mobley, D. L.; Case, D. A. *J. Chem. Phys.* **2007**, *127*, 214108(13).
- (76) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701.
- (77) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978.
- (78) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 4533.
- (79) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350.
- (80) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242.
- (81) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635.
- (82) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
- (83) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.
- (84) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.
- (85) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245.
- (86) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105/1.
- (87) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. *Annu. Rep. Comput. Chem.* **2007**, *3*, 41.
- (88) Khavrutskii, I. V.; Arora, K.; Brooks, C. L., III. *J. Chem. Phys.* **2006**, *125*, 174108.
- (89) Khavrutskii, I. V.; Dzubiella, J.; McCammon, J. A. *J. Chem. Phys.* **2008**, *128*, 044106.
- (90) Khavrutskii, I. V.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 124901.
- (91) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2007**, *127*, 204104/1.
- (92) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118.

- (93) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A. *J. Mol. Biol.* **2009**, *394*, 747.
- (94) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 1231.
- (95) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902(16).
- (96) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583.
- (97) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435.
- (98) Nadler, W.; Hansmann, U. H. E. *Los Alamos Natl. Lab., Prepr. Arch., Quant. Biol.* **2007**, 1.
- (99) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386.
- (100) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903/1.
- (101) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
- (102) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
- (103) Sadowski, J.; Gasteiger, J.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000.
- (104) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. *J. Chem. Inf. Model.* **2006**, *46*, 2324.
- (105) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (106) Stewart, J. J. P. *MOPAC7*.
- (107) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132.
- (108) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623.
- (109) Wang, J. Antechamber. In *Amber Tools*, 1.2 ed.; Case, D. A., Ed.; University of California, San Francisco: San Francisco, CA, 2009.
- (110) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247.
- (111) Morris, G. M.; Huey, R.; Olson, A. J. *Current Protocols in Bioinformatics*; John Wiley and Sons: New York, 2008; Chapter 8, Unit 8, p 14.
- (112) Goodsell, D. S.; Morris, G. M.; Olson, A. J. *J. Mol. Recognit.* **1996**, *9*, 1.
- (113) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293.
- (114) Berendsen, H. J. C.; van, d. S. D.; van, D. R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (115) Hess, B.; Kutzner, C.; van, d. S. D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (116) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.
- (117) van der Spoel, D.; Lindahl, E.; Hess, B.; Kutzner, C.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; Drunen, R. v.; Berendsen, H. J. C. *GROMACS User Manual Version 4.0*; The GROMACS development team: Groningen, The Netherlands, 2006.
- (118) Zhang, W.; Hou, T.; Schafmeister, C.; Ross, W. S.; Case, D. A. LEaP. In *Amber Tools*, 1.2 ed.; Case, D. A., Ed.; University of California, San Francisco: San Francisco, CA, 2009.
- (119) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.
- (120) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. *J. Chem. Theory Comput.* **2008**, *4*, 1669.
- (121) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.
- (122) Boczko, E. M.; Brooks, C. L., III. *J. Phys. Chem.* **1993**, *97*, 4509.
- (123) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048.
- (124) Mezei, M.; Harrison, S. W.; Ravishanker, G.; Beveridge, D. L. *Isr. J. Chem.* **1986**, *27*, 163.
- (125) Spector, T. I.; Kollman, P. A. *J. Phys. Chem. B* **1998**, *102*, 4004.
- (126) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672.
- (127) Jorgensen, W. L.; Gao, J. *J. Am. Chem. Soc.* **1988**, *110*, 4212.
- (128) Gerothanassis, I. P.; Demetropoulos, I. N.; Vakka, C. *Biopolymers* **1995**, *36*, 415.
- (129) Exarchos, K. P.; Papaloukas, C.; Exarchos, T. P.; Troganis, A. N.; Fotiadis, D. I. *J. Biomed. Inf.* **2009**, *42*, 140.
- (130) Troganis, A. N.; Sicilia, E.; Barbarossou, K.; Gerothanassis, I. P.; Russo, N. *J. Phys. Chem. A* **2005**, *109*, 11878.

CT1003302

Stability of Hydrocarbons of the Polyhedrane Family Containing Bridged CH Groups: A Case of Failure of the Colle–Salvetti Correlation Density Functionals

Grigory A. Shamov,^{*,†,‡} Georg Schreckenbach,[†] and Peter H. M. Budzelaar[†]

University of Manitoba, Winnipeg MB, Canada, and Dutch Polymer Institute (DPI),
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received July 13, 2010

Abstract: DFT-computed energies of polyhedral hydrocarbons, such as dodecahedrane $C_{20}H_{20}$, its smaller analogs $C_{16}H_{16}$ and $C_{12}H_{12}$, and the larger $C_{24}H_{24}$, estimated in comparison with corresponding isomeric hydrocarbons, vary widely with the choice of the density functional. In particular, large discrepancies were observed with the functionals that are based on the B88 (as well as G96, B86) exchange and the LYP (as well as OP) correlation parts. The problem is not related to the presence of the smaller cyclopropane rings in the $C_{12}H_{12}$ polyhedrane, for its hydrogenated products do show similar errors; moreover, the larger dodecahedrane that is free from the Bayer strain shows a similar trend. DFT-D corrections that are very useful in fixing long- and medium-range correlation issues with GGA DFT do not help in this case either. We show that these errors stem from the B88 (G96, B86) exchange functionals and are not compensated by Colle–Salvetti-based GGA correlation functionals such as LYP, OP, TCA, etc. However, they can be corrected by the PBE correlation functional based on the PW92 uniform electron gas (UEG) parametrization. Range-separated hybrids (likura and Hirao's LC-BOP, LC-BLYP) perform much better than the parent GGAs. Comparisons of polyhedranes with a well-studied system of similar size, the set of C_nH_n cyclophanes, reveal a completely different performance for the latter—for instance, RHF results are the poorest, and LC-type functionals do not give any improvement, but dispersion-corrected BLYP-D performs very well. We conclude that, while for polyhedranes medium-range delocalization errors from exchange dominate, for cyclophanes, the correlation/overlap-dispersion interactions are more important. The OPTX exchange functional shows significantly lower errors compared to B88 and G96; its combinations like OLYP and especially KT3 perform well for both test sets. The OPTX-based double hybrid, O2PLYP, also outperforms the corresponding B88-based B2PLYP functional for polyhedranes. Our computations also suggest that the $(CH)_{16}$ and $(CH)_{24}$ polyhedranes could be possible synthetic targets.

Introduction

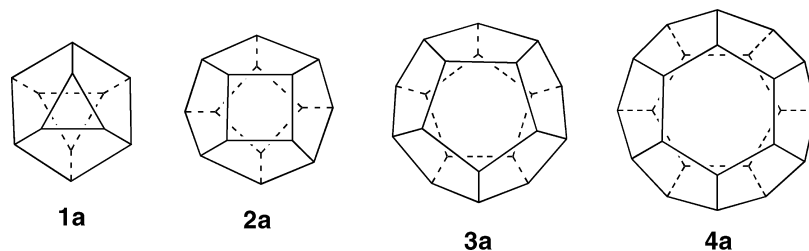
Cage and highly symmetric hydrocarbons are interesting due to their aesthetic appeal. The polycyclic caged hydrocarbons and their derivatives were very instrumental in developing the concepts of the theory of reaction mechanisms in organic

chemistry (see, for example, ref 1). The cage hydrocarbons **1a–4a** that we will discuss in this work (Scheme 1) are built from two parallel, staggered carbocycles linked into a three-dimensional cage by a rim of methyne bridges, so that a “tire” of fused five-membered rings is formed. The dodecahedrane **3a** (two cyclopentane rings) has been known for quite a while and was subject to numerous experimental and theoretical studies.^{2–6} The related compound **1a** was also recently synthesized and studied both experimentally and

* Corresponding author e-mail: gas5x@yahoo.com.

[†] University of Manitoba.

[‡] DPI.

Scheme 1. Polyhedranes (CH)_n, for $n = 12, 16, 20,$ and 24 

theoretically.⁷ Besides their geometrical beauty, there could also be potential applications, for example, for perfluorinated dodecahedrane⁸ which consists of two charged, nearly spherical layers, or using polyhedrane cages to encase atoms and small metal ions for some molecular electronic devices.^{9–11}

There is mounting evidence that the description of inter- and intramolecular interactions in large hydrocarbon systems can be problematic for density functional theory (DFT) methods, which are today the most widely used tools in theoretical organic and organometallic chemistry.^{12–19} This presents a challenge, since for realistic modeling of systems of chemical interest (for example, transformation of hydrocarbons catalyzed by a transition metal complex whose ligands are also hydrocarbons, or an enzyme reaction center, surrounded by several amino acid residues) it is important to have a balanced, reasonably correct description of all parts of the system under study. It would not do if only reaction barriers were correctly described, but not intra- and intermolecular interactions between reactants, ligands, and the surroundings, or vice versa.

Recent developments in the theory have shown that the reason for many problems of approximate DFT lies in its deficiency in the description of dispersion interactions.^{20,21} Also, most of the popular GGA and hybrid density functionals are over-repulsive at medium and large interatomic distances. Various methods were introduced to amend these DFT problems—among the most popular and computationally inexpensive are the CR⁻⁶ corrections by Grimme,^{22,23} named DFT-D, and the pseudopotential modification by the group of Röthlisberger,²⁴ adapted for Gaussian-basis molecular calculations by DiLabio et al.^{25,26} (C-Pot). It was thought¹³ that, while important, dispersion corrections alone cannot be sufficient to accurately describe intramolecular interactions in hydrocarbons and thus their relative stabilities with DFT. However, recent work in the area seems to show^{27,28} that, with the proper parametrization, DFT-D can in most cases successfully describe both intra- and intermolecular interactions in hydrocarbons by correcting the over-repulsive nature of GGA and hybrid functionals at medium-range interatomic distances. This leads to the question of whether the correction can still be called “dispersion”, or whether it is some other exchange-correlation error that is being corrected, perhaps the delocalization error described by Yang and co-workers.¹² The C-Pot method also performs well for some functionals of a repulsive nature, like the popular B3LYP.²⁷ Very recently, a new revision of the DFT-D method, termed DFT-D3, was introduced; this revision tries to restrict the

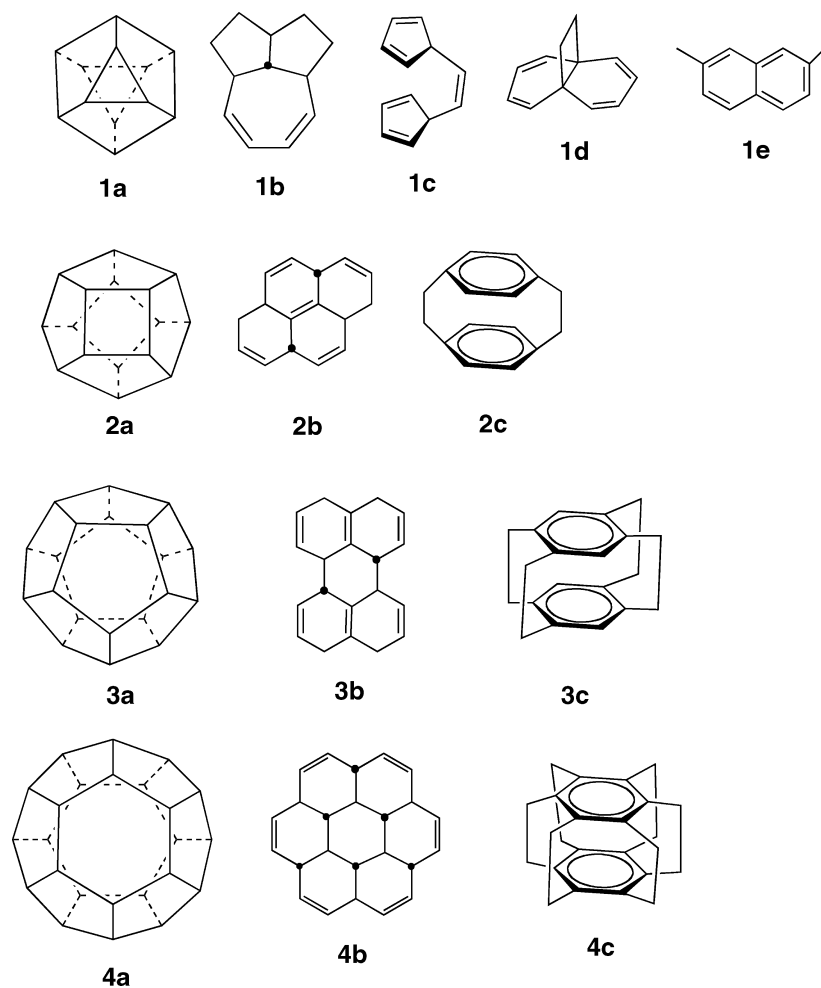
correction to the van der Waals interactions only, without correcting the medium range behavior of the GGA functionals.²⁹

In the recent work of Gill and co-workers,¹⁸ the problems of DFT with reactions that lead to an increased degree of hydrocarbon branching were attributed mainly to errors in the DF exchange functionals (i.e., not to the Coulomb or the kinetic energy, or a correlation functional).

A particularly interesting case however exists where some DFs fail only for some of the combinations of exchange and correlation functionals. It was found previously¹⁴ that for the (CH)₁₂ compound **1a** (Scheme 1), named [D_{3d}]-octahedrane by its creators, the performance of density functionals varies widely; the authors issued a cautionary warning for using DFT for “structures with single bonds only, especially for small rings”. Earlier, we showed²⁷ that for the (CH)₁₂ isomers, some combinations of exchange and correlation functionals, most notably the combination of Becke’s popular B88 exchange³⁰ with the LYP^{31,32} correlation, show large discrepancies from results using other functionals and the MP2 method. Numbers provided in the paper by Schreiner and co-workers¹⁴ also show that some other functional combinations, for example, Gill’s GLYP³³ functional, behave similarly to BLYP. Not all density functionals perform similar to BLYP–PBE, for example, yield results much closer to those by correlated wave function methods. We have shown that the OLYP functional, which is strongly over-repulsive at medium range, describes the stability of the [D_{3d}]-octahedrane well; moreover, the BPBE combination also performs well. DFT-D was found to be insufficient to amend the performance of the combination of Becke’s B88 and LYP (but the C-Pot method with B3LYP has, rather unexpectedly, been successful).

Thus, it is interesting to investigate why a specific combination of exchange and correlation functionals suddenly performs poorly for this specific hydrocarbon case. As was alluded to above, it is always desirable for a density functional to provide a balanced description of the model systems under study. Information on which functional is successful for which kind of system can be essential for possible practical applications. We dedicate this article to answering these two questions for the curious case of octahedrane and related compounds.

To do that, we have applied a systematically chosen set of density functionals, in some cases together with methods that include dispersion interactions (DFT-D by Grimme²³ and C-Pot by DiLabio and Mackie²⁵) in the computation of the stabilities of (CH)_n cage molecules ($n = 12, 16, 20, 24$),

Scheme 2. Isomeric C_nH_n Molecules, $n = 12, 16, 20,$ and 24 , Used to Estimate the Stabilities of **1–4**^a

^a Tertiary carbons of the condensed polycycles **1b–4b** for which hydrogens point up are labeled with black dots; otherwise the hydrogens point down.

shown in Scheme 1. For comparison, we will use isomeric cyclophanes (Scheme 2) that contain stacked π -rings at close distances. For the cyclophanes, the importance of the intramolecular dispersion interactions is well established; these molecules have been thoroughly studied theoretically.^{34–36}

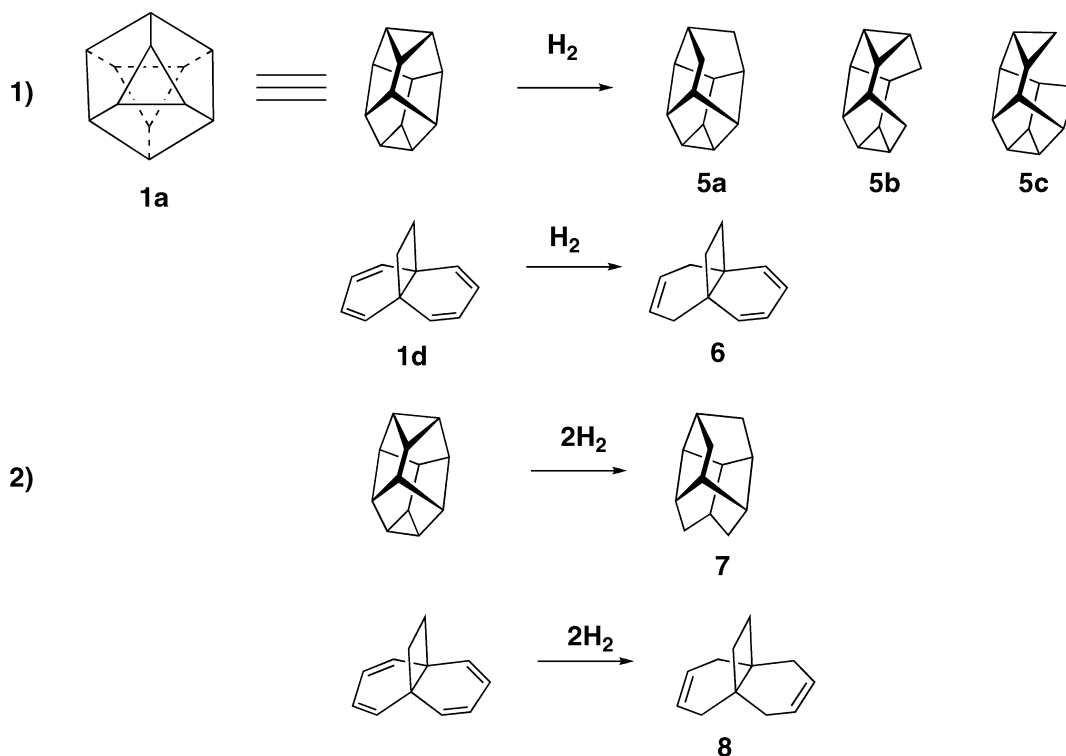
The presence of small carbocycles and sterical strain are the popular culprits when DFT methods happen to fail for molecules like the compound **1a**, $(CH)_{12}$. In the present work, we extend the range of polyhedranes from the highly strained compound **1a**, which contains cyclopropane rings, to the larger polyhedrane cages **2a–4a**. The set includes dodecahedrane **3a**, which is a compound with angles around its carbons close to tetrahedral, and thus free from the Bayer (angular) strain. (We note that in **3a**, as well as in all of the polyhedranes under study, another type of steric strain³⁷ is present, due to the eclipsed conformation of carbon–carbon dihedrals; for dodecahedrane, its value was estimated in refs 4 and 38.) We will investigate whether for these compounds a similar picture of the performance of density functionals will be observed.

Following the current trend³⁹ of DFT benchmarking studies, as a measure of stability, we will consider relative energies of isomeric hydrocarbons rather than heats of their

formations or atomizations, thus avoiding complications with the treatment of open-shell atoms. In addition, in order to assess the stabilities of the $(CH)_n$ cage compounds relative to each other, we will consider their formation energies from acetylene C_2H_2 .

We have selected isomers of C_nH_n , to act as references for evaluating the stabilities of polyhedranes in the following way. First, the isomers must be nonpathological cases for DFT. For instance, they should not have extended conjugated systems, which excludes annulene $(CH)_n$ systems. Second, they should be of approximately the same type of compound for the entire range of n values in the C_nH_n family, to allow for more or less justified comparisons between the stabilities of the different polyhedranes.

On the basis of these criteria, we have constructed several isomers of C_nH_n , borrowing some from our previous work,²⁷ and some from ref 14. The structures are presented in Scheme 2. For $(CH)_{24}$, the partially hydrogenated coronene **4b** was chosen, for it is planar (i.e., has minimal nonbonded C–C contacts), made of six-membered rings only, and yet is a nonconjugated hydrocarbon. Similar types of hydrocarbons for $C_{20}H_{20}$ and $C_{16}H_{16}$, structures **3b** and **2b**, correspondingly, were chosen. (They are not precisely of $(CH)_n$ type, but this

Scheme 3. Hydrogenation Products of the Polyhedrane **1a** and Tricyclohexatetraene **1b**

is not important for our purposes). For $C_{12}H_{12}$, we chose compound **1b** (which is compound **21** from ref 14; note that it is not the most stable, but the most symmetrical one of its isomers). In addition, structures of the tricyclo-dodecatetraene **1d** and 2,7-dimethylnaphtalene **1e** were considered isomers of $C_{12}H_{12}$.

For our second set of compounds, we chose the paracyclophanes **2c**, **3c**, and **4c** as isomers for C_nH_n for n greater than twelve. For the $C_{16}H_{16}$, meta-cyclophane **2d** was also computed. For $C_{12}H_{12}$, the selection of a real cyclophane is impossible; for that reason, we picked the molecule **1c**, *cis*-(bis-cyclopentadienyl)ethylene, in which the unsaturated five-membered rings are roughly parallel. Note that along the series **1c–2c–3c–4c**, the steric strain increases, and the unsaturated/aromatic rings get closer to each other.

In the original paper dedicated to $[D_{3d}]$ -octahedrane **1a** by Schreiner et al.,⁷ the mono- and dihydrogenated species **5a–c** and **7** (Scheme 3) were also considered; we will include them along with the tricyclododecatri- and -dienes **6** and **8**, correspondingly, of which the former are isomers.

In this work, we will investigate the performance of various combinations of exchange and correlation functionals. First, we will apply the BLYP and GLYP functionals that perform poorly, and the OLYP, PBE, and BPBE functionals that perform well for $[D_{3d}]$ -octahedrane,²⁷ to the wider set of test compounds (Schemes 1–3). Second, we will systematically test the density functionals, by using stand-alone exchange, as well as mixing and matching exchange and correlation functionals of various kinds; to apply modifications such as the inclusion of a fraction of the Hartree–Fock exchange, both global and range separated (LC-corrections by Iikura et al.^{40,41}), as well as double hybrids, which in addition to a high fraction of exact exchange contain a portion of the PT2 correlation energy.⁴²

The generalized gradient approximation for exchange–correlation energy is written as follows:

$$E_{XC} = \sum_{\sigma} E_{X\sigma} + E_C \quad (1)$$

$$E_{X\sigma} = \int e_{LDA,\sigma}(\rho_{\sigma}) F(x_{\sigma}) d^3r; x = |\nabla\rho|/\rho^{4/3} \quad (2)$$

Here, F , which depends on the reduced density gradient x , is the GGA enhancement factor, introduced to improve the performance of the local density approximation for systems that differ from the uniform electron gas (UEG; its energy being determined by $e_{LDA}(\rho)$), such as atoms and molecules. Further, $\sigma = \alpha$ and β denotes the spin component. We provide plots of the enhancement factors $F(x)$ for selected functionals in the Supporting Information (Figure S2).

The correlation energy E_C , unlike the exchange energy, depends on both spin densities. Sometimes it can be written with same-spin and opposite-spin correlation contributions separated.

$$E_C = E_{C,\alpha\beta}(\rho_{\alpha}, \rho_{\beta}, x_{\alpha}, x_{\beta}) + \sum_{\sigma} E_{C,\sigma\sigma}(\rho_{\sigma}, x_{\sigma}) \quad (3)$$

The forms of its dependency on the spin densities and their gradients that are used in the literature vary. One of the approaches is to use an approach similar to the exchange functional: take a local correlation energy functional and add gradient corrections to it. There is no analytical solution for the uniform electron gas correlation energy, unlike for the exchange part where such a solution is known. Parameterized forms, interpolating numerical results exist; the most popular are the VWN⁴³ and PW92⁴⁴ functionals, although recently others were proposed also.⁴⁵ Besides GGA-corrected UEG-based correlation functionals

such as PBE, meta-GGA correlation functionals such as VS98⁴⁶ and its modifications, the M06 family of functionals,⁴⁷ as well as many others exist.

Another approach to the DFT correlation energy is to derive a correlation functional *ab initio*, via a series of simplifications—for example, starting from the Colle–Salvetti formula derived from a correlated wave function for the helium atom using a series of approximations. An example is the popular LYP³¹ functional in the form derived by Miehlich et al.³² The older Wigner⁴⁸ formula can be seen as the local part of LYP. Recently, several other Colle–Salvetti type functionals were proposed, most notably the OP by Tsuneda et al.,⁴⁹ Handy and Cohen’s CS1⁵⁰ functional, and the RC local functional⁵¹ and its GGA development called TCA.⁵² Other GGA correlation functionals building on the Wigner formula, such as the Wilson–Levy functional, have been proposed; see the recent work of Thakkar and McCarthy for their extensive evaluation.⁵³

Most of the energy contribution comes from the exchange functional. Thus, its asymptotic behavior is considered the most important. Dynamic correlation is also more short-range; however, nondynamic, or “left–right” correlation energy, is delocalized. It is well established that the inclusion of some fraction of the Hartree–Fock exchange⁵⁴ improves the performance of GGA functionals in many cases, such as the prediction of thermochemistry and reaction barriers. Recently, range separated forms of hybrid density functionals were proposed^{55–58} for which the amount of HF exchange is not constant but varies—for example, from GGA at short distances to HF for long distances. The long-range corrected GGAs show improved performance for properties that depend on the asymptotic behavior of the exchange, which is incorrect for GGA but can be reproduced by HF; at the same time, the electron–electron cusp condition, which is short ranged, is better described by DFT.

Recent attempts were made to include a fraction of the MP2 correlation energy computed with DFT orbitals, leading to the so-called double hybrids.⁴² Below, we will attempt to arrange the functionals to be tested in this work into groups.

Exchange Functionals

Group A. B88 and G96 are repulsive functionals; they are also the ones that perform poorly with the LYP correlation for octahedrane. These functionals exhibit a diverging $F(x)$ for the high limit of the reduced density gradient x . They are purely repulsive at any range of the PES. (See the scans in the Supporting Information, Figure S1.)^{59,60}

We have also included the B86 functional (with the formula taken from the Appendix of ref 59), which is similar to B88 but has a different “asymptotic behavior”, that is, a converging $F(x)$, which results in it showing some attraction at large interatomic distances. We note that for medium and short distances, B86 is over-repulsive, which is typical for all GGAs. (See the potential energy surface scans in the Supporting Information, Figure S1.) We also included the OPTX exchange functional, which has a higher power of x than B86 and violates the UEG limit. It is known^{27,61} that

OPTX is too repulsive at medium and short interatomic distances, even more so than B88. However, for long distances, it shows some attraction, just like B86. This can be traced to the shape of the $F(x)$ enhancement factor.⁶⁰

We will use the B88, G96, and OPTX exchange functionals in a stand-alone manner, naming them HFB, HFG, and HFO, and will match them with a variety of correlation functionals, as described below. We will also augment them with a fraction of HF exchange, globally (B3LYP) and in the long-range only (LC functionals for a few of the combinations). Moreover, we will try global double hybrids that include the PT2 correlation for some of them, namely, B2PLYP and our O2PLYP.

The power series B97 family functionals⁶² are also built on the basis of sums of B86-like expressions up to the second or fourth power. This is combined with a “spin-scaled” UEG PW92 correlation functional with GGA-corrections of a form similar to the B86 exchange. We have included in our set of functionals the original B97 hybrid functional and Grimme’s B97-D²³ GGA functional that has been optimized together with the dispersion correction, to avoid double counting of the long-range attractive interactions.

Recently, Goerigk and Grimme performed a reparameterization of commonly used density functionals, including BLYP, on the basis of their extensive benchmark database GMNKT,⁶³ again with inclusion of the DFT-D corrections. The latter was used in its reparameterized form, with the scaling factor for the R_0 radii increased to 1.34 from 1.1 and a global scaling factor of 1.0, giving it the correct asymptotic behavior at very long range. We include the reoptimized B88 + LYP combination, named oBLYP-D, in our set of functionals.

An interesting modification to the OLYP functional exists in the literature⁶⁴—a same-spin correlation term by Keal and Tozer was added to the OLYP functional, which was reparameterized and called KT3. The local version containing the same term (KT1) was shown to perform well for π – π stacking.⁶¹ We have included the KT3 functional in our set of functionals.

Group B. Group B consists of functionals that are spuriously attractive at the long range and less repulsive in the medium range than those in group A. The most prominent example is the PBE exchange functional. It was proposed⁶⁵ to reparameterize it (as well as the parameters of the PBE correlation counterpart, correspondingly), to give a better description of solids. The modification, named PBEsol, became more attractive than PBE in the medium–long range, which was shown to improve its performance for hydrocarbons, namely, cyclophanes and octahedrane **1a**.⁶⁶ Recently, a further modification of PBEsol including the higher-power x term in the denominator was proposed⁶⁷ under the name regularized gradient expansion (RGE2). In our study, we will include the PBE, PBEsol, and RGE2 functional combinations.

The highly parametrized Minnesota density functionals that are based on the VS98 meta-GGA exchange-correlational functional also claim improved performance for the interactions of interest to this work. We have shown²⁷ previously that the M06-L functional is on the softer side, performing comparably to PBE-D. We will include the local M06-L

Table 1. Combinations of Exchange and Correlation Functionals Employed in This Work

correlation functional		“repulsive” GGA	“softer” GGA and meta-GGA	large amount of exact exchange (hybrids/double hybrids)
none		HFB, HFG, HFO		RHF
LDA	UEG ^a	BPW92		
	CS ^b	BW		
GGA	UEG ^a	BPBE, BPBE-D, OPBE, B97, B97-D	PBE, PBEsol, RGE2, M06-L	M06-2X
	CS ^b	BLYP, BLYP-D, B3LYP, B86LYP, GLYP, BOP, oBLYP-D, OLYP, OLYP-D, O3LYP, KT3, B3LYP, B3LYP-C	TCA	B2PLYP, O2PLYP, BLYP-LC, BOP-LC, GLYP-LC

^a Correlation functionals based on a Uniform Electron Gas parametrization. ^b Correlation functionals based on the Colle–Salvetti formula.

functional and the hybrid that has been recommended by the authors for hydrocarbons, M06-2X, for comparison with the GGAs under study.

Correlation Functionals

Earlier, we showed²⁷ that, for [D_{3d}]-octahedrane, the BLYP combination of LYP with B88 did not work well, while BPBE did. In this paper, we will try to explain this observation and test this and related combinations on similar chemical systems (Schemes 1–3).

LYP is a GGA Colle–Salvetti⁶⁸ functional, while PBE correlation is a GGA based on the PW92 parametrization of the uniform electron gas (UEG). We will try to match the B88 exchange with LYP and PBE correlations; we will also try the B86LYP, OLYP, and OPBE combinations. Besides that, another Colle–Salvetti GGA functional, the one-parameter progressive (OP) functional by Tsuneda and co-workers,⁴⁹ will be combined with B88 and G96, giving the BOP and GOP functionals. Another recently introduced Colle–Salvetti functional, TCA, will be used, as proposed by its authors,⁶⁹ together with the PBE exchange.

To extend our study further, we will use LDA correlation functionals that correspond to the UEG parametrization (PW92) and to the Colle–Salvetti type. For the latter, we will try the Wigner functional, which can be roughly considered as an LDA part of the GGA LYP.^{50,70}

Combinations

Exchange-correlation functional combinations considered in this work are summarized in Table 1. Following our earlier work, to illustrate the “repulsiveness” of these functionals at medium and long interatomic distances, we have performed a PES scan of colliding the methane molecule into the ethane molecule, decreasing the intermolecular carbon–carbon distance from 6.0 to 2.0 Å. The interaction curves and details of the computations are provided in the Supporting Information (Figure S1).

Computational Methods

Geometry optimizations for the BLYP, BPBE, OLYP,⁷¹ and PBE⁷² density functionals were done using the Priroda code^{73–75} in the L11⁷⁴ basis set (a cc-pCVDZ-quality general contracted basis). The RI-JK approximation as implemented in Priroda⁷³ was applied throughout, with corresponding optimized density fitting basis sets. For GGA functionals, analytical second derivatives were computed on the opti-

mized geometries to confirm the nature of the respective minima. Final GGA energies were calculated on the L11 optimized geometries with the L22 basis set⁷⁴ (a cc-CVTZ-quality general contracted basis). RI-MP2 geometry optimizations were done with Priroda as well, in the cc-pVTZ-quality L2 basis.⁷⁴

To extend our study to a broader and more systematic range of functionals, and in particular to include various exchange-correlation functional combinations not available in Priroda, we have used a locally modified version of the MOLCAS code, version 7.4.^{76,77} For some functionals available in the regular version of MOLCAS, parameters were changed so that the optimized version of the BLYP functional by Goerigk and Grimme (oBLYP-D)⁶³ and PBEsol⁶⁵ were also implemented. The exchange functionals G96,³³ B86 (with the formula and parameter values taken from the Appendix of ref 59), OPTX⁷¹ and KT3,⁶⁴ and RGE2⁶⁷ (paired with the PBEsol exchange), as well as the correlation functionals by Wigner^{48,78} and the TCA⁵² correlation functional, were added to the code. On the basis of these additions, the following combinations were tested as single-point calculations on RI-MP2/L2 geometries in the same L22 basis set as for the Priroda computations: (1) pure exchange, with no correlation part—HFB, HFG, HFO—and (2) various GGA functionals named in the text as BLYP/MP2, B86LYP, GLYP, BPBE, BW,⁴⁸ BPW92 (with the local correlation functional of Perdew and Wang⁴⁴), OPBE, KT3, the hybrid functionals B3LYP⁵⁴ and O3LYP,^{71,79} the double-hybrid functional B2PLYP of Grimme,⁴² and its double hybrid analog that we will call O2PLYP using OLYP as the starting point. For the O2PLYP functional, we did not optimize the contributions from the exact exchange and the PT2 correlation but instead took the “physically motivated” values of 50% and 25%, correspondingly. For the GGA part of the O2PLYP exchange, we also took 50% of the original OPTX parameters for its LDA and GGA parts. For both of the double hybrids, in Tables 2–4, we provide energies due to the DFT part only, labeled B2PLYP(0), and with inclusion of the PT2 term (B2PLYP(2), correspondingly, except for Figures 1 and 2, where the functional is simply labeled B2PLYP).

We have also computed canonical MP2 and scaled opposite spin MP2 (SOS-MP2)⁸⁰ energies with the same L22 basis set using MOLCAS. All MOLCAS calculations, both DFT and MP2, were done using the compact atomic Cholesky decomposition technique (acCD) with an accuracy threshold of 1×10^{-4} .^{81–83}

Table 2. Binding Enthalpies Per Acetylene Monomer ΔE_b^b of Polyhedranes **1a–4a**, kcal/mol^a

method	1a	2a	3a	4a	MAD
SOS-MP2	<i>-44.2</i>	<i>-48.7</i>	<i>-53.0</i>	<i>-51.6</i>	
MP2	-4.4	-4.0	-4.2	-4.1	4.2
B2PLYP(0)	2.2	2.7	2.7	3.0	2.6
B2PLYP(2)	1.1	1.4	1.5	1.6	1.4
HFB	27.8	29.3	30.4	30.9	29.6
HFO	19.8	22.2	23.8	24.6	22.6
HFG	27.9	29.5	30.8	31.4	29.9
M06-L	-4.2	-2.7	-2.1	-1.9	2.7
M06-2X	-3.8	-2.9	-2.9	-2.8	3.1
KT3	-10.3	-10.2	-9.9	-9.7	10.0
B86LYP	10.3	11.0	11.4	11.6	11.1
BLYP ^c	9.2	9.8	10.3	10.5	9.9
BPYP-D ^c	5.3	4.7	4.8	4.6	4.8
BPBE ^c	-1.0	0.0	0.7	1.0	0.7
BPBE-D ^c	-4.4	-4.5	-4.2	-4.3	4.3
OLYP ^c	1.9	3.6	4.6	5.1	3.8
OLYP-D ^c	-1.9	-1.6	-1.0	-0.9	1.3
PBE ^c	-5.5	-4.9	-4.5	-4.3	4.8
PBE-D ^c	-7.9	-8.1	-7.9	-8.1	8.0
BW	21.5	22.7	23.6	24.1	23.0
RHF	4.3	4.9	4.9	5.4	4.9
BLYP-LC ^d	-9.9	-8.6	-10.1	-10.0	9.6
GLYP-LC ^d	-10.7	-9.2	-10.9	-10.8	10.4
BOP-LC ^d	-7.6	-7.3	-7.2	-7.0	7.3
BOP ^d	11.6	12.6	13.3	13.7	12.8
B97 ^d	-2.1	-3.0	-1.8	-1.6	2.1
B97-D ^d	3.1	-1.0	3.1	2.9	2.5
O2PLYP(2)	-2.4	-1.7	-1.4	-1.1	1.6
O2PLYP(0)	-1.5	-0.6	-0.4	0.1	0.7
GLYP/MP2	9.6	10.4	11.0	11.3	10.6
oBLYP/MP2	6.5	7.0	7.3	7.5	7.1
oBLYP-D/MP2	2.8	3.3	3.2	3.1	3.1
BPW92	17.4	18.5	19.2	19.6	18.6
OPBE	-10.3	-8.7	-7.6	-7.0	8.4
PBEsol	-14.5	-14.3	-14.1	-14.0	14.2
B2PLYP-D	-1.4	-0.9	-1.1	-1.2	1.1
B3LYP-Cpot ^d	-10.7	-10.7	-12.1	-11.9	11.4
B3LYP	3.2	3.6	3.8	4.1	3.7
RGE2	-5.5	-4.9	-4.5	-4.4	4.8
O3LYP	-9.8	-9.1	-8.7	-8.5	9.0
TCA	9.6	11.4	12.2	12.6	11.4

^a The SOS-MP2 are actual values (*in italics*); for the other methods, differences with respect to SOS-MP2 values are shown.

^b $\Delta E_b = (E(C_2H_2)_m - mE(C_2H_2))/m$. ^c Priroda L22 energies on optimized geometries at the DFT/L11 level. ^d GAMESS-US cc-CVTZ energies on Priroda MP2/L2 geometries.

The MP2 method is the cheapest accurate correlated wave function method. It is well-known that for intramolecular interactions of hydrocarbons (see refs 27, 66, and 84 and references therein), the method errs systematically toward overbinding. It was also found that the various spin-scaled MP2 methods, such as SCS-MP2 and SOS-MP2, perform much better for hydrocarbons (although the description of other systems, such as hydrogen bound complexes, might deteriorate).⁸⁴ Specifically for the polyhedranes, SCS-MP2 was shown to give energies of isomerization of compound **1a**⁶⁶ closer to the CCSD(T) values than MP2. Our preliminary tests for the **1a** system have shown that SOS-MP2 energy differences are closer to CCSD(T) than the SCS-MP2 ones by about 2 kcal/mol. Thus, we used the SOS-MP2 energies computed with the MOLCAS code as the reference energies throughout this paper.

The DFT-D corrections for the dispersion energy for the GGA functionals BLYP, OLYP, BPBE, and PBE were done

Table 3. Relative Energies of Hydrogenated Species of **1a**, **5a–c**, and **7** with Respect to Tricyclododecatetraene **1b** and Its Hydrogenated Derivatives **6** and **8**, Correspondingly, kcal/mol^a

method	5a → 6	5b → 6	5c → 6	7 → 8	MAD
SOS-MP2	<i>35.0</i>	<i>9.9</i>	<i>-5.4</i>	<i>46.7</i>	
MP2	4.3	4.5	3.4	3.1	3.9
B2PLYP(0)	-13.9	-8.7	-13.5	-11.0	11.8
B2PLYP(2)	-10.8	-6.7	-9.8	-8.5	8.9
HFB	-46.7	-35.0	-39.1	-37.6	39.6
HFO	-27.7	-17.7	-21.2	-25.0	22.9
HFG	-43.2	-32.3	-36.2	-34.9	36.6
M06-L	-5.4	3.6	0.8	-9.7	4.8
M06-2X	-4.5	2.8	-0.8	-7.6	3.9
KT3	1.4	4.9	2.5	0.2	2.2
B86LYP	-30.5	-21.2	-25.1	-24.6	25.4
BLYP ^b	-27.4	-37.9	-3.6	-22.5	22.9
BLYP-D ^b	-25.5	-35.7	-1.8	-20.7	20.9
BPBE ^b	-8.7	-20.2	13.1	-9.4	12.8
BPBE-D ^b	-7.1	-18.4	14.7	-8.0	12.1
OLYP ^b	-9.4	-20.7	12.8	-10.7	13.4
OLYP-D ^b	-7.6	-18.6	14.5	-9.0	12.4
PBE ^b	-6.3	-18.3	14.9	-7.3	11.7
PBE-D ^b	-5.3	-17.1	16.0	-6.4	11.2
BW	-39.8	-28.9	-33.2	-32.1	33.5
RHF	-12.5	-8.8	-14.6	-9.2	11.3
BLYP-LC ^c	1.7	5.1	1.9	-0.7	2.4
GLYP-LC ^c	2.9	6.0	2.8	0.3	3.0
BOP-LC ^c	2.1	5.8	2.6	-0.9	2.9
BOP ^c	-27.4	-18.5	-22.0	-22.5	22.6
B97 ^c	-9.6	-5.6	-8.8	-7.8	7.9
B97-D ^c	-21.2	-11.4	-15.2	-18.8	16.6
O2PLYP(2)	-1.8	1.6	-1.2	-2.6	1.8
O2PLYP(0)	-4.5	0.0	-4.4	-4.9	3.5
GLYP/MP2	-23.9	-16.2	-19.7	-19.3	19.8
oBLYP/MP2	-24.3	-16.4	-19.9	-19.7	20.1
oBLYP-D/MP2	-27.3	-16.1	-19.0	-23.0	21.3
BPW92	-35.4	-25.1	-29.4	-28.6	29.6
OPBE	12.8	16.4	14.6	6.0	12.4
PBEsol	5.8	9.2	7.3	2.3	6.2
B2PLYP-D	-12.4	-6.1	-9.0	-10.3	9.4
B3LYP-Cpot ^c	0.9	1.6	-1.4	1.2	1.2
B3LYP	-17.8	-11.1	-15.2	-14.6	14.7
RGE2	-5.9	-0.1	-2.7	-6.8	3.9
O3LYP	2.4	6.7	3.7	-0.3	3.3
TCA	-23.4	-12.6	-16.2	-21.9	18.5

^a The SOS-MP2's are actual values (*in italics*), while for other methods, differences with respect to it are shown. ^b Priroda L22 energies on optimized geometries at the DFT/L11 level. ^c GAMESS-US cc-CVTZ energies on Priroda MP2/L2 geometries.

according to Grimme's 2006 formulation,²³ using the BOptimize code⁸⁵ with the scaling factors being the same as in our previous work,²⁷ as single-point energy computations on the GGA optimized geometries. For the B2PLYP-D and oBLYP-D functionals, dispersion corrections were also computed with BOptimize, using the recommended scaling factors. We have also tried to optimize geometries with GGA DFT-D for some cases, such as the para-cyclophanes **2c** and **3c**. While the geometry changes were not negligible, the relative energy changes were. Hence, we use single-point energies throughout.

In addition, we have investigated the one-parameter progressive correlation functional (OP) by Tsuneda et al.⁴⁹ together with the B88 exchange functional (BOP), Grimme's²³ pure power series GGA optimized with dispersion correction (B97-D), and the original power series hybrid functional B97 by Becke.⁶² Also, the performance of the

Table 4. Calculated Relative Energies of Isomers of Compounds 1–4, kcal/mol^a

method	1a → 1e	1a → 1d	1b → 1a	1b → 1c	2b → 2a	2b → 2c	3b → 3a	3b → 3c	4b → 4a	4b → 4c
SOS-MP2	<i>-34.8</i>	<i>24.6</i>	<i>-35.2</i>	<i>5.4</i>	<i>-56.0</i>	<i>-44.4</i>	<i>-85.4</i>	<i>-41.3</i>	<i>-70.2</i>	<i>-35.7</i>
MP2	-7.3	5.5	-6.4	-0.5	-3.8	-11.0	-3.7	-8.5	-5.0	-9.3
B2PLYP(0)	-32.1	-13.9	13.2	-2.2	22.3	4.1	29.0	28.2	21.7	42.2
B2PLYP(2)	-27.3	-11.2	10.4	-2.5	17.1	-2.9	22.3	12.3	15.7	17.6
HFB	-77.5	-51.9	49.4	-15.4	68.6	-5.8	88.4	30.9	73.2	45.1
HFO	-49.9	-26.8	26.1	-11.7	41.2	1.0	58.0	36.0	42.3	53.2
HFG	-73.4	-47.8	45.7	-14.4	63.6	-3.8	82.9	32.4	67.3	45.5
M06-L	-17.5	-0.5	-0.5	-3.2	11.9	-7.7	24.2	6.0	16.6	5.2
M06-2X	-10.7	-0.3	-0.5	-1.8	10.6	-0.3	15.6	11.6	12.1	19.6
KT3	-9.4	4.6	-3.3	2.3	-5.4	1.8	-4.1	13.8	-14.6	16.0
B86LYP	-54.4	-33.1	31.5	-8.5	43.9	-5.2	56.3	20.3	43.7	28.6
BLYP ^b	-49.4	-30.1	28.0	-9.1	39.9	-4.1	51.1	18.8	39.0	26.7
BLYP-D ^b	-41.7	-28.4	25.3	-5.4	28.6	-10.0	37.2	0.5	30.4	3.6
BPBE ^b	-21.6	-6.1	6.4	-4.6	13.6	1.1	21.1	17.8	9.8	22.8
BPBE-D ^b	-14.9	-4.9	4.1	-1.0	3.7	-4.3	8.7	1.5	2.1	2.6
OLYP ^b	-23.7	-6.1	6.4	-7.3	15.9	2.3	25.5	24.6	13.5	36.8
OLYP-D ^b	-16.1	-4.7	3.9	-2.7	4.3	-3.6	11.0	6.0	4.3	13.7
PBE ^b	-17.9	-3.6	4.1	-2.5	9.8	0.4	15.6	13.9	5.6	17.1
PBE-D ^b	-13.1	-2.8	2.5	-0.5	2.7	-3.7	6.7	2.1	0.0	2.8
BW	-68.5	-43.9	41.5	-12.8	58.0	-5.2	74.8	28.1	59.5	39.6
RHF	-27.8	-12.7	11.9	-0.9	22.2	13.3	29.2	43.2	25.2	66.9
BLYP-LC ^c	-1.5	6.1	-5.4	2.4	8.6	8.6	-0.9	18.6	-4.7	25.4
GLYP-LC ^c	0.1	7.5	-6.6	2.8	8.9	9.0	-3.2	18.5	-6.9	25.2
BOP-LC ^c	-1.0	7.2	-6.4	1.5	-1.9	9.8	-0.7	22.0	-5.9	29.7
BOP ^c	-50.1	-29.0	27.7	-8.6	39.9	-2.6	52.2	23.9	38.1	31.6
B97 ^c	-24.2	-8.8	8.7	-2.1	3.5	2.0	15.8	20.4	5.3	27.6
B97-D ^c	-35.1	-21.6	19.2	-5.7	-8.6	-10.5	30.9	1.0	24.1	6.5
O2PLYP(2)	-14.2	0.8	-0.6	-0.7	4.1	0.8	8.1	15.3	0.9	22.0
O2PLYP(0)	-18.2	-1.4	1.6	-0.4	8.5	7.2	13.6	30.0	5.8	44.9
GLYP/MP2	-46.6	-25.6	24.7	-6.4	35.0	-1.9	46.0	21.8	33.0	27.1
oBLYP/MP2	-46.1	-26.2	25.1	-6.0	35.3	-3.7	45.7	18.2	33.8	23.5
oBLYP-D/MP2	-34.8	-19.6	17.7	-6.0	26.6	-9.8	33.0	4.0	25.6	10.4
BPW92	-63.1	-38.7	36.4	-11.2	51.5	-4.9	66.4	26.3	51.2	36.4
OPBE	8.0	21.5	-18.7	2.4	-17.4	8.9	-13.8	23.1	-25.3	28.8
PBEsol	-0.9	10.6	-8.5	3.6	-7.6	3.0	-5.5	9.3	-13.3	7.3
B2PLYP-D	-19.6	-6.7	5.1	-1.8	11.9	-6.4	15.6	3.1	11.7	7.3
B3LYP-Cpot ^c	-6.7	2.5	-7.0	1.1	-3.0	-4.0	-14.2	-1.1	-16.6	-2.8
B3LYP	-37.7	-18.3	17.5	-4.0	26.8	-0.8	34.9	21.0	24.9	29.0
RGE2	-17.9	-3.0	3.7	-1.0	8.6	0.4	14.3	14.3	4.2	16.5
O3LYP	-8.2	7.1	-5.8	1.8	-2.0	4.0	1.5	17.7	-7.5	22.6
TCA	-42.4	-22.0	21.0	-8.4	36.6	-2.5	50.3	23.9	37.5	34.1

^a See Scheme 2 for a description of the isomers. The SOS-MP2 values are the actual isomerization energies (*in italics*); for the other methods, differences from the SOS-MP2 values are shown. ^b Priroda L22 energies on optimized geometries at the DFT/L11 level. ^c GAMESS-US cc-CVTZ energies on Priroda MP2/L2 geometries.

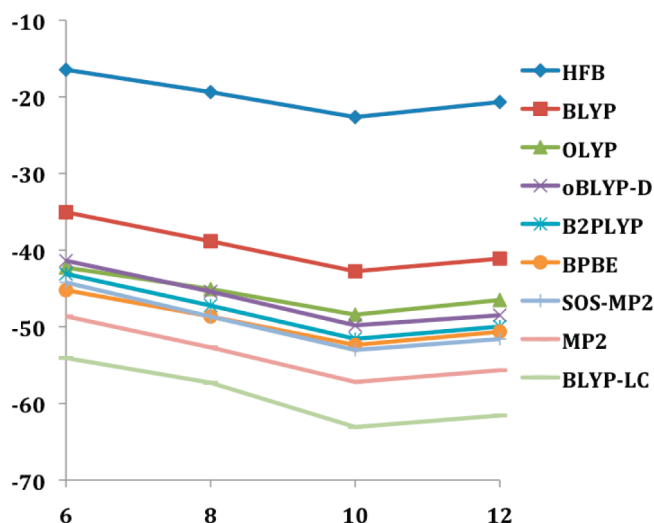


Figure 1. Binding energies of polyhedranes $(\text{CH})_{2n}$ from n molecules of acetylene, per acetylene monomer (eq 1), kcal/mol.

range-separated hybrid functionals by Hirao et al.⁴⁰ (long-range correction, LC) was studied for the BLYP-LC, BOP-LC, and GLYP-LC functionals. Finally, C-Pot computations were done with the B3LYP/cc-CVTZ basis, with C-Pot coefficients optimized for B3LYP/aug-cc-pVTZ taken from ref 25. We have used test calculations to ensure that the difference in the basis sets does not lead to significant changes in MAD values for the carbon-only subset of the S22 set; thus, the pseudopotential parameters are transferable enough. These computations were done with the GAMESS-US⁸⁶ code, version February 2009, using single point energy calculations in the cc-CVTZ basis by Dunning and Hay⁸⁷ on the Priroda RI-MP2/L2 optimized geometries.

Results and Discussion

Geometries of the Hydrocarbons. For the cage molecules 1a–4a and tricyclooctane 5, all GGA DFs for which we did optimize geometries show very similar bond angles. In the dodecahedrane 3a, the CCC angles are equal to 108°, while the CCH angles are 110.9°. Going from 1a to 2a and 4a,

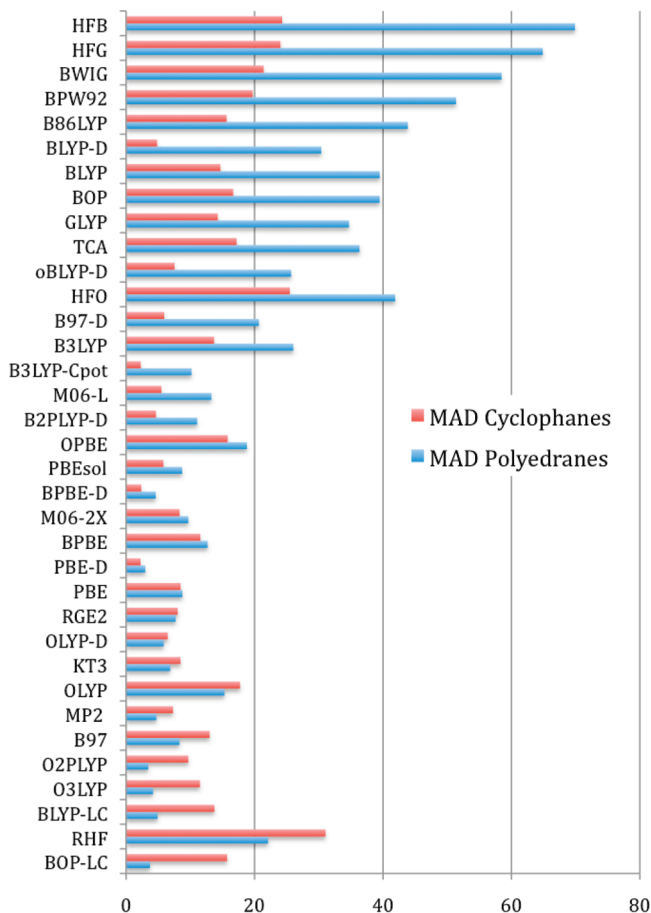


Figure 2. Mean absolute differences (MAD) for cyclophane and polyhedrane isomerization, relative to SOS-MP2 computed values, kcal/mol.

the $C_bC_bC_b$ angle (the angle between the bridge carbons) rises from 100.8° in **1** to 107.8° – 107.9° in **4**; the $C_bC_bC_3$ angle (the angle between two bridge and one top cycle carbon) increases from 106.2° – 106.3° to 108° . The angles including bridge hydrogens decrease from 113.7° to 111.3° in this series. All of these changes indicate that, for the larger polyhedranes, steric strain due to the angular deformation of bridged carbons decreases. The optimized bond lengths are also similar for all of the GGA DFs, except BLYP, which systematically yields longer C_b – C_3 and C_b – C_b bonds in **1a**–**4a**.

The geometries of **1a**–**4a** optimized at the RI-MP2/L2 level show $C_bC_bC_b$ and $C_bC_bC_3$ angles very similar to the DFT ones. Both C_b – C_3 and C_b – C_b bond lengths are usually shorter at the GGA DFT level, OLYP showing the closest agreement in bond length with RI-MP2. All of the geometries of compounds **1a**–**4a**, **1b**–**4b**, and **1c**–**4c** are provided in the Supporting Information as a separate file.

Binding Energies of 1a–4a Relative to Acetylene. The binding energies of the polyhedranes **1a**–**4a** relative to acetylene were weighted per acetylene monomer according to eq 4:

$$E_{\text{bind}} = (1/n) \times E(C_{2n}H_{2n}) - E(C_2H_2) \quad (4)$$

The calculated values are collected in the Table 2 as differences from the SOS-MP2 values. For selected methods, the absolute values are also pictured in Figure 1.

One can see that all methods predict an increase of binding from **1a** to **3a**, which is the most stable polyhedrane, with **4a** slightly less stable than **3a**. This is what one would expect on the basis of the angular strain argument—the angles around the carbon atoms in **4a** and especially **3a** are closest to tetrahedral. The absolute values of the binding energies vary strongly between the computational methods; however, within the series of the polyhedranes, each method performs similarly. That is, the presence of the small cyclopropane and cyclobutane rings in **1a** and **2a** and their absence in **3a** and **4a** does not lead to significant changes in the order of the energies calculated.

In Figure 1, it can be seen that BLYP underestimates the binding energy as compared to SOS-MP2. The exchange-only HFB yields unrealistically small binding. MP2 and BLYP-LC overbind, the latter quite strongly. OLYP, which is quite an over-repulsive functional, is much closer to SOS-MP2 than BLYP; the reoptimized version of the latter, including dispersion corrections, oBLYP-D, performs better. The BPBE combination and the double hybrid B2PLYP functional also perform well.

A more systematic overview, with all of the functionals included in Table 1, shows that, not surprisingly, stand-alone exchange functionals perform much worse than those combined with a correlation functional. HFB and HFG have similarly large differences from the SOS-MP2 values. They are followed by BW and HFO, which are also close to each other, and then comes the BPW92 combination. It is interesting to observe that the pure OPTX exchange functional has much smaller errors than either B88 or G96 and, also, that the Wigner correlation functional (which can roughly be seen as the local part of the LYP correlation functional) is less efficient than the local PW92 correlation that is based on a uniform electron gas (UEG) parametrization.

The RHF method predicts results quite close to those of SOS-MP2 (with MAD under 5 kcal/mol). Many of the GGA combinations, such as BLYP, B86LYP, GLYP, BOP, and OPBE, perform worse. Using the long-range HF hybrids BLYP-LC, BOP-LC, and GLYP-LC does not significantly improve the MADs from SOS-MP2 over the corresponding pure GGAs. However, the sign of the deviations changes—long-range hybrids show overbinding, while pure functionals underbind (see for example BLYP in Figure 1). PBEsol and KT3, as well as the dispersion-corrected PBE-D, also show overbinding, as does the canonical MP2 method. The smallest deviations from SOS-MP2 were obtained for double hybrids; “repulsive” GGAs with DFT-D corrections such as BPBE-D, OLYP-D, oBLYP-D, and B97-D; the hybrid B97 functional; and the M06-family functionals. We note that B88 exchange improves when paired with PBE instead of LYP correlation, while for OPTX exchange, the situation is opposite.

The performance of the exchange functionals of the PBE family, especially PBEsol, is poor compared to B88 and OPTX. In this model system, the result might depend on the accuracy of the description of acetylene by a given method; since acetylene is a small, highly symmetric molecule, one could expect poor performance for functionals that do not do well for atoms, which is the case with PBE.

Thus, below, we will investigate isomerization and hydrogenation reactions that involve large molecules only as another way to assess the stabilities of the polyhedranes. But first, we will look into the hydrogenation reactions for the sterically strained compound **1a**.

Energies of Hydrogenated Compound 1: C₁₂H₁₄ and C₁₂H₁₆. The hydrogenation of compound **1a** had been studied theoretically in the original experimental work, employing MP2 and B3LYP computations in a small basis set.⁷ The authors have considered isomers **5a–5c**; it was also shown that the second stage of the hydrogenation leads preferentially to compound **7**. Considering isomers where some or all of the cyclopropane rings are opened due to hydrogenation, we can test whether the large deviations of the B88 + LYP combination are due to the strain introduced by the small cycles. To assess the relative energies, we have compared them to an isomeric molecule, tricyclododecatetraene **1d** in different degrees of hydrogenation, and **6** and **8** (Scheme 3). We would like to avoid using the actual hydrogenation process for the comparison, preferring to compare isomers. Results are presented in Table 3. One can see that the trends observed in our work (as well as in earlier publications by other authors) for the relative energies of **1a** and its isomers hold for the hydrogenated species **5a–c** and **7** as well: The BLYP functional (as well as the very similar functionals B86LYP, GLYP, and BOP but not OLYP and BPBE) strongly underestimates the stabilities of the hydrogenated derivatives of **1a**. Even for compound **7**, which has no cyclopropane rings at all, there is a perceptible difference in the performance of the different density functionals. Interestingly, going from pure GGAs to the long-range hybrids BLYP-LC and GLYP-LC removes the error. The double-hybrid functional O2PLYP and the pure GGA KT3 perform best among the functionals without dispersion corrections. Overall, the trends in the MADs for different methods are close to those for the isomerizations of polyhedranes **1a–4a**, which are considered in the following section.

Isomerization Energies of Compounds 1–4. We have computed the relative energies of isomerization of our “base reference compounds” **1b–4b** into the two sets of corresponding structures: polyhedranes **1a–4a** and cyclophanes **1c–4c** (considering **1c** as similar to cyclophane). The isomerization energies for **1a** and **1b** were published previously.^{14,27,66} We note that isomer **1b** is not the most stable one studied by Csonka et al.⁶⁶ The cyclophanes C₁₆H₁₆ (**2c** and **2d**) were subjects of several theoretical studies⁸⁸ thanks to their epitomizing of aromatic π – π interactions.⁸⁹

The results for isomerization reactions by different methods are collected in Table 4. Our calculations display a very distinctive feature of the performance of our computational methods: the MADs for two sets (corresponding to the reactions of **1b–4b** to either **1a–4a** or **1c–4c**, respectively) pictured in Figure 2 show that methods performing well for the cyclophanes tend to do poorly for the polyhedranes, and vice versa!

Let us first consider the cyclophanes. In contrast to the results of the previous sections (**1a** and its hydrogenated species) as well as the polyhedranes discussed below, the worst method for the cyclophanes is RHF (Table 3; Figure

2), with a MAD of over 30 kcal/mol. This is followed by the exchange-only functionals, with HFO being marginally worse than HFB and HFG, which have MADs of about 24 kcal/mol. The B88 exchange combined with the local correlation functionals, Wigner and PW92, also has large errors of about 20 kcal/mol; the DFT parts of the double hybrids (labeled O2PLYP(0) and B2PLYP(0), without the PT2 term) perform about as poorly as the former, owing to the large fraction of RHF exchange and the smaller portion of GGA correlation they contain. Next, this is followed by all of the “repulsive” GGA combinations such as OLYP, BOP, BLYP, B86LYP, and GLYP. Interestingly, the long-range hybrids do not improve the performance of these functionals significantly. The reoptimized functional oBLYP without the DFT-D correction is slightly better than the original BLYP but still underestimates the stabilities of the cyclophanes with a MAD of over 13 kcal/mol. Better performance is observed with the softer functionals (PBE family, M06-L, KT3, double hybrids with the PT2 term included) and for DFT-D corrected functionals, including Grimme’s B97-D and B2PLYP-D and oBLYP-D. The PBEsol functional is better than the original PBE (in agreement with previous reports by Csonka et al.⁶⁶). BPBE-D is again one of the best exchange-correlation combinations. MP2 overbinds compared to SOS-MP2.

A general conclusion can be reached that the stabilities of the cyclophanes are governed by the “normal” dispersion interactions that are dominated by correlation energy, with exchange being relatively unimportant. It was suggested⁷¹ that GGA exchange functionals have some contribution of the “left–right” correlation energy that is absent in the Hartree–Fock method; thus stand-alone exchange functionals perform slightly better than the Hartree–Fock method. The importance of the π – π correlation in cyclophanes was also previously noted by Grimme.³⁵ He termed the interactions “overlap–dispersion” and noted the poor performance of RHF as compared to SCS-MP2.

Let us consider the absolute values first for the polyhedranes **1a–4a**. The isomerization of **1a** to the dimethylnaphthalene **1e** is a highly exothermic process according to all of the methods we have used. Thus, while **1a** can be the most stable isomer of (CH)₁₂, it certainly is not the most stable C₁₂H₁₂ compound. Trends in the isomerization energy from **1a** to our reference **1b** are very similar to the tricyclododecatetraene **1d** that was used as the standard for the hydrogenated species **5a–c** and **7**; this confirms that both “standard” molecules are not pathological cases for DFT and/or MP2.

One important finding is that, similar to the hydrogenated **5a–c** and **7**, the polyhedranes with larger basal rings (**3a** and **4a**, which are free from the angular Bayer strain present in **1a** and **2a**) still show a similar picture of differences between the various density functionals: the BLYP, B86LYP, and GLYP combinations have larger errors than BPBE and OLYP; the DFT-D correction is insufficient to amend it. Interestingly, the optimized oBLYP functional, even with the DFT-D correction applied, has errors for polyhedrane stabilities that are smaller than the original BLYP but are

still significant, despite the fact that [D_{3d}]-octahedrane **1a** was included in its training set.⁶³

For the polyhedranes **1a–4a**, the isomerization energies are totally different from the ones for the cyclophanes, as was pointed out above already. By far, the largest differences with the SOS-MP2 isomerization energies, with a MAD of almost 70 kcal/mol, are given by the stand-alone B88 exchange, HFB. HFG fares only slightly better. The combinations of B88 exchange with the local correlation functionals, BW and BPW92, do not do well; the stand-alone OPTX-exchange HFO functional has a much smaller MAD of about 40 kcal/mol, comparable to the B86LYP, BLYP, GLYP, and BOP functionals. The latter functionals, along with oBLYP-D, perform worse than the RHF method (MAD of 22.1 kcal/mol). The double-hybrid B2PLYP functional and the B97-D functional have larger MADs than OLYP (15.3 kcal/mol). The TCA functional, which combines PBE exchange with a new Colle–Salvetti based GGA correlation functional, performs similarly (i.e., quite poorly) to the BLYP and GLYP functionals, despite the less repulsive character of the PBE exchange.

Using PBE correlation with OPTX exchange instead of LYP decreases the accuracy. The MADs for PBE and PBEsol are very close. However, the former systematically overestimates the reaction energies as compared to SOS-MP2, while the latter underestimates them (i.e., makes them more negative). Generally, application of the DFT-D correction for the “repulsive” functional reduces the MADs for the isomerization energies.

In contrast to the cyclophanes, for the polyhedranes, inclusion of the long-range HF corrections to the BLYP, GLYP, and BOP functionals dramatically improves their performance, to the extent that their results rank among those closest to the SOS-MP2 values. The OPTX-based double hybrid, in contrast to B2PLYP, performs very well also. Among the pure, uncorrected GGAs, the KT3 functional is the best, which makes it one of the few methods performing reasonably well for both the polyhedrane and cyclophane sets.

The popular global hybrid B3LYP, while more accurate than the pure BLYP functional, still has quite large errors for the cyclophanes and unacceptably large errors for the polyhedrane stabilities. For both series, it is outperformed by the newer power series B97 hybrid functional that includes a gradient corrected PW92 correlation part. It is interesting that the C-Pot modified B3LYP-C performs much better for the polyhedranes and especially well for the cyclophanes.

Summarizing the results for the polyhedrane series, we conclude that the higher MADs for some combinations, including the B88, B86, and G96 functionals, stem from large errors in the exchange functionals. Some of the GGA correlation functionals, such as the PBE correlation functional, can compensate the errors in the exchange, yielding successful combinations like BPBE; other functionals like LYP and OP do not yield this error compensation. The OLYP combination is better because the OPTX exchange has smaller errors than B88 and the like. The error seems to be related to the behavior of the DFT exchange in the long-range region, because application of range-separated hybrids

like BOP-LC where the long-range exchange is represented by the Hartree–Fock method fixes the problem.

Discussion

Previously, using relaxed PES scans for hydrocarbon collision (methane to neopentane), we²⁷ (as well as others, on similar systems⁶⁰) have shown that GGA functionals, uncorrected for the dispersion, are too repulsive; they can be arranged in the following way: OLYP > BLYP > B3LYP > BPBE > PBE > PBE1, MP2 > VWN5, in decreasing order of repulsiveness. Here, MP2 and VWN5 actually show overbinding. This statement requires a qualification; it was shown⁶⁰ that the character of the enhancement factor $F(x)$ of the exchange functional (eq 2) is responsible for the nonbonded interactions. For LDA exchange, $F(x)$ is constant; this leads to incorrect (different from the HF method) scaling and to spurious attraction at large interatomic distances, where the values of the reduced density gradient are high. Functionals that have $F(x)$ converging to a constant at large x show the same long-range overbinding; in order to get rid of it, $F(x)$ must have the “correct asymptotic behavior” (which, on the other hand, seems to contradict the Lieb–Oxford bound; see discussion in ref 72 and references therein). Of the functionals tested in our present work, G96 has the strongest dependence $F(x) \sim x^{3/2}$; then comes B88. The authors of ref 90 have shown that the optimal $F(x)$ scaling is $\sim x^{2/5}$, which held for the old P86⁹¹ GGA exchange functional. The rest of the exchange density functionals used in our work, including OLYP, are bound. However, at medium interatomic distances, the value and slope of $F(x)$ for intermediate values of x are important; here, the OLYP functional, which shows overbinding at large x , provides the strongest repulsion (which can be seen from Figures S1 and S2 of the Supporting Information). Since problems such as alkane intermolecular interactions fall into medium-range interatomic distances, one can say that OPTX is the most repulsive functional, followed by G96 and then B88. Let us now see how the repulsive character of these functionals affects the results for our model systems.

The dodecahedrane **3** has no small rings. However, it has angles around its carbon atoms of about 108.0°, which is lower than in free tertiary carbons—about 111° for isobutane. We have calculated deformation energies of the latter, optimizing its geometry with all CCC angles fixed to 108.0°. Results for selected GGA methods and MP2 are collected in Table S1 of the Supporting Information. The results show that the deformation energies for the isobutene are in line with the “repulsiveness” of the method—that is, they are highest for OLYP and lowest for VWN5 and MP2; the BLYP and B3LYP values are lower than those of OLYP. Therefore, the underestimation of the stability of dodecahedrane by BLYP and B3LYP cannot be ascribed to the angular deformation energies.

One can also observe that for polyhedranes **1a–4a**, due to their cage nature, all of the C–C dihedrals are in the eclipsed conformation. For that reason, a considerable sterical strain related to dihedral (1,4) interactions is present.^{4,38} Thus, to check whether the differences of the stabilities of these hydrocarbons are related to the description of the 1,4

interactions in saturated hydrocarbons, we have calculated the energy differences for *n*-butane in its *trans* and *cis* conformations (the latter being the rotational transition state). Results, again for the selected methods, are provided in Table S2 of the Supporting Information; they seem to show that for the 1,4 interactions, as well as for the 1,3 interactions described above, the order of the functionals follows their repulsiveness (OLYP > BLYP > PBE \gg VWN5), and the dispersion correction affects the results by stabilizing the *cis* conformer. Thus, the destabilization of the polyhedranes by BLYP as compared to OLYP probably cannot be explained by 1,4-nonbonded interactions similar to the ones in *n*-butane.

As was noted above, the enhancement factor $F(x)$ for the B86, B88, and G96 functionals differs at large values of x (see Figure S2, Supporting Information): while for B86, it converges to a limit, for G96, it diverges rapidly, and B88 shows intermediate behavior. But for small values of x , these exchange functionals are very similar to each other (and different from the OPTX functional). Since these functionals show very close results for the systems under study, the problems of polyhedrane stabilities can be ascribed to the small, not the large, values of the reduced density gradient. As the results show, the problems of the B86, B88, and G96 functionals with the description of polyhedranes can be fixed by matching them with the PBE correlation functional, but not with the LYP or OP ones.

An interesting question arising from this discussion is the following: why is the PBE correlation functional able to compensate for the B88 (B86, G96) errors, and why is this not happening for the LYP and OP functionals? The latter functionals are based on the Colle–Salvetti formula, while the former depends on the Perdew–Wang parametrization of the UEG correlation energy. Thus, various UEG limits are upheld by PBE but not by LYP. Both LYP and OP are also free from correlation self-interaction errors (i.e., yield zero correlation energy for a one-electron system), while PBE is not. Also, implicitly in LYP, and deliberately in the OP correlation functional, opposite-spin and same-spin correlation energies are treated differently—in the OP functional, in fact, the same-spin correlation term is not present; that is, it is said to be described by the matching exchange functional.

It was noticed in the literature⁹² that the Colle–Salvetti (CS) approach, while working well for atoms (however, see a recent study on the heavier atoms⁵³), underestimates “long-range” correlation that might be important for extended systems. UEG-parametrized correlation functionals such as PBE have a long-range correlation hole, which is said to be canceled by the exchange hole at long distances. The cancellation then does not happen if an exchange functional such as B88 is combined with a CS-based correlation counterpart.

It can be noted that reoptimization of the BLYP parameters, like the one done for the oBLYP-D functional, though resulting in some improvement, does not change the picture dramatically. Related functionals—the much simpler G96 and the earlier version, B86, that differs in its asymptotic behavior from B88—perform quite similarly. Interestingly, the original hybrid power series B97 functional performs much better

than B86LYP, BLYP, and B3LYP. The reason could again be that B97 has a correlation part that is based on UEG (PW92 functional,⁴⁴ with B97 gradient corrections, and different scaling of the same-spin and opposite-spin contributions).

Conclusions

The saturated cage molecules of the “polyhedrane” family (CH)_{*n*}, *n* = 12, 16, 20, 24 (**1a–4a**), were studied with a systematic variation of DFT exchange and correlation functionals and compared against the MP2 methods. The values computed with the SOS-MP2 method were used as the main reference. The stabilities of **1a–4a** were assessed by two methods: using the energies of formation from the acetylene (CH)₂ and by comparison with isomeric molecules chosen not to be problematic for DFT. For comparison, isomeric cyclophanes were considered, as systems with well-studied nonbonded intramolecular interaction patterns.

To our knowledge, polyhedranes (CH)₁₆ and (CH)₂₄ have not yet been synthesized, whereas (CH)₂₀ and (CH)₁₂ are known. Our computations predict that the stability of the former two molecules lies between those of the latter two. Thus, **2a** and **4a** could possibly be reasonable targets for synthesis.

The previously reported large discrepancies in the results of BLYP and B3LYP for **1a** cannot be explained by the angular strain or the presence of cyclopropane rings: our calculations have shown that hydrogenated derivatives of **1a** where the rings are opened show similar differences between the DFs; moreover, we have shown that dodecaedrane **3a**, which is free from the angular strain, exhibits the same problems with the B88 and LYP combination as does **1a**. In fact, throughout the entire polyhedrane set, the trend of the discrepancies of the DFT results compared to SOS-MP2 is similar.

The performance of density functional methods for the two sets, the polyhedranes and the cyclophanes, differs strongly. Methods that did well for one set often performed poorly for the other. For the cyclophanes, the correlation energy (long-range van der Waals dispersion and “overlap dispersion”, π – π correlation) is the determining factor of their stabilities: the more important, the closer the unsaturated rings get in the series **1c–4c**. For the polyhedranes, the errors in stability stem from the exchange functional, especially in the long-range region. A systematic study of stand-alone exchange functionals and their combinations with various correlation functionals shows that errors are much larger for the B88 and G96 functionals than for OPTX exchange. For the former two, the UEG-based PBE correlation functional somehow allows for error cancellation, while the Colle–Salvetti based LYP and OP correlation functionals fail to do that. Simple adjustment of the parameters of BLYP, as in the oBLYP-D functional, is insufficient to completely fix its performance. We propose that the difference is due to the ability of the UEG-based GGA correlation functionals to include long-range correlation effects that are absent in the Colle–Salvetti family of functionals.

For the polyhedranes, good results can be achieved by using long-range hybrids like BLYP-LC or BOP-LC. Among

the pure GGA functionals, the KT3 functional shows good performance on both sets of molecules.

Acknowledgment. This work is part of the Research Programme and was performed with financial support from the Dutch Polymer Institute (DPI), Eindhoven, The Netherlands, Project #641. We thank Dr. D. N. Laikov for providing us with the Priroda code. Part of the computations was performed at the high-performance computational facilities of the Canadian WestGrid consortium, which is funded in part by the Canada Foundation for Innovation, Alberta Innovation and Science, BC Advanced Education, and the participating research institutions.

Supporting Information Available: Details of basis sets used, MP2 optimized geometries of molecules **1–8**, and tables discussing isobutane deformation energies and *n*-butane rotational energies. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Schleyer, P. v. R.; Stang, P. J.; Raber, D. J. *J. Am. Chem. Soc.* **1970**, *92*, 4725.
- Paquette, L. A.; Balogh, D. W.; Usha, R.; Kountz, D.; Christoph, G. G. *Science* **1981**, *211*, 575.
- Paquette, L. A.; Ternansky, R. J.; Balogh, D. W.; Kentgen, G. *J. Am. Chem. Soc.* **1983**, *105*, 5446.
- Disch, R. L.; Schulman, J. M. *J. Phys. Chem.* **1996**, *100*, 3504.
- Wahl, F.; Weiler, A.; Landenberger, P.; Sackers, E.; Voss, T.; Haas, A.; Lieb, M.; Hunkler, D.; Worth, J.; Knothe, L.; Prinzbach, H. *Chem.—Eur. J.* **2006**, *12*, 6255.
- Karpushenkava, L. S.; Kabo, G. J.; Bazyleva, A. B. *THEOCHEM* **2009**, *913*, 43.
- de Meijere, A.; Lee, C. H.; Kuznetsov, M. A.; Gusev, D. V.; Kozhushkov, S. I.; Fokin, A. A.; Schreiner, P. R. *Chem.—Eur. J.* **2005**, *11*, 6175.
- Jia, J. F.; Liu, C.; Wu, H. S.; Schleyer, P. v. R.; Jiao, H. J. *J. Phys. Chem. C* **2009**, *113*, 8077.
- Cross, R. J.; Saunders, M.; Prinzbach, H. *Org. Lett.* **1999**, *1*, 1479.
- An, Y. P.; Yang, C. L.; Wang, M. S.; Ma, X. G.; Wang, D. H. *J. Phys. Chem. C* **2009**, *113*, 15756.
- Gapurenko, O. A.; Gribanova, T. N.; Minyaev, R. M.; Minkin, V. I. *Russ. Chem. Bull.* **2007**, *56*, 856.
- Johnson, E. R.; Mori-Sanchez, P.; Cohen, A. J.; Yang, W. T. *J. Chem. Phys.* **2008**, *129*, 204112.
- Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217.
- Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.
- Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R. *Org. Lett.* **2007**, *9*, 1851.
- Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.
- Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42.
- Brittain, D. R. B.; Lin, C. Y.; Gilbert, A. T. B.; Izgorodina, E. I.; Gill, P. M. W.; Coote, M. L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1138.
- Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.
- Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. *J. Phys. Org. Chem.* **2009**, *22*, 1127.
- Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2008**, *41*, 569.
- Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 4.
- Mackie, I. D.; DiLabio, G. A. *J. Phys. Chem. A* **2008**, *112*, 10968.
- Mackie, I. D.; McClure, S. A.; DiLabio, G. A. *J. Phys. Chem. A* **2009**, *113*, 5476.
- Shamov, G. A.; Budzelaar, P. H. M.; Schreckenbach, G. *J. Chem. Theory Comput.* **2010**, *6*, 477.
- Wodrich, M. D.; Jana, D. F.; Schleyer, P. V. R.; Corminboeuf, C. *J. Phys. Chem. A* **2008**, *112*, 11495.
- Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*.
- Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200.
- Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.
- Caramori, G. F.; Galembeck, S. E. *J. Phys. Chem. A* **2008**, *112*, 11784.
- Grimme, S. *Chem.—Eur. J.* **2004**, *10*, 3423.
- Kamya, P. R. N.; Muchall, H. M. *J. Phys. Chem. A* **2008**, *112*, 13691.
- Schleyer, P. v. R.; Williams, J. E.; Blanchard, K. R. *J. Am. Chem. Soc.* **1970**, *92*, 2377.
- Beckhaus, H. D.; Ruchardt, C.; Lagerwall, D. R.; Paquette, L. A.; Wahl, F.; Prinzbach, H. *J. Am. Chem. Soc.* **1994**, *116*, 11775.
- Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, *72*, 2118.
- Ikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540.
- Song, J. W.; Hirose, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 7.
- Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.
- Proynov, E.; Kong, J. *Phys. Rev. A* **2009**, *79*, 4.
- Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- Stewart, P. A.; Gill, P. M. W. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 4337.
- Tsuneda, T.; Suzumura, T.; Hirao, K. *J. Chem. Phys.* **1999**, *110*, 10664.
- Handy, N. C.; Cohen, A. J. *J. Chem. Phys.* **2002**, *116*, 5411.

- (51) Ragot, S.; Cortona, P. *J. Chem. Phys.* **2004**, *121*, 7671.
- (52) Tognetti, V.; Cortona, P.; Adamo, C. *J. Chem. Phys.* **2008**, *128*, 034101.
- (53) Thakkar, A. J.; McCarthy, S. P. *J. Chem. Phys.* **2009**, *131*, 12.
- (54) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (55) Krukau, A. V.; Scuseria, G. E.; Perdew, J. P.; Savin, A. *J. Chem. Phys.* **2008**, *129*, 7.
- (56) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 9.
- (57) Weintraub, E.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Theory Comput.* **2009**, *5*, 754.
- (58) Peach, M. J. G.; Helgaker, T.; Salek, P.; Keal, T. W.; Lutnaes, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558.
- (59) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, 127.
- (60) Zhang, Y. K.; Pan, W.; Yang, W. T. *J. Chem. Phys.* **1997**, *107*, 7921.
- (61) Swart, M.; van der Wijst, T.; Guerra, C. F.; Bickelhaupt, F. M. *J. Mol. Model.* **2007**, *13*, 1245.
- (62) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.
- (63) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107.
- (64) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2004**, *121*, 5654.
- (65) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X. L.; Burke, K. *Phys. Rev. Lett.* **2008**, 100.
- (66) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.
- (67) Ruzsinszky, A.; Csonka, G. I.; Scuseria, G. E. *J. Chem. Theory Comput.* **2009**, *5*, 763.
- (68) Colle, R.; Salvetti, O. *Theor. Chim. Acta* **1975**, *37*, 329.
- (69) Tognetti, V.; Cortona, P.; Adamo, C. *Chem. Phys. Lett.* **2007**, *439*, 381.
- (70) Handy, N. C. *Theor. Chem. Acc.* **2009**, *123*, 165.
- (71) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.
- (72) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (73) Laikov, D. N. *Chem. Phys. Lett.* **1997**, *281*, 151.
- (74) Laikov, D. N. *Chem. Phys. Lett.* **2005**, *416*, 116.
- (75) Laikov, D. N.; Ustynyuk, Y. A. *Russ. Chem. Bull.* **2005**, *54*, 820.
- (76) Aquilante, F.; De Vico, L.; Ferre, N.; Ghigo, G.; Malmqvist, P. A.; Neogrady, P.; Pedersen, T. B.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andres, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (77) Karlstrom, G.; Lindh, R.; Malmqvist, P. A.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P. O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.
- (78) Wigner, E. *Trans. Faraday Soc.* **1938**, *34*, 678.
- (79) Molawi, K.; Cohen, A. J.; Handy, N. C. *Int. J. Quantum Chem.* **2002**, *89*, 86.
- (80) Jung, Y. S.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.
- (81) Pedersen, T. B.; Aquilante, F.; Lindh, R. *Theor. Chem. Acc.* **2009**, *124*, 1.
- (82) Bostrom, J.; Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2009**, *5*, 1545.
- (83) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2009**, 130.
- (84) Pitonak, M.; Neogrady, P.; Cerny, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282.
- (85) Budzelaar, P. H. M. *J. Comput. Chem.* **2007**, *28*, 2226.
- (86) Gordon, M. S.; Schmidt, M. W. *In Theory and Applications of Computational Chemistry: the First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 1167.
- (87) Dunning, T. H.; Hay, P. J. *In Methods of Electronic Structure Theory*; Schaefer, H. F., Ed.; Plenum Press: New York, 1977; p 1.
- (88) Steinmann, S. N.; Csonka, G.; Corminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950.
- (89) Grimme, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 3430.
- (90) Murray, E. D.; Lee, K.; Langreth, D. C. *J. Chem. Theory Comput.* **2009**, *5*, 2754.
- (91) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (92) Tao, J. M.; Gori-Giorgi, P.; Perdew, J. P.; McWeeny, R. *Phys. Rev. A* **2001**, *63*, 032513.

The Dissociated Amorphous Silica Surface: Model Development and Evaluation

Ali A. Hassanali,[†] Hui Zhang,[‡] Chris Knight,[‡] Yun Kyung Shin,[‡] and Sherwin J. Singer^{*,†,‡}

Biophysics Program, and Department of Chemistry, Ohio State University, Columbus, Ohio 43210, United States

Received May 17, 2010

Abstract: At pH 7, amorphous silica has a characteristic negative charge due to the deprotonation of silanol groups on the surface. Electrokinetic phenomena and transport of biomolecules in devices depend sensitively on the surface morphology, distribution of ions and solvent, and adsorption properties of solutes close to the surface in the electrical double layer region. Hence, simulation of these phenomena requires detailed atomistic models of the double layer region. In this Article, we extend our undissociated silica surface model [*J. Phys. Chem. B* 2007, 111, 11181–11193] to include dissociated Si–O[−] groups, which interact with both water and salt (Na⁺ and Cl[−]). We have also conducted ab initio molecular dynamics (AIMD) simulations of a smaller system consisting of a hydrated silica slab. The radial distribution functions predicted by the empirical model are in qualitative agreement with those from the AIMD simulations. The hydrophobic and hydrophilic nature of silanol-poor and silanol-rich regions of the amorphous silica surface observed in our empirical model is reproduced in the AIMD simulations of the smaller slab. In the initial stages of our AIMD simulations, we observe various chemical processes that represent different hydroxylation mechanisms of the surface.

1. Introduction

Interfaces between amorphous silica and water are ubiquitous in chemical, biochemical, and environmental settings.^{1,2} The strong interactions between silica, water, and adsorbates, such as biomolecules,³ make the amorphous silica surface an important and challenging system to model. These strong interactions between silica and analytes are often exploited in chromatographic applications.^{4,5} The interaction of silica and biomolecules is important in several arenas. Silica–DNA interactions are the basis of a standard purification scheme for nucleic acids^{6,7} and novel microfabricated device applications.^{8,9} There has been recent interest in using silica nanochannels to stretch and sequence DNA.^{10,11} Silicates and silica exhibit widely varying health effects, with considerable effort devoted to elucidating mechanisms of toxicity.^{3,12} The importance of biomolecules at the water–silica interface in

a variety of situations has prompted a number of fundamental investigations of the interactions of silica with nucleic acids^{13–19} and proteins.^{20–23} In addition to devices fabricated from silica, the amorphous silica–water interface is important in silicon-based devices where silicon acquires an oxide coating in contact with aqueous solution.^{24,25} At all but the lowest pH values, a silica surface in contact with water is negatively charged. Electrochemical and electrokinetic function of numerous practical devices depends not only on the overall surface charge, but also on the detailed microscopic properties of the silica–water interface. This is because the driving force for electroosmotic flow (EOF), even in fluidic channels of large dimension, originates in the electric double layer,²⁶ the region of excess charge in the fluid that compensates the surface charge. Most of the excess charge of the double layer is typically located within a nanometer from the surface. There is a need for interaction potential models that can describe this region with some realism, and yet will be tractable for simulations that cover large spatial and time scales.

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

[†] Biophysics Program.

[‡] Department of Chemistry.

Recently, we have developed a model for the undissociated amorphous silica interface.^{27–29} In this Article, we describe an extension to that model to include dissociated silanol groups on the surface. The negative charge that develops at the water–silica interface arises from the deprotonation of silanol groups. As one illustration of the applications of this work, we have recently applied the model described herein to study electrokinetic phenomena.³⁰ Using nonequilibrium molecular dynamics (MD) simulations of aqueous solvent containing Na⁺ and Cl[−] ions near a realistically modeled surface, we have critically evaluated the classical Gouy–Chapman–Stern model^{31–34} and the dynamic Stern layer model,^{35–43} used extensively to account for surface ion conduction. The heat of immersion calculated with our model for the undissociated surface is in good agreement with available experimental data.²⁸ Silanol densities typical of common silica materials naturally arose in our annealing and hydroxylating procedure.²⁸ Further calibration of the accuracy of this model, as extended in this work, is supplied in the form of comparisons with ab initio molecular dynamics (AIMD) simulations.

Feuston and Garofalini^{44,45} have developed an empirical potential for the water–amorphous silica interface. In their model, which has been applied and extended,^{46–49} the silicon, oxygen, and hydrogen atoms are not fixed to any molecular unit such as a silanol or water group, but instead can spatially evolve over the course of the simulation based on the interaction potential. Garofalini and co-workers have recently developed an improved dissociative water potential for molecular dynamics simulations⁵⁰ and have applied it to study the chemisorption of water on silica surfaces, yielding important insights into the surface chemistry of water near silica.^{48,49,51,52} To our knowledge, the applicability of this model to dissociated silica surfaces has not been assessed. Rustad and Hay⁵³ have also developed a dissociated model that yields reasonable acid dissociation energies for orthosilicic acid in the gas and aqueous phase. The applicability of this model for very large-scale simulations needed for device applications is however limited. Extending previous work for silicon/silicon oxide systems,⁵⁴ van Duin and co-workers have developed a dissociating force field for the silica–water system⁵⁵ using the ReaxFF force field method.⁵⁶

Schulten and co-workers⁵⁷ have quite recently developed an empirical potential for the water–amorphous silica interface. Their model is exclusively for water interacting with a rigid silica prepared using other models. Their intended fitting procedure was to choose interaction parameters to match the water contact angle reported for quartz surface as a function of the degree of hydroxylation.⁵⁸ The parameters for their surface without silanol groups were adjusted to reproduce the water contact angle for the corresponding quartz surface. Schulten and co-workers report that their model could not reproduce the experimental contact angles of Lamb and Furlong⁵⁸ for any finite degree of hydroxylation. However, contact angle measurements at low degrees of hydroxylation are difficult because the dehydroxylated surface rapidly reacts with water, and other measurements report very different contact angles for dehydroxylated silica.⁵⁹ Jenkins and co-workers^{60–62} simulated

amorphous silica nanoparticles in water starting with initial configurations obtained from DMOL3⁶³ calculations. The simulation was then continued using empirical potentials from the OPLS-AA force field⁶⁴ and borrowed from a simulation of the quartz surface.⁶⁵ Puibasset and Pelenq have modeled the adsorption of water in mesoporous silica from submonolayer coverage to saturation pressures using grand canonical Monte Carlo simulations.^{66–70} Treatment of quartz surfaces involves similar atom types, but as compared to silica is considerably simplified by the regularity of the quartz surface. MacKerell and co-workers⁷¹ have developed a model for the quartz–water interface, which is not designed to characterize dissociated silanol groups. Freund⁷² and Qiao and Aluru^{73,74} have studied electrokinetic transport in electrokinetic channels where the negative charge is distributed on Lennard-Jones wall particles. In these simple models, the heterogeneity and surface roughness are not captured. Aluru and co-workers have also performed similar studies for the interface between water and a more realistic, electrically neutral quartz surface.⁷⁵ Lorenz et al.^{76,77} have employed a modified version of the silica–water potential developed by Schulten and co-workers⁵⁷ to study charge inversion in the presence of divalent cations and electrokinetic phenomena associated with mono- and divalent cations.

In addition to developing an empirical potential, we have used ab initio molecular dynamics (AIMD) simulations in this work, to test several features of our empirical model, including the strong variability in relative hydrophobicity/hydrophilicity⁷⁸ with surface silanol density reported in our earlier work.^{27,28} The physisorption and chemisorption of small numbers of water molecules near silica has also been examined previously by ab initio methods. Physisorption has been studied using quantum chemical fragment calculations by Saengsawang et al.⁸¹ Cheng et al. studied the reaction of an SiO₂ molecule in a cluster with up to six water molecules.⁸² Sutton and co-workers have studied the hydroxylation mechanisms of different silica clusters with ab initio optimization methods using DFT.⁸³ These calculations shed important insight into the activation barriers for chemical reactions between water molecules and silica clusters, although thermal effects are not included in these calculations.⁸³ Similar types of calculations have recently been performed by Konečný and Doren⁸⁴ and Ugliengo and co-workers.⁸⁵ With an eye toward dissolution of silica, Criscenti et al. studied the reaction of a silanol-containing cluster with a hydronium ion in the presence of four water molecules.⁸⁶ Ma et al.⁸⁷ have conducted ab initio MD simulations of water near different defects found in silica clusters and find that several waters are needed for the hydroxylation mechanisms they observe. More recently, Hamad and Bromley⁸⁸ have conducted longer thermal ab initio simulations to study the hydroxylation mechanisms of nonbridging oxygens (NBOs), oxygen atoms that have a single covalent bond to a silicon atom, in small silica clusters. These results suggest that the NBOs are a negatively charged, closed-shell species, whose hydroxylation is quite sensitive to the surface morphology. Du et al. have investigated the reaction of one or two water molecules with an amorphous silica surface using QM/MM techniques,⁸⁹ yielding information about the initial steps of

silica surface hydroxylation. Ab initio molecular dynamics have been used to study the reactions of water with the amorphous silica surface by Masini and Bernasconi⁹⁰ and Mischler et al.,⁹¹ and the pK_a of silanol groups on crystalline silica surfaces by Leung et al.⁹² Tielens et al.⁹³ initiated an AIMD simulation of a hydroxylated silica slab using a starting point from Garofalini's empirical potential models and studied the surface structure, making detailed comparisons with experiment. They calculated the deprotonation energy of silanol groups and the binding energy of individual water molecules. Trilocca and Cormack have reported AIMD studies of liquid water near the related water–bioglass interface,⁹⁴ and Leung et al. have reconstructed the bonding arrangement on crystalline silicate surfaces to mimic the amorphous surface.⁹² To our knowledge, ab initio molecular dynamics of the amorphous silica–water interface, along the lines of what we provide in section 3, have not been reported.

We have performed AIMD simulations on a small hydrated silica slab in which one surface does not contain silanols and the other surface contains several silanols, one of which is dissociated. The relative wetting properties of silanol-rich and silanol-poor regions are confirmed, although the AIMD system was too small to quantitatively compare radial density distributions. We also compared AIMD and empirical results for a single orthosilicic acid molecule in water. Our potential was not designed to fit the properties of orthosilicic acid, but at least the comparison is much more direct because the size of the silicate system is not an issue. We find that the number of waters in the first and second solvent shells compares well for orthosilicic acid, although the radial densities of waters surrounding the silanol groups of orthosilicic acid are more structured in the AIMD simulations. A direct comparison between AIMD and empirical potential results is clouded by several factors. The empirical potential simulations of a silica surface encompass much more surface variability than the smaller, AIMD system. A further limit on our ability to benchmark our empirical potential with AIMD is the limitation on the accuracy of currently available density functionals and the numerical methods available for their implementation. The BLYP^{95,96} and PBE⁹⁷ functionals, as normally implemented, lead to overstructuring of the water, lead to peaks in the radial distribution function of liquid water that are too high and narrow, and underestimate the diffusion constant.^{98–100} The degree of overstructuring is sensitive to the details of the implementation and system size.¹⁰⁰ Lee and Tuckerman have shown that very accurate basis sets are needed to converge liquid state properties,^{101,102} including the diffusion constant.¹⁰³ Exceptionally large charge-density cut-offs are needed to obtain a converged water density.¹⁰⁴

In section 2, we describe the potential form that is used to model our interface between the dissociated silica surface, water, and salt. Within this section, we provide all the parameters of the model for the use of interested readers. We also compare the quality of our empirical potentials to the ab initio quantum chemistry cluster calculations that are used to generate our potentials. A comparison of some properties from the ab initio MD simulations on the smaller

system to those predicted from our empirical model is reported in section 3. We also describe specific chemical events that occur during the first few picoseconds of our AIMD simulations. Finally, we end with a conclusion in section 4.

2. Development of a Model for Dissociated Amorphous Silica

2.1. Formulation of the Potential. Our goal is to model the amorphous silica/water interface using a computationally inexpensive model with sufficient realism to capture essential features of the electrical double layer, electrokinetic phenomena, and adsorbate binding. These phenomena require simulations that extend to large spatial and temporal scales. Models that can describe surface chemistry, either so-called dissociating potentials^{44–46,53,105–107} or AIMD methods, are extremely valuable, but they are not sufficiently tractable for the large-scale calculations we envision. The chemical bond structure will be fixed in the model we propose, although the surface is not constrained to be rigid. For example, the equilibrium between undissociated and dissociated silanols will not be dynamic. For many purposes, the effects of the chemically diverse features of the silica surface can be captured by a distribution of various species on a surface of sufficient size, or a collection of different surface realizations.¹⁰⁸

Like our empirical potential for the undissociated amorphous silica surface,^{27–29} our model for the hydrated dissociated amorphous silica surface extends the BKS model for bulk silica¹⁰⁹ and SPC/E model for water¹¹⁰ to describe the water–silica interface. Hence, the bulk silica and water regions are described by models that have been tested in a variety of physical situations. The BKS and SPC/E potentials succeed in describing physical properties of silicates, both crystalline and amorphous, and water, respectively, but they also have limitations. Among the limitations, especially for the BKS model, is unrealistically large partial charges, which are in place to mimic other physical effects. Concern about the large partial charges was a major motivation for us to compare with AIMD results. The simplicity of these models is both a virtue and limitation: a virtue in the sense that large-scale simulations are possible, but a limitation in that, even with the extensions put in place for surface species, the functional form cannot fit all the available data obtained in fragment ab initio calculations.

In keeping with the BKS model,¹⁰⁹ all silicon and oxygen atoms of the silica are assigned charges of +2.4 and -1.2 . The BKS potential is a sum of pairwise Coulomb and Buckingham (exponential repulsion + r^{-6} attraction) interactions. To avoid excessive overall attraction between water and our undissociated silica surface, we found it necessary to distinguish between silanol-type and siloxane-type oxygens, for the short-range potentials.^{27,28} As shown before, an acceptable fit to ab initio data could be achieved with these fixed charges²⁸ for the water–silanol clusters. Charge neutrality of the undissociated surface requires that the charge on hydrogen atoms be +0.6. In principle, it is possible to introduce silanol groups on the surface with different partial

charges from those used in the BKS potential. In the dissociated surface model, we have added a dissociated oxygen type, O^- , to include dissociated groups on the surface. The charge of the O^- group was required to have a value of -1.6 to maintain a net charge of -1 for each dissociated silanol generated in our simulations.

Without adjusting the charge on silicon atoms, the partial charges on silanol oxygen and hydrogen atoms could be adjusted away from their BKS values of -1.2 and $+0.6$, respectively, as long as their sum is kept at -0.6 , required to maintain charge neutrality of the undissociated surface.²⁸ Further adjustment of the silanol group partial charges, or modification of the O^- charge, would require distinguishing surface silicon atoms with charges different from interior silicons. However, retaining the BKS charges for all atoms greatly simplifies the construction of the hydroxylated amorphous silica (section 3.2) because the atoms that will eventually anneal to surface positions after cleavage of pure silica are not known in advance. Furthermore, it should be noted that it is the sum of the Coulomb and short-range interactions (Buckingham or Lennard-Jones) that is ultimately fit to ab initio quantum chemistry calculations. While realistic partial charges are desirable, it is possible to compensate the unrealistic BKS partial charges with other parts of the complete interaction potential.

As mentioned above, our potential is designed to maintain a given bonded configuration, and not to predict making or breaking of chemical bonds. In some cases, the model has to be extended to prevent unintentional, and sometimes unphysical, bond formation driven by strong Coulomb interactions. Our potential for the undissociated surface included 3-body components that prevented the hydrogen atoms of silanols from binding to more than one oxygen atom.²⁸ We refer to these components as “blocking potentials”. After the introduction of O^- groups, thermal simulations without the additional blocking potentials yielded species where the dissociated oxygen group (O^-) formed new bonds, either bonding to more than one silicon, mimicking a newly formed siloxane bond, or forming another (unphysical) species where the dissociated oxygen is divalently bonded to a silanol hydrogen and a silicon atom. We introduced two blocking potentials in addition to those used for the undissociated surface, given in eqs 1 and 2, which stabilized the chemical bond structure.

$$u_{SiO_D Si}(r_{SiO_D}, r_{O_D Si'}) = k \exp[-\rho^{-4}(r_{SiO_D}^4 + r_{O_D Si'}^4)] \quad (1)$$

$$u_{SiO_D H}(r_{SiO_D}, r_{O_D H}) = k \exp[-\rho^{-4}(r_{SiO_D}^4 + r_{O_D H}^4)] \quad (2)$$

Equation 1 is a 3-body interaction between the O^- of a dissociated silanol group (O_D) and two Si atoms (Si, Si'). Equation 2 is a similar potential involving a Si atom, O^- , and the hydrogen of a nearby silanol group. The values of k and ρ in eqs 1 and 2, which control the strength and range, respectively, of the 3-body blocking potentials are given in Table 1.

Because the BKS potential¹⁰⁹ consists of Buckingham potentials between the oxygen atoms, we also included three other Buckingham interaction potentials between each pair

Table 1. Interaction Parameters for the 3-Body Potentials, Eqs 1 and 2

atom triplet	k (eV)	ρ (Å)
Si- O_D -Si	1000.000	1.6
Si- O_D -H	100.000	1.6

Table 2. Pairwise Lennard-Jones Parameters, Eq 5

ion/atom pair	ϵ (eV)	σ (Å)
Na ⁺ - O_W	0.005406	2.8760000
Cl ⁻ -Si	0.010823	3.8805759
Cl ⁻ - O_W	0.005406	3.7840000
Cl ⁻ -Na ⁺	0.004336	3.4920000
Cl ⁻ -H	0.006378	3.6253307

Table 3. Pairwise Buckingham Potential Parameters, Eq 6

atom pair	A (eV)	ρ (Å)	C_6 (eV Å ⁶)
Si- O_D	13 536.40	0.219247	128.344
O_D - O_D	1388.773	0.3623188	175.0000
O_D -H	5907.000	0.1254160	0.0000
O_D - O_W	6533.490	0.284692	336.7540
O_D - H_W	70.795	0.306200	0.0000
Na ⁺ - O_H	5151.600	0.2725004	60.44529
Na ⁺ - O_X	5151.600	0.2763590	60.00000
Na ⁺ - O_D	40 286.40	0.2133588	70.40000

of O^- groups, a silanol oxygen and O^- group, and finally an O^- and siloxane oxygen. The total silica interaction potential including dissociated groups on the surface is now of the following form:

$$U = U_{\text{pair}} + U_{\text{3-body}} \quad (3)$$

$$U_{\text{pair}} = \sum_{i=2}^N \sum_{j=1}^{N-1} \left(\frac{q_i q_j}{r_{ij}} + u_{ij}(r_{ij}) \right) \quad (4)$$

While Buckingham potentials are often used to describe inorganic solids like silica, Lennard-Jones potentials

$$u_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (5)$$

are commonly used for water, as for SPC/E, ions in water, and biomolecules. Because we need to interface with common force fields, we employed Lennard-Jones potentials, denoted by the symbol “LJ”, for the salt, water, and silica interactions described below. The procedure used to fit the potentials and the quality of fits are described in the following section. The Lennard-Jones potential parameters are collected in Table 2.

In Table 3 are obtained the parameters A , ρ , and C_6 by fitting the ab initio data to a Buckingham potential of the form:

$$u_B(r) = A e^{-r/\rho} - \frac{C_6}{r^6} \quad (6)$$

The Buckingham potential as shown in eq 6 has an unphysical maximum at small r , and then plunges to $-\infty$ at even smaller r . It is thus standard to modify the Buckingham potential to eliminate the unphysical maximum and make it

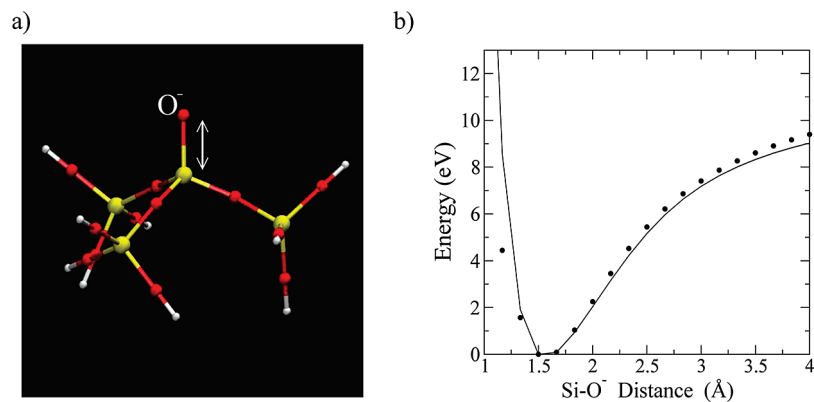


Figure 1. Fitted (solid line) versus ab initio energies (●) for (b) Si–O[−] stretch for fragment shown in (a).

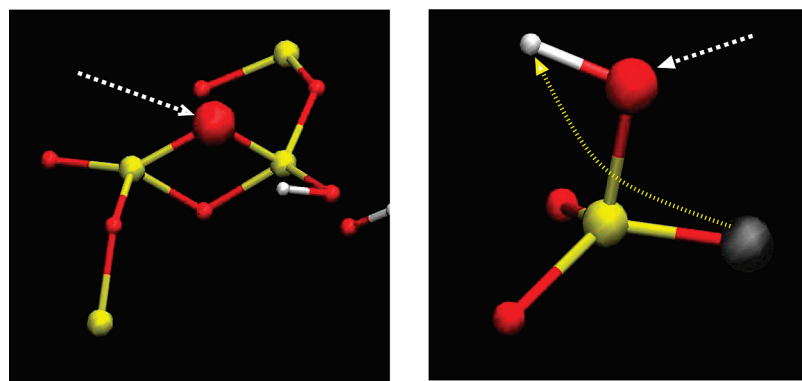


Figure 2. Left panel shows the formation of a two-membered ring involving a dissociated oxygen that is divalently bonded to two silicons. The dotted white arrow points in both cases to the dissociated oxygen involved in the interaction. In the right panel, the dotted yellow arrow shows the hydrogen that originates from one of the hydroxyl groups of a geminal silanol pair, which are attached to the same silicon atom.

smoothly repulsive at small r . Details of the small r patch are described in our previous work.²⁸

For the convenience of those wishing to use the potentials, the final parameters forming the interaction potentials U_{pair} and $U_{3\text{-body}}$ for the extension of our undissociated surface are collected in Tables 1–3. Readers are referred to our previous work for the other interaction parameters that were unchanged in modeling the dissociated surface.^{28,29} In the tables, the subscript “X” refers to oxygen types that are not connected to hydrogens but different from dissociated types (i.e., O[−]), the subscript “D” refers to oxygen types that are dissociated, and the subscript “H” refers to oxygens that are part of silanol OH groups.

2.2. Parameter Adjustment To Match Ab Initio Data. Using fragments excised from a silica surface, ab initio quantum mechanical calculations were performed from which we fit our empirical potentials. As in our previous work,²⁸ the parameters of empirical potentials are adjusted to match ab initio data generated for coordinate grids. This section is devoted to describing the quality of the match between our empirical potentials and ab initio data. The Born–Oppenheimer energy surface was obtained using MP2 perturbation theory^{111–115} to account for electron correlation. We use the 6-311++G** basis set, and all electronic structure calculations were performed using Gaussian 03.¹¹⁶

2.2.1. Dissociated Groups. The potential for the dissociated groups Si–O[−] consists of a single Si–O[−] Buckingham plus Coulomb interactions. The Si–O[−] distance in the

fragment shown in Figure 1a is varied between 0.7 and 4.0 Å, and the resulting energies are used to fit the Buckingham potential form. The quality of the fit for the Si–O[−] stretch is shown in Figure 1b.

As indicated earlier, we found that thermal simulations without blocking potentials yielded configurations as seen in the left panel Figure 2 where an O[−] is divalently bonded to two silicons. We also found that the hydrogens of silanol oxygens can transfer to dissociated oxygens during the simulations as seen in the right panel of Figure 2. In this defect, the hydrogen on the dissociated oxygen group originated via the deprotonation of the silanol oxygen shown in translucent red. Our model is not intended to capture realistic deprotonation and protonation of oxygen species on the surface. To prevent new chemical structures from forming on the surface, we inserted 3-body blocking potentials as described in eqs 1 and 2. These were not fit to ab initio data. Instead, the parameters were adjusted by trial and error until these defects no longer occurred. Our potential form also includes the BKS Buckingham potential between the dissociated oxygen type and all other oxygens and silicon atoms of the silica surface to ensure that the surface bond structure remains intact.

2.2.2. Water-Dissociated Silica Interactions. The water-dissociated silica potentials were fit using a path of approach for a water molecule near a dissociated group shown in the left panel of Figure 3. The water-dissociated oxygen separation was varied from 1.8 to approximately 13 Å to obtain a

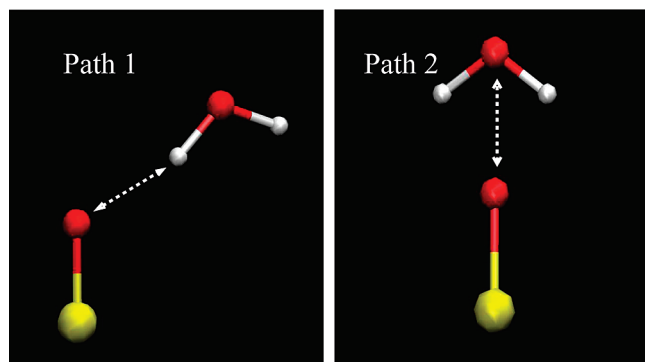


Figure 3. Paths of approach for water near a dissociated silanol.

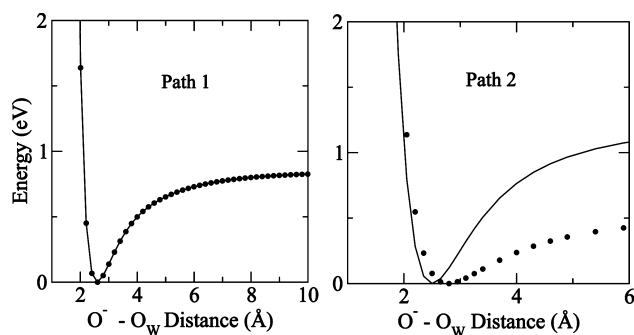


Figure 4. Fitted (solid line) versus ab initio (filled symbols) energies for the paths of approach shown in Figure 3. On the left are path 1 energies, and on the right path 2 energies.

comprehensive sweep of the potential energy surface. The water geometry was fixed to that of the SPC/E model¹¹⁰ and was not allowed to relax as it approached the dissociated silanol. This scheme is similar to our design of a potential between water and undissociated silanols, for which the reader is referred to our previous work.^{27,28}

In most common water–water and water–ion potentials, only the water oxygen carries a non-Coulombic interaction, like a Lennard-Jones potential. The potential surface for these common water potentials contains basins where the potential diverges to $-\infty$ when a hydrogen is superimposed on an oxygen. These unphysical regions are never discovered in simulations. However, the dissociated silanol oxygen carries a large negative charge, and the attractive Coulomb infinity was discovered in preliminary simulations. We found it necessary to include an exponential repulsive Buckingham between the dissociated oxygen and water hydrogen (O_D-H_W potential).

To test the transferability of our water-dissociated silica potential to another path of approach, a water molecule was brought toward a dissociated oxygen shown in the right panel of Figure 3. Ab initio calculations indicate that path 1 (Figure 3) is the more favorable path for approach of a water to a dissociated silanol. We were not successful in reproducing this feature in our empirical potential using the available flexibility in the Buckingham + Coulomb form of our potential, together with the many constraints on partial charges (BKS charges for silica Si and O, SPC/E charges for H_W and O_W). Figure 4 shows all the ab initio energies for the two paths of approach, and the predictions of our

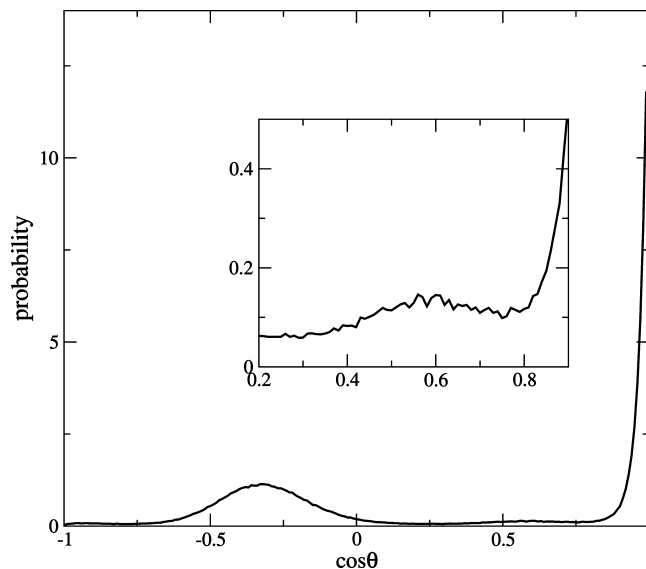


Figure 5. Orientational distribution of water near dissociated oxygens. The distribution is given for the dot product between unit vectors linking water oxygens to a dissociated surface oxygen (O_W-O_D) within a distance of 3.5 Å, and vectors along each of the two hydrogen bonds. The data confirm that in the thermal simulations, path 1 (Figure 3) is favored over path 2. The inset shows a small peak near the cosine of one-half the tetrahedral angle, the angle of the bifurcated structure of path 2.

empirical potential for each of these paths. Rather than “split the difference”, we chose to fit the ab initio results along path 1 because preliminary simulations indicated that the binding geometry of path 1 was preferred, even when the most stable orientation of an isolated water was the bifurcated structure of path 2. This tendency is confirmed for the final form of the potential. Figure 5 shows the distribution of the angle between the vectors O_W-O_D and O_W-H_W for waters within 3.5 Å of each O^- group on the dissociated surface. These results show that in the thermalized simulations the cosine of the angle of one of the two water hydrogens is close to 1, and hence path 1 is the most likely orientation on the surface despite the fact that our empirical potential predicts a higher binding energy for a single water molecule along path 2. In Figure 5, there is a second peak near $\cos \theta = -1/3$ for the hydrogen not hydrogen bonding to the dissociated oxygen. The inset in Figure 5 reveals a very small peak corresponding to the bifurcated structure.

2.2.3. Na^+ and Cl^- Silica and Water Interactions. Initial simulations indicated that the Na^+ ions can strongly interact not only with dissociated silanols, but also silanol and siloxane oxygens. For this reason, we selected fragments from our simulations with and without a dissociated silanol, where a Na^+ ion interacts with all species to generate empirical potentials for these interactions.

Shown in Figure 6 are the three fragments that were used for the ab initio calculations. The first fragment consists of a sodium ion interacting with a single dissociated oxygen and siloxane oxygens. The second fragment consists of a sodium ion interacting with the silanol and siloxane oxygens, while the third fragment consists of a sodium ion interacting with all three species of oxygens. The free parameters of

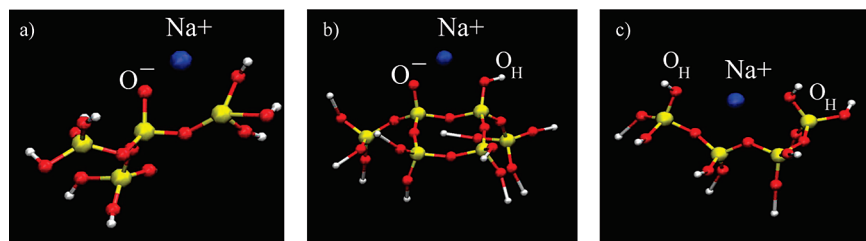


Figure 6. The three fragments used for silica–sodium interactions labeled (a) with a single O_D , (b) with a single O_D and silanol OH , and (c) with three silanol OH 's. Unlabeled red spheres correspond to siloxane oxygens in our empirical potential. The blue sphere is the sodium ion.

Table 4. Fitted and Ab Initio Binding Energies for Silica– Na^+ Fragments

fragment	ab initio (eV)	empirical (eV)
frag 1	−6.44	−5.59
frag 2	−6.85	−7.35
frag 3	−3.05	−2.98

the sodium–oxygen Buckingham potential given in Table 3 were adjusted to match the ab initio binding energies.

For the fragments with a dissociated oxygen, it was not possible to generate a full potential energy surface representing the approach of a Na^+ ion to a surface fragment using ground-state ab initio methods. Without solvent, the ground state of this system at large separations is a neutral sodium atom and neutral fragment. The ground state becomes ionic in character as the sodium approaches the fragment. Hence, we relied on the binding energy to fix the non-Coulombic interaction parameters between the Na^+ ion and silica, the partial charges being constrained to the BKS values. The binding energy was calculated by separately calculating the energy of the ion–fragment complex, and then the energies of the isolated ion and fragment.

$$\Delta E_{\text{bind}} = E(\text{fragment}/Na^+) - E(\text{fragment}) - E(Na^+) \quad (7)$$

The comparison of the empirical and ab initio binding energies of a Na^+ ion to the three fragments shown in Figure 6 are reported in Table 4. The Si–O bond length and the position of the sodium ion were initially optimized with an ab initio calculation for the first fragment shown in Figure 6a. The Si–O bond length used in the other two fragments was very close to the optimized bond length obtained for the first fragment.

Chloride ions are repelled from a negatively charged silica surface, and, as compared to sodium, chloride ions will have considerably weaker interactions with the silica surface. Hence, the interactions between the chloride ions and the silica surface were calibrated without further quantum chemical calculations using available potentials in the GROMOS96 force field.¹¹⁷ Our preliminary simulations also indicated that the Cl^- ions quickly explore the Coulomb infinity at the positively charged silicon and silanol hydrogen atoms. For this reason, short-range LJ potentials were inserted between these species. These potentials were derived using combining rules with potentials from the GROMOS96 force field. The interaction parameters between the sodium and chloride ions and water were obtained from previous work

by Dang.¹¹⁸ The parameters used in our simulations are shown in Table 2.

3. Comparison of Ab Initio and Empirical Results

3.1. Simulation Methods. In this section, we describe the evaluation of our empirical potential using ab initio MD simulations on a smaller hydrated slab system. Because of computational feasibility, AIMD simulations were limited to rather small systems. The starting configuration for the AIMD simulations was generated by annealing a bulk crystalline silica tridymite structure with no free surfaces measuring $10.13 \text{ \AA} \times 17.55 \text{ \AA} \times 8.275 \text{ \AA}$, and consisting of 32 silicon atoms and 64 oxygen atoms using the BKS potential¹⁰⁹ with dispersion interactions truncated to one-half of the shortest side of the box. The protocol that was used to generate amorphous silica from the starting material was adopted from cycle I–IV of Huff and co-workers,¹¹⁹ as in our previous work. We cleaved the surface by opening a gap in the z -dimension, followed by annealing for 5 ps at 300 K. This small surface yielded one two-membered ring on one surface and no structural features on the other surface that would lead to silanol groups. The two-membered ring was opened with the addition of water to form two nearby silanol groups in a process described below in section 3.2, initially yielding a total of 2 silanols on one of the surfaces and no silanols on the other. After hydroxylation, a total of 71 waters were added to the system. This hydrated silica slab was then used as input for the ab initio MD simulations. Further reaction of water with the surface during the ab initio MD simulations led to the appearance of more silanol groups on the side that already contained silanols, and none on the other side of the slab. This is appealing because it furnished a means to test our findings²⁸ that silanol groups make the silica surface hydrophilic, while surfaces depleted in silanols are relatively hydrophobic.⁷⁸ Unfortunately, we were unsuccessful in conducting a classical simulation of systems of the hydroxylated surface using our empirical potential due to technical limitations implementing 3-body interactions in the DLPOLY package for a small system size. For this reason, we could not use a starting configuration for the ab initio MD simulations obtained by first equilibrating with our empirical potential, or compare data for AIMD and empirical simulations of exactly the same system. We emphasize that the 3-body interactions were omitted only for the small system size that is used to generate a starting configuration for the ab initio MD simulations. The 3-body

Table 5. Summary of Systems Simulated in This Work Using Both Our Empirical Model and the Ab Initio MD Method^a

system	method	simulation length	cell dimensions
large dissociated silica slab	empirical	9 ns	32.83 × 32.84 × 127.59 Å ³
small silica slab (initial configuration for AIMD)	empirical	145 ps	10.13 × 17.55 × 8.275 Å ³
small silica slab + excess proton	AIMD (DZVP)	22 ps	8.837 × 15.309 × 27.219 Å ³
small silica slab + Na ⁺	AIMD (DZVP)	15 ps	8.837 × 15.309 × 27.219 Å ³
orthosilicic acid	empirical	2 ns	12.75 × 12.75 × 12.75 Å ³
orthosilicic acid	AIMD (DZVP)	20 ps	12.75 × 12.75 × 12.75 Å ³
orthosilicic acid	AIMD (TZV2P)	15 ps	12.75 × 12.75 × 12.75 Å ³

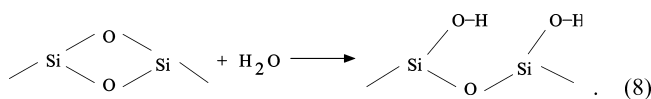
^a The 9 ns empirical model simulation was used to obtain the results presented in section 3. The small, 145 ps empirical simulation provided the starting configuration for the AIMD simulations of a silica slab.

interactions were implemented for the larger systems that are used in our production runs to compare the empirical and ab initio MD results.

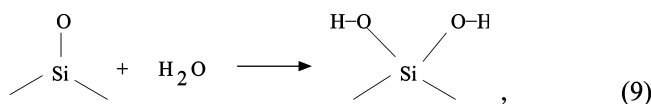
The AIMD simulations were conducted using Quickstep, which is part of the CP2K package.^{120,121} In these calculations, ab initio Born–Oppenheimer molecular dynamics is used for propagation of the classical nuclei. The electronic orbitals are converged to the Born–Oppenheimer surface at every step in the molecular dynamics simulation. The wave function was optimized using an orbital transformation method¹²² in conjunction with the DIIS scheme,^{123,124} as described in ref 121. The convergence criterion for optimization of the wave function was set to 10⁻⁶. Using the Gaussian and plane waves (GPW) method, the wave function was expanded in the Gaussian DZVP basis set. While a triple- ζ basis set was not feasible for the silica slab, simulations using the larger TZV2P basis set were used to check convergence of our radial distribution functions with respect to basis set size for orthosilicic acid in water, as reported below. An auxiliary basis set of plane waves was used to expand the electron density up to a plane wave cutoff of 300 Ry. We used the Becke–Lee–Yang–Parr gradient correction^{95,96} to the local density approximation and Goddecker–Tetter–Hutter (GTH) pseudopotentials.¹²⁵ A time step of 0.5 fs was used in all ab initio simulations. During the first 3 ps of the simulation, we observe several chemical processes that occur on the surface. These chemical processes will be documented in detail later. At this stage, reaction with water adds 3 silanols to the surface that already contained 2 silanols. Exchange of hydrogen and oxygen atoms between the water molecules and atoms on the silica surface results in the formation of an extra proton that, during the length of our simulations, fluctuates between two Eigen structures and samples a Zundel complex during the fluctuations. The presence of the excess proton in the solvent near the surface during the course of the simulation suggests that the silica slab is negatively charged, which is confirmed below. The simulation with the proton in the bulk was then run for a total of approximately 22 ps. For data analysis, the first 6 ps of this simulation was treated as equilibration. An additional simulation was begun from a configuration chosen from the first 3 ps of this simulation, where the proton was replaced by a Na⁺ ion. This simulation was run for a total of approximately 15 ps. For the data analysis, the first 4.5 ps of the run with Na⁺ was treated as equilibration. All the AIMD simulations described were conducted within the NVE ensemble.

We also conducted AIMD simulations of orthosilicic acid using the DZVP and TZV2P basis sets. These simulations consist of a single orthosilicic acid molecule (Si(OH)₄) surrounded by 66 water molecules in a box of side length 12.75 Å. Using the same methodology described above, AIMD simulations of length 20 ps were conducted using the DZVP basis set and 15 ps using the TZV2P basis set. Simulations of this system were also performed using our empirical model. All the simulations conducted with our empirical model and DFT ab initio method are summarized in Table 5.

3.2. Construction of a Dissociated Silica Surface. In our previous work,²⁸ we described the method used to generate a hydroxylated silica surface. We will briefly review some of those procedures in this section. The starting configuration begins with crystalline tridymite with no free surfaces. The amorphous silica slab is then generated using a protocol adopted from cycle I–IV of Huff and co-workers.¹¹⁹ The slab is cleaved by opening a gap in the *z* dimension followed by annealing for a limited time at 300 K. The silica surface consists of various species such as two-membered (2M) rings, nonbridging oxygens (NBOs), and three coordinated silicons. In Table 4 of our earlier work,²⁸ we showed that the length of annealing time in the simulations controls the population of these species on the silica surface and consequently the silanol density of the silica surface. Thus, the concentration of silanols on the surface can be controlled in a systematic way to reproduce reported experimental values of the silanol density.^{1,126} The 2M rings were manually converted into a pair of vicinal silanols, that is, two close by silanols, as seen below:



while the NBOs were transformed into geminal silanols:



which are attached to the same silicon atom. These procedures follow likely hydroxylation reactions identified by previous ab initio calculations^{90,91,127,128} as described in our earlier work.²⁸ The detailed structure of the aqueous silica surface, its mechanism of hydroxylation, and the distribution of isolated, vicinal, and geminal silanols have not yet been conclusively established theoretically or experimentally, and certainly more work is needed in this area.

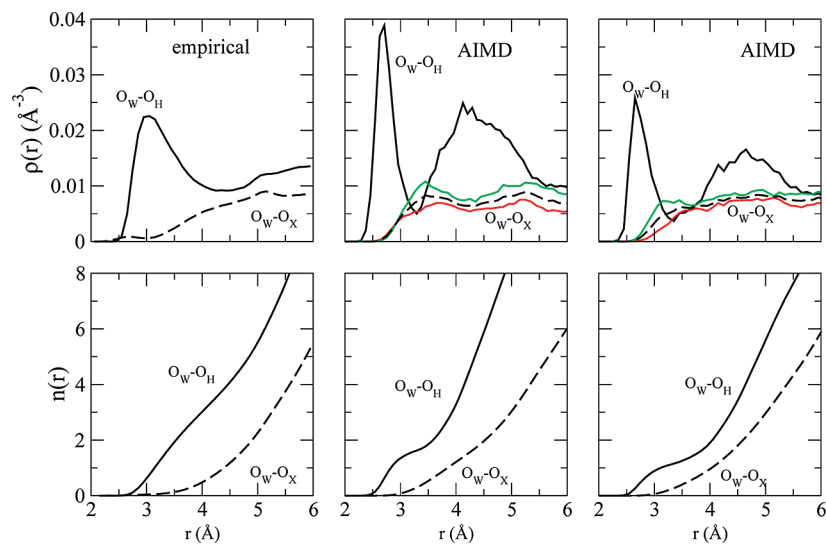


Figure 7. Radial density of water (top row), and cumulative number of neighboring waters (bottom row) near silanol (solid line) and siloxanes (thick dashed lines) for the empirical model shown in left panel, AIMD simulation with proton in middle panel, and AIMD simulation with Na^+ shown in right panel. The green and red thick lines are radial densities near siloxanes in only hydrophilic and only hydrophobic regions, respectively.

Once the hydroxylated surface is formed, silanols on the surface need to be selected for deprotonation to form O^- groups. Because of the lack of any experimental or theoretical insight, the silanols are deprotonated in a random fashion except that we never deprotonate both hydroxyl groups of geminal silanols because we expect that charge repulsion will make doubly dissociated geminals a high energy species. For the simulations reported here using our empirical model, 18 dissociated sites, 9 on each surface of the silica slab, are chosen so that they are separated by at least 6 Å. Our empirical simulations for the dissociated surface consist of a box measuring $32.83 \text{ \AA} \times 32.84 \text{ \AA} \times 127.59 \text{ \AA}$. Along the box height of 127.59 Å, water occupies approximately 92 Å, and the silica slab fills the remainder. The charge density of the surface was 0.835 e nm^{-2} .

3.3. Radial Densities Near Silanol Groups of Silica and Orthosilicic Acid. Shown in Figure 7 is a comparison of the radial distribution functions for water near the silanol oxygens and siloxanes from our empirical model, and for the ab initio simulations for the two systems described above. We note that the striking difference in water density near silanol and siloxane oxygens, originally noted in our study of the undissociated surface,²⁸ is confirmed by AIMD. Previously, we had reported²⁸ that our empirical model did not perform favorably in reproducing the paths of approach of a single water molecule to a siloxane group as compared to ab initio results. This discrepancy does not appear to directly affect the hydrophobic/hydrophilic property of the surface. Water is depleted near siloxanes regardless of the proximity of silanol groups. It occurs on both sides of the AIMD slab, the side with no silanols and the side with 5 silanol groups (red and green curves in Figure 7).

The comparison between empirical and AIMD radial water densities in Figure 7 shows under-structuring by our empirical potential as compared to AIMD in the form of lower and wider density peaks. Also, the radial distribution function (RDF) for the water oxygen–silanol oxygen peaks at about 3.0 Å in our empirical model, but is peaked at about 2.7 Å

in the ab initio simulations. Studies of orthosilicic acid reported below confirm that part of this trend can be attributed to the empirical potential parameters, which tend to under-structure water near silanol groups. However, very limited sampling of silanol group environments in the AIMD runs also contributes significantly to the difference between AIMD and empirical results in Figure 7. As noted earlier, it was not possible to carry out empirical and AIMD simulations on exactly the same system. The empirical potential simulations were performed on a much larger sample. The surface area of the slab used in the empirical potential simulations is approximately 8 times that used in our AIMD simulations. Thus, the empirical potential simulations encompass a much larger range of silanol environments. The variety of local environments is illustrated by radial density plots for water near four individual silanol groups in our empirical potential simulations shown in Figure 8. A buried silanol group is found in the left-most plot, and the cumulative population plot beneath it shows that relatively few waters are surrounding it. All four examples exhibit a sharper first peak than the overall average for the empirical potential surface in Figure 7, where such features are washed out. The right-most plot in Figure 8 shows a silanol group with abundant exposure to water and a pronounced minimum between the first and second peaks. In contrast to the diversity encompassed by the empirical simulation, the environments for the five silanol groups in the ab initio simulations were relatively similar, and individual radial water densities near individual silanol groups for the AIMD runs did not show strong variation. To obtain a quantitative comparison of the radial distribution functions, larger AIMD system sizes over multiple realizations that have surface morphology similar to that used in our larger empirical model would be required.

In the left and right panels of Figure 9, we compare the radial densities of sodium ions and water about dissociated groups and the radial density of water about sodium ions from our empirical model and AIMD simulations, respectively. As mentioned before, the AIMD simulations consist

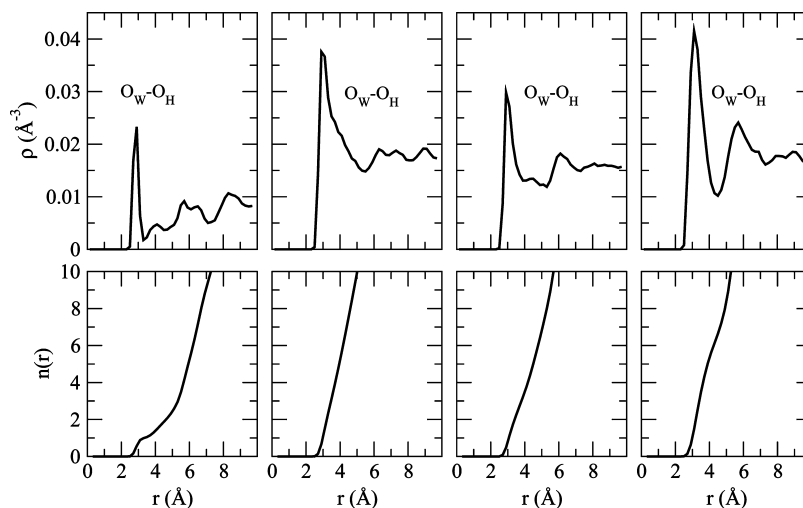


Figure 8. Radial density of water (top row), and cumulative number of neighboring waters (bottom row) near four individual silanol groups from empirical potential simulations.

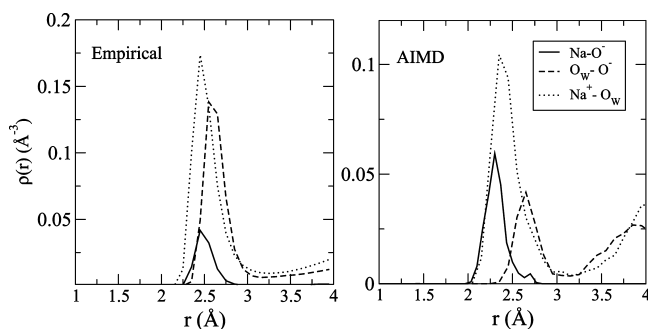


Figure 9. Radial density of Na^+ ions near O^- groups (solid thick line), O_w near O^- groups (solid dashed lines), and Na^+ near O_w (solid dotted line) for our empirical model in the left panel and AIMD simulations in the right panel.

of only a single sodium ion, while the empirical model simulations are averaged over many sodium ions and dissociated O^- groups. As for silanol groups (see discussion of Figure 7), we cannot make a direct comparison of O^-

groups averaged over the entire empirical potential surfaces with a single O^- group in the AIMD simulation. With these caveats, a comparison of the positions of the maxima of the various densities suggests that our empirical model for the dissociated surface at least does not contradict the AIMD results. The $\text{O}^- - \text{Na}^+$ density peaks at about 2.3 Å for the single dissociated oxygen in the AIMD simulations and 2.45 Å in our empirical model over 18 dissociated oxygens. The $\text{O}_w - \text{O}^-$ density peaks at about 2.65 Å in the AIMD simulations and 2.6 Å in the empirical model, and finally the $\text{Na}^+ - \text{O}_w$ density peaks at about 2.4 Å in the AIMD and at 2.45 Å in our empirical model. The empirical model qualitatively reproduces the trend observed in the AIMD simulations of a larger first peak position in the $\text{O}^- - \text{O}_w$ radial density as compared to the first peak positions in the $\text{O}^- - \text{Na}^+$ and $\text{Na}^+ - \text{O}_w$ densities.

The radial densities of the water oxygens and atoms on orthosilicic acid are shown in Figure 10. The radial densities show good agreement between the DZVP and TZV2P basis

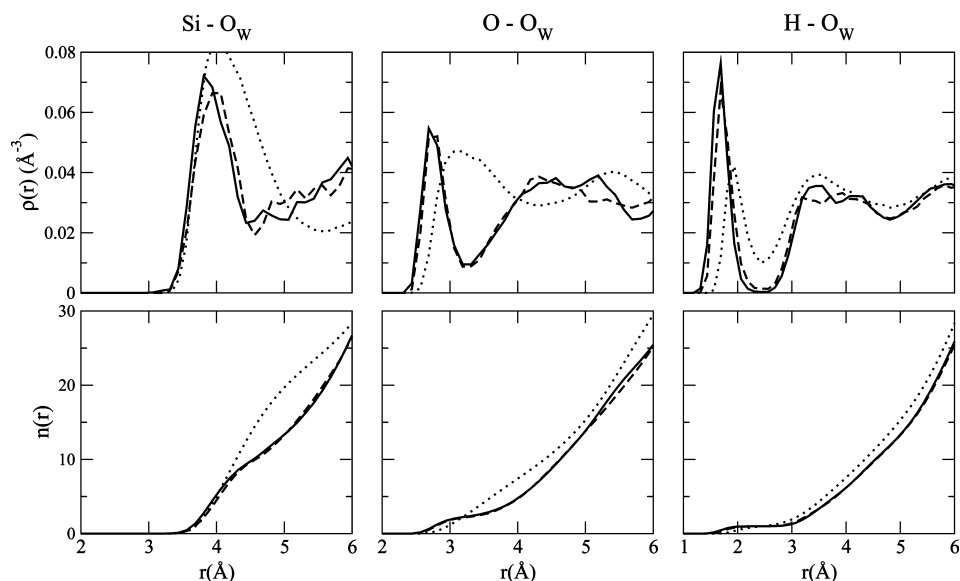


Figure 10. Radial density of water (top row) and cumulative number of neighboring waters (bottom row) near silicon, oxygen, and hydrogen atoms of orthosilicic acid shown for DZVP (solid black), TZVP (solid dashed), and empirical model (solid dotted).

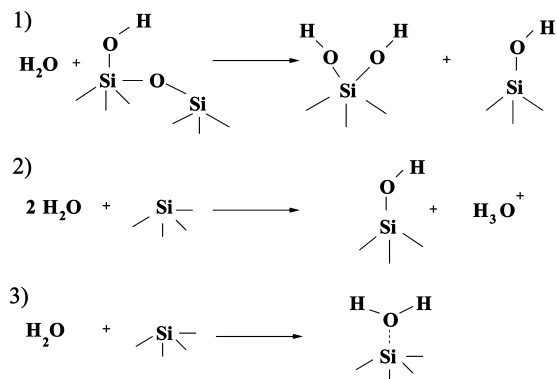


Figure 11. Three different chemical processes that occur within the first 3 ps of our AIMD simulations are labeled 1–3. The first shows the conversion of an isolated silanol (silicon atom with a single hydroxyl group) to a geminal silanol and another isolated silanol. The second shows the formation of an isolated silanol and a hydronium ion, and the third shows a single water molecule sticking to an exposed silicon atom on our hydrophobic surface.

sets, suggesting that the radial densities in Figure 7 that confirm the hydrophilic (hydrophobic) property of silanols (siloxanes) are at the very least qualitatively converged with respect to the basis set. Convergence with respect to the charge density cutoff was not explored.¹⁰⁴ We find that our empirical model, which was not designed to model orthosilicic acid, under-structures the water near the orthosilicic acid. However, the cumulative neighbor populations shown in the bottom row of Figure 10 demonstrate that the number of waters near orthosilicic acid matches well between AIMD and our empirical model.

3.4. Description of Surface Chemistry in AIMD Simulations. We observed interesting chemical processes occurring on the surface within the first 3 ps after a freshly cleaved silica surface was exposed to water in *ab initio* simulations. This resulted in the formation of three more silanol groups on one surface. Some of these important events will be reviewed in this section. These results should be viewed as preliminary information on processes that deserve much further study. It should be noted that the starting configuration for the AIMD simulations was not pre-equilibrated using our empirical potential and hence represents a high energy starting configuration that is subject to significant surface relaxation. Further work is needed to show

how sensitive the occurrence of these chemical events is to larger system sizes, and different initial surface morphologies that may arise from annealing and equilibration protocols using our empirical model.

The schematic shown in Figure 11 illustrates three separate processes that occur on our surface. The first process shows the conversion of an isolated silanol to a geminal silanol and another isolated silanol, resulting in the addition of two more silanols to the surface. The intermediate steps are seen more clearly in Figure 12. As a water molecule binds strongly to the silicon atom of the isolated silanol, one of the Si–O bonds associated with the silicon atom lengthens, resulting in a nonbridging oxygen (Figure 12i). At the same time, the water molecule splits, donating an OH group to the silicon, transiently forming a hydronium ion H_3O^+ (Figure 12ii). The proton that forms the transient hydronium is then recaptured by the newly formed silanol. Simultaneously, the hydrogen that was originally added as part of the OH group to the silicon is transferred to the nonbridging oxygen (Figure 12iii). The resulting product is a geminal silanol and an isolated silanol (Figure 12iv). The role of the formation of transient hydronium ions during the chemisorption of small numbers of waters onto the silica surface has previously been observed by Du et al. in QM/MM simulations, by Mahadevan and Garofalini using a dissociating potential model,^{48,50} and, for large silica clusters, by Ma et al. with *ab initio* MD simulations.⁸⁷

Snapshots from the second hydroxylation scheme of Figure 11 are shown in Figure 13. A two-coordinate silicon is transformed into an isolated silanol along with the transfer of a proton to the water. Ma and co-workers,⁸⁷ who have also conducted *ab initio* MD simulations of water near silica clusters, found that the two-coordinate silicons in their simulations were highly unreactive with water. In our simulations, we observe that as the water molecule strongly binds to the under-coordinated silicon, the OH group of the water is transferred to the silicon, and the proton is transferred to the surrounding water forming a hydronium ion H_3O^+ . During the transfer of the OH group from the water to the silicon, one of the Si–O bond lengths associated with the silicon increases by approximately 0.2 Å. This process involves charge separation between the slab and the solvent, as illustrated by the charges in Table 6 obtained using the DDAP¹²⁹ charge partitioning method from 192 configurations

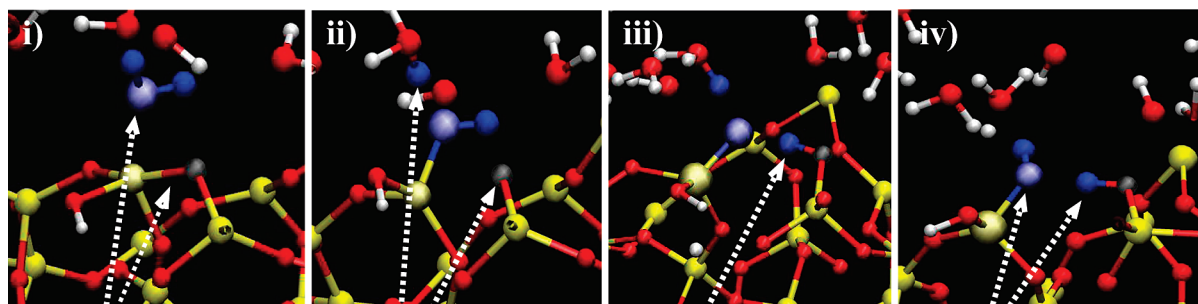


Figure 12. Steps in the formation of a geminal and isolated silanol from an initial single silanol, process (1) in Figure 11. The atoms of the water molecule that reacts with the surface are shown in blue in all four frames. (i) Arrows point to water that attacks silicon atom and Si–O bond that begins to break. (ii) Arrows point to the non bridging oxygen formed after the Si–O bond breaks and the hydronium ion transiently formed. (iii) Arrow points to the proton transferred from the newly formed silanol to the NBO. (iv) Final products are a geminal and a single silanol.

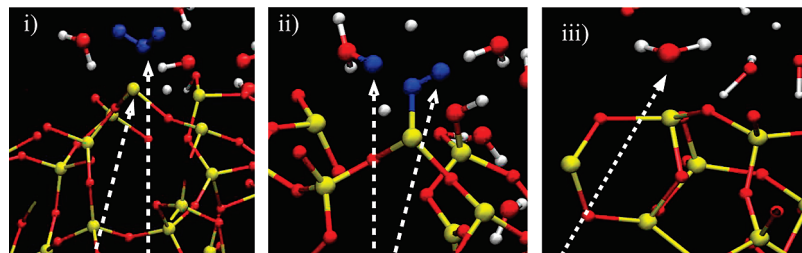


Figure 13. The left and middle frames illustrate steps in the formation of an isolated silanol from an under-coordinated silicon, scheme 2 of Figure 11. The atoms of the water molecule that reacts with the surface are shown in blue in these frames. (i) Water attacks an undercoordinated silicon atom. (ii) An OH group is added to silicon and hydronium ion is formed. (iii) The frame on the right shows a water molecule on the hydrophobic surface that binds to an exposed silicon atom without further reaction during the length of the AIMD simulation.

Table 6. Average Charges of Species in AIMD Simulations^a

species	slab + excess proton	slab + Na ⁺
slab	-0.923	-0.807
O (silanols)	-0.659	-0.690
O (siloxanes, hydrophilic side)	-0.830	-0.792
O (siloxanes, hydrophobic side)	-0.782	-0.794
NBO	-0.832	-0.904

^aCharges were obtained according to the DDAP¹²⁹ charge partitioning scheme, which is based on the electron density.

after an excess proton is transferred to the solvent. The silica slab carries close to a full electron charge (-0.923), and the solvent carries a corresponding positive charge. All the oxygens of the silica slab have a charge close to -0.8 , including NBOs. After a proton was transferred to the solvent and the surface was stabilized, we could identify a buried NBO within a cleft on the surface, which was solvated by a single water molecule. This could be considered as the location of the excess charge on the silica slab.

The initial configuration for the AIMD simulation with the Na⁺ was constructed by replacing a proton from the newly formed H₃O⁺ (Figure 13ii) with a Na⁺. The newly formed silanol (Figure 13ii) was deprotonated to form a nonbridging oxygen (NBO), and the proton was transferred to a nearby siloxane oxygen, which was converted thereafter into a silanol. The charge partitioning analysis was conducted for 192 configurations sampled from our simulations with the Na⁺. The magnitude of the charges is shown in Table 6. The average charge of the Na⁺ in our simulations is $+0.726$, indicating that a significant amount of the positive charge has leaked to its surroundings. Furthermore, the data also suggest that a significant proportion of the formal -1 charge of a singly dissociated silica surface leaks out on to its environment. Charge transfer between solvated ions and solvent has already been reported in several systems. Klein and co-workers¹³⁰ conducted ab initio simulations of a zwitterionic peptide, halide anions, and alkali cations in water. They observed a substantial amount of charge transfer between the carboxylate terminus and the solvent ($0.1e$). Similar values were also found for K⁺ and Na⁺ ions. Chloride and bromine anions were found to transfer more electronic charge ($0.26e$) onto the surrounding solvent.

It is interesting that the charges of the oxygens of hydrophilic silanol groups are less than the oxygen from

hydrophobic siloxane bonds in Table 6, once again confirming that hydrogen-bonding interactions are not purely electrostatic. The magnitude of the density-derived charges from the AIMD simulations, especially for silica atoms that follow the BKS model, is less than the charges used in our model. However, these differences do not preclude qualitative agreement between the empirical and AIMD predictions of the hydrophobic/hydrophilic property of the silica surface. Future work in the development and improvement of empirical models for amorphous silica may require adjustment of the charges in the empirical potential, guided by data shown in Table 6.

Finally, we find that the oxygen of a single water molecule interacts intimately with a 4-coordinated silicon atom throughout the length of our simulation on the hydrophobic side of the silica surface (third scheme of Figure 11 and right panel of Figure 13). The Si–O_w distance for this water molecule fluctuates between 1.8 and 2.0 Å. Although the silicon atom is quite exposed to solvent, we do not observe any chemistry occurring at this site within the time scales of the simulations conducted. Perhaps longer simulations would lead to a hydroxylation event.

4. Conclusion

In contact with aqueous solution with pH greater than ~ 3 , the silica surface is negatively charged because a fraction of the silanol groups are dissociated and an electrical double layer is present.¹ In this work, we extended our empirical model for the undissociated surface. The silica surface in this model is flexible, so it can be used to model heat transport in nanoscale devices. The extended model now includes deprotonated silanol groups, and specifies water and sodium and chloride ion interactions with the surface. We employed the same strategy that was used in the previous model for the undissociated surface,^{27,28} fitting the parameters of our empirical potential to accurate ab initio cluster calculations.

We have made comparisons with experimental quantities, such as the heat of immersion, where measurements are available.²⁸ Otherwise, we have relied on comparisons with quantum chemical and ab initio molecular dynamics results. We used well-calibrated and widely used models, BKS¹⁰⁹ and SPC/E,¹¹⁰ to describe bulk silica and water, respectively, away from the interface. However, these models employ somewhat unrealistic partial charges to mimic other physical

interactions. Therefore, the benchmarking against quantum chemical and AIMD data is essential. We have confirmed a key prediction of our empirical potential, the striking difference in water density near silanol and siloxane groups. Despite the fact that the partial charges of the empirical model, especially the BKS parameters, are larger than those predicted by electron density calculations, the data for orthosilicic acid suggest that, if anything, our silica surface slightly under-structures nearby water. Populations of neighboring water near the surface are in good agreement with AIMD results. The evident under-structuring of our empirical potential as compared to AIMD may be, in part, attributed to the tendency of the BLYP density functional to over-structure hydrogen-bonding fluids, as discussed in the Introduction. The hydroxylation of a freshly cleaved silica surface is poorly understood, as is the distribution of various chemical groups on the surface.² Further tests of the convergence of AIMD methodologies to this system are needed before quantitative benchmarks can be established. We have provided an account of the reactions that occur in our AIMD simulations as water comes in contact with silica, and general agreement is found with previous theoretical studies of water chemisorption using small numbers of water near silica.^{48,87,89} However, given the experimental difficulties in disentangling the complex structure of the silica surface, this is an area where extensive further AIMD investigations are needed.

In this work, we have compared the distribution of water near silanol groups on a smaller amorphous silica slab using ab initio MD simulations, and using our empirical potential. We have also compared the radial density of water near orthosilicic acid. For the slab systems, the surface heterogeneity of the larger system simulations performed with our empirical model complicates the comparison with the smaller ab initio MD simulations (see Figures 7 and 8 and the corresponding discussion). When our model is applied to simulating orthosilicic acid, we find that it results in the understructuring of water near the silanols. At this point, we cannot conclusively determine whether this behavior is an artifact of the density functional used in the ab initio MD simulations or whether our empirical model under-structures water near the silanols, or a combination of both.

An interface between water and amorphous silica is found in many applications, ranging from chromatography to novel nanofluidic devices. Our potential model is designed to enable atomistic device simulations at large length and time scales, which is desirable for several reasons. Large length scales are needed to go beyond straight channels and model features like nozzles or bends. Long time scales are needed to model fluid flow. Typical flow velocities are many orders of magnitude smaller than a typical thermal velocity, $(k_B T/m)^{1/2}$, where m is the molecular mass. As a result, flow velocities in nonequilibrium simulations are obtained by averaging over much larger numbers. The strategy in such simulations is to induce unrealistically large flow velocities to improve signal-to-noise, yet to keep the flow velocity below the point where nonlinear effects arise. The largest tolerable flow rates are on the order of 10% of a thermal velocity,¹³¹ which means that extensive averaging is needed.

Use of equilibrium simulations to evaluate Green–Kubo expressions has recently been advocated as an efficient alternative to nonequilibrium simulations for the calculations of transport coefficient^{132,133} and flow profiles.^{134,135} In either case, these calculations are demanding, underscoring the need for interaction models for the amorphous silica–water interface that combine computational efficiency with an acceptable level of realism, which is addressed in this work.

Acknowledgment. This material is based upon work supported by the National Science Foundation under Grant No. EEC-0914790. The calculations reported here were made possible by a grant of resources from the Ohio Supercomputer Center.

References

- (1) Iler, R. K. *The Chemistry of Silica*; Wiley: New York, 1979.
- (2) Legrand, A. P., Ed. *The Surface Properties of Silicas*; Wiley: New York, 1998.
- (3) Fubini, B. In *The Surface Properties of Silicas*; Legrand, A. P., Ed.; Wiley: New York, 1998; pp 415–464.
- (4) Unger, K. K. *Porous Silica, Its Properties and Use as Support in Column Liquid Chromatography*; Elsevier: New York, 1979.
- (5) Jennings, W. *Comparisons of Fused Silica and Other Glass Columns in Gas Chromatography*; Alfred Huthig Verlag: New York, 1981.
- (6) Towner, P. In *Essential Molecular Biology*; Brown, T. A., Ed.; Oxford: London, 2000; Vol. 1, Chapter 3, pp 55–67.
- (7) Nanassy, O. Z.; Haydock, P. V.; Reed, M. W. *Anal. Biochem.* **2007**, *365*, 240–245.
- (8) Cady, N. C.; Stelick, S.; Batt, C. A. *Biosens. Bioelectron.* **2003**, *19*, 59–66.
- (9) Rech, I.; Cova, S.; Restelli, A.; Ghioni, M.; Chiari, M.; Cretich, M. *Electrophoresis* **2006**, *27*, 3797–3804.
- (10) Liang, X.; Morton, K. J.; Austin, R. H.; Chou, S. Y. *Nano Lett.* **2007**, *7*, 3774–3780.
- (11) Zhang, B.; Wood, M.; Lee, H. *Anal. Chem.* **2009**, *81*, 5541–5548.
- (12) Saffiotti, U. *Acta Bio-Med. Ateneo Parmense* **2005**, *76*, 30–37.
- (13) Melzak, K. A.; Sherwood, C. S.; Turner, R. F. B.; Haynes, C. A. *J. Colloid Interface Sci.* **1996**, *181*, 635–644.
- (14) Mercier, P.; Savoie, R. *Biospectroscopy* **1997**, *3*, 299–306.
- (15) Fujiwara, M.; Yamamoto, F.; Okamoto, K.; Shiokawa, K.; Nomura, R. *Anal. Chem.* **2005**, *77*, 8138–8145.
- (16) Kang, S. H.; Shortreed, M. R.; Yeung, E. S. *Anal. Chem.* **2001**, *73*, 1091–1099.
- (17) Li, H.-W.; Park, H.-Y.; Porter, M. D.; Yeung, E. S. *Anal. Chem.* **2005**, *77*, 3256–3260.
- (18) Isailovic, S.; Li, H.-W.; Yeung, E. S. *J. Chromatogr.* **2007**, *A1150*, 259–266.
- (19) Walter, S. R.; Geiger, F. M. *J. Phys. Chem. Lett.* **2010**, *1*, 9–15.
- (20) Sui, J.; Tleugabulova, D.; Brennan, J. D. *Langmuir* **2005**, *21*, 4996–5001.

- (21) Lundqvist, M.; Andresen, C.; Christensson, S.; Johansson, S.; Karlsson, M.; Broo, K.; Jonsson, B.-H. *Langmuir* **2005**, *21*, 11903–11906.
- (22) Valle-Delgado, J. J.; Molina-Bolívar, J. A.; Galisteo-González, F.; Gálvez-Ruiz, M. J.; Feiler, A.; Rutland, M. W. *Langmuir* **2005**, *21*, 9544–9554.
- (23) van der Veen, M.; Norde, W.; Stuart, M. C. *Colloids Surf.* **2004**, *B35*, 33–40.
- (24) Dabrowski, J.; Müssig, H.-J. *Silicon Surfaces and Formation of Interfaces: Basic Science in the Industrial World*; World Scientific: River Edge, NJ, 2000.
- (25) Yao, S.; Myers, A. M.; Posner, J. D.; Rose, K. A.; Santiago, J. G. *J. Microelectromech. Syst.* **2006**, *15*, 717–728.
- (26) Lyklema, J. *Fundamentals of Interface and Colloid Science*; Academic: San Diego, 1991.
- (27) Hassanali, A. A.; Singer, S. J. *J. Comput.-Aided Mater. Des.* **2007**, *14*, 53–63.
- (28) Hassanali, A. A.; Singer, S. J. *J. Phys. Chem.* **2007**, *B111*, 11181–11193.
- (29) As written, the first two terms on the right side of eq 7 in ref 28 involve sums over the silanol oxygens of the silica. They should include a sum over all oxygens of the silica, including both silanol and siloxane types.
- (30) Zhang, H.; Hassanali, A. A.; Shin, Y. K.; Knight, C.; Singer, S. J. *J. Chem. Phys.*, in press.
- (31) Gouy, M. *Comptes Rendu* **1909**, *149*, 654–657.
- (32) Gouy, M. *J. Phys. (Paris)* **1910**, *9*, 457–468.
- (33) Chapman, D. L. *Philos. Mag.* **1913**, *25*, 475–481.
- (34) Stern, O. *Z. Elektrochem.* **1924**, *30*, 508–516.
- (35) Put, A. G. V. D.; Bijsterbosch, B. H. *J. Colloid Interface Sci.* **1983**, *92*, 499–507.
- (36) Zukoski, C. F., IV; Saville, D. A. *J. Colloid Interface Sci.* **1986**, *114*, 32–44.
- (37) Zukoski, C. F., IV; Saville, D. A. *J. Colloid Interface Sci.* **1986**, *114*, 45–53.
- (38) Kijlstra, J.; van Leeuwen, H. P.; Lyklema, J. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 3441–3449.
- (39) Dukhin, S. S. *Adv. Colloid Interface Sci.* **1995**, *61*, 17–49.
- (40) Mangelsdorf, C. S.; White, L. R. *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 2441–2452.
- (41) Mangelsdorf, C. S.; White, L. R. *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 2583–2593.
- (42) Lyklema, J.; Minor, M. *Colloids Surf.* **1998**, *A140*, 33–41.
- (43) Dukhin, S. S.; Zimmermann, R.; Werner, C. *Colloids Surf.* **2001**, *A195*, 103–112.
- (44) Feuston, B. P.; Garofalini, S. H. *J. Chem. Phys.* **1988**, *89*, 5818–5824.
- (45) Feuston, B. P.; Garofalini, S. H. *J. Appl. Phys.* **1990**, *68*, 4830–4836.
- (46) Litton, D. A.; Garofalini, S. H. *J. Appl. Phys.* **2001**, *89*, 6013–6023.
- (47) Ma, Y.; Garofalini, S. H. *J. Chem. Phys.* **2006**, *124*, 234102.
- (48) Mahadevan, T. S.; Garofalini, S. H. *J. Phys. Chem.* **2008**, *C112*, 1507–1515.
- (49) Lockwood, G. K.; Garofalini, S. H. *J. Chem. Phys.* **2009**, *131*, 074703.
- (50) Mahadevan, T. S.; Garofalini, S. H. *J. Phys. Chem.* **2007**, *B111*, 8919–8927.
- (51) Mahadevan, T. S.; Garofalini, S. H. *J. Phys. Chem.* **2008**, *C112*, 5694.
- (52) Mahadevan, T.; Garofalini, S. *J. Phys. Chem.* **2009**, *C113*, 11177.
- (53) Rustad, J. R.; Hay, B. P. *Geochim. Cosmochim. Acta* **1995**, *59*, 1251–1257.
- (54) van Duin, A. C. T.; Strachan, A.; Stewman, S.; Zhang, Q.; Xu, X.; Goddard, W. A. *J. Phys. Chem.* **2003**, *A107*, 3803–3811.
- (55) Fogarty, J. C.; Aktulga, H. M.; Grama, A. Y.; van Duin, A. C. T.; Pandit, S. A. *J. Chem. Phys.* **2010**, *132*, 174704.
- (56) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. *J. Phys. Chem.* **2001**, *A105*, 9396–9409.
- (57) Cruz-Chu, E. R.; Aksimentiev, A.; Schulten, K. *J. Phys. Chem.* **2006**, *B110*, 21497–21508.
- (58) Lamb, R. N.; Furlong, D. N. *J. Chem. Soc., Faraday Trans. I* **1982**, *78*, 61–73.
- (59) Muster, T. H.; Prestidge, C. A.; Hayes, R. A. *Colloids Surf.* **2001**, *A176*, 253–266.
- (60) Jenkins, S.; Kirk, S. R.; Persson, M.; Carlen, J.; Abbas, Z. *J. Chem. Phys.* **2007**, *127*, 224711.
- (61) Jenkins, S.; Kirk, S. R.; Persson, M.; Carlen, J.; Abbas, Z. *J. Chem. Phys.* **2008**, *128*, 164711.
- (62) Jenkins, S.; Kirk, S. R.; Persson, M.; Carlen, J.; Abbas, Z. *J. Chem. Phys.* **2009**, *130*, 134702.
- (63) Delley, B. *J. Chem. Phys.* **1990**, *92*, 508–517.
- (64) Jorgensen, W. L. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 5, pp 3281–3285.
- (65) Wensink, E. J. W.; Hoffmann, A. C.; Apol, M. E. F.; Berendsen, H. J. C. *Langmuir* **2000**, *16*, 7392–7400.
- (66) Puibasset, J.; Pellenq, R. J.-M. *J. Chem. Phys.* **2003**, *119*, 9226–9232.
- (67) Puibasset, J.; Pellenq, R. J.-M. *Phys. Chem. Chem. Phys.* **2004**, *6*, 9226–9232.
- (68) Puibasset, J.; Pellenq, R. J.-M. *J. Phys.: Condens. Matter* **2004**, *16*, S5329–S5343.
- (69) Puibasset, J.; Pellenq, R. J.-M. *J. Chem. Phys.* **2005**, *122*, 094704.
- (70) Puibasset, J.; Pellenq, R. J.-M. *Eur. Phys. J.* **2007**, *ST141*, 41–44.
- (71) Lopes, P. E. M.; Murashov, V.; Tazi, M.; Demchuk, E.; Mackerell, A. D., Jr. *J. Phys. Chem.* **2006**, *110*, 2782–2792.
- (72) Freund, J. B. *J. Chem. Phys.* **2002**, *116*, 2194–2200.
- (73) Qiao, R.; Aluru, N. R. *J. Chem. Phys.* **2003**, *118*, 4692–4701.
- (74) Qiao, R.; Aluru, N. R. *Phys. Rev. Lett.* **2004**, *92*, 198301.
- (75) Joseph, S.; Aluru, N. R. *Langmuir* **2006**, *22*, 9041–9051.
- (76) Lorenz, C. D.; Travasset, A. *Phys. Rev.* **2007**, *E75*, 061202.
- (77) Lorenz, C. D.; Crozier, P. S.; Anderson, J. A.; Travasset, A. *J. Phys. Chem.* **2008**, *C112*, 10222–10232.
- (78) Some authors (e.g., ref 79) use a water contact angle of 90° as a dividing line between hydrophobic and hydrophilic

- surfaces. Others (e.g., ref 80) do not follow a strict convention. Because the experiments of which we are aware^{58,59} always report contact angles for unhydroxylated quartz or silica less than 90°, the unhydroxylated silica surface by the former convention would still be considered hydrophilic. In this work, we follow the latter convention, and perhaps we should refer to the relative hydrophobicity of siloxanes as compared to silanols.
- (79) Giovambattista, N.; Debenedetti, P. G.; Rossky, P. J. *J. Phys. Chem.* **2007**, *111*, 9581–9587.
- (80) Israelachvili, J. *Intermolecular and Surface Forces*, 2nd ed.; Academic: San Diego, CA, 1991.
- (81) Saengsawang, O.; Remsungnen, T.; Frizsche, S.; Haberlandt, R.; Hannongbua, S. *J. Phys. Chem.* **2005**, *B109*, 5684–5690.
- (82) Cheng, H.-P.; Barnett, R. N.; Landman, U. *J. Chem. Phys.* **2002**, *116*, 9300–9304.
- (83) Walsh, T. R.; Wilson, M.; Sutton, A. P. *J. Chem. Phys.* **2000**, *113*, 9191–9201.
- (84) Konený, R.; Doren, D. J. *J. Chem. Phys.* **1997**, *106*, 2426–2435.
- (85) Rimola, A.; Ugliengo, P. *J. Chem. Phys.* **2008**, *128*, 204702.
- (86) Criscenti, L. J.; Kubicki, J. D.; Brantley, S. L. *J. Phys. Chem.* **2006**, *A100*, 198–206.
- (87) Ma, Y.; Foster, A. S.; Nieminen, R. M. *J. Chem. Phys.* **2005**, *122*, 144709.
- (88) Hamad, S.; Bromley, S. T. *Chem. Commun.* **2008**, 4156–4158.
- (89) Du, M.-H.; Kolchin, A.; Cheng, H.-P. *J. Chem. Phys.* **2003**, *119*, 6418–6422.
- (90) Masini, P.; Bernasconi, M. *J. Phys.: Condens. Matter* **2002**, *14*, 4133–4144.
- (91) Mischler, C.; Horbach, J.; Kob, W.; Binder, K. *J. Phys.: Condens. Matter* **2005**, *17* (26), 4005–4013.
- (92) Leung, K.; Nielsen, I. M. B.; Criscenti, L. J. *J. Am. Chem. Soc.* **131**, *131*, 18358–18365.
- (93) Tielens, F.; Gervais, C.; Lambert, J. F.; Mauri, F.; Costa, D. *Chem. Mater.* **2008**, *20*, 3336–3344.
- (94) Tilocca, A.; Cormack, A. N. *ACS Appl. Mater. Interfaces* **2009**, *1*, 1324–1333.
- (95) Becke, A. D. *Phys. Rev.* **1988**, *A38*, 3098–3100.
- (96) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785–789.
- (97) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (98) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300–311.
- (99) Schwegler, E.; Grossman, J. C.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *121*, 5400–5409.
- (100) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem.* **2004**, *B108*, 12990–12998.
- (101) Lee, H.-S.; Tuckerman, M. E. *J. Phys. Chem.* **2006**, *A110*, 5549–5560.
- (102) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154507.
- (103) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.
- (104) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *ChemPhysChem* **2005**, *6*, 1894–1901.
- (105) Feuston, B. P.; Garofalini, S. H. *J. Chem. Phys.* **1989**, *91*, 564–570.
- (106) Feuston, B. P.; Garofalini, S. H. *J. Phys. Chem.* **1990**, *94*, 5351–5356.
- (107) Rustad, J. R.; Wasserman, E.; Felmy, A. R.; Wilke, C. J. *Colloid Interface Sci.* **1998**, *198*, 119–129.
- (108) Van Ginhoven, R. M.; Jónsson, H.; Corrales, L. R. *Phys. Rev.* **2005**, *B71*, 24208.
- (109) van Beest, B. W. H.; Kramer, G. J.; van Santen, R. A. *Phys. Rev. Lett.* **1990**, *64*, 1955–1958.
- (110) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (111) Møller, C.; Plesset, M. *Phys. Rev.* **1934**, *46*, 618–622.
- (112) Bartlett, R. J.; Silver, D. M. *Int. J. Quantum Chem., Symp.* **1974**, *8*, 271–276.
- (113) Binkley, J. S.; Pople, J. A. *Int. J. Quantum Chem.* **1975**, *9*, 229–236.
- (114) Bartlett, R. J.; Silver, D. M. *Int. J. Quantum Chem., Symp.* **1975**, *9*, 183–198.
- (115) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem., Symp.* **1976**, *10*, 1–19.
- (116) 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.: Pittsburgh, PA, 2003.
- (117) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (118) Dang, L. X. *J. Am. Chem. Soc.* **1995**, *117*, 6954–6960.
- (119) Huff, N. T.; Demiralp, E.; Cagin, T.; Goddard, W. A., III. *J. Non-Cryst. Solids* **1999**, *253*, 133–142.
- (120) CP2K website, <http://cp2k.berlios.de>.
- (121) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.
- (122) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2003**, *118*, 4365–4369.
- (123) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.

- (124) Pulay, P. *J. Comput. Chem.* **1982**, 3, 556–560.
- (125) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev.* **1996**, B54, 1703–1710.
- (126) Zhuravlev, L. T. *Pure Appl. Chem.* **1989**, 61, 1969–1976.
- (127) Bakos, T.; Rashkeev, S. N.; Pantiledes, S. T. *Phys. Rev. Lett.* **2002**, 88, 55508.
- (128) Rignanese, G.-M.; Charlier, J.-C.; Gonze, X. *Phys. Chem. Chem. Phys.* **2004**, 6, 1920–1925.
- (129) Blöchl, P. E. *J. Chem. Phys.* **1995**, 103, 7422–7428.
- (130) Peraro, M. D.; Raugei, S.; Carloni, P.; Klein, M. L. *ChemPhysChem* **2005**, 6, 1715–1718.
- (131) Zhu, W.; Singer, S. J.; Zheng, Z.; Conlisk, A. T. *Phys. Rev.* **2004**, E71, 041501.
- (132) Van de Ven-Lucassen, I. M. J. J.; Vlugt, T. J. H.; Van der Zanden, A. J. J.; Kerkhof, P. J. A. M. *Mol. Phys.* **1998**, 94, 495–503.
- (133) Chen, T.; Smit, B.; Bell, A. T. *J. Chem. Phys.* **2009**, 131, 246101.
- (134) Marry, V.; Dufreche, J.-F.; Jardat, M.; Meriguet, G.; Turq, P.; Grun, F. *Colloids Surf.* **2003**, A222, 147–153.
- (135) Dufreche, J.-F.; Marry, V.; Malíková, N.; Turq, P. *J. Mol. Liq.* **2005**, 118, 145–153.

CT100260Z

Automatic Structure Analysis in High-Throughput Characterization of Porous Materials

Maciej Haranczyk^{*,†} and James A. Sethian[‡]

Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States and Department of Mathematics, University of California, Berkeley, California 94720, United States

Received August 4, 2010

Abstract: Inspection of the structure and the void space of a porous material is a critical step in most computational studies involving guest molecules. Some sections of the void space, like inaccessible pockets, have to be identified and blocked in molecular simulations. These pockets are typically detected by visual analysis of the geometry, potential or free energy landscapes, or a histogram of an initial molecular simulation. Such visual analysis is time-consuming and inhibits characterization of large sets of materials required in studies focused on identification of the best materials for a given application. We present an automatic approach that bypasses manual visual analysis of this kind, thereby enabling execution of molecular simulations in an unsupervised, high-throughput manner. In our approach, we used a partial differential equations-based front propagation technique to segment out channels and inaccessible pockets of a periodic unit cell of a material. We cast the problem as a path planning problem in 3D space representing a periodic fragment of porous material, and solve the resulting Eikonal equation by using Fast Marching Methods. One attractive feature of this approach is that the to-be-analyzed data can be of varying types, including, for example, a 3D grid representing the distance to the material's surface, the potential or free energy of a molecule inside the material, or even a histogram (a set of snapshots) from a molecular simulation showing areas which were visited by the molecule during the simulation.

1. Introduction

1.1. Background. Porous materials contain complex networks of void channels and cages that are exploited in many different industrial applications. Zeolites, probably the most recognized class of crystalline porous materials, have found wide use in industry since the late 1950s. They are commonly used as chemical catalysts, in particular as cracking catalysts in oil refinement, membranes for separations and water softeners.^{1–4} There is an increasing interest in utilizing zeolites as membranes or adsorbents for CO₂ capture applications. In addition to zeolites, other classes of nanoporous materials, such as metal organic frameworks

(MOF)^{5,6} and zeolitic imidazolate frameworks (ZIFS),⁷ have enormous potential for gas separations or storage.^{8,9} A key aspect for the success of any nanoporous material is that the chemical composition along with pore topology is optimal at the given conditions for a particular application. It should be noted that the number of possible pore topologies is extremely large. There are approximately 190 unique zeolite frameworks known to exist today¹⁰ in more than 1400 zeolite crystals of various chemical composition and different geometrical parameters.¹¹ These experimentally known zeolites constitute only a very small fraction of more than 2.7 million structures that are feasible on theoretical grounds,^{12,13} of which between 314k and 585k are predicted to be thermodynamically accessible as aluminosilicates, with the remainder potentially accessible via elemental substitution.¹⁴ Databases of similar or greater magnitude can be developed for other nanoporous materials such as MOFs or ZIFs.

* To whom correspondence should be addressed E-mail: mharanczyk@lbl.gov.

[†] Lawrence Berkeley National Laboratory.

[‡] University of California, Berkeley.

Development of such databases brings great promise for discovery of new materials for many applications, particularly urgent ones like CO₂ capture. However, in order to make such discoveries possible, new computational and cheminformatics techniques have to be developed to characterize, categorize, and screen such large databases, bypassing manual and/or visual analysis in order to process and characterize a large number of structures in a high-throughput manner.

An important aspect of analysis of porous materials and their void space is the detection of inaccessible pockets, which can be occupied by guest molecules in computer calculations, even though such pockets are inaccessible in adsorption experiments. It is important to account for, and often artificially block or exclude, these pockets in the calculation of guest-accessible volumes, surface areas or prediction of guest-related properties using molecular simulation techniques. For example, in Monte Carlo (MC) simulations of adsorption, the blocking procedure can be a simple distance-check from the center of the small pockets and a rejection of all Monte Carlo trial moves that would place a molecule inside a certain radius. Alternatively, such pockets can be filled with dummy atoms. The importance of pore blocking in Monte Carlo simulations has been recently reemphasized by Krishna and van Baten.¹⁵ (see for further examples and discussion.) Similar to MC simulations, molecular dynamics simulations have to account for such pockets in order to ensure that initial positions are chosen in the main channel system and not in such pockets.

Detection of inaccessible pockets is usually performed by visual analysis of so-called pore landscapes. The pore landscapes are isosurfaces corresponding to the maximum accessible free energy level (see Keffer et al.¹⁶ for explanation on how isopotential energy surfaces are constructed.) Surfaces spreading across a periodic unit cell of a porous material encapsulate channels, whereas isolated surfaces correspond to inaccessible pockets. Similar approaches involve either visualization of histograms from molecular simulation which highlight all positions visited by a probe molecule during the time of simulation or simple visualization of isosurfaces corresponding to a distance from the material's surface equal to probe radii (Figure 1). Detection of inaccessible pockets in the last two cases is essentially the same as in the case of pore landscapes represented in terms of potential or free energy. An alternative approach involves analysis of abstract structure representations such as chemical hieroglyphs.¹⁷ The latter presents structural building blocks such as cages and segments of channels, detailing their sizes and their connectivity. The notation used to represent connections between building blocks highlight the possibility of a guest molecule moving between the building blocks. Inaccessible pockets can be easily identified in such representation as building blocks without "valid" connections.

Although visual analysis is powerful, it can be error prone and its throughput is limited by the resources and abilities of a researcher, making it impractical for a high-throughput computational characterization pipeline. Instead, efficient computational techniques which automatically detect and

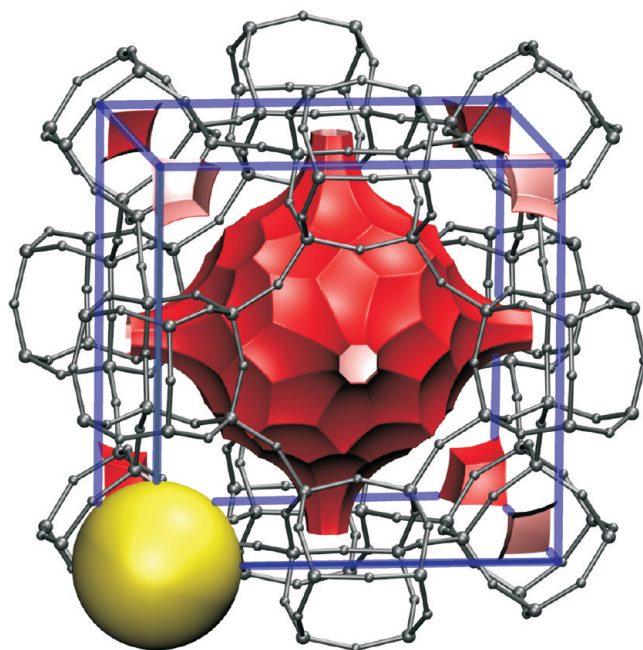


Figure 1. A periodic unit cell of LTA zeolite with highlighted probe accessible surface. Fragments of inaccessible pockets are visible in seven corners of the unit cell. One fragment is contained (blocked) by the yellow sphere.

characterize void space in porous materials are attractive alternatives. Detecting internal cavities has been explored in the context of proteins,¹⁸ and the more general question of finding possible pathways through chemical system has been addressed in both proteins¹⁹ and materials,^{20–23} while detection of inaccessible pockets in porous materials seems not to have been pursued. Regardless, the vast majority of these studies have been based on geometrical considerations. All but one study attempted to study paths of a spherical probe representing the molecule inside a convex hull constructed from atoms of a protein or materials framework. However, most molecules of interest are rarely spherical and a spherical approximation significantly overestimates the required channel diameter for a molecular sieve.

1.2. A Different Approach: Molecular Worms, Path Planning, and Hamilton-Jacobi Equations. To address this issue, in previous work we built a more advanced approach, in which a spherical probe is replaced with one resembling the shape and flexibility of a "real" molecule.²⁴ In this approach, complex objects are built from solid blocks connected by flexible links, called "molecular worms". Such worms are able to change orientation and/or shape during the traversal of chemical structure, allowing them to reach areas not accessible to either a single large spherical probe or rigid real-shape probes. This converts the problem into one of path planning, which we characterize by an Hamilton-Jacobi-type Eikonal equation in configuration space; here, the cost of entering each point in configuration space corresponds to the local geometry. The equation is solved by using a variant of Fast Marching Methods,^{25,26} which are Dijkstra²⁷-like methods to solve the boundary value problems of the form of the Eikonal equation. Starting with an initial position for the front, the method systematically marches the

front outward one grid point at a time, exploring all continuous pathways in the configuration space.

1.3. Current Work: A Comprehensive Approach to the Analysis of Porous Materials. In the current work, we present a comprehensive approach to analysis of porous materials, which allows detection and blocking of inaccessible pockets inside porous materials. Our approach does not require visualization and was developed with the goal to become part of a high-throughput pipeline for computational characterization of porous materials. In its essence, our approach mimics the visual analysis typically performed by researchers involving investigation of images representing pore landscapes. Unlike previously developed approaches to analyze accessibility, our framework is general. It allows analysis on the basis of geometry, potential or free energy profiles or histograms from molecular simulations. In particular, we use Fast Marching Methods to segment channels and inaccessible pockets from the 3D data representing any of the above. Additionally, we provide an algorithm to obtain blocking spheres for inaccessible pockets which can then be taken into account in molecular simulations, or excluded from calculations of accessible surfaces and volumes.

In this work, we present algorithms which examine accessibility issues based on data which can represent multiple characteristics, including the distance to the material's surface, the potential or free energy of a molecule inside a material, or the results of a molecular dynamics simulation showing areas which were visited by a molecule. For ease of exposition, in this work we only consider a three-dimensional configuration space, and hence our molecular worms can change position but not orientation or conformation. For this work, this translates into assuming spherical probes: it is important to stress that this is not a limitation of the method, and a full approach using molecular worms in higher dimensional configuration space is in progress and will be reported on elsewhere.

2. Methods

Our approach to automatic characterization of porous materials consist of two main tasks: (1) detection of channels and inaccessible pockets, and (2) generation of blocking spheres for inaccessible pockets. The detection of the latter relies on their segmentation from a guest molecule-dependent 3D data representing a repeating fragment of a porous material (e.g., 3D grid representing free energy of a molecule inside a periodic unit cell of a material). In the following subsections, we outline mathematical tools used for segmentation of space, their implementation to detect channels and inaccessible pockets, and the procedure of generating blocking spheres.

2.1. Mathematical Background. We cast the problem of segmentation of space in a periodic unit cell of a material as an Eikonal equation in domain representing this fragment and in which the cost of entering each point in the domain corresponds to its ability to be occupied:

$$|\nabla U| = C(x)$$

Here, U is the minimal total cost and $C(x)$ is a cost function defined at each point x in the domain. Abstractly, this cost

function is defined at the beginning of the problem, and the solution $U(x)$ to the above problem represents the total cost, which is the smallest obtainable integral of $C(x)$, considered over all possible trajectories throughout the computational domain from a start point to finish point. In practice, the above equation is solved for this total cost $U(x)$ first, and the actual cheapest path is obtained by starting at the finish and integrating a trajectory backward along the gradient field ∇U .

The Eikonal equation is an example of the general static Hamilton-Jacobi equation, and applies in the case of a convex, non-negative isotropic cost function. In order to provide a numerical solution to the above equation, we first lay down a computational mesh in three-dimensional space: this grid need not be uniform in the coordinate directions. The next step is to define an approximation u_{ijk} defined at each mesh point, and approximate the gradient in the above equation is by an upwind approximant of the form:

$$\left[\begin{array}{l} \max(D^{-x}u_{ijk}, -D^{+x}u_{ijk}, 0)^2 + \\ \max(D^{-y}u_{ijk}, -D^{+y}u_{ijk}, 0)^2 + \\ \max(D^{-z}u_{ijk}, -D^{+z}u_{ijk}, 0)^2 \end{array} \right]^{1/2} = c_{ijk} \quad (1)$$

where we have used standard finite difference notation.

Our goal now is to efficiently compute the solution to this equation on a mesh, starting with given boundary values. This equation is solved by using a variant of Fast Marching Methods, which systematically marches the solution outward from the known boundary values one grid point at a time. The key is to exploit the essential idea of Dijkstra's method: the causal nature of Bellman's optimality principle means that the value u_{ijk} that satisfies the above approximation can be constructed using only smaller values of u_{ijk} . Thus, one can start with a single boundary value, solve the above approximation for trial values at each neighbor one mesh point away, and then be assured that the smallest such trial value must be correct. Repeating this process systematically computes the solution at all mesh points in the computational domain: using a heap to extract the smallest such trial values produces the solution in $O(N \log N)$ time, where N is the total number of mesh points in the computational domain.

In more detail, the Fast Marching Method operates as follows. Suppose at some time the Eikonal solution is known at a set of *Accepted* points. For every not-yet accepted grid point with an *Accepted* neighbor, we compute a trial solution to the above quadratic eq 1, using the given values for u at accepted points, and values of ∞ at all other points. We now observe that the smallest of these trial solutions must be correct, since it depends only on accepted values which are themselves smaller. This "causality" relationship can be exploited to efficiently and systematically compute the solution as follows:

First, tag points in the initial conditions as *Accepted*. Then tag as *Considered* all points one grid point away and compute values at those points by solving eq 1. The latter points form a boundary (a front) around *Accepted* grid points, which will move outward as procedure continues. Finally, tag as *Far* all other grid points. Then the loop is as follows:

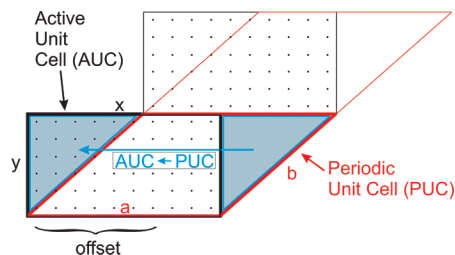


Figure 2. Geometrical relation between periodic unit cell and active unit cell. Faces of neighboring AUCs, unlike faces of neighboring PUCs, do not have to be fully aligned. Instead, they can be shifted by an offset value.

(1) Begin Loop: Let *Trial* be the *Considered* point with smallest value of u .

(2) Tag as *Considered* all neighbors of *Trial* that are not *Accepted*. If the neighbor is in *Far*, then remove it from that set and add it to the set *Considered*.

(3) Recompute the values of u at all *Considered* neighbors of *Trial* by solving the piecewise quadratic equation according to eq 1.

(4) Add point *Trial* to *Accepted*; remove from *Considered*.

(5) Return to top until the *Considered* set is empty.

This is the Fast Marching Method given in ref 25: the key to an efficient implementation of the above technique lies in a fast heap algorithm to locate the grid point in set of trial values with the smallest value for u . These methods have been successfully applied to problems in such topics as robotic navigation, fluid mechanics, and image analysis.²⁶ For more details on Fast Marching Methods we refer the Readers to ref 26.

In the following three sections, we describe components of Fast Marching Methods that have to be implemented and adopted to allow detection of inaccessible pockets, mainly: discretization of the space, definition of the cost function, and sequence of execution to segment these pockets.

2.2. Discretization of Space. The structure of crystalline porous material is typically stored as atomic coordinates of a repeating fragment of a structure (asymmetric unit), symmetry information and periodic box parameters. This information enables reconstruction of a periodic unit cell (PUC), which may or may not have rectangular shape; however, a rectangular shape is preferred here because of implementation issues.

At the first step, PUC is used to construct the active unit cell (AUC), which will be used to discretize space and build a grid on which Fast Marching Methods will be executed. AUC and the to-be-constructed grid for solving Fast Marching Methods is always rectangular (the cell vectors that define AUC are orthogonal) to facilitate implementation of the Fast Marching Methods procedure and ensure minimal memory overhead compared to nonrectangular PUC embedded inside a rectangular box.

For rectangular PUC, AUC is essentially the same as PUC. For remaining types of PUCs, AUC has to be constructed from PUC by translating sections of the PUC that are outside AUC. Two neighboring AUCs do not necessarily have common faces fully overlapped (see Figure 2). For example, faces aligned in the x direction on the y level have an offset

compared to the $y + 1$ level, which is important in handling periodicity of the chemical system represented by AUC.

The 3D space encapsulated by AUC is then discretized using specified step sizes v_x , v_y , and v_z along each of orthogonal directions (x , y , z , respectively). These sizes have to be small enough to achieve the desired accuracy. Very small steps may, however, result in a very large computational grid. The maximum number of points in a grid is limited by the available resources, in particular the memory size of a computer. Each grid point can be associated with a bin, which surrounds it, and has volume equal to $v_x \cdot v_y \cdot v_z$. Bins are used to divide the system into smaller volumes and sample its properties (e.g., free energy) in each of these volumes. Averages from such sampling for each bin are later assigned to the corresponding grid points.

2.3. Constructing the Cost Function. Given a discretization of space, the most straightforward cost function $C(x)$, where x is a grid point, is defined as follows: $C = 1$ for each point x which can be occupied, $C = \infty$ otherwise. Here, an “occupiable” point corresponds to a position in which the probe can reside. Its exact definition depends on the source data used to represent the investigated chemical system (see the summary in Table 1). For example, if considering atoms as hard spheres (geometry consideration), the occupiable point means a position in which the probe is not colliding with any atom of material’s framework: here, a collision occurs when the distance between centers of the probe and any of structure’s atoms is smaller than the sum of their specified radii.

When analysis is performed using free energy profiles, the occupiable point means that the average free energy of the probe inside the bin corresponding to the point is equal-to or lower than a specified threshold corresponding to accessible energy levels. Finally, when analyzing histograms from molecular simulations, “occupiable” point means that a position pointed by a point (or associated with the corresponding bin) was visited t number of times (for the time t) by a molecule during the time of simulation, where t is a selected threshold value.

Besides the simple definition of the cost function discussed above, it is also possible to define a continuous cost function. A number of examples of such definitions are presented in Table 2. It must be noted that the definitions collected in Table 2 are not the only ones possible and the cost function can be tuned accordingly to the application. From a practical point of view, the cost function controls the direction and speed at which a front propagates. When considering grids with distance to atomic surface, one may define cost to be inverse proportional to this distance (Table 2). In such definition, the front will propagate faster through centers of cavities rather than areas close to surfaces. Similarly, one may define the cost proportional to the energy experienced by a probe. In such a manner, the front will explore low energy regions sooner than higher energy regions. From a practical point of view, it is worth setting the cost function to infinity for regions which are occupied by the framework atoms and the probability of visiting these regions is practically zero. The latter can be controlled by specifying threshold values (for example r_p , e_t , and t_t in Table 2). From

Table 1. Definition of Cost Function at Point x Depending on the Type of Guest-Molecule-Dependent 3D Data Representing a Porous Material

cost	distance-to-surface	potential (or free) energy	histogram
1 (occupiable)	Probe placed at x does not overlap with any atom of the framework	The value of averaged potential (or free) energy inside a bin associated with the point x is lower than the specified threshold corresponding to accessible potential (or free) energy level at a given temperature.	Bin associated with the point x was visited by a probe during the time of molecular simulation t number of times (or for time longer than t). (The probe can be also defined by any atom of any guest molecule included in the simulation.)
∞ (not-occupiable)	otherwise	otherwise	otherwise

Table 2. Example Definitions of a Continuous Cost Function at Point x Depending on the Type of Guest-Molecule-Dependent 3D Data Representing a Porous Material

cost	distance-to-surface	potential (or free) energy	histogram
Value associated with point x	The distance d from the center of probe with the radii of r_p placed at x to the surface of any atom of the framework	The value of averaged potential (or free) energy, e , inside a bin associated with the point x	Number of times, t , the bin associated with the point x was visited by a probe during the time of molecular simulation (t can also be defined as integrated time a probe was visiting the bin)
Definition of occupiable point	$d > r_p$	$e < e_t$, where e_t is the specified threshold corresponding to accessible potential (or free) energy level at a given temperature	$t > t_t$, where t_t is the specified threshold value
Cost at an occupiable point	$1/(d - r_p)$	e	$1/t$
Cost at a not-occupiable point	∞	∞	∞

the point of view of the work presented here, the threshold defining occupiable/not-occupiable points is more important than the definition of cost function itself. This is because the threshold to determine the stop criteria for front propagation is directly correlated with the volume to be explored, which is essentially the idea behind this work. In turn, the definition of cost function has a significant effect on the shape of the moving front and in typical cases, like the one to be discussed here, has no effect on accessible volume. However, the ability to tune the cost function may be used to develop other approaches (for example, finding the shortest/lower energy diffusion paths), and is therefore highlighted here.

2.4. Segmentation of Topological Features. Having defined the value of the cost function at each point in the domain, we execute the Fast Marching Method to detect topological features on the void space. There are only two types of features that can be associated with occupiable points (ones with cost lower than infinity): channels and inaccessible pockets (Figure 3A). Channels are defined as features manifested on the face of a periodic unit cell that provide an accessible path throughout the unit cell from one face to the opposite face. Our procedure focuses on detection of channels: pockets are then defined as features consisting of occupiable points that are not identified as channels.

A porous material can have a number of channels in each of three principle directions. Detection along each direction is performed sequentially and consists of two steps illustrated in Figure 3B. In the first step, the periodic boundary conditions along the investigated direction are turned off and then all occupiable points on one face of the periodic unit cell are accepted. The Fast Marching Method is then executed until it finds occupiable points to accept. The important

information gained from this algorithm is the identification of points on the opposite-to-initial face of the unit cell that belong to channels in the considered direction. These points are then used in Step 2 to initiate the front in the subsequent passes of the algorithm. All points accepted in the second execution are points corresponding to channels. The implementation of this procedure is fairly straightforward. However, one should note that in the case of nonrectangular unit cells, faces of PUC may not overlap with faces of AUC used to store the 3D data and therefore one needs to identify points in AUC that correspond to faces of the original PUC.

After the algorithm finishes with detection of channels, it proceeds to the detection of inaccessible pockets. Any occupiable point that has not been assigned to a channel corresponds to a pocket. There may be, however, more than one pocket, and therefore the algorithm runs until all such points have been assigned to either channels or pockets. In a loop, the first not-yet-assigned point is taken as a starting point for a Fast Marching Methods procedure, and then the front propagates through the pocket until all occupiable points have been visited and assigned to the pocket. Then, it proceeds to detect the next pocket unless all points have been assigned.

2.5. Inaccessible Pore Blocking. Execution of the previous steps of our approach segments the occupiable space into two types of features: channels and inaccessible pockets, where each of these features is represented as a set of grid points. Typically, molecular simulation packages have the option to exclude particular areas of the system (e.g., inaccessible pockets) by placing “blocking spheres”. Therefore, the final part of our approach is focused on calculation of these blocking spheres from the points corresponding to

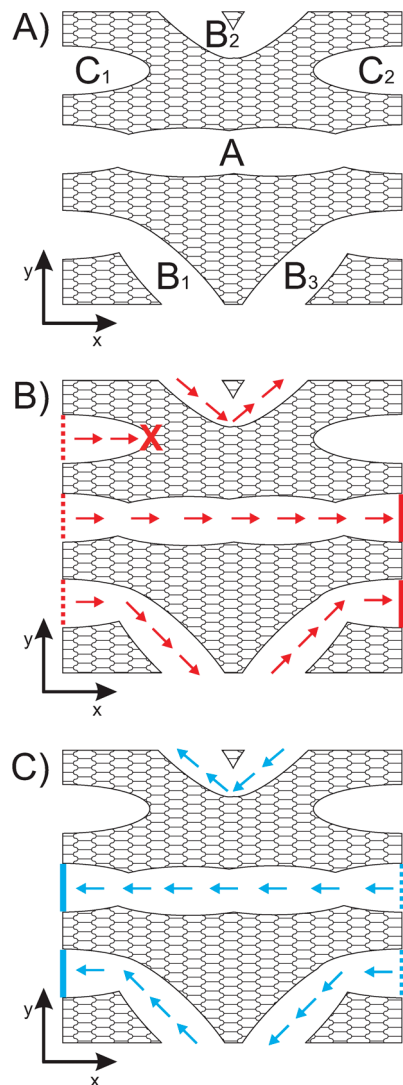


Figure 3. A sequence of front propagations to segment void space. (A) Three features of the void space in a periodic unit cell: straight channel A, topsy turvy channel B build of fragments $B_1 - B_3$ and inaccessible pocket C build of fragments C_1 and C_2 ; (B) "Forward" front propagation; (C) "Back" propagation. In B-C, dashed colored lines mark points on a face used to initiate the front, solid lines mark points on the opposite face that were reached by a front and X mark the front that has not reached the opposite face (inside a pocket).

identified features of this kind. It is important that all pockets are fully covered by the blocking spheres but at the same time, channels cannot overlap with any of such spheres.

Our procedure for blocking a pocket with spheres is executed in a loop over all not-yet-blocked pockets. Each pocket has an assigned number of grid points n . The procedure highlighted in Figure 4 is as follows:

(1) For each grid point assigned to the inaccessible pocket in consideration, n , calculate its density given by a Gaussian filter: $d_a = \sum_{k=1}^n \exp((-d_{ak}^2)/\langle d \rangle)$, where d_{ak} is the distance between points a and k, and $\langle d \rangle$ is the averaged distance between the points forming a pocket;

(2) Select the most dense point of 1 as pocket center, P_c ;

(3) Find the distance from P_c to the farthest point in the pocket, the radius of the pocket, r_p ;

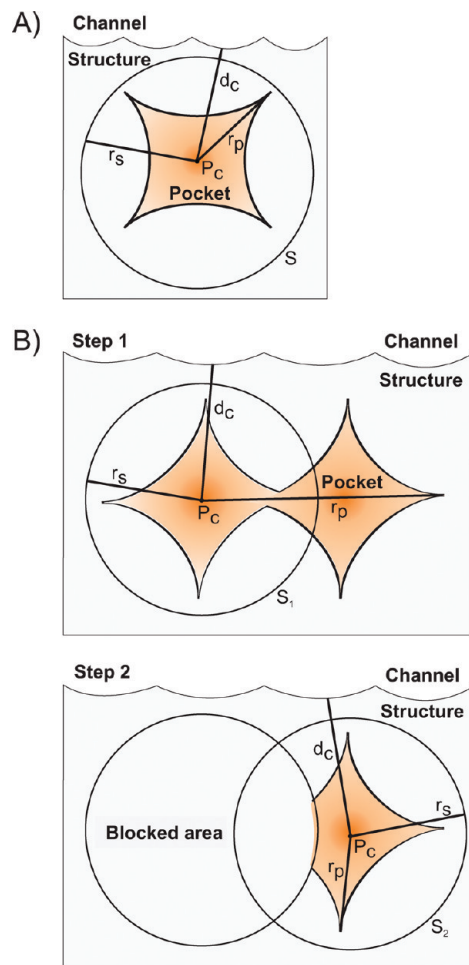


Figure 4. Blocking of inaccessible pockets. (A) Simple case with one blocking sphere. (B) Complex case with iterative blocking (Example with two iterations shown).

(4) Find the distance from P_c to the nearest grid point assigned to any channel, the distance to a channel, d_c ;

(5) Check $r_p < d_c$.

• If *true*, then the pocket is blocked with one sphere placed at P_c and with a radius of $(1/2)(r_p + d_c)$;

• If *false*, then the pocket has to be blocked iteratively with a number of spheres. The first sphere is placed at P_c and its radius is set to $d_c - \gamma$, where γ is a small fraction of grid step. All points that belong to the current pocket and that have not been blocked with the newly defined blocking sphere are now defined as a new pocket, which is added to the end of the list of pockets.

The algorithm presented here deserves comment: The advantage of using density calculation in Pt. (1) as a way to detect the center of a pocket over, for example, taking a geometrical center of a pocket is that it always return a point within a pocket, which is desirable in the case of complex-shaped pockets. The obvious drawback is the cost of density calculations, which scales as the square of number of points forming a pocket. For large pockets, we can speed up this calculation by using an approximation to the Gaussian density filter. For example, we can apply an MC sampling of points used in the density calculation of Pt. 1.²⁸ In our implementation, we use such an approximation for pockets larger than 10000 grid points.

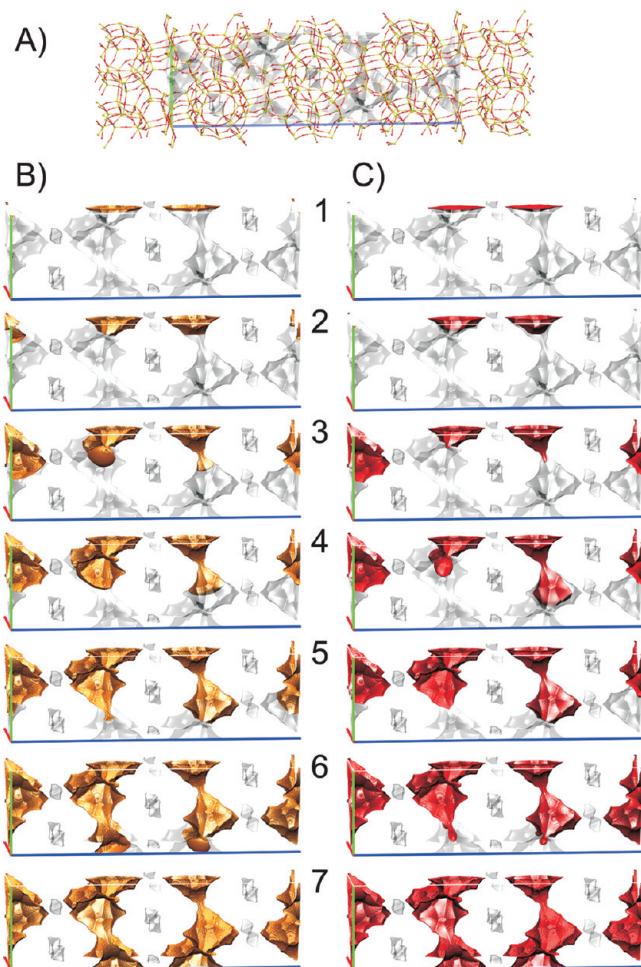


Figure 5. Example of DDR zeolite. (A) $1.5 \times 1.5 \times 1.5$ supercell of DDR zeolite with probe accessible surface in a periodic unit cell. (B) A number of snapshots presenting a propagating front inside a porous material using binary cost function. (C) A number of snapshots presenting a propagating front inside a porous material using continuous cost function.

3. Results and Discussion

We demonstrate an application of our approach to automatic analysis of porous materials using the framework of DDR zeolite as an example. DDR was selected because of its common characteristics. As stored in the IZA database, it has a medium size and nonrectangular periodic unit cell of 6716 \AA^3 volume. Figure 5A presents a fragment of the DDR zeolite and highlights an isosurface corresponding to 1.6 \AA distance from the atomic surface. This surface corresponds to the boundary value of how close the center of a nitrogen molecule can approach the surface of the zeolite. Our analysis starts with the detection of channels followed by the identification of inaccessible pockets. Figure 5B demonstrates a number of snapshots from the front propagation procedure (a simple binary cost function is used). The front initiated on the face of the unit cell marches its way through the channels. The final snapshot highlights all fragments of space corresponding to channels. It can be noticed that these channels are manifested on two pairs of opposite faces and therefore they form a 2D channel system. The channel detection procedure is repeated for two other directions but,

in the case of DDR, no additional channels are detected. The algorithm moves on to detection of inaccessible pockets. Nine pockets are identified. The relatively simple geometry of the material enable the blocking of pockets with nine blocking spheres with radii from 3.6 to 3.9 \AA . The presented procedure was repeated for other probe molecules. For example, for a probe of the radius of 1.8 \AA , which would roughly correspond to methane, we did not detect any channel in the DDR zeolite. We note that geometry-based considerations based on treating guest molecules and framework atoms as hard spheres may not always be valid. For example, the kinetic energy of guest molecules may allow them to diffuse through narrow passages of higher potential energy leading to a quantitatively different picture of accessible pore topology. Analysis of histograms from molecular simulations do provide much more confidence in this regard.

The time required to perform analysis with our algorithm depends most strongly on the number of grid points that represent the system being investigated. The size of the grid depends on the size of the periodic unit cell and the requested accuracy. In our benchmark case of the DDR zeolite, it took about 5 s to perform analysis of a grid of 0.85 M points. This grid was obtained for steps of 0.2 \AA . Similar analysis involving grids with smaller spacing of 0.1 and 0.05 \AA required 40 s and $5 \text{ min } 20 \text{ s}$, respectively. We note that the reported analysis times represent only a small fraction of the time required to generate the to-be-analyzed grids. For example, when using distances to the material's surface, the last two grids required, respectively, 4 and 30 min to compute. These times are orders of magnitude larger when calculating averaged properties such as average free energy in a bin corresponding to each grid point. Nevertheless, the times briefly discussed here were obtained using our prototype implementation, and they possibly can be improved by code optimization.

As mentioned in the Section 2.3, it is also possible to construct a continuous cost function. In comparison to the previously presented binary cost function, such a continuous cost function provides more flexibility on controlling the behavior of the propagating front. Figure 5C illustrates a propagating front in the DDR zeolite with the cost function defined as inverse of the distance to atomic surface (as discussed in Table 2). Comparing Figure 5, parts B and C, side by side exposes a different shape of the front: essentially in the continuous case, the front propagates faster in open areas of the structure (away from the atomic surfaces) creating tongues rather than spherical cups like in the case of binary cost function. The bottom line, however, is that all occupiable points of the grid are explored in both cases, and therefore for the application discussed in this article, binary definition of the cost function is sufficient.

Finally, we demonstrate the effect of selecting a threshold value used to define an occupiable point. In this case, we perform analysis using the free energy grid of IRMOF-10 metal organic framework. The free energy grid of $343 \times 343 \times 343$ size was prepared following the procedure discussed in ref 16 and the used simulation parameters and forcefield followed ref 29. Figure 6 presents the IRMOF-10 structure with two isosurfaces corresponding to accessed

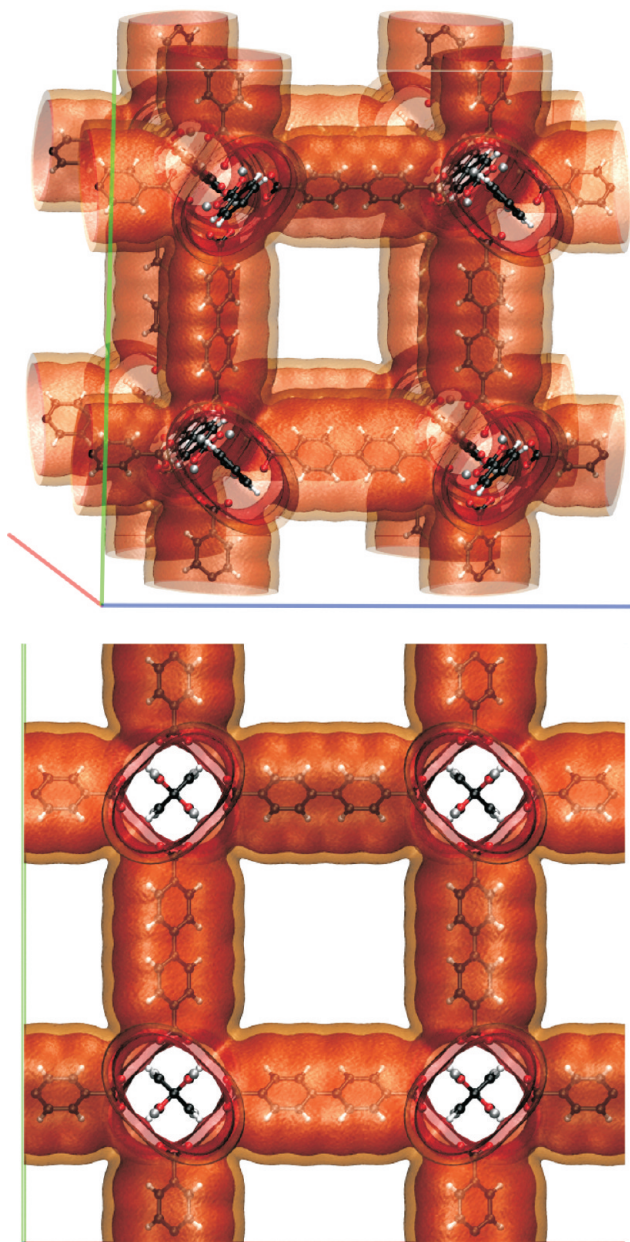


Figure 6. Example of IRMOF-10 metal organic framework. Surfaces highlight accessible volume corresponding to free energy threshold of $10k_B T$ (orange) and $150k_B T$ (red).

volume at free energy level of $10k_B T$ and $150k_B T$. As seen in the Figure 6, the isosurface corresponding to higher energy levels is closer to atoms, and therefore corresponding to larger accessible volume. Continuous cost function allows one not to treat guest molecules as hard spheres but rather allow them to interpenetrate the framework atoms at a higher cost (in our initial formulation we called the accessible interpenetration region “rubber walls”²⁴). In general, in case of structures with inaccessible pockets, it may be sometimes possible that setting a higher accessible energy level will lead to exploration of new cavities separated by tight windows with high energy barriers.

4. Conclusions

Inspection of the structure and the void space of a porous material is a critical step in most computational studies

involving guest molecules. For example, detection of channels and inaccessible pockets is performed, depending on application, before or after performing a molecular simulation.

We presented an automatic approach to perform such structure inspection. In our approach, we used a partial differential equations-based front propagation technique to segment out features of the void space present in a periodic unit cell of a material. We cast the problem as a path planning problem in 3D space representing a periodic fragment of porous material, and solve the resulting Eikonal equation by using Fast Marching Methods. One attractive feature of this approach is that the to-be-analyzed data can be of varying types, including, for example, a 3D grid representing the distance to the material’s surface, the potential or free energy of a molecule inside the material, or even a histogram (a set of snapshots) from a molecular simulation showing areas which were visited by the molecule during the simulation.

The ultimate worth of the algorithm proposed here rests in the ability to perform automatic unsupervised structure analysis and thereby bypass the time-consuming manual analysis. The time savings and reliability are substantial, and further efficiency will come from parallel implementations, leading to an automated high-throughput characterization environment.

We are currently using our approach to characterize all 191 zeolites in the IZA database. Our initial results confirm the need for the automatic structure analysis tools, especially the presented application for detection and blocking of inaccessible pockets in molecular simulations. For example, when considering a probe with radii of 1.4 \AA that mimics the He atom, we found out that 22 zeolites have no valid channels, 66 have just channel systems, and 103 (more than 50%!) have both channel systems and inaccessible pockets that have to be blocked in molecular simulations. Similar calculations conducted for the methane molecule ($r = 1.8 \text{ \AA}$) indicated that 60 out of 191 zeolites have no valid channels, 79 have just channel systems, and 52 have both channel systems and inaccessible pockets.

Acknowledgment. M.H. is a 2008 Glenn T. Seaborg Fellow at Lawrence Berkeley National Laboratory. M.H. is supported by the U.S. Department of Energy under Contract DE-AC02-05CH11231. M.H. is also supported jointly by DOE Office of Basic Energy Sciences and the Office of Advanced Scientific Computing Research through SciDAC project CSNEW918 entitled “Knowledge guided screening tools for identification of porous materials for CO₂ separations”. J.A.S. is supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract DE-AC03-76SF00098, and the Division of Mathematical Sciences of the National Science Foundation. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

- (1) Auerbach, S. M.; Carrado, K. A.; Dutta, P. K. *Handbook of Zeolite Science and Technology*; Marcel Dekker: New York, USA, 2004.
- (2) Smit, B.; Maesen, T. L. M. *Nature* **2008**, *457*, 671–677.
- (3) Smit, B.; Maesen, T. L. M. *Chem. Rev.* **2008**, *108*, 4125–4184.
- (4) Krishna, R.; van Baten, J. M. *Chem. Eng. J.* **2007**, *133*, 121–131.
- (5) Millward, A. R.; Yaghi, O. M. *J. Am. Chem. Soc.* **2005**, *127*, 17998–17999.
- (6) Walton, K. S.; Millward, A. R.; Dubbeldam, D.; Frost, H.; Low, J. J.; Yaghi, O. M.; Snurr, R. Q. *J. Am. Chem. Soc.* **2008**, *130*, 406–407.
- (7) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O’Keeffe, M.; Yaghi, O. M. *Science* **2008**, *319*, 939–94.
- (8) Sumida, K.; Hill, M. R.; Horike, S.; Dailly, A.; Long, J. R. *J. Am. Chem. Soc.* **2009**, *131*, 15120–15121.
- (9) Choi, H. J.; Dinca, M.; Long, J. R. *J. Am. Chem. Soc.* **2008**, *130*, 7848–7850.
- (10) (a) Baerlocher, C.; Meier, W. M.; Olson, D. H. *Atlas of Zeolite Framework Types*, 7th ed.; Elsevier: Amsterdam, NL, 2007. (b) <http://www.iza-online.org/> (accessed Jan 1, 2010).
- (11) Zimmermann, N. E. R.; Haranczyk, M.; Sharma, M.; Liu, B.; Smit, B.; Keil, F. J. *Mol. Sim.*, submitted for publication.
- (12) Foster, M. D.; Treacy, M. M. J. <http://www.hypotheticalzeolites.net> (accessed Nov 13, 2009).
- (13) Earl, D. J.; Deem, M. W. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449–5454.
- (14) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. *J. Phys. Chem. C* **2009**, *113*, 21353–21360.
- (15) Krishna, R.; van Baten, J. M. *Langmuir* **2010**, *26*, 2975–2978.
- (16) Keffer, D.; Gupta, V.; Kim, D.; Lenz, E.; Davis, H. T.; McCormick, A. V. *J. Mol. Graph.* **1996**, *14*, 108–116.
- (17) Theisen, K.; Smit, B.; Haranczyk, M. *J. Chem. Inf. Model.* **2010**, *50*, 461–469.
- (18) Till, M. S.; Ullmann, G. M. *J. Mol. Model.* **2010**, *16*, 419–429.
- (19) Petrek, M.; Otyepka, M.; Banas, P.; Kosinova, P.; Koca, J.; Damborsky, J. *BMC Bioinformatics* **2006**, *7*, 316–325.
- (20) Foster, M. D.; Rivin, I.; Treacy, M. M. J.; Friedrichs, O. D. *Micropor. Mesopor. Mater.* **2006**, *90*, 32–38.
- (21) Blatov, V. A.; Ilyushin, G. D.; Blatova, O. A.; Anurova, N. A.; Ivanov-Schits, A. K.; Dem’yanets, L. N. *Acta Crystallogr.* **2006**, *B62*, 1010–1018.
- (22) Haldoupis, E.; Nair, S.; Sholl, D. S. *J. Am. Chem. Soc.* **2010**, *132*, 7528–7539.
- (23) Haranczyk, M.; Rycroft, C. H.; Willems, T. F. *Mol. Inform.*, submitted for publication.
- (24) Haranczyk, M.; Sethian, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 21472–21477.
- (25) Sethian, J. A. *Proc. Nat. Acad. Sci.* **1996**, *93*, 1591–1595.
- (26) Sethian, J. A. *Level Set Methods and Fast Marching Methods*, 2nd ed.; Cambridge University Press: New York, 1999, 86–99.
- (27) Dijkstra, E. W. *Numer. Math.* **1959**, *1*, 269–271.
- (28) de Silva, V.; Carlsson, G. *Topological Estimation Using Witness Complexes*; Alexa, M., Rusinkiewicz, S., Eds.; Eurographics Symposium on Point-Based Graphics, 2004.
- (29) Liu, B.; Smit, B. *Langmuir* **2009**, *25*, 5918–5926.

CT100433Z

JCTC

Journal of Chemical Theory and Computation

Analysis of Bonding between Conjugated Organic Molecules and Noble Metal Surfaces Using Orbital Overlap Populations

Gerold M. Rangger,[†] Lorenz Romaner,[‡] Oliver T. Hofmann,[†] Georg Heimel,[§]
Michael G. Ramsey,^{||} and Egbert Zojer^{*†}

*Institut für Festkörperphysik, Technische Universität Graz, Petersgasse 16,
A-8010 Graz, Austria, Department Materialphysik, Montanuniversität Leoben,
Franz-Josef-Strasse 18, A-8700 Leoben, Austria, Institut für Physik,
Humboldt-Universität zu Berlin, Brook-Taylor-Strasse 6, D-12489 Berlin, Germany,
and Institut für Physik, Karl Franzens Universität Graz, Universitätsplatz 5,
A-8010 Graz, Austria*

Received August 11, 2010

Abstract: The electronic structure of metal–organic interfaces is of paramount importance for the properties of organic electronic and single-molecule devices. Here, we use so-called orbital overlap populations derived from slab-type band-structure calculations to analyze the covalent contribution to the bonding between an adsorbate layer and a metal. Using two prototypical molecules, the strong acceptor 2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane (F4TCNQ) on Ag(111) and the strong donor 1*H*,1'*H*-[4,4']bipyridinylidene (HVO) on Au(111), we present overlap populations as particularly versatile tools for describing the metal–organic interaction. Going beyond traditional approaches, in which overlap populations are represented in an atomic orbital basis, we also explore the use of a molecular orbital basis to gain significant additional insight. On the basis of the derived quantities, it is possible to identify the parts of the molecules responsible for the bonding and to analyze which of the molecular orbitals and metal bands most strongly contribute to the interaction and where on the energy scale they interact in bonding or antibonding fashion.

1. Introduction

Over the past decade it has been increasingly acknowledged that the understanding of metal–organic interfaces is crucial for further improving organic (opto)electronic^{1,2} and single-molecule devices.^{3–5} In this context, adsorbing organic acceptors or donors onto electrodes provides a convenient tool for modifying metal work functions and consequently for tuning the alignment between the metal Fermi level and the organic semiconductor states.^{6–14} A prototypical organic acceptor is 2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane

(F4TCNQ) (see right inset in Figure 1). F4TCNQ is a highly promising candidate for decreasing hole-injection barriers and has been thoroughly investigated in several spectroscopic and theoretical studies.^{7–16} For efficiently decreasing electron-injection barriers, doubly reduced viologens have recently been suggested as particularly potent materials. This is especially true for 1*H*,1'*H*-[4,4']bipyridinylidene, HVO (see left inset in Figure 1), for which a very strong charge transfer to the Au(111) surface has been predicted theoretically.¹⁷ This has later been confirmed experimentally for the more stable doubly methylated derivative 1,1'-dimethyl-[4,4']bipyridinylidene. This material has been found to decrease the work function of a Au(111) surface by 2.2 eV¹² and to concomitantly decrease also the electron injection barrier into subsequently deposited organic electron transport layers such as C60 and Alq₃.¹²

* Corresponding author e-mail: egbert.zojer@tugraz.at.

[†] Technische Universität Graz.

[‡] Montanuniversität Leoben.

[§] Humboldt-Universität zu Berlin.

^{||} Karl Franzens Universität Graz.

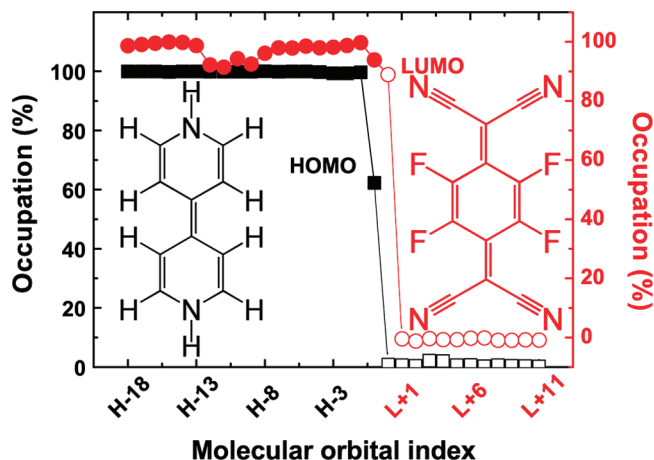


Figure 1. Comparison of the molecular orbital occupation of F4TCNQ on Ag(111) (circles), and HV0 on Au(111) (squares). The filled (open) symbols represent molecular orbitals that are occupied (unoccupied) in the isolated molecules. The right y-axis representing the occupations for F4TCNQ orbitals has been shifted by a constant to ease the comparison. The chemical structures of HV0 (left, 1*H*,1'*H*-[4,4']bipyridinylidene) and F4TCNQ (right, 2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane) are shown as insets. A more in-depth discussion of these data can be found in refs 16 and 17.

For both systems examined throughout this study, F4TCNQ on Ag(111) and HV0 on Au(111), the induced work-function modifications have been found to primarily originate from interfacial charge transfer. For F4TCNQ it is dominated by a nearly complete filling of the LUMO and for HV0 by a partial emptying of the HOMO.^{12,16} A more in-depth understanding of the interfacial charge-transfer processes can be obtained from the density of states (DOS) projected onto the states of the monolayer without the metal present (i.e., essentially onto the orbitals of the noninteracting molecules). This quantity has been referred to as the “molecular orbital DOS”.¹⁸ Integrating the individual molecular orbital projected DOSs up to the Fermi level yields the occupation of the associated molecular states in the interacting system. The respective results for F4TCNQ on Ag(111) and HV0 on Au(111) are shown in Figure 1; an in-depth discussion of the molecular orbital DOSs of these systems can be found in refs 16 and 17. They indicate which molecular orbitals effectively gain or lose electrons upon adsorption. The underlying process that allows for fractional occupations is a hybridization of the molecular states with the metal states and the filling of only part of the resulting hybrid orbitals. The reduced occupation of the HOMO in HV0 on Au(111) and the nearly complete filling of the molecular LUMO for F4TCNQ on Ag(111) are clearly visible; for the latter, also a back-donation of electrons from deeper lying F4TCNQ orbitals to the Ag(111) surface is visible as a reduced occupation of the HOMO-12 to the HOMO-9. These states have been identified as being localized on the terminal CN substituents,^{15,16} which is a first indication for the strong contribution of these groups to the metal–molecule bonding.

Their important role for the interaction can, furthermore, be deduced from the strong molecular distortions occurring upon adsorption that have been observed by X-ray standing

wave experiments for F4TCNQ on Cu(111) and by density functional theory-based modeling on all coinage metals.^{15,16} The F4TCNQ molecule, which is planar in gas phase, bends in a way that the terminal CN groups are closer to the surface by several tenths of an Angstrom than the central ring (with the absolute magnitude depending on the substrate metal). A bent adsorption geometry has recently been confirmed by scanning tunneling microscopy experiments on TCNQ (the nonfluorinated analogue to F4TCNQ) on Cu(100).¹⁹ A nonvanishing (albeit much smaller) distortion has been calculated also for HV0 on Au(111).¹⁷ In both cases, the bending gives rise to an intramolecular dipole which, in addition to the above-discussed charge transfer, determines the induced change in the substrate work function. The latter, amounting to +0.85 eV for F4TCNQ on Ag(111) and −1.21 eV HV0 on Au(111), is another indication for a strong metal–molecule interaction.^{16,17}

Strong chemisorption usually leads to the establishment of (partial) bonds between the adsorbates and the substrate that goes hand in hand with a change of the bonding pattern within the adsorbed molecule. A technique to analyze these processes based on theoretical modeling was first introduced by Hoffmann et al.^{18,20,21} They defined a crystal orbital overlap population (COOP) to interrogate the bonding and antibonding behavior of the atomic orbitals and their overlap with the metal bands resolved on the energy scale. In the present contribution, we describe what can be learned from such atomic orbital-based overlap populations about the covalent contributions to the bonding between organic adsorbates and metal surfaces by performing an in-depth analysis of the F4TCNQ/Ag(111) and HV0/Au(111) interfaces. Moreover, we expand the analysis tool to bonding and antibonding contributions arising from the interaction between certain molecular orbitals and the metal. These analyses allow identification of the individual atoms, the metal bands, and the molecular orbitals dominating the bonding process.

2. Methodology

All calculations presented here are based on optimized adsorption geometries obtained from the plane-wave density functional theory (DFT) code VASP, version 4.6.^{22,23} Details on the applied methodology can be found in the respective papers dealing with the electronic structure of F4TCNQ and HV0 adsorbates.^{15–17} Here, the same unit cells as in those contributions are used, i.e.,

$$\begin{pmatrix} 4 & -1 \\ 4 & 0 \end{pmatrix}$$

for F4TCNQ on Ag(111) and $(3 \times 3^{1/2} \times 5)$ for HV0 on Au(111); the respective structures are included in the Supporting Information. To ease the projection onto atomic and molecular orbitals, we used the VASP geometries as an input to perform a single self-consistency cycle in the atomic orbital based code SIESTA²⁴ version 2.0. There, the PBE exchange–correlation functional, norm-conserving pseudopotentials based on relativistic calculations using the Troullier–Martins scheme, as well as the same $(3 \times 3 \times 1)$ Monkhorst–Pack²⁵ *k*-point grids (as

in the VASP calculations) were used. A double- ζ polarized (DZP) basis set with 15 atomic orbitals (AOs) for the metal atoms, 5 AOs for the hydrogen atoms, and 13 AOs for all other atoms has been used as implemented in the SIESTA code.^{24,26,27}

Test calculations applying a single- ζ polarized (SZP) basis as implemented in SIESTA²⁴ and a user-generated²⁸ triple- ζ polarized (TZP) basis with 21 AOs for the metal and 17 for the molecule (6 for H) were made to test the influence of the basis-set size on the obtained results. We found virtually identical results for the double- and triple- ζ basis sets. The single- ζ calculations yielded equivalent trends but resulted in some quantitative deviations, as discussed in the Supporting Information.

In this context it should be noted that, naturally, well-known shortcomings of density functional theory will to some extent adversely impact the full quantitative validity of the results in any such calculations. To minimize their role, we have, however, carefully chosen the test systems so that their impact should be comparably small. Moreover, they affect in no way the main purpose of the present paper, which is showing the versatility of orbital overlap populations for describing metal–molecule bonding. They can, of course, also be applied in future calculations that contain corrections to the flaws of current (semi)local DFT implementations.

The above-mentioned shortcomings include the neglect of van der Waals interactions by (semi)local functionals, which can lead to an overestimation of the bonding distance. During the past few years, several remedies to that problem have been discussed.²⁹ Here, to minimize its impact, we chose to study strongly bonded systems, where at least for F4TCNQ on Cu(111) a good agreement between theory and experiment has been observed.¹⁵ Also the studied metal substrates were deliberately chosen among the coinage metals, as the charge transfer and interactions between a donor and Au (with the largest work function) and an acceptor and Ag (with the smallest work function) are particularly strong. Other effects including the lack of derivative discontinuity of the functionals,³⁰ the occurrence of self-interaction errors,³⁰ and improperly captured correlation-screening at the metal–organic interface^{31,32} affect the relative alignment between the adsorbate and the metal states. In our test systems, this does at least not impact the positions of the frontier orbitals, for which it has been established theoretically and experimentally that they are pinned at the Fermi energy.^{12,15–17} Also the positions and widths of the metal bands are known to be reasonably well described using relativistic pseudopotentials as in our calculations.^{33–35}

3. Analysis of the Metal–Adsorbate Bonding

To elucidate the details of the bonding process, three different types of quantities are used: (i) densities of states (DOSs), (ii) orbital overlap populations (OoPs), and (iii) total overlap populations (ToPs).

(i) Densities of states in various forms are frequently applied to analyze interactions between adsorbate layers and metals. They give the numbers of states per energy interval either of the total system or projected onto a certain region of space (e.g., molecular density of states, metal density of states) or onto a volume element (local density of states). Alternatively, the projection can also be onto individual

molecular orbitals of the noninteracting adsorbate (molecular-orbital-DOS), as discussed in Introduction (cf. Figure 1). A detailed DOS-based analysis of the HV0– and F4TCNQ–metal bonding can be found in refs 16 and 17.

(ii) Orbital overlap populations (OoPs) for analyzing bonds have originally been introduced by Hughbanks and Hoffmann²⁰ in the form of a crystal orbital overlap population (COOP). Its general definition is given by

$$\text{OoP}_{X,Y}(E) = \sum_{m \in X, l \in Y, i, \bar{k}} c_{im\bar{k}}^* c_{i\bar{k}l} S_{ml\bar{k}} G(E - \varepsilon_{i\bar{k}}) \quad (1)$$

Here, X and Y denote two groups of atoms and m and l the corresponding orbitals representing the basis set; the $c_{im\bar{k}}$ denote the linear combination of atomic orbital (LCAO) coefficients for state $\varepsilon_{i\bar{k}}$ and the $S_{ml\bar{k}}$ indicate the overlap matrices, where i is the band index. G is a line shape function, which in our case is a normalized Gaussian.

One can distinguish between different types of OoPs depending on which orbitals are included in the above summation; furthermore, one can sum over OoPs between different pairs of atoms or use different types of consistent basis sets, with respect to which the above expansion coefficients are defined. The latter results in different values for the S and c 's in eq 1. The different types of OoPs used throughout this work are summarized in Table 1.

First, we will use atomic orbitals in line with the work by Hoffmann et al.,¹⁸ which contains also a number of examples for the application of such OoPs to instructive test cases. Subsequently, we will combine the molecular orbitals of the isolated monolayer and the atomic orbitals of the metal slab to create a new basis³⁶ that is particularly useful for analyzing the interaction in a chemically intuitive way. A schematic illustration to explain this kind of OoP is provided in Figure 2, which shows how a molecular orbital interacts with a metal band. Among other effects, this causes a broadening of the orbital and the formation of hybrid states. Those that are of bonding character are typically located at energies below the original orbital, while antibonding states are located above. Hence, the OoP of the derived hybrid band is positive at low energies, cuts through zero at the position of the molecular orbital, and is negative at higher energies. This characteristic pattern is observed here for all molecular orbital-related OoPs (see section 4.3 and the corresponding Supporting Information), although the detailed shapes of the OoPs are usually more complex than in the schematic picture in Figure 2. Note that a very strong interaction with the substrate, such as the formation of covalent bonds, could result in deviations from this picture, if it causes a rehybridization of the states within the molecule. This is because in such a situation, the assumption of a single broadened molecular orbital is no longer appropriate.

(iii) To analyze the total contribution of the interaction between X and Y to the bonding in the investigated system, a total overlap population (ToP) can be introduced. It is defined as the integral over the corresponding OoP up to the Fermi energy; i.e., it is the integral over the “occupied” part of the respective OoP (cf., Figure 2). This quantity has been shown to scale with the bond order¹⁸ and, consequently, is closely related to the bond length and the bonding strength.

Table 1. Definition of Orbital Overlap Populations Used throughout This Manuscript

intramolecular OoP	used in Supporting Information
Here, X and Y in eq 1 denote selected pairs of atoms within the molecule; <i>m</i> and <i>l</i> contain all AOs of the selected atoms. The resulting OoP then serves to analyze the bond between X and Y.	
metal–molecule OoP	used primarily in section 4.1
In this case, X refers to all atoms of the top two metal layers (the lower-lying layers do not contribute to the OoP) and Y to all atoms of the adsorbed molecule(s); <i>m</i> and <i>l</i> again include all related AOs. This quantity describes the overall bonding between the adsorbate layer and the metal substrate.	
metal- <i>z</i> -band–molecule OoP	used primarily in section 4.2
This quantity is equivalent to the metal–molecule OoP, with the exception that only specific orbitals are included in the summation over the metal atoms. In the present case, “z” = “s” and “z” = “d” will be studied to analyze the contributions from the metal s- and d-bands. Equivalently, one could also define an OoP in which only s- or p-type orbitals on the molecule are included; such an analysis would, however, be inferior to the metal–“molecular orbital” OoP discussed below.	
metal–molecule _{part} OoP	used primarily in section 4.1
Here, the summation over Y runs only over a few selected atoms of the molecule(s). These can, for example, be the CN groups of F4TCNQ (“part” = “CN”). <i>m</i> and <i>l</i> again include all metal AOs and the AOs localized on the chosen part of the molecule, respectively. This quantity allows the study of bonding and antibonding interactions between a specific part of the molecule and the metal.	
metal–“molecular orbital” OoP	used primarily in section 4.3
This OoP allows partitioning of the metal–molecule OoP into contributions from specific molecular orbitals. It is defined such that X and <i>m</i> include all metal AOs and the $c_{im\bar{k}}$ are the corresponding LCAO coefficients. Y defines a single or a range of molecular orbitals, and the $C_{i\bar{l}k}$ correspond to the linear combination of molecular orbital coefficients (LCMO). ³⁶ This quantity allows identification of the contributions of chosen molecular orbital(s) to the metal–molecule OoP. For example, for “molecular orbital” = “HOMO”, the bonding and antibonding overlap population between the molecular HOMO and the metal is obtained as a function of energy. In this context, it should be mentioned that summing over all the metal–“molecular orbital” OoPs recovers the metal–molecule OoP.	

ToPs can be calculated for all of the above-described OoPs and will be shown in the following whenever they provide additional insight. They will simply be denoted by replacing OoP with ToP in the quantities described in Table 1.

To avoid confusion, we will refrain from an excessive use of acronyms; i.e., we will write the “full” names of the above quantities apart from the acronyms OoP and ToP.

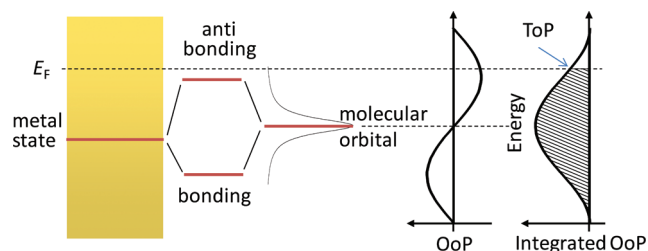


Figure 2. Schematic illustration of the interaction between a molecular orbital and a metal band giving rise to the metal–“molecular orbital” overlap population (OoP). The rightmost panel shows the OoP integrated over energy with the value of that curve at E_F corresponding to the total overlap population (ToP) associated with that orbital.

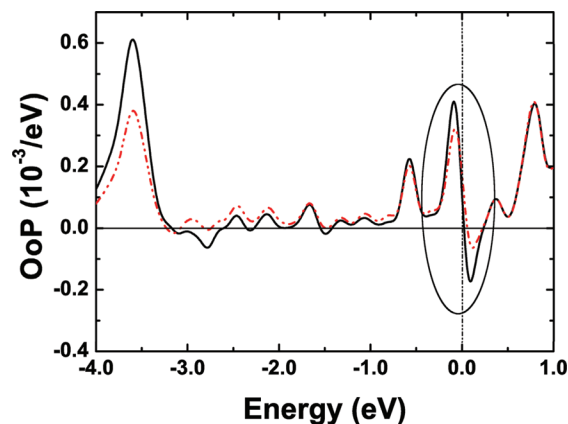


Figure 3. Metal–molecule OoP of HV0 on Au(111), black line. The dash-dot-dot line represents the metal–molecule_{part} OoP where the part of the molecule consists of only the outer secondary amine groups with their hydrogen atoms. The Fermi edge is set to zero, and the horizontal line divides bonding (positive) and antibonding (negative) areas. The ellipse marks the area around E_F that is discussed in the main text.

4. Results and Discussion

The focus of the current manuscript is on using various types of overlap populations for analyzing the metal–molecule interactions. Therefore, in the main manuscript, we will refrain from discussing changes in the intramolecular OoP that occur due to adsorption-induced geometric deformations of the molecules.^{15,17} The latter can, however, be found in the Supporting Information.

4.1. Orbital Overlap Population (OoP) between the Molecules and the Metal. *Bonding between HV0 and a Au(111) Surface.* As a first step, the metal–molecule OoP will be used to investigate the bonding between the molecules and the metal. The metal–molecule OoP of HV0 on Au(111) is shown as a black, solid line in Figure 3. Its shape around the Fermi energy is reminiscent of the “classical” situation described by Hoffmann for CO on Ni(100)¹⁸ that is also sketched in Figure 2: The lower energy part of the band adopts a bonding and the upper energy part an antibonding character, as seen in the region marked by an ellipse in Figure 3. As the HOMO-derived band of the HV0 layer is only partially occupied due to the charge-transfer (cf., Introduction), it is not surprising that the Fermi energy depicted as a vertical dash-dotted line cuts right through the bonding to antibonding wave. At this point, only the significantly larger magnitude of the bonding feature might appear somewhat surprising. We will return to that and also identify the origin of the strongly bonding features at -0.6 eV and -3.6 eV, when discussing the molecular orbital–metal OoP for this system.

To identify the parts of the molecules that most strongly contribute to the metal–molecule interaction, metal–molecule “part” OoPs were calculated. The most relevant is the one associated with the two outer secondary amine groups together with their hydrogen atoms, the metal–molecule_{amine} OoP. It is shown as a dash-dot-dotted line in Figure 3 and reproduces the positions of the main features of the metal–molecule OoP. This indicates that the secondary

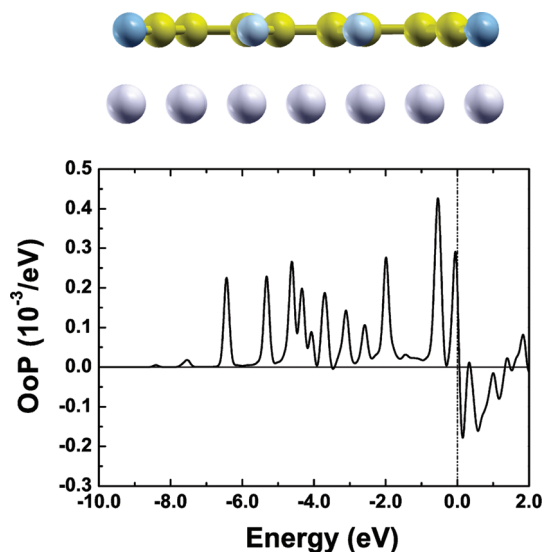


Figure 4. Metal–molecule OoP for the hypothetical case of a planar F4TCNQ monolayer adsorbed on Ag(111). The Fermi edge is set to zero, and the horizontal line divides bonding (positive) and antibonding (negative) areas. On top of the graph, a side view of an adsorbed planar F4TCNQ molecule is shown together with the top metal layer.

amines are to a large extent responsible for the covalent contribution to the bonding between HV0 and Au. This can at least to some degree be associated with the calculated slight bending of these groups toward the surface.¹⁷

OoP of F4TCNQ on Ag(111): The Impact of Bending. As the bending of the F4TCNQ molecules adsorbed on Ag(111) is much stronger,¹⁶ it is advisable to first study the hypothetical case of a planar adsorbate (see structure in the top of Figure 4) and only as a second step investigate the impact of bending. The metal–molecule OoP for planar F4TCNQ on Ag(111) with the central ring at the same adsorption distance as in the fully relaxed structure (i.e., 3.21 Å above the metal)¹⁶ is shown in Figure 4. A side view of the adsorbed molecule with the top metal layer is shown above the graph. The region around E_F looks qualitatively similar to HV0, with a bonding to antibonding transition. Also the absolute magnitude of the OoP peaks is comparable. Here, the features around E_F are related to the former LUMO which, due to the accepting nature of F4TCNQ, becomes slightly occupied. However, the amount of charge transfer is significantly lower than for the fully relaxed, bent adsorption geometry where a nearly complete filling of the LUMO has been observed as discussed in the introduction section and in ref 16.

The bending down of the terminal CN groups has dramatic consequences for the metal–molecule OoP. The result for the fully relaxed (i.e., strongly bent) adsorption geometry is shown in Figure 5 as a solid black line. Compared to the planar case in Figure 4, the whole region around E_F has become antibonding, and there is a strong negative peak at -3.3 eV. These are more than compensated by strongly bonding OoP contributions between -4.0 eV and -7.0 eV, which results in an increase of the corresponding total overlap population from 0.09 for the planarized adsorbate layer to 0.26 for the fully optimized case. It should be noted that the

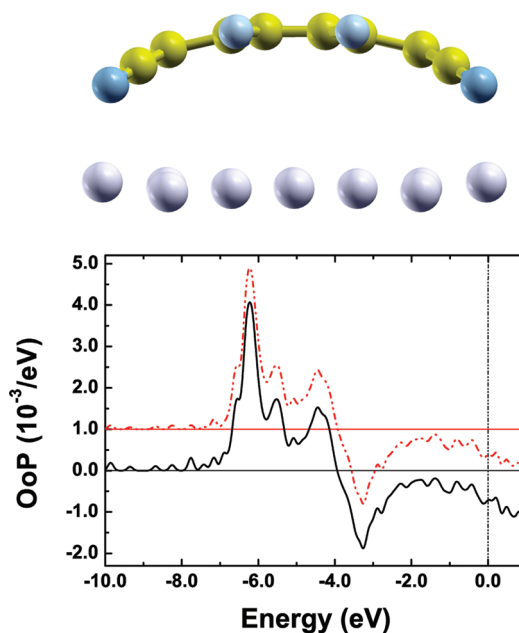


Figure 5. Metal–molecule OoP of F4TCNQ on Ag(111), black line, and metal–molecule_{part} OoP of only the CN atoms of the molecule with the metal, red-shifted curve. The Fermi edge is set to zero and the horizontal line divides bonding (positive) and antibonding (negative) areas. On top of the plot, a side view of an adsorbed fully relaxed F4TCNQ molecule is shown together with the top metal layer.

y-scales in Figures 3 and 4, on the one hand, and Figure 5, on the other hand, differ by a factor of 10(!).

Considering that the main difference between the planar and the fully optimized geometry is that in the latter the CN groups are bent down by 1.23 Å (position of the N-atoms), it is reasonable to assume that they must also be responsible for the huge changes in the OoP. This can be checked, by calculating the metal–molecule_{CN} OoP in which only the interaction between the N and C atoms of the four terminal CN groups with the metal are considered. The corresponding OoP curve is shown as a dash, dot-dotted gray line in Figure 5. To be visible in the plot, the curve had to be shifted because the result is virtually identical to the full metal–molecule OoP. This confirms the leading contribution of the CN groups in the covalent part of the metal–molecule bonding in the fully relaxed geometry. This finding also implies that the prominent features in Figure 5 are related to molecular orbitals largely localized on the CN groups. A more in-depth analysis of the origin of the peaks addressing this question will be provided in section 4.3.

4.2. The Role of the Metal s- and d-Bands. In the next step, the role of the metal in the bonding process is investigated. The metal_{s band}–molecule and metal_{d band}–molecule OoPs for HV0 on Au(111) are shown in the top part of Figure 6. The corresponding ToPs are contained in the bottom part of Figure 6 for an extended energy range.

In this context it needs to be mentioned that in our calculations, the Au(111) d-band starts around -2 eV below E_F and has a width of 4.5 eV consistent with the calculations in ref 37. Interestingly, all strong d-band contributions to the overlap population have a positive sign below the Fermi edge. The metal_{s band}–molecule OoP starts contributing much

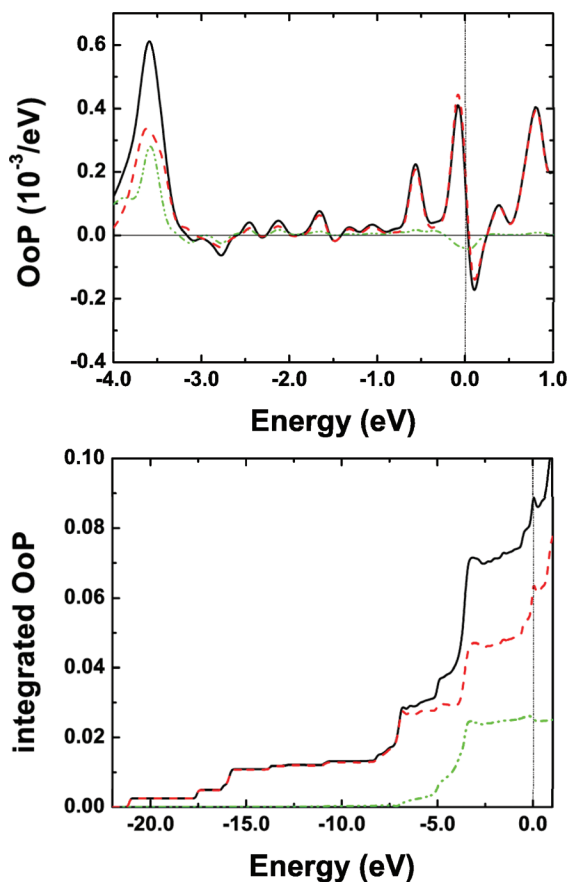


Figure 6. (top) Metal–molecule OoP (black, solid line) and $\text{metal}_{s,d\text{-band}}$ –molecule OoP of HV0 on Au(111) (top graph) and integrated metal–molecule and $\text{metal}_{s,d\text{-band}}$ –molecule OoP of HV0 on Au(111) (bottom graph). The red, dashed line represents the $\text{metal}_{s\text{-band}}$ –molecule interaction, the green, dash, dot-dotted line the $\text{metal}_{d\text{-band}}$ –molecule interaction. The vertical dash, dot-dotted line represents the Fermi energy, which is set to zero.

lower in energy (approximately -22 eV below E_F , which is attributed to the hybridization of low-lying molecular orbitals with parts of the metal s-band at higher energies). Its contributions to the overlap population are essentially bonding below the Fermi edge as well. As a result, about 2/3 of the metal–molecule ToP originates from the bonding to the Au s-electrons.

Qualitatively, a different picture arises when studying the influence of the metal bands for the bonding process of F4TCNQ on Ag(111). In Ag, we calculate the onset of the d-band at ca. -3 eV (i.e., 1 eV lower than for Au) and the d-bandwidth as ca. 4 eV. Here, the interaction with the Ag d-band plays a strong role for the bonding OoP between -7 and -4 eV and even dominates the antibonding feature between -4 and -2 eV, as shown in Figure 7 (top), i.e., the interaction of the bent-down CN groups with the metal has a strong contribution from the Ag d-orbitals. The integral over the metal_d –molecule OoP in Figure 7 (bottom) reveals that the bonding and antibonding contributions largely cancel. In fact, in contrast to HV0 on Au(111) with its small positive metal_d –molecule ToP, here the antibonding d-band contributions slightly outweigh the bonding ones, resulting in a small negative metal_d –molecule ToP. This is far outweighed by

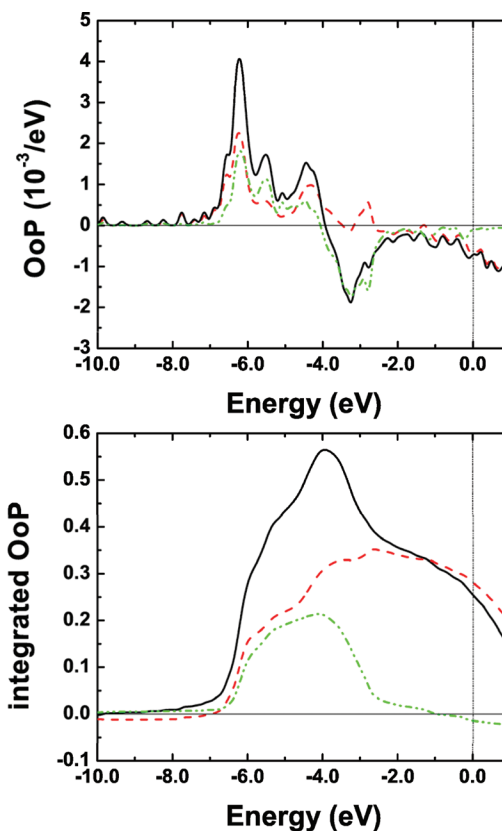


Figure 7. (top) Metal–molecule OoP (black, solid line) and $\text{metal}_{s,d\text{-band}}$ –molecule OoP of F4TCNQ on Ag(111) and (bottom) integrated metal–molecule (black, solid line) and $\text{metal}_{s,d\text{-band}}$ –molecule OoP of F4TCNQ on Ag(111). The red, dashed line represents the $\text{metal}_{s\text{-band}}$ –molecule interaction, the green, dash, dot-dotted line the $\text{metal}_{d\text{-band}}$ –molecule interaction. The vertical dash, dot-dotted line represents the Fermi energy, set to zero. The horizontal line in the top part divides bonding and antibonding contributions.

the contribution from the s-band, for which especially the interaction with the CN groups is nearly exclusively bonding (Figure 7 (top)).

4.3. Molecular Orbital–Metal OoP. Finally, the molecular orbitals need to be identified, whose hybridization with the metal states gives rise to the various bonding and antibonding features in the above OoPs; i.e., we will discuss the metal– “molecular orbital” OoPs for HV0 on Au(111) and F4TCNQ on Ag(111). As in the present adsorbates, one frequently encounters groups of orbitals with similar character, we will “group” them in the following discussion for the sake of clarity. The corresponding OoPs for the individual orbitals can be found in the Supporting Information. The results for the OoPs associated with the most relevant (groups of) orbitals for HV0 on Au(111) are displayed in the left column of Figure 8. It is shown that the strongly bonding OoP of HV0 on Au(111) peaking at 3.6 eV is mostly a superposition of contributions from HOMO-3 and HOMO-2 derived states. Interestingly, an integration of the metal–HOMO-2 and metal–HOMO-3 OoPs shows that for these orbitals this strongly bonding peak is completely compensated by antibonding contributions closer to E_F (see right column of Figure 8). Also for the HOMO-1 (shown only in the Supporting Information), bonding and antibonding

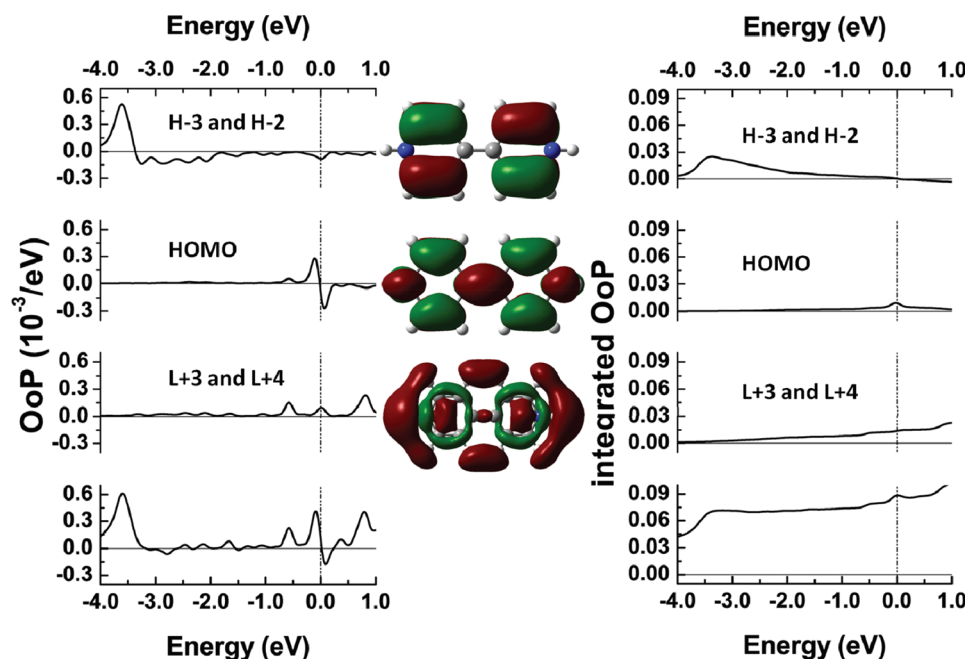


Figure 8. Left column: Metal–“molecular orbital” OoP analysis for HV0 on Au(111). The most significant (groups of) molecular orbitals and their OoPs with the metal are shown. From top to bottom: HOMO-2 plus HOMO-3, HOMO, LUMO+3 plus LUMO+4, and the total molecule–metal OoP; central column: shapes of the HOMO-2, the HOMO, and the LUMO+3 as representative examples; right column: corresponding integrated OoPs. ToPs and OoPs for all individual orbitals can be found in the Supporting Information.

features cancel and the ToP becomes zero. Of all occupied orbitals considered in Figure 8, only the metal–HOMO ToP does not vanish. The corresponding OoP, i.e., the metal–HOMO OoP is also responsible for the bonding to antibonding transition at E_F . However, it cannot explain its asymmetry as well as the magnitude of the peak at -0.6 eV. As shown in Figure 8, these features can only be rationalized by the partial occupation of unoccupied orbitals, namely the LUMO+3 and LUMO+4 (cf., Figure 1). Each of these orbitals displays a ToP contribution comparable to that of the HOMO. This is surprising, as the molecular orbital DOS of these two orbitals in Figure 1 shows that they do not bear significant electron density because they are filled to only about 2%¹⁷ and, therefore, have received no attention in previous studies.¹⁷ An orbital representation of these two states reveals the orbitals’ σ -character and their large amplitudes around the secondary amine parts of the molecule. Exactly these parts of the molecule have been identified above to be of particular importance, when comparing the metal–molecule_{amine} OoP to the metal–molecule OoP. Thus, as a net effect, the hybridization between metal states and unoccupied molecular states contributes more than 1/3 of the overall metal–molecule ToP, which is surprising for the adsorption of a strong electron donor. The rest of the ToP stems from deeper lying orbitals that are not included in Figure 8.

The analysis of the metal–“molecular orbital” OoP for F4TCNQ on Ag(111) confirms the notion that deep lying orbitals with large amplitudes on the CN groups are crucial for the development of the strong bond.^{15,16} The metal–“molecular orbital” OoPs of the HOMO-12 to HOMO-9 (cf., Introduction) are responsible for the most strongly bonding feature at about -6.3 eV, as shown in the left column of

Figure 9. However, they do not completely explain the bonding to antibonding transition at somewhat higher energies, which is mostly due to the contributions from the hybrid states formed by the HOMO-8 to HOMO-5 and the metal. Albeit the details of the shapes of the HOMO-12 to HOMO-9, the HOMO-8/HOMO-7, and the HOMO-6/HOMO-5 orbitals differ (see central column in Figure 9), they are all strongly localized on the terminal $-\text{CN}$ group.

The LUMO-derived OoPs display a bonding to antibonding transition at -0.2 eV (similar to the HOMO-derived feature giving rise to a bonding to antibonding transition at about -1.2 eV; see Supporting Information). Compared to the OoPs related to HOMO-12 to HOMO-5, the respective contributions are, however, so weak that they are hardly visible in the metal–molecule OoP.

An analysis of the corresponding ToPs displayed in the right column of Figure 9 shows that, not unexpectedly, the dominant contribution to the covalent part of the metal–molecule bonding comes from the HOMO-12 to HOMO-9 contributions.

4.4. Comparison between Orbital Overlap Populations and the DOS Projected onto the Molecular Region. Finally, the relation between the metal–molecule OoP and the DOS projected onto the molecular region need to be discussed. Here, one has to keep in mind that the prerequisite for a nonvanishing metal–molecule OoP in a certain energy region is that there metal–molecule hybrid orbitals must exist, i.e., the region will typically be close to the respective bands of the isolated molecular layer and metal slab. Moreover, the orbitals on the molecule and the metal need to overlap, which finds its mathematical manifestation in the overlap integral in eq 1. Keeping that in mind, the similarities and differences between the molecular DOS and

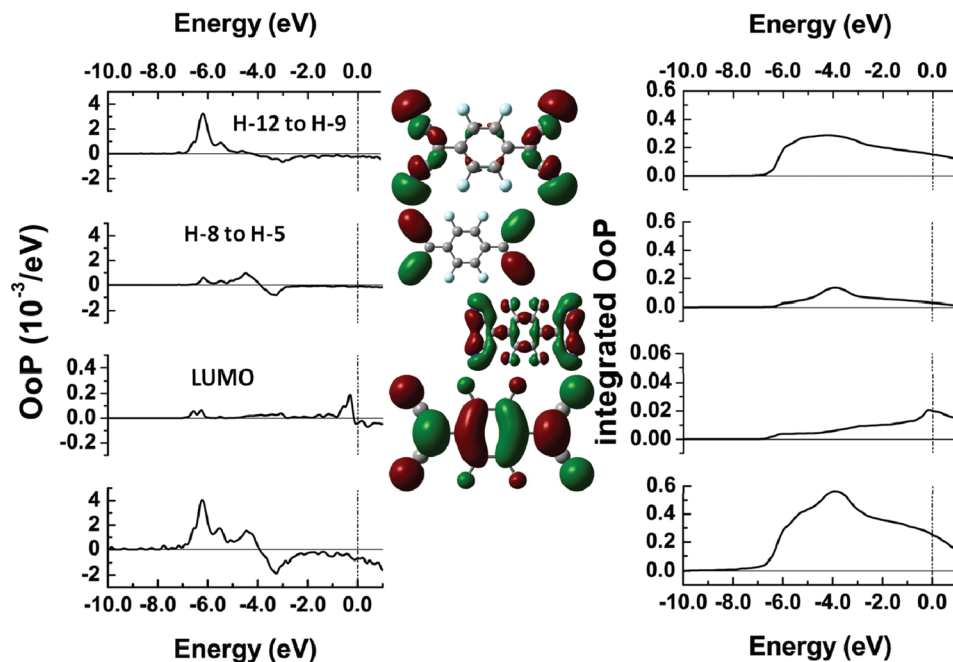


Figure 9. Left column: Metal–“molecular orbital” OoP analysis for F4TCNQ on Au(111). The most significant (groups of) molecular orbitals and their OoPs with the metal are shown. From top to bottom: Sum of HOMO-12 to HOMO-9, sum of HOMO-8 to HOMO-5, LUMO and the total molecule–metal OoP. Note that the scale for the metal–LUMO OoP differs from the others by a factor of 10. Central column: shapes of the HOMO-9, the HOMO-7, the HOMO-5, and the LUMO shown as representative examples; right column: corresponding integrated OoPs. ToPs and OoPs for all individual orbitals can be found in the Supporting Information.

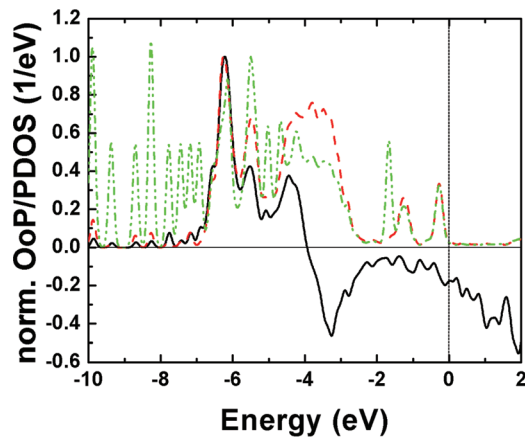


Figure 10. Normalized metal–molecule OoP (black solid line), DOS projected onto the molecular region (green dash-dotted line), and DOS projected onto the region around the N-atoms (red dashed line) for F4TCNQ adsorbed on Ag(111).

the metal–molecule OoP can be well understood and are shown in Figure 10 for the example of F4TCNQ on Ag(111).

The dominant metal–molecule OoP features between -7.0 and -2.5 eV correlate well with an energy range of increased DOS in the molecular region, especially in the region around the downward bent N-atoms (green dash-dotted and red-dashed line in Figure 10, respectively) and a very large DOS in the metal in exactly that energy range due to the d-bands (vide supra). Also for the features at higher energies, a correlation with the DOS projected on the N-atoms and the metal–molecule OoP is well visible. The overall magnitude of the OoP is, however, somewhat lower due to the lack of

metal-d-band contributions. The bonding to antibonding transitions in the OoP are, of course, not visible in the DOS. Figure 10 also shows that the OoP is clearly different from a DOS simply separated into bonding to antibonding contributions. This is evidenced, for example, by only very small OoP peaks associated with the DOS maxima below -6.8 eV and at -1.8 eV that are not associated with the downward-bent N-atoms, which manifests itself in small overlap integrals and, thus, in a reduced OoP.

5. Conclusions

We introduce a number of versatile tools for analyzing the covalent contribution to the bonding between organic adsorbates and metal surfaces that are derived from the crystal orbital overlap population introduced by Hoffmann et al.¹⁸ These tools allow the identification of energies at which metal–molecule hybrid states have bonding and antibonding character, respectively, and together with the calculation of a total overlap population they provide a measure for the total covalent bonding strength. Overlap populations enable the identification of the atoms and groups of the molecule that most strongly contribute to the bonding and also make it possible to quantify the role of different metal bands as well as (sets of) molecular orbitals in the bonding process.

As instructive examples, we apply various overlap populations to explain the bonding between a particularly strong electron donor, HV0, and Au(111) and the prototypical acceptor F4TCNQ and Ag(111). For the former example, they highlight the contribution of otherwise easily overlooked unoccupied orbitals that become only slightly occupied in the course of the bonding but have large amplitudes on those

parts of the molecules that are closest to the metal. For F4TCNQ on Ag(111), the pivotal contribution of the terminal CN groups and the orbitals localized on those groups is identified.

Acknowledgment. The authors acknowledge the financial support by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) project P20972-N20 and the European Commission project “IControl” (EC-STREP-033197). The authors thank H. Sormann for performing test-calculations of metal band structures.

Supporting Information Available: Figure showing the structure of the unit-cell for F4TCNQ on Ag(111) and HV0 on Au(111). Comparison of the metal–molecule OoP for these systems calculated with single-, double-, and triple- ζ basis sets. Discussion of changes in the intramolecular OOP; metal–“molecular orbital” OoPs for individual molecular orbitals and corresponding shapes of orbitals. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Ishii, H.; Sugiyama, K.; Ito, E.; Seki, K. *Adv. Mater.* **1999**, *11*, 605.
- (2) Koch, N. *ChemPhysChem* **2007**, *8*, 1438.
- (3) Patrone, L.; Palacin, S.; Charlier, J.; Armand, F.; Bourgoin, J. P.; Tang, H.; Gauthier, S. *Phys. Rev. Lett.* **2003**, *91*, 096802.
- (4) Yaliraki, S. N.; Kemp, M.; Ratner, M. A. *J. Am. Chem. Soc.* **1999**, *121*, 3428.
- (5) Seminario, J. M.; De La Cruz, C. E.; Derosa, P. A. *J. Am. Chem. Soc.* **2001**, *123*, 5616.
- (6) Werner, A. G.; Li, F.; Harada, K.; Pfeiffer, M.; Fritz, T.; Leo, K. *Appl. Phys. Lett.* **2003**, *82*, 4495.
- (7) Koch, N.; Duhm, S.; Rabe, J. P.; Vollmer, A.; Johnson, R. L. *Phys. Rev. Lett.* **2005**, *95*, 237601.
- (8) Chan, C. K.; Kim, E. G.; Brédas, J. L.; Kahn, A. *Adv. Funct. Mater.* **2006**, *16*, 831.
- (9) Duhm, S.; Glowatzki, H.; Cimpeanu, V.; Klankermayer, J.; Rabe, J. P.; Johnson, R. L.; Koch, N. *J. Phys. Chem B* **2006**, *110*, 21069.
- (10) Jackel, F.; Perera, U. G. E.; Iancu, V.; Braun, K. F.; Koch, N.; Rabe, J. P.; Hla, S. W. *Phys. Rev. Lett.* **2008**, *100*, 126102.
- (11) Lindell, L.; Unge, M.; Osikowicz, W.; Stafstrom, S.; Salaneck, W. R.; Crispin, X.; de Jong, M. P. *Appl. Phys. Lett.* **2008**, *92*, 163302.
- (12) Bröker, B.; Blum, R. P.; Frisch, J.; Vollmer, A.; Hofmann, O. T.; Rieger, R.; Mullen, K.; Rabe, J. P.; Zojer, E.; Koch, N. *Appl. Phys. Lett.* **2008**, *93*, 243303.
- (13) Bröker, B.; Blum, R.-P.; Beverina, L.; Hofmann, O. T.; Sassi, M.; Ruffo, R.; Pagani, G. A.; Heimel, G.; Vollmer, A.; Frisch, J.; Rabe, J. P.; Zojer, E.; Koch, N. *ChemPhysChem* **2009**, *10*, 2947.
- (14) Mukai, K.; Yoshinobu, J. *J. Electron Spectrosc.* **2009**, *174*, 55.
- (15) Romaner, L.; Heimel, G.; Bredas, J.-L.; Gerlach, A.; Schreiber, F.; Johnson, R. L.; Zegenhagen, J.; Duhm, S.; Koch, N.; Zojer, E. *Phys. Rev. Lett.* **2007**, *99*, 256801.
- (16) Rangger, G. M.; Hofmann, O. T.; Romaner, L.; Heimel, G.; Bröker, B.; Blum, R.-P.; Johnson, R. L.; Koch, N.; Zojer, E. *Phys. Rev. B* **2009**, *79*, 12.
- (17) Hofmann, O. T.; Rangger, G. M.; Zojer, E. *J. Phys. Chem. C* **2008**, *112*, 20357.
- (18) Hoffmann, R. *Rev. Mod. Phys.* **1988**, *60*, 601.
- (19) Santato, C.; Rosei, F. *Nat. Chem.* **2010**, *2*, 344.
- (20) Hughbanks, T.; Hoffmann, R. *J. Am. Chem. Soc.* **1983**, *105*, 1150.
- (21) Wijeyesekera, S. D.; Hoffmann, R. *Organometallics* **1984**, *3*, 949.
- (22) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15.
- (23) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 11169.
- (24) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejon, P.; Sanchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745.
- (25) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188.
- (26) Junquera, J.; Paz, Ó.; Sánchez-Portal, D.; Artacho, E. *Phys. Rev. B* **2001**, *64*, 235111.
- (27) García-Gil, S.; García, A.; Lorente, N.; Ordejón, P. *Phys. Rev. B* **2009**, 79–075441.
- (28) Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Martin, R. L.; Ordejon, P.; Sanchez-Portal, D.; Soler, J. M. User’s Guide SIESTA 2.0; 2006, p 82.
- (29) Tkatchenko, A.; Romaner, L.; Hofmann, O. T.; Zojer, E.; Ambrosch-Draxl, C.; Scheffler, M. *MRS Bull.* **2010**, *35*, 435.
- (30) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3.
- (31) Newns, D. M. *Phys. Rev.* **1969**, *178*, 1123.
- (32) Neaton, J.; Hybertsen, M.; Louie, S. *Phys. Rev. Lett.* **2006**, *97*, 216405. Li, Y.; Lu, D.; Galli, G. *J. Chem. Theory Comput.* **2009**, *5*, 881.
- (33) Marini, A.; Onida, G.; Del Sole, R. *Phys. Rev. Lett.* **2001**, *88*, 016403.
- (34) Marini, A.; Onida, G.; Del Sole, R. *Phys. Rev. B* **2001**, *64*, 195125.
- (35) Speer, N. J.; Brinkley, M. K.; Liu, Y.; Wei, C. M.; Miller, T.; Chiang, T.-C. *Europhys. Lett.* **2009**, *88*, 67004.
- (36) Rangger, G. M.; Romaner, L.; Heimel, G. *Zojer Surf. Interface Anal.* **2008**, *40*, 371.
- (37) Di Felice, R.; Selloni, A.; Molinari, E. *J. Phys. Chem. B* **2002**, *107*, 1151.

Theoretical Design by First Principles Molecular Dynamics of a Bioinspired Electrode–Catalyst System for Electrocatalytic Hydrogen Production from Acidified Water

Federico Zipoli,^{*,†} Roberto Car,^{†,‡} Morrel H. Cohen,^{†,§} and Annabella Selloni^{*,†}

Department of Chemistry and Princeton Institute for the Science and Technology of Materials, Princeton University, Princeton, New Jersey 08544, and Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854

Received June 11, 2010

Abstract: Bacterial di-iron hydrogenases produce hydrogen efficiently from water. Accordingly, we have studied by first-principles molecular-dynamics simulations (FPMD) electrocatalytic hydrogen production from acidified water by their common active site, the $[\text{FeFe}]_{\text{H}}$ cluster, extracted from the enzyme and linked directly to the (100) surface of a pyrite electrode. We found that the cluster could not be attached stably to the surface via a thiol link analogous to that which attaches it to the rest of the enzyme, despite the similarity of the (100) pyrite surface to the Fe_4S_4 cubane to which it is linked in the enzyme. We report here a systematic sequence of modifications of the structure and composition of the cluster devised to maintain the structural stability of the pyrite/cluster complex in water throughout its hydrogen production cycle, an example of the molecular design of a complex system by FPMD.

1. Introduction

There is currently a great deal of scientific and technological interest in the photo- or electrocatalytic production of hydrogen from water. Metallic platinum has suitably high catalytic activity with a turnover number on the order of 10^3 per second but is too rare and expensive an element for large-scale deployment. A correspondingly active catalyst comprised of earth-abundant elements is needed. Now, the hydrogen evolution reaction $2\text{H}^+ + 2\text{e}^- \rightarrow \text{H}_2$ is a central process in the global biological energy cycle. It is mediated by three different classes of enzymes with multiple subgroups.^{1,2} Of these, the di-iron hydrogenases found in hydrogen-producing microorganisms have the highest turnover frequency, over 9000 H_2 molecules per second at room temperature, almost an order of magnitude faster than that

of Pt.² These enzymes have been extensively studied both theoretically^{3–13} and experimentally.^{14–24} Removed from the protein environment, their active site, the $[\text{FeFe}]_{\text{H}}$ cluster, could be an attractive candidate for a catalyst for hydrogen production from water by electro- or photocatalysis.²⁵ It is composed of abundant elements and is small enough to pack densely. However, it seems unlikely that the $[\text{FeFe}]_{\text{H}}$ cluster could, without changes in its structure and composition, retain its hydrogen production activity and remain structurally stable throughout its catalytic production cycle when removed from its specialized, protected environment within the enzyme and exposed to water. To explore the needed changes, we have carried out a theoretical study of hydrogen production by the active center of the di-iron enzyme, starting with the bare $[\text{FeFe}]_{\text{H}}$ cluster detached from the enzyme without a change in structure or composition, linked to a pyrite surface acting as an electrode, and immersed in acidified water. We modeled this functionalized surface and investigated its stability and catalytic properties using static density functional theory (DFT) calculations and first-principles molecular-dynamics (FPMD) simulations.²⁶ The pyrite surface was chosen because its atomic structure is compatible with that

* To whom correspondence should be addressed. E-mail: fzipoli@princeton.edu (F.Z.).

[†] Department of Chemistry, Princeton University.

[‡] Princeton Institute for the Science and Technology of Materials, Princeton University.

[§] Rutgers University.

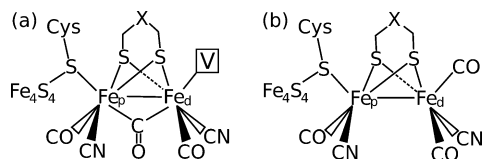


Figure 1. Sketches of the active center of di-iron hydrogenase in the bridging (a) and in the terminal (b) configurations. In a, the configuration in the enzyme, the distal iron has a vacant site, indicated by **V**, which can coordinate ligands such as CO, H₂O, etc. The “X” in the chelating S–CH₂–X–CH₂–S group stands for CH₂, NH, or O.¹⁸

of the [FeFe]_H cluster and *a priori* suggests the possibility of stable linkage and easy electron transfer to the cluster. Not surprisingly, we found that the [FeFe]_H cluster did not remain structurally stable in contact with acidified water during the course of its hydrogen production cycle. We then initiated a systematic search for structural and compositional modifications which could establish stability while maintaining suitable hydrogen production activity. We found that it is indeed possible to realize a strongly linked cluster configuration on the pyrite surface able to produce hydrogen efficiently and stably from acidified water. The details of the hydrogen production cycle of that configuration are reported elsewhere.²⁷ Here, we report the detailed numerical studies and analyses of the sequence of changes leading to the successful configuration, a process of scientific interest in its own right and, more generally, an illustration of the potential of FPMD for the molecular design of quite complicated chemical systems.

In the [FeFe]_H cluster, two iron atoms are coordinated with CO and CN ligands and bridged by a chelating group, S–CH₂–X–CH₂–S, where X can be a NH (DTMA) or CH₂ (PDT) group, cf. Figure 1a.²⁰ In the enzyme, one of the two iron atoms, the proximal iron (Fe_p), is connected to an iron–sulfur cluster (a cubane) via the *sulfur* of a cysteine.²⁰ The other iron is the distal iron (Fe_d). Two important groups of isomers have been identified,^{4,28} CO-bridging (μ -CO) and CO-terminal (CO_T). They differ in the position of one of the CO ligands. In CO_T, each CO is connected to only one of the two iron atoms, while in μ -CO, the configuration of the active ready state of the enzyme,¹⁸ there is one CO bridging the two iron atoms leaving a vacant coordination site **V** on Fe_d. In previous work,²⁹ we showed that the CO-bridging configuration had to be stable for the [FeFe]_H cluster of hydrogenase to function as an efficient hydrogen-production catalyst. We found that in vacuo, however, the CO-terminal configuration was slightly more stable and would effectively stop catalytic action. Since in practical applications the cluster would be immersed in acidified water, we subsequently studied³⁰ the effects of a water environment on its structure and reactivity. The electrons were added without considering the electrode explicitly. The main advantages of the model used in ref 30 were that the electrostatic effects, the dynamics of the proton diffusion in water via the Grotthuss-shuttle mechanism, and the proton sharing between the cluster and the water molecules were all taken into account in the FPMD computations. The main results were that (i) interconversion between the bridging and terminal configurations occurs without significant activa-

tion energies; (ii) the presence of terminal isomers does not stop the catalytic activity because a local hydrophobicity kinetically prevents the formation of a low-energy CO_T isomer in which a proton bridges the two Fe atoms (μ -H); (iii) there are at least three different pathways for H₂ production by the bridging configuration which involve a sequential protonation of Fe_d only or of both Fe_d and DTMA; and (iv) the bare active center of the di-iron hydrogenases can be an efficient catalyst for H₂ production provided that electrons are transferred to the cluster. However, we also found (v) that when the cluster is detached from the enzyme and immersed in acidified water, there are configurations in which the bond between Fe_p and the linking sulfur is weak and can break. The stability of the link between the electrode and the di-iron cluster thus emerged as a key issue in the design of a viable system. Accordingly, in the present work, we first focused our simulations on the linkage between the catalyst and FeS₂(100), the most stable surface of pyrite. We searched for structural and minor compositional modifications of [FeFe]_H which would lead to stable attachment while preserving the vacant coordination site on Fe_d in a configuration which exposed it to the water. We found such a configuration, one which formed a strong tridentate link to the surface, as described in section 3, and produced hydrogen successfully, as described in section 4.2. However, during the course of the cycle, the link of Fe_p to the chelating bridge was susceptible to breaking upon protonation of the bridging sulfur, which is exposed to the water, as described in section 5.1. Substitution of a PH group for that S eliminated the instability and maintained the hydrogen production capability, as discussed in section 5.2. We denoted the resulting cluster [FeFe]_P because of the importance of the phosphorus substitution, and we give an account here of our systematic transformation of [FeFe]_H from its in-the-enzyme configuration to [FeFe]_P linked to FeS₂(100).

2. Methods

Our study is based on Car–Parrinello (CP)²⁶ FPMD simulations of the [FeFe]_H cluster in a liquid-water environment with or without hydronium ions. The FPMD simulations were performed within the framework of density-functional theory in the local-spin-density approximation supplemented by generalized-gradient corrections³¹ as implemented in the CP code of the Quantum-ESPRESSO package.³² We employed ultrasoft pseudopotentials^{33,34} with plane-wave expansions of the Kohn–Sham orbitals and the augmented density up to kinetic energy cutoffs of 30 and 240 Ry, respectively. The pyrite electrode was modeled in a slab geometry with (100) surfaces and 3D periodic boundary conditions at the experimental lattice constant, 5.428 Å.³⁵ (Our theoretical lattice constant, computed with fully converged Brillouin zone (BZ) sampling, is 5.404 Å, only 0.4% shorter than the experimental one.) We used a 2 × 2 supercell having the ideal bulk termination of the slab. The slab was nine-atomic-layers-thick (24 FeS₂ units, cf. Figure 2) in a tetragonal simulation box with *a* = 10.856 Å and *c* = 21.712 Å with a vacuum width between slabs of 15 Å. The atoms of the three bottom layers were fixed to their positions in the clean

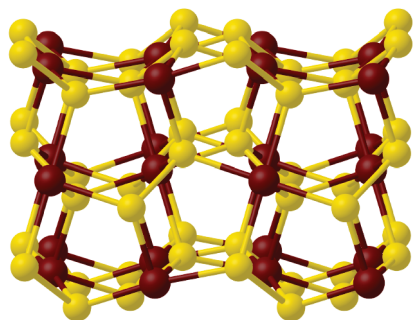


Figure 2. Side view of the FeS_2 (100) surface modeled in a geometry having a slab nine-atomic-layers-thick separated by 15 Å of vacuum in a tetragonal simulation box with $a = 10.856$ Å and $c = 21.712$ Å. Fe atoms are dark red, and S atoms are yellow.

surface during geometric optimizations of the supported catalyst, whereas all atoms were free to move during the FPMD simulations. We used only the Γ point to sample the surface BZ for both FPMD and structural optimizations. Test calculations comparing the results of the Γ -point integration with those of a k -point mesh showed differences in relative energies on the order of $k_B T$. We attached a sequence of progressively further modified di-iron $[\text{FeFe}]_H$ clusters (see section 3) to the $\text{FeS}_2(100)$ surface and added 37 water molecules between slabs to simulate immersion of the supported cluster in water of real density. A fictitious electronic mass of 350 au and a time step of 0.072 fs were used in the FPMD simulations. The deuterium mass was used for hydrogen to allow for a longer time step. Constant temperature was imposed on the ions by a Nosé–Hoover thermostat.³⁶ In situations where the system’s HOMO–LUMO energy gap was “small” (less than 0.1 eV), adiabaticity of the FPMD trajectory was maintained by coupling two separate Nosé–Hoover thermostats to the nuclear and electronic subsystems.³⁷ Static calculations on the isolated $[\text{FeFe}]_H$ cluster in vacuo were performed using the PW code of the Quantum-ESPRESSO package.³² Energy barriers were calculated using the climbing-image nudged-elastic-band (NEB)³⁸ and string methods.³⁹ For geometry optimizations and energy-barrier determinations, the norm of the force vector was required to be smaller than 5 meV/Å at convergence. Different pseudopotentials were used for the iron atoms of the cluster and for those of the pyrite slab, with 16 and 8 electrons explicitly treated as valence electrons, respectively. This choice reduced the computational cost without affecting the accuracy of our description of the delicate chemistry involving the two iron atoms of the $[\text{FeFe}]_H$ or $[\text{FeFe}]_P$ cluster. The validity of DFT for addressing the electronic and chemical properties of the FeS_2 surface^{40–43} and the di-iron cluster in different environments is well established.^{3–12,29,30} Because the duration of our simulations is limited to a few picoseconds, each computed trajectory may remain sensitive to the initial conditions imposed. To overcome this possible source of bias, we normally compute several trajectories spanning a representative range of initial conditions.

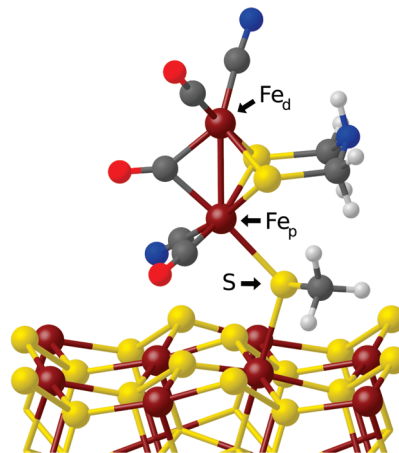


Figure 3. Hypothetical geometry of the $[\text{FeFe}]_H$ cluster linked to a (100) FeS_2 surface via a linking S atom, labeled S. The distal and the proximal iron are indicated by the labels Fe_d and Fe_p , respectively. Fe_p and Fe_d are bridged by di(thiomethyl)amine (DTMA, $\text{S}-\text{CH}_2-\text{NH}-\text{CH}_2-\text{S}$). Fe atoms are dark red. S atoms are yellow. C atoms are dark gray. N atoms are blue. O atoms are red. H atoms are light gray. The cluster atoms and the surface atoms are represented by balls and sticks. The remaining slab atoms are represented by cylinders.

3. Model

3.1. $\text{FeS}_2(100)$ Surface. The stoichiometric (100) surface of pyrite is formed by breaking the Fe–S bonds between adjacent (100) atomic layers while keeping the sulfur-dimer units intact. The S and Fe atoms on the FeS_2 surface are three- and 5-fold coordinated, respectively, whereas they are four- and 6-fold coordinated in the bulk. The LUMO of the clean defect-free surface derives from the d_z^2 orbitals of the surface iron atoms, and the computed band gap is 0.50 eV (0.42 eV for bulk pyrite, compared to the experimental value of ~ 0.9 eV,⁴⁴ a typical DFT underestimate, not of concern in the present studies).

Water adsorption on the FeS_2 (100) surface has been extensively studied both theoretically^{40,43,45–47} and experimentally,^{48,49} see refs 50 and 51 and references therein. In agreement with these previous studies, we found that water adsorbs in molecular form via a coordinative covalent bond between a lone-pair orbital of its oxygen and an empty d_z^2 orbital of a surface iron atom. The computed adsorption energy is 0.57 eV at low (1/8 ML) coverage. Due to intermolecular H-bond formation, this binding energy increases slightly with increasing coverage and becomes 0.62 eV at full (1 ML) coverage.

3.2. Supported Catalyst. Failure of Terminating Thiol Linkages. In modeling the $[\text{FeFe}]_H$ cluster in our earlier works,^{29,30} we followed the common procedure of replacing the cysteine linking Fe_p to the cubane within the enzyme by a methylthiol as in Figure 1b. In analogy with the link between the $[\text{FeFe}]_H$ cluster and the cubane in the enzyme, it is possible to attach the catalyst to the pyrite surface using the sulfur atom (S) of that SCH_3 to connect Fe_p to one of the surface Fe atoms ($\text{Fe}_{\text{surf}1}$). A sketch of this configuration is shown in Figure 3. We optimized its geometry in vacuo and found in addition that the CN group connected to Fe_p

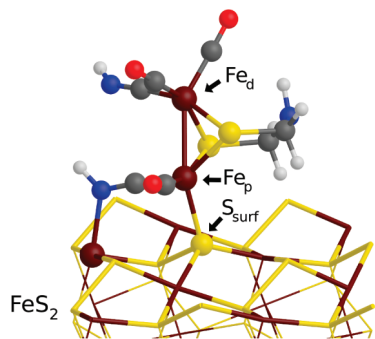


Figure 4. Optimized geometry of the active center of the $[\text{FeFe}]_H$ subcluster of the hydrogenase connected to the pyrite surface via a $\text{Fe}_p\text{-S}_{\text{surf}}$ bond. The SCH_3 group connected to Fe_p has been removed so that Fe_p can be connected directly to S_{surf} . The N atom of $(\text{CN})_p$ is connected to a surface Fe atom via a dative bond. The cluster is in the terminal configuration. The color code and the labels are those of Figure 3.

$((\text{CN})_p)$, which lies close to the surface, can easily make a dative bond with a nearby Fe atom ($\text{Fe}_{\text{surf}2}$). Binding the SCH_3 to $\text{Fe}_{\text{surf}1}$ restores the 6-fold coordination the latter would have had in the bulk, which leads to a relatively short $\text{Fe}_{\text{surf}1}\text{-S}$ distance, 2.4 Å. Upon protonation of Fe_d during the course of the H_2 production cycle, however, the $\text{Fe}_p\text{-S}$ bond breaks. This is analogous to the weakening of the $\text{Fe}_p\text{-SCH}_3$ link observed on protonation of the methylthiolate in water in the absence of the electrode,³⁰ with the $\text{Fe}_{\text{surf}1}\text{-S}$ bond playing the role of the thiolate protonation. To increase the stability of the $\text{Fe}_p\text{-S}$ bond, we removed the CH_3 group connected to the linking sulfur, which is then 2-fold coordinated. We found that this sulfur is usually deprotonated in water;³⁰ nevertheless, the $\text{Fe}_p\text{-S}$ bond breaks when the sulfur occasionally does become protonated. We conclude that an $\text{Fe}_p\text{-S-Fe}_{\text{surf}1}$ linkage is not stable. The fact that an analogous thiol linkage does occur in the enzyme suggests that the enzyme pocket provides the stabilization.

Dispensing with the Added Thiol; Protonation of the CN's. To circumvent the above problems, we removed the SCH_3 group entirely and connected the Fe_p directly to a sulfur atom on the pyrite surface, S_{surf} , in effect substituting S_{surf} for the linking sulfur of the enzyme. As the S atoms on the surface are undercoordinated, this link to Fe_p restores the bulk 4-fold coordination of S_{surf} . A further stabilization of 0.38 eV is provided by the interaction between the N atom of $(\text{CN})_p$ and an Fe atom on the surface. In addition, the CN groups connected to both Fe_p and Fe_d , $(\text{CN})_p$ and $(\text{CN})_d$, are protonated, as usually found in water.³⁰ The optimized geometry is shown in Figure 4. The $\text{Fe}_p\text{-S}_{\text{surf}}$ bond length is 2.243 Å, even shorter than the bulk Fe-S separation, 2.264 Å. However, only CO_T is stable in this linking geometry; $\mu\text{-CO}$ is unstable and converts to CO_T . The strong $\text{Fe}_p\text{-S}_{\text{surf}}$ bond destabilizes the bond between the Fe_p and the $\mu\text{-CO}$ group bridging the two iron atoms by withdrawing electron density from the $\text{Fe}_p\text{-}(\mu\text{-CO})$ bond. We conclude that a further modification of the cluster is needed as there is no vacant coordination site exposed to the water and available for hydrogen production.

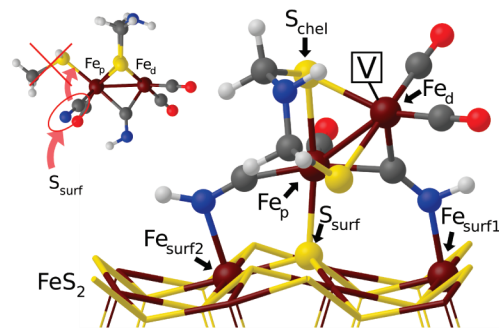


Figure 5. Optimized geometry of the $\mu\text{-(CN)}_d\text{-H}$ configuration (inset) when the N atom of $(\text{CN})_d$ is linked to a surface Fe atom ($\text{Fe}_{\text{surf}1}$) via a dative bond. $(\text{CN})_p\text{-H}$ is linked to another surface iron ($\text{Fe}_{\text{surf}2}$) via another dative bond. To allow linkage of both $(\text{CN})_d$ and Fe_p to the surface, the coordination of $(\text{CN})_p$ to Fe_p has been changed, as indicated by the upper red arrow in the inset. Fe_p is linked to a surface S atom (S_{surf}) as indicated by the lower red arrow in the inset. The bonds of Fe_p and $(\text{CN})_d$ with the surface stabilize the bridging configuration with V-up. The result is a tridentate anchoring of the cluster to the surface. S_{chel} is the upper bridging thiol, that further from the surface. Note that S_{chel} is opposite S_{surf} along an approximate octahedral coordination axis of Fe_p . The color code and the labels are those of Figure 3.

Adopting the $\mu\text{-(CN)}_d\text{-H}$ Configuration. In ref 30, we reported observing an unusual structure, one with the protonated $(\text{CN})_d$ bridging the two iron atoms (the $\mu\text{-(CN)}_d\text{-H}$ configuration), during an FPMD simulation at $T = 300\text{--}350$ K in water, cf. Figure 6d of ref 30 and the inset in Figure 5. The spontaneous formation of this isomer suggests that its energy is close to those of the $\mu\text{-CO}$ and CO_T isomers. We find that this novel isomer can form a stable link with the $\text{FeS}_2(100)$ surface. In the resulting adsorption geometry, the dative bond between the N of the $\mu\text{-(CN)}_d\text{-H}$ and a surface Fe atom (Fe_{surf}) locks the system into the bridging configuration with the vacancy on Fe_d in the “up” position (V-up) as needed for H_2 production, Figure 5, eliminating the lability found for the isolated cluster in water.⁴ To bring the $\mu\text{-(CN)}_d\text{-H}$ close enough to the surface for that bond to form, we had to move the $(\text{CN})_p\text{-H}$ to the Fe_p coordination site vacated by the methylthiol. An $\text{Fe}_p\text{-S}_{\text{surf}}$ bond then formed at the coordination site vacated by CN_p , while an additional dative bond forms between the N of $(\text{CN})_p\text{-H}$ and another surface Fe. This tridentate adsorption geometry is more stable by 0.51 eV than that of the CO_T isomer shown in Figure 4. The bonding of Fe_p to S_{surf} restores the bulk coordination of S_{surf} and is consequently strong, with a bond length of 2.271 Å close to the 2.264 Å Fe-S bond length of the bulk. Fe_p has an approximately octahedral coordination shell in which S_{surf} and that S of the dithiol bridge which is further from the surface, S_{chel} , are at opposite ends of one of the three octahedral axes, cf. Figure 5. Elementary ligand theory tells us that introducing asymmetry by strengthening one bond on an axis weakens the opposite bond on that axis. The $\text{Fe}_p\text{-S}_{\text{chel}}$ bond is indeed longer, 2.381 Å, a significant result we return to in section 5.

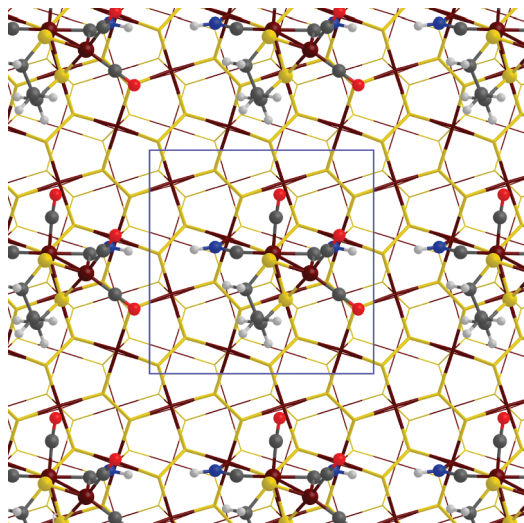


Figure 6. Top view of the μ -(CN)₄-H isomer described in Figure 5. But here, the chelating group DTMA has been replaced by PDT (see text). The structure has been replicated in the (x,y) plane to show the periodicity of the slab and the distance between the images. The edges of the tetragonal cell in the (x,y) plane are indicated by blue lines. The color code is that of Figure 3. Only the atoms of the cluster are indicated by balls and sticks.

Modifying the Chelating Bridge. To improve further the stability of the supported cluster, we have also changed the chelating group from di(thiomethyl)amine (DTMA, S-CH₂-NH-CH₂-S)^{29,30} to 1,3-propanedithiolate (PDT, S-CH₂-CH₂-CH₂-S). In vacuo, we found that protonation of DTMA modifies the HOMO charge density. When DTMA is not protonated, or analogously DTMA is replaced by PDT, the HOMO is mainly localized on the iron atoms on the pyrite surface. On the other hand, after the protonation of DTMA, 6% and 10% of the HOMO charge density is localized on Fe_p and S_{chel}, respectively. This modification of the charge weakens the Fe_p-S_{chel} bond, which increases in length from 2.37 Å to 2.48 Å. Since DTMA is protonated easily and stably,³⁰ by replacing DTMA with PDT, we both make the Fe_p-S_{chel} bond more stable and eliminate the competition for the first protonation between the amine of DTMA and Fe_d.

A Model Suitable for Simulating H₂ Production. Top and side views of the resulting model, a modified [FeFe]_H cluster with a PDT bridge supported on a (100) surface of the pyrite slab, are shown in Figures 5, 6, and 7a. The modified di-iron cluster occupies roughly half of the surface of the simulation cell used here, thus modeling a densely functionalized surface, cf. Figure 6. The distance of closest approach between the cluster replicas is 3.2 Å, the separation between the H atoms of the (CN)₄ of one cluster and of the (CN)_p of its neighbor. The use of this simulation cell allows studying H₂ formation at high coverage within an affordable computation time.

3.3. Water Environment. The water environment was modeled using 37 water molecules, either without an extra proton (“neutral water”) or with one (“acidified water”). Excluding the volume of the slab, estimated from the pyrite bulk density, and that of the cluster, estimated from tabulated

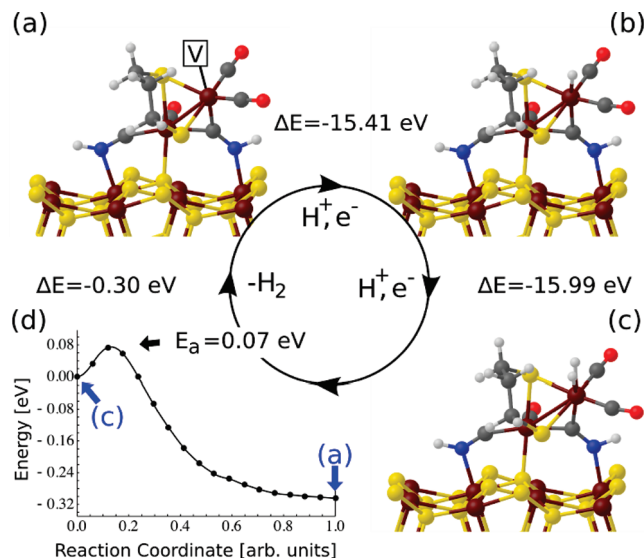


Figure 7. Catalytic cycle in vacuo for H₂ production by the μ -(CN)₄-H isomer with PDT. (a) Starting configuration with a vacant site on Fe_d. (b) The first H atom has been added to Fe_d. (c) The second H atom has been added to Fe_d. ΔE indicates the energy change (in eV), and (d) E_a is the energy barrier encountered during desorption of a H₂ molecule along the minimum-energy path computed via the NEB method. The continuous black line represents a cubic interpolation. The color code is that of Figure 3.

van der Waals radii, the density of 37 H₂O molecules within our simulation cell is close to the density of water at standard conditions. Our explicit quantum treatment of the water molecules not only takes into account important electrostatic effects but also provides a description of proton diffusion in water⁵² and of proton sharing between the cluster and water. As demonstrated previously,³⁰ all of these aspects are important for a more realistic representation of the system and its dynamics.

4. H₂ Production by the Model-Supported Catalyst

4.1. Production of H₂ in Vacuo. We addressed the pathway for hydrogen production by the FeS₂(100)-supported μ -(CN)₄-H model cluster in vacuo as preliminary to our investigation of H₂ production in water. In our previous studies,^{29,30} H₂ production occurred by two sequential reductions and protonations of the distal iron. To analyze the electron flow within the system and the effect of protonation upon that flow, we continue this step by step procedure here. In particular, we analyze changes of the electron density upon the same sequential reductions and protonations. However, we compute energy differences only between the neutral systems resulting after both the electron and proton have been added, the energies of which are well-defined in unbounded periodic systems. In the next section, where the water environment is taken into account, we discuss how electrons are transferred dynamically to Fe_d from the electrode during coordinated dynamic proton transfers from the acidified water to the cluster.

There are three main stages of the cycle: after desorption of H₂ and before hydrogenation of Fe_d, singly hydrogenated

Table 1. Relevant Bond Lengths (Å) during the Cycle in Vacuo for [FeFe]_H, H[FeFe]_H, and HH[FeFe]_H, cf. Figs. 7a–c, respectively

	Fe _p –Fe _d	N–Fe _{surf}	Fe _d –H	Fe _p –S _{surf}	Fe _p –C of (<i>μ</i> –(CN) _d)	Fe _d –C of (<i>μ</i> –(CN) _d)
[FeFe] _H	2.52	2.03		2.27	1.99	1.84
H[FeFe] _H	2.57	2.03	1.54	2.32	1.94	1.95
HH[FeFe] _H	2.89	2.05	1.73, 1.74	2.28	1.94	1.98

Table 2. Relevant Charges (In Units of the Proton Charge $e = 1.6 \times 10^{-19}$ C) and Electron-Number Changes in the [FeFe]_H, H[FeFe]_H, and HH[FeFe]_H Intermediates of the Cycle in Vacuo, cf. Figure 8^a

	change of electron number		charge			
	ΔN_{slab}	$\Delta N_{\text{cluster}}$	Fe _d	Fe _p	1st H	2nd H
[FeFe] _H , Figure 8a			–0.56	–0.66		
[FeFe] _H ^{–1} , Figure 8b	+0.74	+0.19	–0.59	–0.66		
H[FeFe] _H , Figure 8c	–0.66	+0.72	–0.62	–0.64	0.04	
H[FeFe] _H ^{–1} , Figure 8d	+0.43	+0.56	–0.65	–0.66	–0.01	
HH[FeFe] _H , Figure 8e	–0.44	+0.43	–0.55	–0.66	0.12	0.10

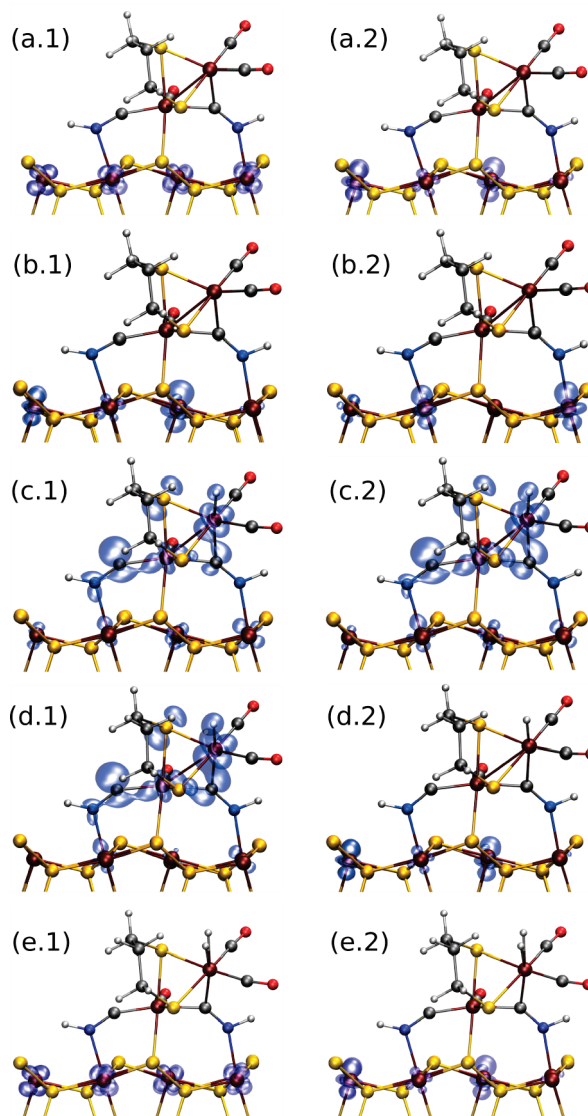
^a ΔN_{slab} ($\Delta N_{\text{cluster}}$) is the difference in electron number between adjacent configurations in the path, following the order of the table. The charges have been estimated by projection of the Kohn–Sham orbitals onto atomic wavefunctions.

Fe_d, and doubly hydrogenated Fe_d before desorption. The configuration at each of these stages is shown in Figure 7a–c, and values of the most relevant bond lengths are listed in Table 1. As electrons and protons are added during the course of the cycle, the electron numbers and the charge and spin distributions of course change. To quantify the changes, we projected the Kohn–Sham orbitals of the system onto appropriate atomic orbitals. The relevant atomic charges and electron number changes so obtained are listed in Table 2. Spin density contours are displayed in section S1 of the Supporting Information (SI) for the two odd-electron stages of the cycle.

The starting configuration is the neutral *μ*–(CN)_d–H with V-up described in the previous section and shown in Figures 6 and 7a. The HOMO of this configuration, Figure 8a.1, is formed by nonbonding *t*_{2g}-type d orbitals on the surface Fe atoms, the same as the HOMO of the clean surface.⁴³ On the other hand, the LUMO, *d*_{z²} on each clean-surface Fe, is affected by the presence of the cluster. After functionalization, it is mostly localized on those surface Fe atoms, which are not linked to the cluster, cf. Figure 8a.2. The HOMO–LUMO gap is 0.16 eV. Fe_d and Fe_p have projected charges of $-0.56e$ and $-0.66e$, respectively, cf. Table 2. The total number of electrons is even, and there is no net spin.

We used the GGA+U method⁵³ to test the sensitivity of the results in Table 2 to the GGA functional we have used. We computed values of *U* for Fe in the pyrite slab and for Fe_d and Fe_p in the [FeFe]_H cluster, both in vacuo, via the linear response method of Cococcioni and de Gironcoli,⁵⁴ finding 7.0 eV for the slab and 8.4 eV for both Fe_d and Fe_p. However, a value of 7.0 eV yields a poor value for the gap (much too large) and for the lattice constant of the clean

slab. A value of 3 eV instead yields an acceptable gap for the slab of 0.98 eV and does not affect the GGA lattice constant. We used this value of *U* for all Fe atoms in the functionalized slab. To obtain a value of *U* of 3.6 eV for both Fe_d and Fe_p in the supported cluster which would be consistent with that 3 eV value for the slab, we scaled up the latter by a factor of 1.2, the ratio of isolated cluster to

**Figure 8.** Optimized structures and HOMO (left panels, a–e.1) and LUMO (right panels, a–e.2) charge densities for the structures involved in the catalytic cycle of Table 2. (a) [FeFe]_H neutral and (b) [FeFe]_H^{–1} charged –1 isomers with a vacant coordination site on Fe_d. (c) H[FeFe]_H neutral and (d) H[FeFe]_H^{–1} charged –1 isomers with a H atom on Fe_d. (e) HH[FeFe]_H neutral with Fe_d doubly hydrogenated. The contour value for plotting the electron density was 0.003. The color code is that of Figure 3.

clean slab values obtained by the linear response method.⁵⁴ Using these values of U , we recomputed the relevant atomic charges and electron-number changes entered in Table 2, which contains the GGA results from which we have inferred delocalization within the functionalized slab. The resulting charges and number changes are reported in Table S1 of the Supporting Information. They do not differ significantly (at most by 10%) from the entries in Table 2, demonstrating that the GGA is sufficiently accurate in this respect for our purposes.

Following the scheme of ref 29, we added one electron and one proton to the Fe_d of Figure 7a, resulting in the geometry of Figure 7b, $\text{H}[\text{FeFe}]_H$. From the computed energy difference between neutral systems $[\text{FeFe}]_H$ and the resulting $\text{H}[\text{FeFe}]_H$, we obtained a hydrogen affinity for Fe_d of 15.41 eV. This addition is energetically less favorable than $1/2\text{H}_2$ by $\Delta E_1 = 0.44$ eV. Both the HOMO, which is singly occupied, and the spin-opposite LUMO are partially localized on the cluster, with 14% on Fe_d and 7% on Fe_p , cf. Figures 8c.1,2. With the HOMO density being roughly comparable on Fe_p and Fe_d , the Fe_p – Fe_d separation increases relatively little upon hydrogenation of Fe_d , from 2.52 to 2.57 Å. Before adding the proton to $[\text{FeFe}]_H^{-1}$, 74% of the added electron density resides on the slab and 19% on the cluster. [The shortfall of 7% is due to the shortcomings of the projection process.] Only 0.03 of an electron is added to Fe_d and none to Fe_p (Table 2). The remaining 16% is added primarily to the most electronegative ligands on Fe_d . The HOMO (LUMO) charge density is shown in Figure 8b.1 (Figure 8b.2). This delocalization of the electron distribution both within the complex formed by the Fe atoms and their ligands within the cluster and between the cluster and the slab mitigates against the use of the conventional notion of oxidation states and their changes for the iron atoms in this complex system. Upon adding the proton, only 2/3 of an electron is transferred from the slab to the cluster, while the proton picks up 0.96 of an electron, becoming a nearly neutral H atom, the remaining third coming largely from the electronegative ligands of Fe_d .

We next added a second proton and electron to the Fe_d –H of Figure 7b; the resulting optimized geometry is shown in Figure 7c. The hydrogen affinity is 15.99 eV, which means that this addition is energetically more favorable than $1/2\text{H}_2$ by $\Delta E_2 = -0.14$ eV. Upon addition of the second hydrogen to Fe_d , the Fe_p – Fe_d bond distance increases from 2.57 to 2.89 Å (cf. Table 1). Even though the individual Fe_d –H distances are 0.2 Å longer than that in the singly hydrogenated case, the η bond of the dihydrogen to Fe_d is strong and substantially breaks the Fe_p – Fe_d resonance, leading to this large additional opening of the Fe_p – Fe_d bond. The H–H distance of the two hydrogen atoms coordinated to Fe_d is 0.82 Å, which is only 0.07 Å larger than the computed distance of the isolated H_2 molecule. The HOMO and the LUMO are localized on the pyrite surface, essentially the same as for the configuration with a vacancy on Fe_d , cf. Figure 8e. Before adding the proton, cf. Figure 8d, 0.43 and 0.56 of the added electron goes to the slab and the cluster, respectively. Within the cluster, the added 0.56 is diffusely distributed over Fe_p , Fe_d , their electronegative ligands, and

the H atom. Upon adding the proton, only 0.44 of an electron transfers from the slab to the cluster, while nearly twice that much is transferred to the resulting second H atom. The balance is transferred from the first H on Fe_d (13%), Fe_d (10%), and its electronegative ligands. Both the H atoms on Fe_d are weakly positively charged. As above, within our computational framework, there is little evidence for the utility of the concept of classical oxidation states.

The last step of the cycle is H_2 desorption. We found that for this step no additional electron is needed: H_2 desorption from the neutral Fe_d – H_2 isomer of Figure 7c is an exothermic process with a $\Delta E = -(\Delta E_1 + \Delta E_2) = -0.30$ eV. Its activation energy barrier as computed with NEB³⁸ is only 0.07 eV, cf. Figure 7d. The barrier is low because of a near balance between the energy cost of breaking the η bond and the energy benefit from molecular bonding arising from decreasing the initially close approach of the H atoms. After desorption, the initial state with a vacancy on Fe_d is restored, Figure 7a, and a new cycle can start again.

4.2. H_2 Formation in Water. To investigate hydrogen production in water, we started by adding 37 H_2O molecules to the neutral isomer of Figure 7a. We then relaxed the system via damped dynamics, followed by an FPMD simulation at room temperature (RT) lasting 1.5 ps. We analyzed several configurations extracted from the simulation by computing the Kohn–Sham (KS) orbitals and by projecting them onto atomic wave functions. The HOMO–LUMO energy gaps in these configurations are in the range 0.34–0.49 eV, and both the HOMO and LUMO are almost entirely ($\sim 80\%$) localized on the Fe atoms of the pyrite surface. On the functionalized face of the slab, two of the eight surface Fe atoms within the unit cell are linked to the cluster; two under the cluster are inaccessible to the water, and water molecules are absorbed on the remaining four sites after the relaxation. They remain absorbed throughout the hydrogen production cycle described in the following.

First Protonation of Fe_d . After the equilibration described above, we added an electron to the electrode-cluster complex and a proton to a nearby water molecule close to the distal iron, forming a hydronium. The system remains globally neutral. Earlier,³⁰ we found that in dense water there is a competition between the diffusion of an H_3O^+ toward Fe_d and the transfer of its proton away via the Grothuss-shuttle to the surrounding water molecules that solvate it. The diffusion toward Fe_d was energetically favorable, but proton exchange between water molecules was entropically favorable. The proton transfer away from the Fe_d was so rapid that we could not observe the less-frequent proton transfer to the site on the brief time scale of our simulations. This is the so-called “wandering proton” problem.⁵⁵ To simulate protonation of Fe_d , we biased the competition by constraining two of the three O–H bonds of the hydronium, leaving only one free to move, either to transfer away from or to protonate the Fe_d . Here, we constrained the hydronium in the same way. As the present focus is on learning how to modify the cluster for stable hydrogen production, we do not attempt here to determine the free energy barrier for protonation; we address that issue elsewhere.²⁷ During the 2.6 ps of an FPMD simulation at RT, the hydronium stayed close to

the distal iron, but the proton was not transferred to Fe_d . The distance between Fe_d and the unconstrained proton of the H_3O^+ fluctuated within 1.9–2.8 Å. This behavior suggests not only an entropic barrier but also the presence of an energy barrier that could not be overcome during our short simulation.

The HOMO is singly occupied and localized on the pyrite surface. However, the LUMO is partially localized on Fe_d . In a snapshot in which the distance between Fe_d and the closest H atom of the hydronium is 1.94 Å, 22% of the LUMO is on Fe_d . Accordingly, to enhance the rate of proton transfer to Fe_d , we added a second electron to a configuration taken from this FPMD simulation, changing the total charge to -1 . We kept the constraints on two of the three O–H bonds of the hydronium and observed proton transfer to Fe_d after 1.5 ps of an FPMD simulation at RT lasting 3.6 ps. After protonation, the system was stable; the proton remained on Fe_d . We found that the presence of the hydronium close to the cluster stabilized an increased negative charge of Fe_d before the protonation. From an analysis of five configurations extracted from this simulation with distances of the nearest H of the hydronium to Fe_d in the range 2.36–2.68 Å, we found that about 25% of the HOMO was on Fe_d . The HOMO (doubly occupied) is in the gap, about 0.2 eV above the next valence state and 0.1 eV below the pyrite conduction bands. We also analyzed several configurations extracted from an FPMD simulation at RT lasting 2.2 ps with the supported catalyst charged -2 and immersed in *neutral* water. Without the hydronium ion, the extra charge was delocalized on the Fe atoms of the surface. From all of the above evidence, we conclude that the hydrogenation of Fe_d occurs via coordinated electron and proton transfers to Fe_d .

We repeated the simulation starting from a slightly different initial configuration with charge -1 and observed the protonation of Fe_d after 0.6 ps of a simulation lasting 1.1 ps. However, without constraining the bond lengths of the H_3O^+ ion, the proton was instead transferred to a nearby water molecule and solvated. The hydronium stayed 4–5 Å away from Fe_d for the duration of the simulation, 3.1 ps. At this larger distance, the hydronium did not stabilize the charge on Fe_d . Occasionally, depending on the configuration, a small contribution of Fe_p and Fe_d (less than 5% each) to the HOMO was observed. In contrast, when the hydronium is close to Fe_d , the HOMO amplitude on the Fe_d is significant. This observation suggests that the catalytic reaction will be constrained by proton diffusion and that the diffusion rate will be significantly affected by the potential energy surface experienced by the proton near the surface of the functionalized electrode.

To investigate further the effects produced by the presence of the hydronium on the charge distribution in the cluster, we removed the hydronium from a configuration selected before the proton transfer to Fe_d and computed the KS orbitals of the resulting system (which has a charge of -2). The HOMO was partially located on Fe_d (13%), with its remainder mostly on the surface. We equilibrated the system starting from this configuration and observed a rapid decrease of the contribution of Fe_d to the HOMO. In less than 0.6 ps,

the HOMO was localized entirely on the pyrite surface. This suggests that the presence of a H_3O^+ close to Fe_d induces, in addition to its attractive electrostatic potential, a structural modification of the supported cluster which contributes to stabilization of the charge on Fe_d . After the removal of the hydronium, time is required for the system to relax and for the charge to transfer back to the slab. We found the same effect in vacuo by removing the water molecules as well as the hydronium ion from the snapshot discussed above. In summary, we found that *two* additional electrons are required to lower the energy barrier involved in the proton transfer, and that the presence of the hydronium is required to stabilize the charge on the distal iron.

Second Protonation of Fe_d and H_2 Production. After protonation of Fe_d , we equilibrated the system for 2.1 ps at RT and then introduced a second proton into the water, neutralizing the system. In the starting configuration, the distance between the O atom of the H_3O^+ and Fe_d was 4.4 Å. The proton diffused very rapidly toward the Fe_d –H. Transfer of the H^+ to the Fe_d –H was followed by H_2 desorption, the supported catalyst remaining neutral. When Fe_d is singly protonated and the system is in the neutral $\text{H}[\text{FeFe}]_{\text{H}}$ configuration, the H atom on Fe_d is essentially neutral, cf. Table 2 for the vacuum case. When a single electron is added prior to the addition of a second proton to the water, creating the $\text{H}[\text{FeFe}]_{\text{H}}^{-1}$ configuration, that H atom remains essentially neutral (compare Table 2). Nevertheless, in the presence of a chain of water molecules which connects the solvated hydronium to the Fe_d –H (with the final molecule of the chain having one of its H atoms close to and spontaneously coordinated with the protruding H atom), the protonation occurs via the fast Grotthuss-like diffusion of H^+ directly toward Fe_d driven by a transfer of an electron which is coordinated with the motion of the proton. The transfer comes in part from the electrode and in part from the electronegative ligands bonded to Fe_d . The resulting H_2 molecule is η -coordinated with Fe_d in the $\text{HH}[\text{FeFe}]_{\text{H}}$ configuration and is weakly positively charged (compare Table 2). The total charge on Fe_d and Fe_p remains unchanged, the electron deficiency on the η - H_2 arising from charge flowing from it to the electronegative ligands of Fe_d . Conversely, the bare Fe_d , either in the neutral $[\text{FeFe}]_{\text{H}}$ configuration or in the negatively charged $[\text{FeFe}]_{\text{H}}^{-1}$ configuration, does not coordinate with an H_2O molecule because of (i) Coulomb repulsion between the lone pair of the O atom and the negative charge on Fe_d in both configurations (cf. Table 2), (ii) the Pauli repulsion associated with formation and occupation of an antibonding σ orbital between the d_z^2 orbital of the vacant coordination site and the overlapping sp^3 -like lone pair orbitals of the oxygen, and (iii) the cost of modifying the H-bond network of the water. This reduces the rate of the first protonation of Fe_d , which, as pointed out in ref 30, is the bottleneck of the cycle because it occurs only after the diffusion of a partially unsolvated H_3O^+ toward Fe_d to a distance short enough for a proton transfer (<2 Å). Once the Fe_d –H bond is formed in $\text{H}[\text{FeFe}]_{\text{H}}$, however, effect ii is no longer in play, and the lowering in energy arising from sharing the electron of that H with an H atom of a water molecule together with that from the transfer of

the additional electron is strong enough to overcome effect iii, as we have observed in the 2.1 ps simulation referred to above.

We repeated the simulation of the second protonation of Fe_d by changing the initial position of the additional proton introduced in water. Here, the initial distance between the O atom of the H_3O^+ and Fe_d was 5.6 Å. From this configuration, the proton diffused away from Fe_d . After about 0.7 ps, the hydronium was within 4 Å of the S_{chel} exposed to the water. The proton remained in the water as H_3O^+ at distances shorter than 4 Å from this sulfur for the remaining 2.3 ps of an FPMD at RT lasting 3.0 ps. From these results, we can draw two conclusions. First, there is only a small subset of reaction pathways along which barrier-free adsorption of a second proton and subsequent hydrogen production can occur. Second, there are local configurations in which the second proton can be at least temporarily trapped, e.g., near S_{chel} in the second simulation, reducing the overall catalytic efficiency, as is the case also for the first protonation. This apparent attraction of S_{chel} for the proton turns out to have serious consequences, as discussed below.

5. An Instability of the Supported $[\text{FeFe}]_{\text{H}}$ Cluster: $[\text{FeFe}]_{\text{P}}$ as a Remedy

Thus far, we have focused on hydrogen production at a vacant coordination site on Fe_d . However, the S atom of PDT further from the surface, S_{chel} , turns out to be readily protonated. We explore in subsection 5.1 the negative consequences of that protonation for the stability of the cluster both *in vacuo* and in water and consequently for hydrogen production. In subsection 5.2, we describe a further modification of the cluster which eliminates that difficulty.

5.1. Protonation of S_{chel} Opens the PDT Bridge. Calculations *in vacuo* show that protonation of the S atom of PDT furthest from the surface (S_{chel}) breaks the $\text{Fe}_p\text{--S}_{\text{chel}}$ bond, opening the bridge and exposing a bridging site between the two Fe's of the cluster. The resulting configuration, Figure 9a, is lower in energy than that of $\text{Fe}_d\text{--H}$ by 0.07 eV when the system is neutral. Transfer of the H atom from S_{chel} to the vacant site on Fe_d would restore the catalytic cycle discussed in the previous section, but the corresponding energy barrier computed *in vacuo* via NEB is quite high, 0.68 eV. Starting from this configuration with S_{chel} protonated, we added one electron and one proton to Fe_d , producing the configuration in Figure 9b. This configuration is less stable than that with both protons on Fe_d by 0.26 eV. However, the energy barrier to transferring a H atom from S_{chel} to Fe_d when a proton is already present on Fe_d is 0.61 eV. Such a high barrier would make H_2 production very inefficient.

We also added an electron and a proton to Fe_p in the open configuration of Figure 9a. The optimized geometry, Figure 9c, is similar to the $\mu\text{-H}$ configuration explored previously.^{29,30} A H atom bridges Fe_p and Fe_d . The energy is lower than for two H atoms on Fe_d by 0.52 eV, Figure 7c. The presence of $\mu\text{-H}$ would make H_2 production after adding another H to Fe_p very inefficient because of a large H_2 desorption barrier arising from strong dihydrogen bonding (see below), as found earlier.^{29,30}

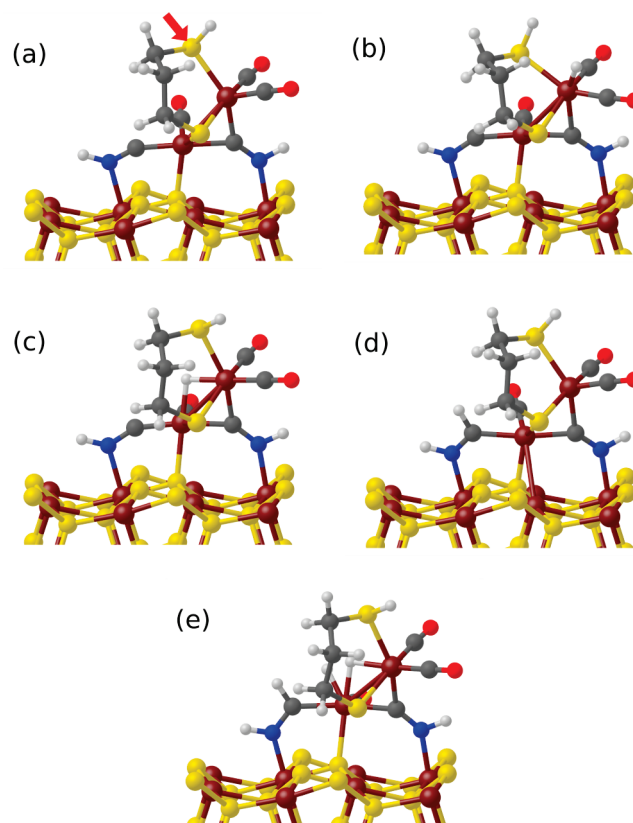


Figure 9. Configurations corresponding to alternative pathways involving the protonation of the S_{chel} atom of the PDT. These neutral isomers were optimized *in vacuo*. (a) Geometry of the supported catalyst with a hydrogen atom connected to the S_{chel} , indicated by the red arrow. The protonation breaks the $\text{Fe}_p\text{--S}_{\text{chel}}$ bond. The resulting configuration is lower in energy than that of the $\text{Fe}_d\text{--H}$ of Figure 7b by 0.07 eV. (b) Geometry with S_{chel} and Fe_d hydrogenated. (c) Geometry with an H atom bridging Fe_p and Fe_d ($\mu\text{-H}$), and S_{chel} hydrogenated. (d) Geometry with the C atom of the $(\text{CN})_p\text{--H}$ group protonated. (e) Geometry with one H atom bridging Fe_p and Fe_d ($\mu\text{-H}$), a second H atom on Fe_p , and as in d the C atom of the $(\text{CN})_p\text{--H}$ group protonated. The color code is that of Figure 3.

In water, protonation of S_{chel} was observed for a configuration where the supported catalyst was charged -2 and had a vacancy on Fe_d . The O atom of the hydronium ion was located at 4.85 Å and 3.93 Å from Fe_d and S_{chel} , respectively. We constrained all three O–H bond lengths of the H_3O^+ to prevent proton diffusion and observed displacement of S_{chel} toward the hydronium driven by electrostatic interaction. After removal of one of the three constraints in the H_3O^+ , displacement of S_{chel} breaks the $\text{Fe}_p\text{--S}_{\text{chel}}$ bond, whose length increases up to about 4 Å. Simultaneously, the proton approaches S_{chel} along the line linking it with Fe_p . The main bond lengths involved in the reaction are plotted as function of time in Figure 10. S_{chel} becomes 2-fold coordinated and accepts a proton from the hydronium. The outcome of this reaction is an opened structure with the chelating group displaced toward Fe_d , resulting in reduction of the empty space associated with the vacant site. The corresponding structure optimized *in vacuo* (cf. Figure 9a) has a lower energy than that with a

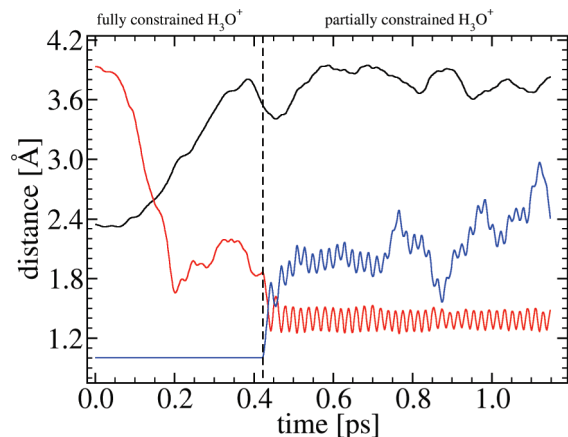


Figure 10. Plot of the $\text{Fe}_p\text{--S}_{\text{chel}}$ (black line), $\text{S}_{\text{chel}}\text{--H}$ (red line), O--H (blue line) bond length [Å] as a function of time [ps]. The FPMD simulation at RT was divided into two parts separated by the vertical dotted line at about 0.4 ps. In the first part, lasting about 0.4 ps, all three O--H bond lengths of the hydronium were constrained (fully constrained H_3O^+). The $\text{Fe}_p\text{--S}_{\text{chel}}$ bond broke (black line), and the hydronium moved toward S_{chel} (red line). When the constraint on the closest H atom of the hydronium to S_{chel} was removed (partially constrained H_3O^+), the proton transferred to S_{chel} .

proton on Fe_d (cf. Figure 7b and section 4) by 0.67 eV (neutral) and 0.21 eV (charged -1).

The protonation of S_{chel} strongly affects the structure of the catalyst. It opens the structure and creates an undercoordinated Fe_p that could be protonated. It also reduces the empty space available to the vacancy on Fe_d , making its protonation more difficult. To examine the consequences of these potentially unfavorable changes, we explored the reactivity of the cluster with S_{chel} protonated in acidified water. We performed two FPMD simulations at RT in water containing one hydronium ion differing in its initial position. In one case, the H_3O^+ was close to Fe_p , and in the other, it was close to Fe_d . In neither case did a proton transfer to the Fe atom when the charge was -1 or -2 . For Fe_p , the C atom of the $(\text{CN})_p\text{--H}$ was protonated instead, producing the isomer in Figure 9d. Nevertheless, we optimized the geometry for a dihydrogen at Fe_p in vacuo, Figure 9e, resulting in one H in the μ -bridging position and one H on Fe_p . The resulting energy is lower by 0.31 eV than that after desorption, suggesting that this channel for H_2 production would be inefficient in water. Adding an additional electron makes desorption of H_2 an exothermic process by 0.20 eV, but the activation energy barrier of 0.47 eV is too large for an efficient turnover. For Fe_d , we attribute the reduction of the reactivity of its vacant site to the compression of the space around it and the consequent decrease in electron density.

In summary, S_{chel} can be protonated along very specific pathways in vacuo, pathways consistent with elementary ligand-field theory. In water, the $\text{Fe}_p\text{--S}_{\text{chel}}$ bond opens in the process of the protonation of S_{chel} along similar pathways. With S_{chel} protonated, all H_2 production pathways *either* are very inefficient in themselves and would reduce the overall efficiency of the catalyst due to large barriers *or* would lead to destruction of the catalyst. As we demonstrated in section

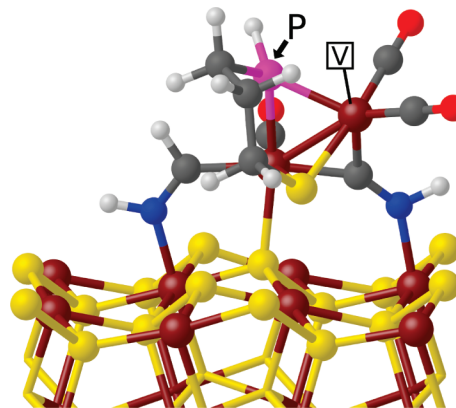


Figure 11. Side view of the $[\text{FeFe}]_p$ cluster linked to the FeS_2 (100) surface in the same way as the $[\text{FeFe}]_H$ cluster in Figure 5. The S_{chel} of $[\text{FeFe}]_H$ is replaced by a PH group. The distal iron has a vacant site, indicated by V. The color code is that of Figure 3.

4.2 that the supported catalyst produces hydrogen efficiently when its chelating bridge is intact, the next task is to find a modification of the cluster which stabilizes its chelating bridge.

5.2. Introducing $[\text{FeFe}]_p$. Opening of the chelating bridge must be avoided. The opening is a consequence of the inability of the sulfur atom S_{chel} to maintain four strong covalent bonds after the protonation that will inevitably occur upon exposure to acidified water. The weakest of the four, its bond to Fe_p , breaks and opens the bridge. The $\text{S}_{\text{chel}}\text{--Fe}_p$ bond is weak because it lies opposite of the strong $\text{Fe}_p\text{--S}_{\text{surf}}$ bond along one of the quasi-octahedral coordination axes of Fe_p , which withdraws electron density from it. The $\text{Fe}_p\text{--S}_{\text{surf}}$ bond is strong because it restores the bulk 4-fold coordination of S_{surf} . Clearly, one needs to replace S_{chel} by a group stereochemically similar to a sulfur to minimize structural distortion of the cluster and distortion of its electron density, yet capable of sustaining or being protected from protonation as well as forming the three strong bonds required for anchoring the bridge to the di-iron. The group P--H can be regarded as an approximation to sulfur and could satisfy the minimum distortion requirement. Its phosphorus atom is also capable of forming the three strong bonds needed to anchor the PDT bridge to the di-irons. Substitution of P--H for S_{chel} yields the $[\text{FeFe}]_p$ cluster shown in Figure 11 stably attached to the pyrite surface. We have carried out computations on hydrogen production in vacuo and in water by $[\text{FeFe}]_p$. We find that it remains stably attached to the electrode and intact throughout a successful hydrogen production cycle, as reported in detail elsewhere.²⁷

6. Discussion and Conclusions

Our goal in this work was to find a conformation of the $[\text{FeFe}]_H$ cluster which would (1) link stably to the (100) surface of pyrite while (2) retaining a vacant coordination site on Fe_d at which H_2 could be produced readily in acidified H_2O . We started by mimicking the sulfur link between Fe_p and an Fe of the Fe_4S_4 cubane in the enzyme by inserting a sulfur between the Fe_p of the naked cluster and a surface Fe. However, the $\text{Fe}_p\text{--S}$ link broke in water upon proton-

ation, just as the methylthiol-Fe_p link did in water in the absence of the surface.³⁰ Next, we linked Fe_p directly to a surface S atom, eliminating the S intermediary but still mimicking the linkage in the enzyme. The resulting complex was stable only in the CO_T configuration, which could not produce hydrogen efficiently. We inferred that the difference between the enzyme environment and that of a pyrite surface exposed to water was too great for the twin requirements 1 and 2 to continue to be met by a linking sulfur, and we abandoned the notion of that biomimetic linkage.

One possible solution would be to use for coupling to the surface a ligand of the di-irons already in the μ -bridging position. The μ -CO of the native configuration could not serve as that ligand, but we had observed an unusual novel structure with the protonated (CN)_d moved into the bridging position during an FPMD simulation at $T = 300\text{--}350$ K in water,³⁰ Figure 5 (inset). To allow the N of that μ -(CN)_d-H to come close enough to the surface to form a dative bond with a surface Fe atom, we shifted the (CN)_p-H into the coordination site formerly occupied by the methylthiol in the cluster of refs 29 and 30, freeing Fe_p to bond to a surface S atom at the vacated coordination site. The N of the (CN)_p-H forms a dative bond with another surface Fe atom, creating a strong, stable tridentate support on the surface, Figure 5, with the vacant coordination site preserved at Fe_d, meeting requirement 1 and opening the possibility that requirement 2 could be met. The DTMA bridge used in refs 29 and 30 was replaced by a PDT bridge, with a further increase in stability.

In our first study of hydrogen production in vacuo,²⁹ we simply added both electrons and protons to the cluster by fiat. In our second study,³⁰ protons were introduced more realistically, migrating through the water via the Grotthuss-shuttle mechanism,⁵² while electrons were still simply added by fiat. In the present work, the electrons were added to the entire surface/ μ -(CN)_d-H cluster complex. The relevant LUMO before addition and HOMO after addition of an electron at the various stages of the hydrogen producing cycle were localized primarily on the pyrite surface before the approach of a proton to Fe_d. As the proton-Fe_d distance shrank, the HOMO shifted primarily to Fe_d. After protonation, the first hydrogen projected outward into the water and coordinated with the H of the closest water molecule, providing a channel along which the second proton could approach by the Grotthuss mechanism to form the second hydrogen. This demonstrates that both electron transfer from the FeS₂ slab to the μ -(CN)_d-H cluster and proton transfer from the water can take place, the essential steps in modeling hydrogen production.

As found in our previous study,³⁰ transfer of the first proton from the water to Fe_d is the bottleneck in the H₂-production cycle because of its competition with rapid proton diffusion away via the Grotthuss-shuttle. To observe that proton transfer within the picosecond time scale of our simulations, we constrained two of the OH bonds of the neighboring hydronium. No constraint was needed for the second because the projecting H of the first Fe_d-H complex coordinated with the H of a neighboring water molecule, providing a path for transfer of the second proton to Fe_d via

the Grotthuss mechanism. Desorption of the resulting H₂ was then observed during our picosecond-scale simulations, demonstrating H₂ production by the supported cluster, meeting requirements 1 and 2.

That did not complete the story, however. We observed very specific configurations of a hydronium ion neighboring S_{chel} which both weaken the Fe_p-S_{chel} bond, breaking it and allowing spontaneous proton transfer to S_{chel}. When that H of the hydronium which is closest to S_{chel} approaches it along or near the direction of the bond joining it to Fe_p, the Fe_p-S_{chel} bond lengthens and then opens as the bond switches from Fe_p to H, forming a thiol. These pathways are consistent with expectations from elementary valence-bond theory. The PDT chelating bridge decouples from Fe_p, opening a channel of proton access to the di-iron from the water and allowing for the possibility of the binding of two hydrogen atoms in a bridging configuration there, in agreement with calculations by Felton et al.^{23,24} on an analogous open-bridged cluster in a vacuum. These authors also observed hydrogen production experimentally with that cluster as a catalyst and, on the basis of their computations, attributed it to hydrogen formation at and desorption from the di-iron bridging site. They did not determine the barrier to desorption, however. We have found the barrier to desorption of H₂ in vacuo from the di-iron bridging site of a CO_T configuration to be 0.47 eV–0.54 eV depending on the cluster configuration. This should be a reasonable estimate for the barrier to desorption of a di-iron bridging, H-atom pair within our μ -(CN)_d-H cluster when its chelating PDT bridge is opened by protonation of S_{chel}. Moreover, we have found the desorption of H₂ from the open configuration when a dihydrogen is at Fe_p to be inhibited by an activation energy barrier of 0.47 eV. Should the cluster remain in the open configuration for a significant fraction of the time, its performance as a catalyst would be degraded or the catalyst would even be destroyed.

This opening of the chelating bridge must be avoided. The opening is a consequence of the inability of the sulfur atom S_{chel} to maintain four strong covalent bonds after the protonation that will inevitably occur upon exposure to acidified water. The weakest of the four, its bond to Fe_p, breaks and opens the bridge. As discussed briefly in section 5.2, substituting a P-H group for S_{chel} solves that problem, creating the [FeFe]_p cluster as a viable model catalyst for the electrocatalytic production of hydrogen. We report elsewhere a detailed analysis of hydrogen production by [FeFe]_p.²⁷ Moreover, substitution of other moieties for the H of the P-H allows for fine-tuning of the catalytic activity. In particular, it opens the possibility of reduction of the free-energy barrier to first protonation, the bottleneck in the production cycle.

In conclusion, we have presented extensive FPMD simulations of hydrogen production by a coupled catalyst–electrode system consisting of a modified [FeFe]_H cluster supported on a FeS₂(100) surface in acidified water, a system far more complex than those typically treated by first principles methods. The principal findings of the present study are as follows: (i) A stable thiol link analogous to that in the enzyme cannot be made between the unmodified [FeFe]_H cluster and the pyrite surface. (ii) A modified μ -(CN)_d-H isomer of the

[FeFe]_H cluster forms a stable, tridentate link to the surface. (iii) There is a low-activation-energy pathway for hydrogen production from acidified water by the functionalized pyrite surface. (iv) The electrons in the electrode and the protons in the water attract each other to the catalyst in the first hydrogenation of Fe_a. (v) That di-iron bridging sulfur which is exposed to the water develops a weak link to Fe_p upon its protonation, but replacing it with a P–H group eliminates this instability²⁷ (vi) Facile electrocatalytic hydrogen production is feasible by this modified-cluster/pyrite system. To achieve these results, we have employed the first principles molecular dynamics method at an advanced level of system complexity and have illustrated thereby the utility and power of FPMD for molecular design.

Acknowledgment. This work was supported by the Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award DE-FG02-06ER-46344. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The simulations presented in this article were in part performed on computational resources supported by the PICSciE-OIT High Performance Computing Center and Visualization Laboratory. We thank Patrick Sit for discussion of the GGA+U method.

Supporting Information Available: Figures S1 and S2 display the electron-spin density contours in the two odd-electron configurations of the cycle, [FeFe]_H^{−1} and H[FeFe]_H, respectively. Table S1 contains the relevant charges and electron number changes in the intermediates of the H₂ production cycle in vacuo obtained via GGA+U for comparison with the corresponding GGA results in Table 2. This material is available free of charges via the Internet at <http://pubs.acs.org/>.

References

- Vignais, P. M.; Billoud, B.; Meyer, J. *FEMS Microbiol. Rev.* **2001**, *25*, 455.
- Fontecilla-Camps, J. C.; Volbeda, A.; Cavazza, C.; Nicolet, Y. *Chem. Rev.* **2007**, *107*, 4273.
- Bruschi, M.; Greco, C.; Kaukonen, M.; Fantucci, P.; Ryde, U.; De Gioia, L. *Angew. Chem., Int. Ed.* **2009**, *48*, 3503.
- Cao, Z. X.; Hall, M. B. *J. Am. Chem. Soc.* **2001**, *123*, 3734.
- Liu, Z. P.; Hu, P. *J. Am. Chem. Soc.* **2002**, *124*, 5175.
- Bruschi, M.; Fantucci, P.; De Gioia, L. *Inorg. Chem.* **2002**, *41*, 1421.
- Greco, C.; Bruschi, M.; De Gioia, L.; Ryde, U. *Inorg. Chem.* **2007**, *46*, 5911.
- Fiedler, A. T.; Brunold, T. C. *Inorg. Chem.* **2005**, *44*, 9322.
- Zilberman, S.; Stiefel, E. I.; Cohen, M. H.; Car, R. *J. Phys. Chem. B* **2006**, *110*, 7049.
- Szilagyi, R. K.; Winslow, M. A. *J. Comput. Chem.* **2006**, *27*, 1385.
- Liu, T. B.; Li, B.; Singleton, M. L.; Hall, M. B.; Darensbourg, M. Y. *J. Am. Chem. Soc.* **2009**, *131*, 8296.
- Siegbahn, P. E. M.; Tye, J. W.; Hall, M. B. *Chem. Rev.* **2007**, *107*, 4414.
- Fan, H. J.; Hall, M. B. *J. Am. Chem. Soc.* **2001**, *123*, 3828.
- Cracknell, J. A.; Vincent, K. A.; Armstrong, F. A. *Chem. Rev.* **2008**, *108*, 2439.
- Tye, J. W.; Hall, M. B.; Darensbourg, M. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16911.
- Tard, C.; Liu, X. M.; Ibrahim, S. K.; Bruschi, M.; De Gioia, L.; Davies, S. C.; Yang, X.; Wang, L. S.; Sawers, G.; Pickett, C. J. *Nature* **2005**, *433*, 610.
- Peters, J. W.; Lanzilotta, W. N.; Lemon, B. J.; Seefeldt, L. C. *Science* **1998**, *282*, 1853.
- Nicolet, Y.; de Lacey, A. L.; Venede, X.; Fernandez, V. M.; Hatchikian, E. C.; Fontecilla-Camps, J. C. *J. Am. Chem. Soc.* **2001**, *123*, 1596.
- Vincent, K. A.; Cracknell, J. A.; Lenz, O.; Zebger, I.; Friedrich, B.; Armstrong, F. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16951.
- Nicolet, Y.; Piras, C.; Legrand, P.; Hatchikian, C. E.; Fontecilla-Camps, J. C. *Struct. Fold. Des.* **1999**, *7*, 13.
- Lubitz, W.; Reijerse, E.; van Gestel, M. *Chem. Rev.* **2007**, *107*, 4331.
- Reisner, E.; Fontecilla-Camps, J. C.; Armstrong, F. A. *Chem. Commun.* **2009**, 550.
- Felton, G. A. N.; Vannucci, A. K.; Chen, J. Z.; Lockett, L. T.; Okumura, N.; Petro, B. J.; Zakai, U. I.; Evans, D. H.; Glass, R. S.; Lichtenberger, D. L. *J. Am. Chem. Soc.* **2007**, *129*, 12521.
- Felton, G. A. N.; Vannucci, A. K.; Okumura, N.; Lockett, L. T.; Evans, D. H.; Glass, R. S.; Lichtenberger, D. L. *Organometallics* **2008**, *27*, 4671.
- Hambourger, M.; Gervaldo, M.; Svedruzic, D.; King, P. W.; Gust, D.; Ghirardi, M.; Moore, A. L.; Moore, T. A. *J. Am. Chem. Soc.* **2008**, *130*, 2015.
- Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- Zipoli, F.; Car, R.; Cohen, M. H.; Selloni, A. *J. Am. Chem. Soc.* **2010**, *132*, 8593.
- Liu, Z. P.; Hu, P. *J. Am. Chem. Soc.* **2002**, *124*, 11568.
- Sbraccia, C.; Zipoli, F.; Car, R.; Cohen, M. H.; Dismukes, G. C.; Selloni, A. *J. Phys. Chem. B* **2008**, *112*, 13381.
- Zipoli, F.; Car, R.; Cohen, M. H.; Selloni, A. *J. Phys. Chem. B* **2009**, *113*, 13096.
- Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M. *J. Phys.: Condens. Matter* **2009**, *21*, 395502.
- Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892.
- Laasonen, K.; Pasquarello, A.; Car, R.; Lee, C.; Vanderblit, D. *Phys. Rev. B* **1993**, *47*, 10142.
- Finklea, S. L.; Cathey, L.; Amma, E. L. *Acta Crystallogr., Sect. A* **1976**, *32*, 529.

- (36) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.
- (37) Blöchl, P. E.; Parrinello, M. *Phys. Rev. B* **1992**, *45*, 9413.
- (38) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.
- (39) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301.
- (40) Nair, N. N.; Schreiner, E.; Marx, D. *J. Am. Chem. Soc.* **2006**, *128*, 13815.
- (41) Stirling, A.; Bernasconi, M.; Parrinello, M. *Phys. Rev. B* **2007**, *75*, 165406.
- (42) Stirling, A.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *119*, 4934.
- (43) Stirling, A.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 8917.
- (44) Ferrer, I. J.; Nevskaja, D. M.; Delasheras, C.; Sánchez, C. *Solid State Commun.* **1990**, *74*, 913.
- (45) Schreiner, E.; Nair, N. N.; Marx, D. *J. Am. Chem. Soc.* **2008**, *130*, 2768.
- (46) Boehme, C.; Marx, D. *J. Am. Chem. Soc.* **2003**, *125*, 13362.
- (47) Rodriguez, J. A.; Abreu, I. A. *J. Phys. Chem. B* **2005**, *109*, 2754.
- (48) Kendelewicz, T.; Doyle, C. S.; Bostick, B. C.; Brown, G. E. *Surf. Sci.* **2004**, *558*, 80.
- (49) Rosso, K. M.; Becker, U.; Hochella, M. F. *Am. Mineral.* **1999**, *84*, 1549.
- (50) Rickard, D.; Luther, G. W. *Chem. Rev.* **2007**, *107*, 514.
- (51) Murphy, R.; Strongin, D. R. *Surf. Sci. Rep.* **2009**, *64*, 1.
- (52) Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601.
- (53) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. *Phys. Rev. B* **1991**, *44*, 943.
- (54) Cococcioni, M.; de Gironcoli, S. *Phys. Rev. B* **2005**, *71*, 35105.
- (55) Leung, K.; Nielsen, I. M. B.; Criscenti, L. J. *J. Am. Chem. Soc.* **2009**, *131*, 18358.

CT100319B

JCTC

Journal of Chemical Theory and Computation

Computational Study of the DNA-Binding Protein *Helicobacter pylori* NikR: The Role of Ni²⁺

Francesco Musiani,^{†,∇} Branimir Bertoša,^{‡,§,∇} Alessandra Magistrato,[‡]
Barbara Zambelli,[†] Paola Turano,^{⊥,‡} Valeria Losasso,[‡] Cristian Micheletti,[‡]
Stefano Ciurli,^{*,†,⊥} and Paolo Carloni^{‡,||}

Laboratory of Bioinorganic Chemistry, University of Bologna, Viale G. Fanin 40,
40127 Bologna, Italy, International School for Advanced Studies (SISSA) and
CNR-IOM-DEMOCRITOS National Simulation Center, via Bonomea 265,
34136 Trieste, Italy, Ruder Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia, German
Research School for Simulation Science, FZ-Jülich and RWTH, Wilhelm-Johnen-Strasse,
52428 Jülich, Germany, Center for Magnetic Resonance (CERM), University of Florence, Via
Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy, and Department of Chemistry, University of
Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Received November 30, 2009

Abstract: An integrated approach, combining atomistic molecular dynamics simulations, coarse-grained models, and solution NMR, was used to characterize the internal dynamics of *Hp*NikR, a Ni-dependent transcription factor. Specifically, these methods were used to ascertain how the presence of bound Ni²⁺ ions affects the stability of the known open, cis, and trans forms observed in the crystal structures of this protein as well as their interconversion capability. The consensus picture emerging from all the collected data hints at the interconversion of NikR among the three types of conformations, regardless of the content of bound Ni²⁺. On the basis of atomistic and coarse-grained simulations, we deduce that the interconversion capability is particularly effective between the cis and the open forms and appreciably less so between the trans conformer and the other two forms. The presence of the bound Ni²⁺ ions does, however, affect significantly the degree of the correlations on the two DNA-binding domains of NikR, which is significantly suppressed as compared to the apo form. Overall, the findings suggest that the binding of *Hp*NikR to DNA occurs through a sophisticated multistep process involving both a conformational selection and an induced fit.

Introduction

Nickel is an essential cofactor for a number of microorganisms.^{1,2} The toxicity of Ni²⁺ imposes control on its homeo-

stasis and cellular trafficking through the regulation of genes expressing proteins involved in nickel metabolism.³ Several human bacterial pathogens rely upon the expression of nickel-dependent enzymes such as urease and hydrogenase to survive in the host organisms,¹ suggesting that a comprehension of the molecular determinants of the regulations of these genes might help in developing efficient drugs. In this biological framework, a key regulator is NikR, a highly homologous protein found across ca. 30 species of bacteria and archaea. Several crystal structures of NikR have consistently established that this protein is a homotetramer made of a dimer of dimers, constituted by a central metal-binding domain (MBD) and two peripheral DNA-binding domains (DBD) (Figure 1, Table 1SI, Supporting Information).^{4–8}

* Corresponding author phone: (+39)-051-2096204; fax: (+39)-051-2096203; e-mail: stefano.ciurli@unibo.it.

[†] University of Bologna.

[‡] International School for Advanced Studies (SISSA) and CNR-IOM-DEMOCRITOS National Simulation Center.

[§] Ruder Bošković Institute.

^{||} German Research School for Simulation Science.

[⊥] Center for Magnetic Resonance (CERM), University of Florence.

[∇] Department of Chemistry, University of Florence.

[∇] Francesco Musiani and Branimir Bertoša contributed equally to the simulations presented here.

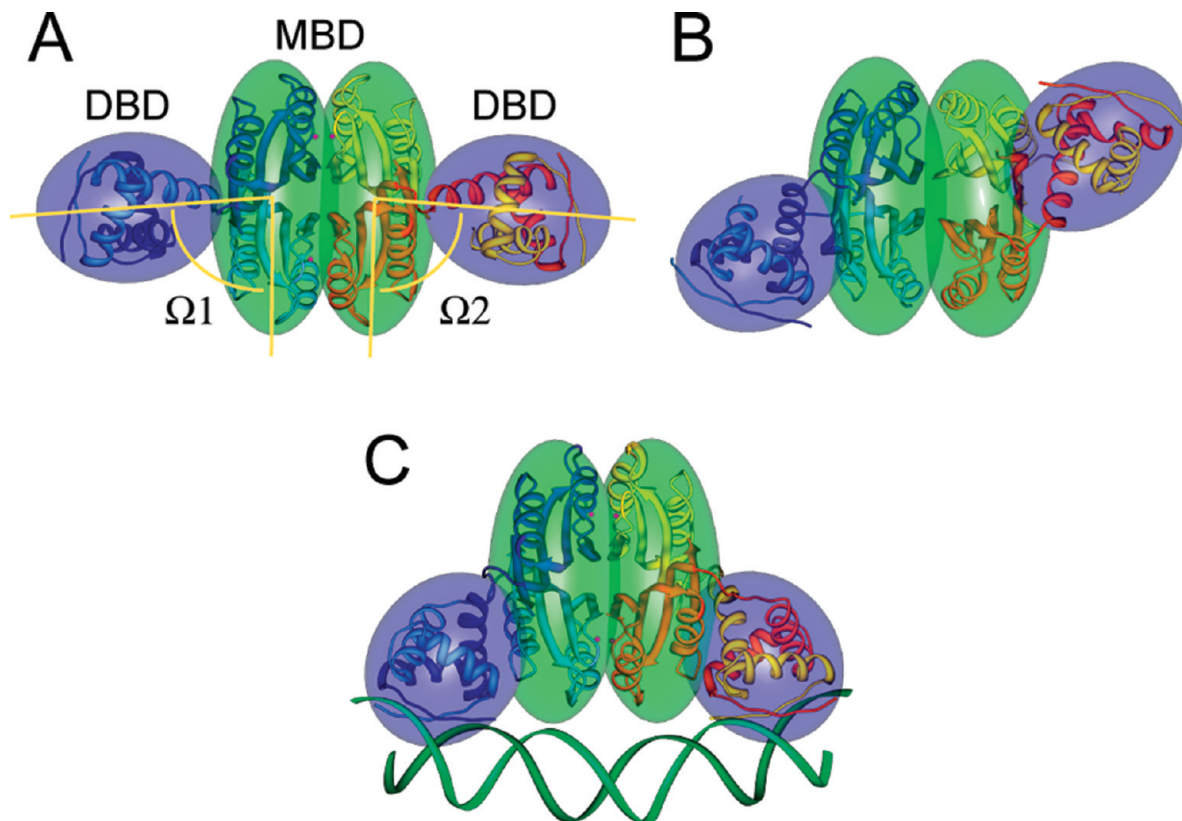


Figure 1. Ribbon diagrams and inertia ellipsoids of NikR in (A) open, (B) trans, and (C) cis conformation (PBD codes 2HZA, 2HZV, 2CA9). In C, the DNA fragment is bound to the protein. Ribbons are colored according to monomer chains. Ellipsoids for DBD and MBD are colored in blue and green, respectively.

The DBD features a ribbon–helix–helix motif typical of prokaryotic transcription factors.⁹ The MBD hosts four metal-binding sites located at the tetramerization interface, where square planar Ni^{2+} ions bind three fully conserved His and one Cys residues.

Three major conformations of NikR, open, trans, and cis, have been observed in the solid-state structures (Figure 1, Table 1SI, Supporting Information). Such conformations are characterized by the position of the peripheral DBDs with respect to the central MBD, separated by unstructured linkers, and are here defined in terms of the two Ω angles between the principal axes of inertial ellipsoids approximating such domains: as a result from the crystal structures, in the open conformation (Figure 1A) Ω_1 and Ω_2 are in the $90 \pm 20^\circ$ range, for the trans conformation (Figure 1B) $\Omega_1 = 45 \pm 20^\circ$ and $\Omega_2 = 135 \pm 20^\circ$, and for the cis conformation (Figure 1C) Ω_1 and $\Omega_2 = 35 \pm 1^\circ$.

Several key studies have identified which protein form binds to DNA in vitro. First, the crystal structure of the *Escherichia coli* (*Ec*) NikR–DNA complex has shown that the protein must be in its cis conformation (Figure 1C) to interact with DNA.⁶ This conformation, indeed, is complementary to that of the double-stranded nucleic acid and allows for optimal interactions between the peripheral DBDs and DNA.⁶ Second, electrophoretic mobility shift assays, fluorescence anisotropy, DNaseI footprinting, and calorimetric DNA titrations established that only the Ni^{2+} -bound form of NikR specifically binds target DNA sequences, thus providing a direct Ni^{2+} -dependent metabolic response.^{10–13} Third, calorimetric metal-binding titrations on NikR from

the human pathogen *Helicobacter pylori* (*Hp*NikR) revealed two steps for Ni^{2+} binding, each involving a pair of metal ions.¹⁴ This indicates the existence of three metal-bound forms of the protein characterized by the presence of 0, 2, or 4 Ni^{2+} ions. The 4Ni-*Hp*NikR form binds to DNA, while apo-*Hp*NikR does not.¹³ The DNA-binding capability of *Hp*NikR bound to 2 Ni^{2+} ions has so far not been determined due to the too small difference (ca. 1 order of magnitude) in the dissociation constants for the two steps of Ni^{2+} binding.

Overall, these observations indicate that the Ni-bound form of *Hp*NikR binds to DNA in the cis conformation. The key question is therefore what is the role of Ni^{2+} content for modulating the protein affinity for DNA by favoring its reactive cis conformation? This question is far from being answered. The cis conformation has been characterized in the solid state only with 4 bound Ni^{2+} ions in the complex with DNA (Table 1SI, Supporting Information). This contrasts with the open and trans conformations, for which crystal structures without and with 2 or 4 bound Ni^{2+} ions are available (Table 1SI, Supporting Information). It should also be noted that the insights offered from X-ray structures are expectedly limited by crystal packing energetics affecting the conformational properties of DNA-binding proteins like NikR, as suggested by molecular dynamics (MD) simulations.¹⁵

In the absence of solution structural information on the role of Ni^{2+} in the reactivity of NikR toward DNA, insights have been obtained, in recent years, by molecular simulations. A coarse-grained elastic-network model applied to the apo-open, 4Ni-open, and 4Ni-trans conformations of *Pyro*-

coccus horikoshii (*Ph*) NikR (PDB codes 2BJ3, 2BJ1, and 2BJ7, respectively) suggested the presence of intrinsic fluctuations that are encoded in the protein architecture capable of favoring the interconversion among the conformations observed in the solid-state structures of NikR.¹⁶ Subsequently, the same group reported 100 ns implicit solvent MD simulations on the same systems, concluding that the secondary structure segments involved in binding DNA and Ni²⁺ tend to influence motion across domains much more strongly than other residues and that portions with high flexibility are correlated with biologically relevant regions.¹⁷ In another instance, atomistic 80 ns MD simulations in explicit solvent were analyzed using a correlation-matrix-based approach and hinted to the presence of an allosteric communication pathway between residues in the nickel-binding sites and the DNA-binding region.¹⁸ However, the metal-bound state was not investigated, preventing a comparative study. It should also be noted that this latter atomistic simulation was started from a conformation of the apo form of *Ec*NikR that is somewhat atypical in that the Ω angles observed only in a single-crystal structure (PDB code 1Q5V, Table 1SI, Supporting Information) are between the open and trans conformations. A recent paper describing shorter (ca. 3 ns) MD and Poisson–Boltzmann electrostatic calculations on 4Ni-*Ec*NikR in the open and cis conformations (PDB codes 2HZA and 2HZV) appeared to indicate that the protein–DNA complex is stabilized when 4 Ni²⁺ ions are bound to the MBD.¹⁹

In order to contribute to the understanding of the role of Ni²⁺ on the reactivity of NikR toward DNA binding, we carried out a comparative analysis of atomistic MD simulations (overall 540 ns) of three structure-based homology models of the protein from *H. pylori* in the open, trans, and cis conformations in explicit aqueous solution at finite ionic strength as a function of the absence or presence of 2 or 4 Ni²⁺ ions. The MD simulations were complemented by a state-of-the-art coarse-grained analysis and modeling of the molecule internal dynamics. ¹H–¹⁵N TROSY-HSQC and ¹³C–¹³C NOESY NMR experiments on apo- and 4Ni-*Hp*NikR were also carried out in order to investigate experimentally the structural properties of *Hp*NikR in solution and to link them to the theoretical analysis.

The overall results of this multifaceted approach indicate that NikR is present in solution as an ensemble of interconverting structures spanning the entire conformational space from cis to open to trans forms. The interconversion capability appears to depend on the content of bound Ni²⁺ ions, which in turn affects the motion of the DBDs. The conformational fluctuations are, in addition, finely tuned by the Ni²⁺ content. The NMR solution studies indicate that Ni²⁺ ion binding does not induce, by itself, the stabilization of the cis conformation competent for DNA binding. Overall, these results support the view that the likely mechanism of interaction of the protein with its operator DNA sequence

involves a selection of the correct conformation coupled with an induced fit mechanism facilitated by the presence of bound Ni²⁺.

Materials and Methods

Structural Models. Initial structural models of *Hp*NikR were obtained using homology modeling based on the available X-ray structures of NikR. The alignment of *Hp*NikR with *Ph*NikR and *Ec*NikR was produced using ClustalW²⁰ and subsequently manually adjusted in order to match up the primary and secondary structure of the proteins. The calculated sequence identity (similarity) between *Hp*NikR and *Ec*NikR is 29% (55%), between *Hp*NikR and *Ph*NikR is 34% (62%), and between *Ph*NikR and *Ec*NikR is 33% (63%). The final alignment (Figure 1SI, Supporting Information) was used to calculate 50 structural models of the open, trans, or cis conformations of tetrameric *Hp*NikR using the program MODELER 9v5.²¹ The structural templates used in each of the conformations were (i) the NikR structures from *P. horikoshii* (PDB code 1BJ1, resolution 3.00 Å) and *E. coli* (PDB code 2HZA, resolution 2.10 Å) for the open conformation, (ii) the DNA-bound *Ec*NikR structure (PDB code 2HZV, resolution 3.00 Å) for the cis conformation, and (iii) the structures of *Hp*NikR (PDB codes 2CA9 and 2CAD, resolution 2.05 and 2.30 Å, respectively) and *Ph*NikR (PDB 2BJ8, resolution 2.10 Å) for the trans conformation.

The models were selected on the basis of the lowest value of the DOPE score in MODELER.²² The structural identity of monomers pairs, depending on the conformation, was imposed. The stereochemical quality of the structures was evaluated using PROCHECK,²³ and the distribution of residual energy was evaluated in ProSA.²⁴ The results of this analysis, reported in Table 2SI, Supporting Information, indicate that the models are highly reliable. Polar hydrogen atoms were added using WHATIF²⁵ by optimizing the hydrogen-bonding networks. During this procedure, few amino acid side chains were allowed to change their orientation. Nonpolar hydrogen atoms were added with TLEAP (AMBER 10).²⁶ Four square planar Ni²⁺ ions were initially included in the model, bound in the known metal binding sites. These metal ions were partially or totally removed from the model depending on whether the apo form, the 2Ni-*Hp*NikR, or the 4Ni-*Hp*NikR was considered. In the case of the 2Ni-bound form, the selection of the two loaded binding sites was made according to the crystal structure of the partially metal-bound form of *Hp*NikR (PDB code 2CAD) independent of the conformation considered. The two nickel ions are located on subunits, giving rise to distinct DBDs, and in opposite positions within the tetrameric cluster.

The models were placed in the center of a water box (113–124, 77–84, and 75–80 Å) using a 10 Å buffer zone of solvent around the protein. The systems were neutralized by adding Na⁺ and Cl[−] ions using TLEAP (AMBER10).²⁶ The counterions were placed at the points of lowest or highest electrostatic potential, according to a Coulombic potential calculated on a 1.0 Å grid. Analogously, additional Na⁺ and Cl[−] ions were placed in the water box to achieve the ionic

strength used in the calorimetric binding experiments (150 mM).¹⁴ The number of additional salt ions (61–68 Na⁺ and Cl⁻ ions) was calculated according to the number and density of the solvent molecules in each water box. The possible formation of NaCl aggregates during the simulation²⁷ was excluded by calculation of the radial distributions of the ions using the PTRAJ module of AMBER 10.²⁶ The resulting systems consisted of ca. 70 000 atoms.

Force Field. The force field charges for the metal binding sites were obtained by applying DFT calculations on a model consisting of a Ni²⁺ ion coordinated by a methyl-thiolate and three imidazole rings, mimicking cysteine and histidine side chains, respectively. The initial coordinates for this model were taken from the *Ec*NikR structure in the cis conformation bound to its operator DNA (PDB code 2HZV).⁶ The geometry of the initial model was optimized at the B3LYP/6-31G(d) level using tight convergence criteria in the GAUSSIAN software.²⁸ The optimized structure of this model was subjected to single-point calculations in order to derive ESP charges according to the Merz–Kollman scheme.²⁹ RESP charges were obtained with the RESP module of the AMBER 10 software package.²⁶ Force field values for distances, angles, dihedral angles, and partial charges were obtained from the optimized structure (Tables 3–5SI, Supporting Information). Appropriate force field constants were found in the literature.^{30,31} The Amber ff99SB³² and TIP3P³³ force fields were used for the protein frame and water, respectively, while known parameters were used for Na⁺³⁴ and Cl⁻.³⁵

Molecular Dynamics. The time step was 1 fs, and the structures were sampled every picosecond. Periodic boundary conditions (PBC) were applied. Particle mesh Ewald (PME) was used to calculate electrostatic interactions. The cutoff values for the real part of the electrostatic interactions and for the van der Waals interactions were set to 10 Å.

Each system was geometry optimized in five cycles of 5000 steps of conjugate gradient. In the first cycle, water molecules were relaxed while the protein was constrained using a harmonic potential with a force constant of 50 kcal mol⁻¹ Å⁻². In the second cycle, the same constraint was applied to all non-hydrogen atoms. In the third and fourth cycle, only the position of the Ni²⁺ coordination polyhedron was constrained using a harmonic potential with a force constant of 50 and 25 kcal mol⁻¹ Å⁻², respectively. Finally, in the fifth cycle no constraints were applied.

The systems were subjected to 40 or 100 ns of molecular dynamics (MD) simulations. During the first 240 ps, the temperature was raised from 0 to 300 K at a regular rate of 1.25 K ps⁻¹ and keeping the pressure constant (1 atm) using a Berendsen barostat.³⁶ During this time, positional constraints were applied on the protein atoms and Ni²⁺ ions (force constant of 32 kcal mol⁻¹ Å⁻²). Subsequently, 160 ps of MD simulation was run at 300 K using a Berendsen thermostat³⁶ using the same positional constraints. Then, 200 ps of MD simulation was run at 300 K applying the same harmonic potential as above to the protein backbone and Ni²⁺ ions. Finally, in the following 200 ps of MD simulation, only the Ni²⁺ ions were constrained, using the same harmonic

potential. In the subsequent simulations, no constraints were applied and the Berendsen thermostat was used.³⁶

These calculations were performed using NAMD 2.7b1³⁷ running on the IBM Blue Gene/P JUGENE supercomputer (Jülich Supercomputing Centre, JSC, Jülich, Germany) for a total of more than 1 300 000 h of calculation.

Calculated Properties. The Cα root-mean-square deviation (rmsd) was calculated using the minimized structure as reference. On the basis of rmsd vs time plots (see Results section), we calculated averaged properties for the cis and trans conformations after the first 15 ns of simulation. The average number of hydrogen bonds was calculated using the Chimera program³⁸ with its standard parameters. Representative conformations during the dynamics were identified by the clustering analysis of the PTRAJ module of AMBER 10.²⁶ The same module was used to calculate the Na–O, Cl–O, and Na–Cl radial distribution functions, allowing us to exclude formation of NaCl aggregates during the simulation.²⁷ The covariance matrices of pair correlations of the displacements of the Cα atoms were calculated as

$$C_{ij,\mu\nu} = \langle (r_{i,\mu} - \langle r_{i,\mu} \rangle) (r_{j,\nu} - \langle r_{j,\nu} \rangle) \rangle$$

where $r_{i,\mu}$ is the μ th Cartesian coordinate of the Cα atom of the i th amino acid and $\langle \rangle$ represents the MD average. The calculation was preceded by removal of the rigid-body motions (translations and rotations) accomplished by optimally superposing the MBDs onto the initial reference structure. This procedure, as opposed to aligning the whole molecular complex, removes possible artifactual covariance signals resulting from compensating the displacements of the mobile DBDs with reorientations of the whole NikR molecule. From the covariance matrix we calculated (i) the principal components after matrix diagonalization, (ii) the normalized correlation matrix³⁹

$$C_{ij} = \frac{\sum_{\mu} C_{ij,\mu\mu}}{\sqrt{\sum_{\mu} C_{ii,\mu\mu} \sum_{\nu} C_{jj,\nu\nu}}}$$

(iii) the cREL parameter,¹⁷ which provides an estimate of the relevance of specific residues in inter-residue dynamical relationships

$$\text{cREL}_i = \sum_{\text{chains}} \sum_{j=1}^N |C_{ij}|$$

and (iv) the cFLX parameter,¹⁷ which estimates flexibility as a function of the standard deviation of the distance between consecutive Cα atoms summed over chains

$$\text{cFLX}_i = \sum_{\text{chains}} \sigma(|x_{i-1} - x_{i+1}|)$$

The protein subdomains were identified by submitting the top 10 principal components to the PiSQRD server (available at <http://pisqrd.escience-lab.org/>).⁴⁰

Coarse-Grained Modeling. The *Hp*NikR internal dynamics were investigated using the β -Gaussian elastic network model,⁴¹ as implemented in an in-house computer program

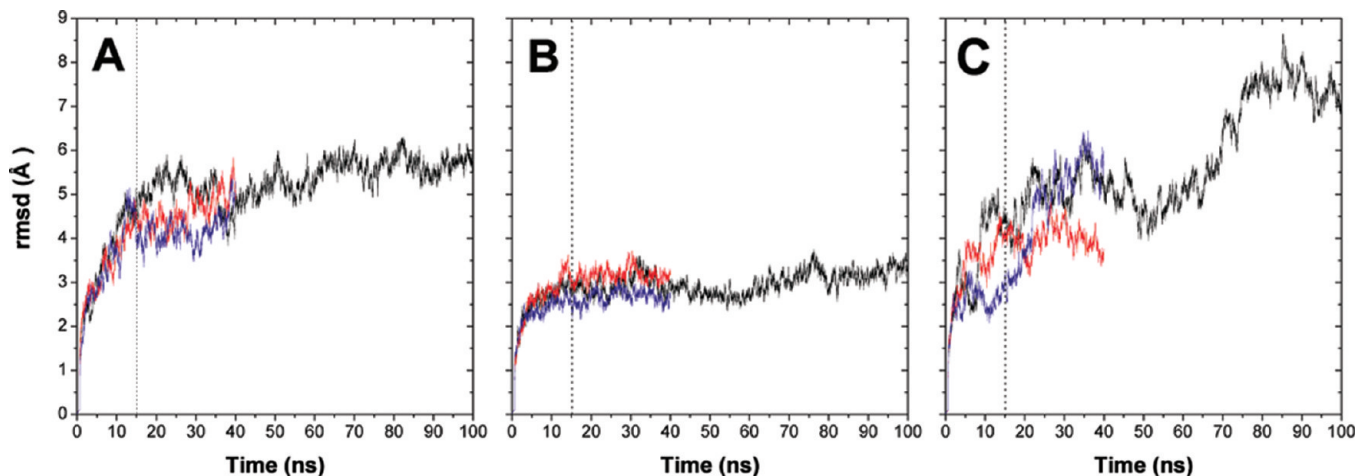


Figure 2. rmsd vs time plots for C α atoms of *HpNikR* in (A) cis, (B) trans, and (C) open conformations. The metalation states of the protein are reported using the black line (apo), red line (2 Ni²⁺ ions), and blue line (4 Ni²⁺ ions). The dotted vertical line in each panel indicates the 15 ns time limit.

freely available for academic use (requests should be directed to C.M.). In the model used, all amino acids are represented by one centroid for the main chain and another for the side chain, except for glycines for which only the former is used. In the case of ENMs based on single-centroid representations, it is necessary to set the range of the harmonic interaction between pairs of interacting centroids to ca. 13–15 Å in order to avoid the appearance of zero-energy modes other than the global translations and rotations of the molecule. The presence of the side chain centroid, while not impacting on the computational complexity of the model, allows us to limit the range of the contact interactions to the more realistic values of ca. 7.5 Å without generating spurious zero-energy modes. In the present context, each bound Ni²⁺ ion was treated as a single effective centroid in the network (as if it were the main chain centroid of a glycine).

NMR Spectroscopy. Triply labeled [98% ²H/¹⁵N/¹³C] apo-*HpNikR* was prepared using *E. coli* BL21(DE3) cells, harboring the pET15b-nikR construct.¹⁴ The cells were grown at 37 °C in Silantes-OD2 medium (Spectra 2000), enriched with 98% of ²H, ¹³C, and ¹⁵N. The expression was induced using IPTG (isopropyl β -thiogalactopyranoside) and carried out at 28 °C for 16 h after induction. Protein purification was carried out as previously reported,¹⁴ yielding 35 mg/L of pure labeled protein. The NMR samples contained 0.5 mM of tetrameric *HpNikR*, diluted in 20 mM HEPES, 150 mM NaCl, at pH 7.0. Solutions of 4Ni-*HpNikR* were prepared by adding a 100 mM solution of NiSO₄ in the same buffer to a final concentration of 2.0 mM.

2D ¹H–¹⁵N TROSY-HSQC experiments⁴² for the apo- and 4Ni-bound protein were carried out at 298 and 315 K on a 21.1-T Bruker AVANCE 900 spectrometer equipped with a TCI cryoprobe. Experiments were acquired with a relaxation delay of 1.2 s and an acquisition time of 87 ms. Typical acquisitions were done with 4 scans for each FID for a matrix of 1024 \times 256 data points.

¹³C–¹³C NOESY maps⁴³ for the apo- and 4Ni-bound protein were acquired on a 16.4 T Bruker AVANCE 700 spectrometer equipped with a triple-resonance observe cryoprobe optimized for ¹³C direct detection at 298 and 315 K. Experiments were acquired with a relaxation delay of 2 s



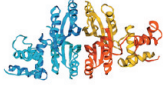
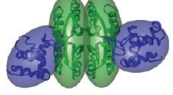

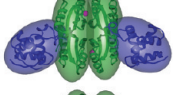
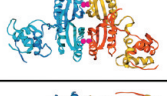
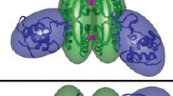
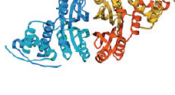

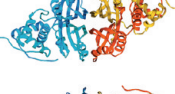
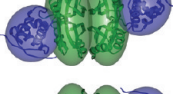

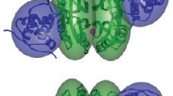
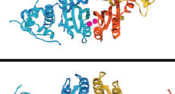
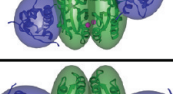



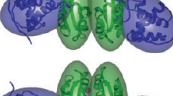

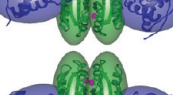
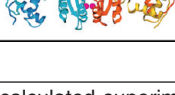

and an acquisition time of 29 ms. A mixing time of 1.5 ms was used to allow spin diffusion and detection of side chain complete spin patterns. Composite pulse decoupling on ¹H and ²H was applied during the whole duration of the experiments. ¹³C–¹³C NOESY maps were recorded on the full spectral width (200 ppm) with 64 scans per increment and with 2048 \times 1024 data points.

Results

I. Molecular Dynamics. Hereafter we report on the results of atomistic MD simulations of the three conformers of *HpNikR* in each of the three different metal-bound states, carried out in explicit aqueous solution at 150 mM NaCl ionic strength. The calculations involved a well-established approach^{44–49} that entailed molecular dynamics simulations starting from structures obtained by homology modeling carried out using known structures of NikR from *H. pylori*, *E. coli*, and *P. horikoshii*. This allowed us to prepare a set of consistently built and equally validated structures representing the three conformations present in solutions as sampled by X-ray crystallography. In all cases investigated, both the overall protein fold and the structure of each domain were largely maintained. The rmsds of the cis and trans conformations appear to be converged (Figure 2A and 2B) as opposed to the open conformation, for which convergence is not attained for any of the metal-bound forms (Figure 2C).

Simulation of the *cis*-apo form (Figure 2A) reveals a structural evolution characterized by a considerable widening of one interdomain angle, while the other angle does not change significantly (Table 1). The conformation thus attained peculiarly resembles that of the atypical X-ray structure of apo-*EcNikR* (PDB code 1Q5V). The *cis*-2Ni form tends to evolve toward the open conformation found in the X-ray structures, with a widening of the Ω angles reaching \sim 50°. In the presence of 4 Ni²⁺ ions, the *cis* form evolves toward an asymmetric structure characterized by a ca. 35° increase of one interdomain angle to values typical of the open form (see Table 1SI, Supporting Information) while the other angle decreases by ca. 10°. These data indicate that the *cis* conformer is not dynamically stable independent of the presence of Ni²⁺ ions. In turn, this

Table 1. HpNikR Model Structures and Most Representative Structures During the MD Simulations

Model	Persistence of the most populated cluster (ns)	Average backbone rmsd (Å) (a)	Average Ω_1/Ω_2 angles (°) (b)	Ribbon diagram (c)	Inertia ellipsoid (d)
cis starting model	-	-	34 / 34		
cis - 0 Ni ²⁺	39/85	4.4 (2.1)	41 ± 3 / 78 ± 10		
cis - 2 Ni ²⁺	25/25	3.7 (1.7)	50 ± 4 / 52 ± 3		
cis - 4 Ni ²⁺	18/25	3.7 (1.8)	68 ± 5 / 26 ± 3		
trans starting model	-	-	45 / 135		
trans - 0 Ni ²⁺	55/85	2.9 (1.6)	52 ± 4 / 131 ± 4		
trans - 2 Ni ²⁺	25/25	2.6 (1.5)	45 ± 4 / 143 ± 4		
trans - 4 Ni ²⁺	25/25	2.2 (1.7)	36 ± 5 / 156 ± 5		
open starting model	-	-	88 / 88		
open - 0 Ni ²⁺	10/85	5.5 (2.1)	75 ± 4 / 48 ± 3		
open - 2 Ni ²⁺	7/25	3.5 (1.7)	85 ± 4 / 77 ± 4		
open - 4 Ni ²⁺	10/25	4.5 (3.2)	64 ± 4 / 71 ± 3		

^a rmsd calculated for the most representative cluster of structures. The rmsd calculated superimposing separately the DBD and the MBD domains in order to eliminate the rmsd contribution due to domain movement is reported in parentheses. ^b Ω angles defined as in Figure 1; the average is calculated considering only the most representative cluster of structures \pm one standard deviation. ^c The structure reported is representative of the most populated cluster of structures. ^d The inertia ellipsoids were calculated using the UCSF Chimera software.

suggests that the presence of DNA is necessary to stabilize the correct conformation of NikR for the interaction with its operator.

Simulations of the trans conformation in the presence of 0, 2, or 4 Ni²⁺ ions indicate an overall change in the rmsd (Figure 2B) that is about one-half of that observed for the cis case (~ 3 vs ~ 6 Å) with the Ω angles remaining significantly closer to the initial values independent of the

Ni²⁺ content (Table 1). This observation indicates that this conformer is relatively more resilient to change the inter-domain orientation as compared to the cis form. The large movements observed for the open conformation, overall spanning a range of 4–8 Å (Figure 2C) and leading to the lack of convergence of the rmsd independent of the Ni²⁺ content, suggests that in this case the protein is able to explore a shallower free energy landscape. At present, a much

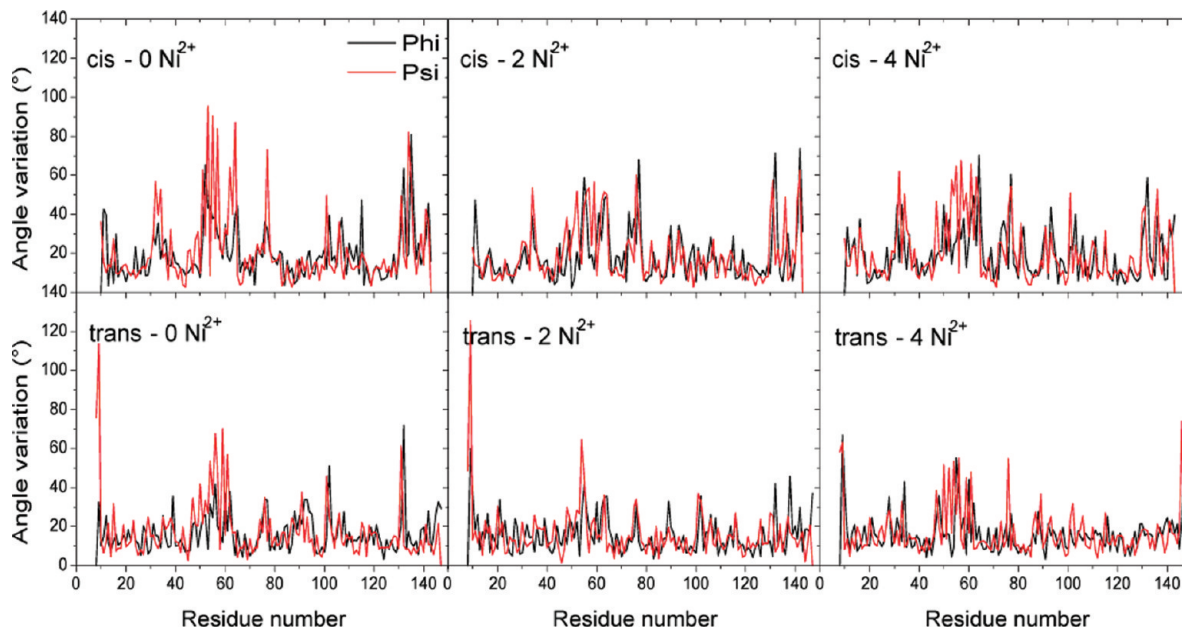


Figure 3. Backbone torsional angle analysis for *HpNikR* in cis (top panels) and trans (bottom panels) conformations in different metalation states: apo (left panels), 2Ni^{2+} ions (central panels), and 4Ni^{2+} ions (right panel). The plots report the absolute value of the difference between the angles in the structure representative of the most populated cluster and those in the minimized starting structure. Φ angles are reported with black lines, while Ψ angles are reported with red lines.

longer time scale simulation of such large systems appears not to be feasible even in the largest supercomputers available, such as the one used for this work. The lack of convergence for the open conformations induced us not to carry out a detailed analysis of the structural parameters for this case.

An analysis of the average number of H bonds in the various metal-bound states of all structures belonging to the most representative cluster revealed that the trans conformation is characterized by ca. 10% more H bonds as compared to the cis conformation, independent of the Ni^{2+} content. This is contrasted by the ca. 15% decrease of the number of H bonds within four shells around the metal-binding sites observed in the case of the trans vs cis conformer. These observations support the view that, in the case of the trans form, the nickel binding relaxes the structure in the MBD while enhancing, on the other hand, the rigidity of the overall protein architecture.

To pinpoint the protein regions involved in the observed structural changes during the MD simulations of the cis and trans conformers, we carried out an analysis of the backbone dihedral angles Φ and Ψ as a function of residue number (Figure 3). Two regions were detected as undergoing variations consistently among all cis and trans conformers: residues 50–60 (the linker between the MBD and the DBD) and residues 30–35 (a loop connecting the first and second helix in the DBD domain). This reveals the presence of two subdomains in the DBDs. No significant differences of the variations of dihedral angles were observed either upon changing the Ni^{2+} content or between the two conformations.

The correlation and flexibility parameters cREL and cFLX can be used to estimate, respectively, the structural stability and instability of proteins regions.¹⁷ Plots of cREL as a function of residue number for the cis and trans conformations (Figure 4A and 4B, respectively) allow us to propose

that the presence of 2 or 4 Ni^{2+} ions causes an increase of disorder of the MBD region in contact with the DBD (residues ca. 110–140). Moreover, addition of Ni^{2+} ions in the trans conformation also increases disorder in the protein region starting from the linker between the MDB and DBD and extending to the metal-binding portion (residues 55–110). Plots of cFLX for the same conformers (Figure 4C and 4D) show the presence of a very flexible region corresponding to the linker between the DBD and the MBD as well as a flexible portion that divides the DBD into two subdomains in correspondence of the loop between the first and the second α -helix, consistent with the analysis of the dihedral angles. Other minor regions of high flexibility correspond to loops between elements of secondary structure throughout the protein.

Motion correlations between various subparts of the protein can be identified by a calculation of the covariance matrices of the amino acids displacements. Visual inspection of the corresponding maps (Figure 5) immediately conveys the remarkable fact that in the cis conformation the motion of the two DBDs is anticorrelated, meaning that they move along opposite directions, with both Ω angles tending to concomitantly increase or decrease. In contrast, in the trans conformation the two DBDs have a positive motion correlation. The results, supported by inspection of the covariance principal components, indicate that the internal fluctuations of the cis conformation favor a scissor-like movement of the DBDs while the trans conformation tends to maintain a collinear geometry of the DBDs. Hence, the internal dynamics of both the cis and the trans forms can facilitate their conversion to the open conformer. For the considered conformations, the simulations with bound Ni^{2+} ions display an appreciable decrease of the absolute magnitude of the DBDs correlation with respect to the simulations of the apo forms. (We recall that the apo simulations have a longer

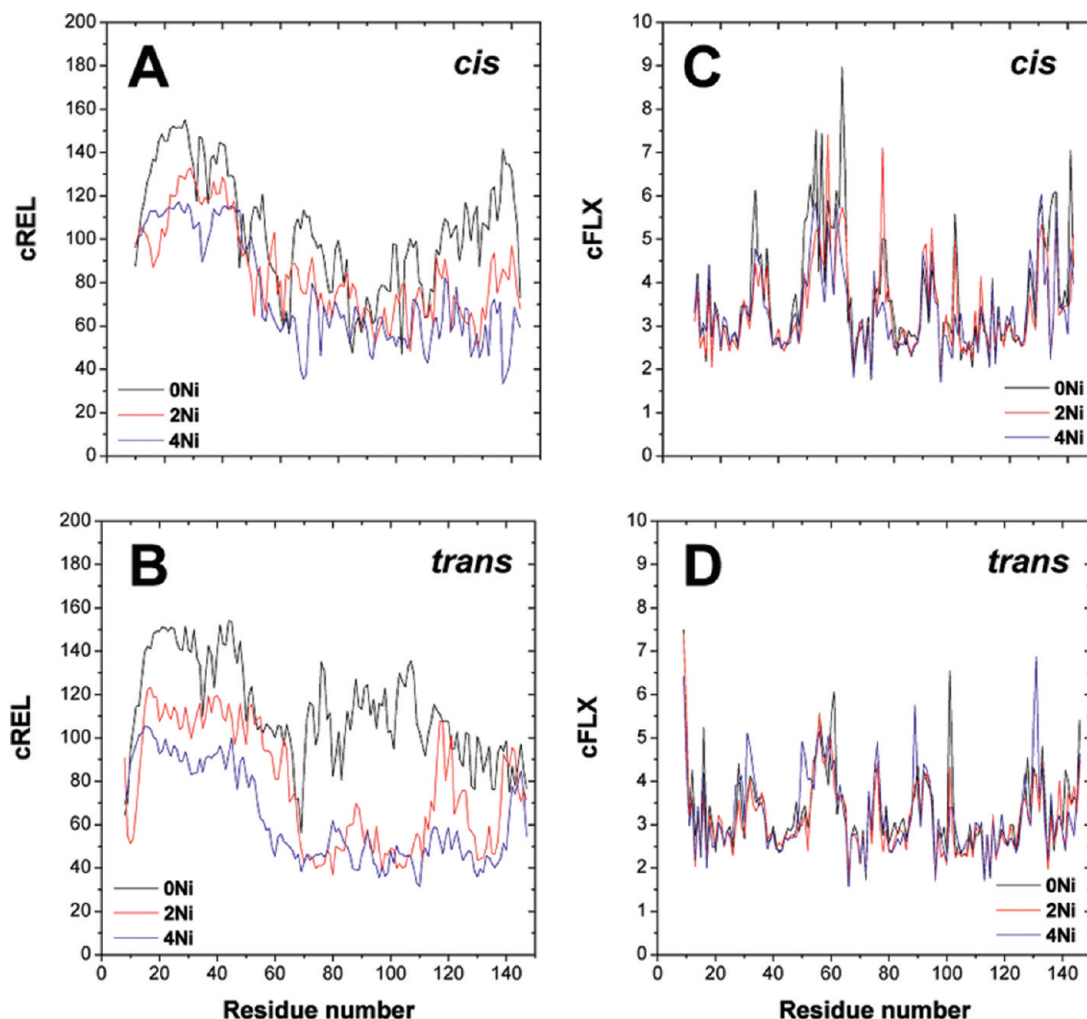


Figure 4. Correlation relevance summed over chains (cREL, A and B) and residue flexibility (cFLX, C and D) for the cis and trans simulations. The metalation states of the protein are reported using black lines (apo), red lines (2 Ni²⁺ ions), and blue lines (4 Ni²⁺ ions).

duration (100 ns) compared to those with the bound metal ions (40 ns), yet the described DBD correlations are maintained independent of the different simulation durations. This was ascertained by verifying the consistency of the correlation matrix shown in Figure 5 with that restricted to the 40 ns simulation time.) These results indicate that the degree of correlation of the DBDs motion depends on the content of Ni²⁺ ions. This fact suggests that the presence of nickel affects the interconversion capability across the various NikR conformers, causing a loss of correlated motion of the DBD domains. Interestingly, for both the trans and the cis trajectories the loss of correlation is not accompanied by an appreciable decrease of the overall mobility of the domains. In fact, the total mean square fluctuation of the molecule, calculated over the 15–40 ns interval, remains approximately equal to 1100 Å² for the cis simulations and about 660 Å² for the trans case irrespective of the number of bound Ni²⁺ ions.

Further insights into the large-scale internal motions of *Hp*NikR can be obtained by a suitable analysis of the most representative conformations for the various trajectories. In particular, the principal components of the covariance matrix computed over the nine MD most representative structures were processed by the PiSQRD online tool⁴⁰ with the aim

of identifying the protein quasi-rigid dynamical subdomains. The number of optimal quasi-rigid domains depends on the amount of overall structural fluctuations of the system (the mean square deviation of the nine MD most representative structures) that one wishes to capture in terms of the relative motion of quasi-rigid subparts.⁵⁰ By considering simultaneously the ensemble of all such conformers visited by all the trajectories it is possible to identify the main dynamical subdomains whose relative, rigid-like motion is sufficient to reproduce the breadth of the heterogeneous conformational space visited by all the trajectories, including the configurational space spanned by the slowly converging open simulations.

Nearly 50% of the mean square deviation (MSD) across the structure representatives of the nine trajectories is ascribable to the relative movements of the two DBDs with respect to the central MDB. The subdivision indicates that the linkers between the MDB and the DBDs act as retractable hinges for the DBD motion. In fact, on the one hand, the two DBDs are capable of rotating around the ideal hinge axis connecting the two linkers and, on the other hand, the separation of the DBD from the MDB depends on the degree of stretching of the linkers themselves. Throughout the various trajectories, the central core maintains its structure, with only minor

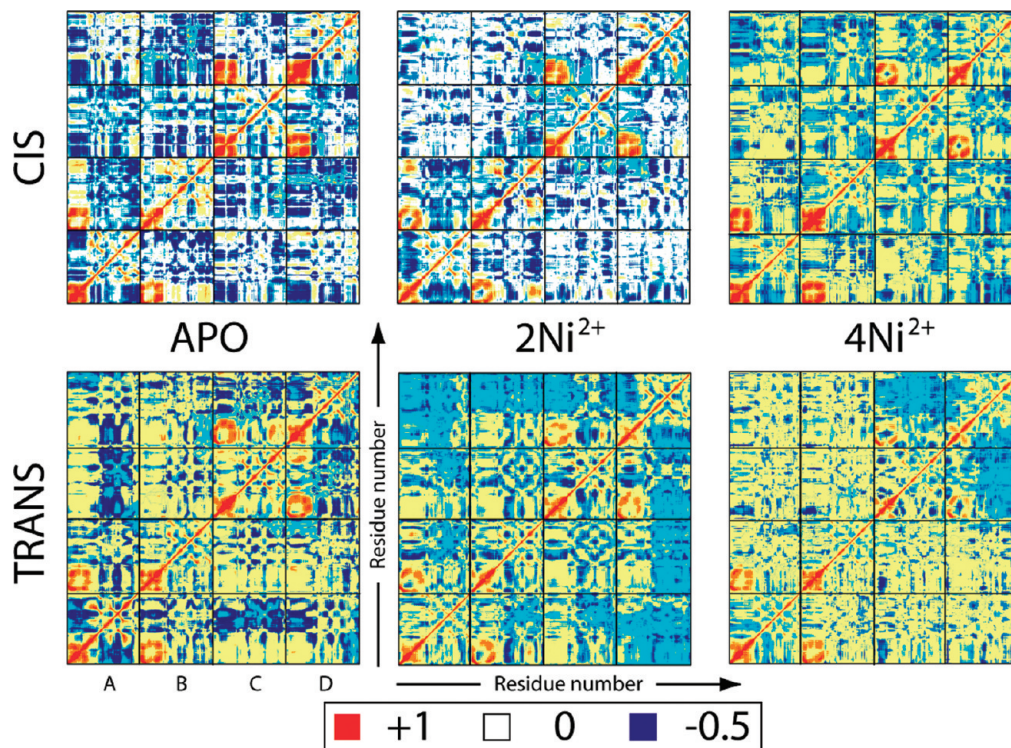


Figure 5. Residue–residue-based map of α correlation matrices of *HpNikR* in cis (top panels) and trans (bottom panels) conformations. Red and orange regions show positive correlation, while dark blue regions indicate anticorrelation. The black lines indicate the border between one chain and the subsequent chain.

internal deformations. Indeed, upon increasing to five the number of quasi-rigid domains, so to capture an even larger fraction of the ensemble MSD, it is found that the central core is still recognized as a single rigid unit while each of the DBD is split in two subunits. A minor but still sizable fraction of the *HpNikR* conformational variability resides in the internal mobility of each of the two DBD, as signaled by the fact that the subdivision in five units captures ca. 70% of the ensemble MSD. Consistent with the analysis of dihedral angles and the cFLX parameter, the internal dynamical boundary of the DBDs separates the part closest to the MBD from the most peripheral protein portion constituted by amino acids 8–42 on one chain and 8–21 on the other chain.

The results from this coarse-grained perspective indicate that the structural variability of *HpNikR* observed across the nine trajectories is accountable by the relative roto-translational motion of very few quasi-rigid subparts that are connected by motion hinges, fully consistent with the indications of the dihedral angles analysis.

II. Coarse-Grained Modeling. In light of previous studies^{17,18} it appears most interesting to ascertain whether the large-scale internal motions of *HpNikR* are encoded in the overall structural organization of the protein or if they are influenced by fine chemical aspects, such as the presence of bound Ni^{2+} ions. We tackled these questions using the β -Gaussian elastic network model (ENM).⁴¹ Because of its minimalistic character, the model highlights the internal dynamic properties that are robustly encoded in the overall structural organization of a protein. The specific model employed here differs from that previously used for *NikR*¹⁸ by the fact that the amino acids are represented by more

than one interaction center (or centroid, see Methods for details). This has been shown to improve the consistency of the model predictions (covariance and correlation matrices, essential dynamical spaces) with results from extensive atomistic MD simulations.^{41,51–53}

The model was accordingly used to compute the 10 lowest energy modes of fluctuations of the main structural representatives for each of the six cis and trans trajectories; because of the lack of convergence of the open MD trajectories, the three initial energy-minimized open structures were used. In a similar spirit to previous work,¹⁸ we ascertained whether the low-energy modes can assist the open, cis, and trans interconversions at a fixed number of bound Ni ions by projecting the structure difference vector of various pairs of representatives (aligned over the common set of amino acids) on the space of the 10 lowest energy modes of either representative. We found that irrespective of the number of bound Ni^{2+} ions, the 10 lowest energy modes of the open structures could account for $48 \pm 9\%$ of the difference vectors with either the cis and trans conformations. Comparable results were found for the cis structures, whose modes captured about $43 \pm 9\%$ of the difference vectors with either the open or trans form. The worst compliance of the lowest energy modes with the difference vectors was observed for the trans form. Indeed, only about $36 \pm 12\%$ of the difference vectors with the open or cis form was projected on the lowest energy modes of the trans structure. This result appears to be highly consistent with the analysis of the MD covariance matrices described above, which indicates a facilitated interconversion from the cis to the open form, while the internal dynamics of the trans

conformer is not particularly prone to interconvert toward the two other conformations.

A notable difference between the results obtained using the elastic-network model as compared to the observations derived from the MD analysis is that, in the first case, no significant differences are observed as a function of the number of bound Ni^{2+} . Overall, the results of ENM calculations confirm previous observations^{17,18} that the structural architecture of NikR is predisposed to sustain low-energy fluctuations that can assist the interconversion between different NikR forms. In addition, the MD covariance analysis reveals that the presence of the bound metal ions can significantly alter the internal dynamics and hence impact on the interconversion dynamics.

III. NMR Spectroscopy. To investigate the structural determinants of the protein at longer time scales than those studied by MD, we performed NMR measurements of the protein in solution at the same ionic strength as that of the simulations and at temperatures (298–315 K) that include that of our MD simulations. The ^1H – ^{15}N TROSY-HSQC spectrum of apo *HpNikR*, which provides information on the structural features of the backbone, shows a number of peaks consistent with the number of amino acids in the protein sequence (Figure 6A, black trace). This observation points to the presence, in solution, of the apo form of the protein either in the single open symmetric conformation or undergoing conformational equilibria with submillisecond interconversion among various conformers. Concomitantly, these results exclude the presence of only the rigid cis or rigid trans conformations in solution, for which two sets of resonances would be expected because of the different structures of the linker regions between the MBD and the DBDs for each monomer. The same situation holds for the ^1H – ^{15}N TROSY-HSQC spectrum of the 4Ni-bound *HpNikR* (Figure 6A, red trace), indicating that the presence of 4 Ni^{2+} ions bound to the protein does not induce a conformational change to a rigid cis form. Over the investigated range of temperatures (298–315 K), signals remain sharp and no signal splitting is observed. While this does not allow us to distinguish between the rigid symmetric or the fluxional behaviors described above using NMR, the presence of the rigid open conformation in solution can be excluded on the basis of our MD and coarse-grained calculations.

^{13}C – ^{13}C NOESY maps of both apo- and 4Ni-bound *HpNikR* provide information on the structural features of the side chains. In these maps, intraresidue cross peaks could be easily detected for most residues due to the spin-diffusion effects operative at the long mixing time used in these experiments (Figure 6B). By taking advantage of the residue-specific chemical shifts of carbon nuclei resonances and using intraresidue NOESY connectivities, we were able to identify the spin patterns of most aliphatic residues and of some aromatic residues. This allowed us to verify that the number of spin patterns observed for a given type of amino acid corresponds to the frequency of that residue type in the protein sequence (this is, for example, the case of isoleucine (a) and alanine (b) residues indicated in Figure 6B). Therefore, no distinct conformations, on the chemical shift

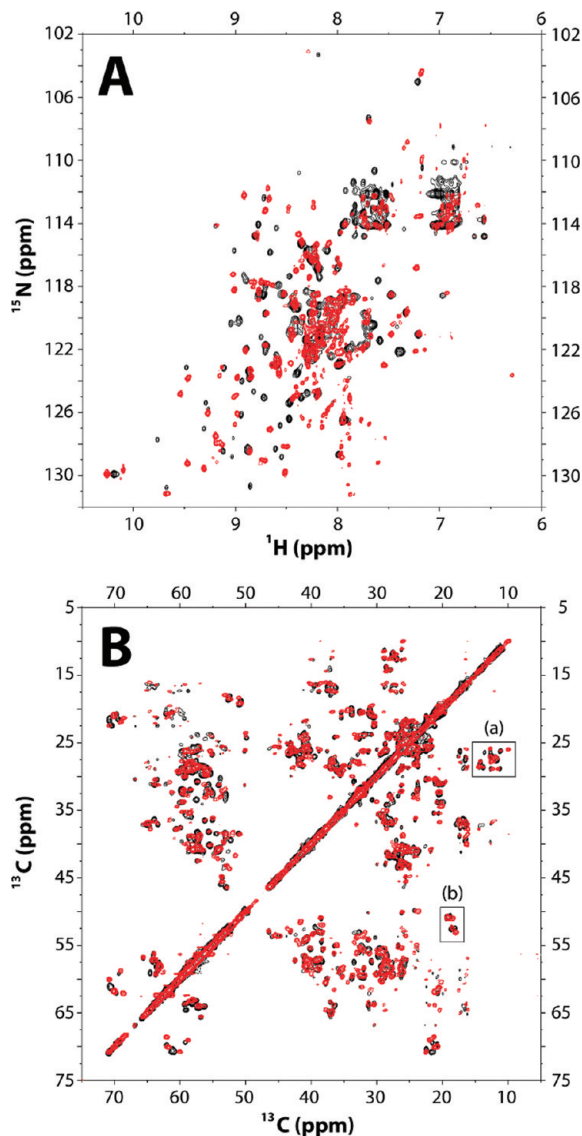


Figure 6. Superimposition of the NMR spectra for [^2H , ^{13}C , ^{15}N] apo-*HpNikR* (black trace) and 4Ni-*HpNikR* (red trace): (A) ^1H – ^{15}N TROSY-HSQC spectra at 315 K; (B) aliphatic region of the ^{13}C – ^{13}C NOESY at 315 K. Insets a and b in the NOESY spectrum correspond to the region of the $\text{C}\gamma_1$ – $\text{C}\delta$ connectivities for Ile residues and to the region of the $\text{C}\alpha$ – $\text{C}\beta$ connectivities for Ala residues, respectively.

NMR time scale, are observed for any side chains of the identified amino acids.

Comparison of the spectra of the apo- and 4Ni-protein shows significant differences in the chemical shift of the backbone amides (in the TROSY-HSQC) and side chain carbon nuclei (in the ^{13}C – ^{13}C NOESY) for the two metalation states. The ^1H – ^{15}N TROSY-HSQC patterns can be taken as a fingerprint for backbone structure of the differently metal-bound protein forms. The existence of a backbone structure signature for each metal-bound state translates into different spectral patterns for the side chains in ^{13}C – ^{13}C NOESY maps. Our NMR data indicate that the presence of Ni^{2+} does affect the position of the conformational equilibria occurring on the submillisecond time scale. On the other hand, MD simulations do not show Ni-dependent changes in the structural interconversion on the multiananosecond

time scale. Taken together, the two independent sets of information provide a time range for the equilibria among different structures.

Discussion

Several issues involving the role of Ni^{2+} ions in the structure–function relationships of *H. pylori* NikR, a nickel-dependent transcription regulator, have been addressed in this study using computational and theoretical tools as well as solution NMR experiments. The techniques were combined to gain insight into the capability of *Hp*NikR to interconvert between the main distinct structural states found in crystal structures, namely, the open, cis, and trans states. In particular, the questions that we addressed concern (i) the structure and conformational fluctuations of the protein in a relatively short time scale (tens of nanoseconds) using atomistic MD simulations in explicit solvent, and at longer time scales, using coarse-grained MD analysis and modeling, as well as NMR experiments, (ii) the role of Ni^{2+} in promoting the cis conformation, known to bind DNA, as opposed to the open or trans conformations, (iii) the mechanism by which the active cis conformation interacts with DNA, investigating whether the NikR–DNA interaction occurs through an induced fit mechanism or a conformation selection in solution or a combination of both.

Our MD simulations allow us to suggest that within the time scale of few tens of nanoseconds, the cis conformation in solution evolves toward the open form found in the solid state.⁶ The trans conformation in solution features a significantly higher resilience to change its morphology. A key region of the protein undergoing local changes upon Ni^{2+} binding is the loop acting as a hinge between two subdomains of each DBD.

In particular, the open conformation did not reach convergence in a relatively long time scale for this large system (up to 0.1 μs for the apo form), pointing to a larger degree of structural mobility than that of the other two forms. This hampered any analysis of the structural parameters. However, this finding does suggest that the observation of a specific conformation for this protein form in the solid state (Table 1SI, Supporting Information) might be the result of crystal packing effects, which may not be maintained in solution. Most likely, also the Ni-bound forms do not converge in a similar time scale, although we were able to explore only up to 0.04 μs for these forms. NMR experiments, performed here in the same conditions as those of the MD simulations, indicate that a much longer time scale, up to a millisecond, might be required for the system to reach equilibration. In fact, the symmetric average structure found by NMR combined with the known fact that *Hp*NikR in solution binds DNA in an asymmetric (cis) conformation indicate not only that conformational states with different symmetry exist in solution but that they also interconvert on the submillisecond time scale.

The tendency of Ni^{2+} to increase the motional disorder of flexible protein regions at the interface between the MBD and the DBD was observed in both the cis and trans conformers by MD simulations. In the absence of metal ions, the motion of the DBDs with respect to the MBD is

anticorrelated in the case of the cis conformer: this indicates that the protein architecture is intimately designed to move in a scissor-like mode, producing a concerted modulation of the distance of the DNA-binding regions and therefore favoring an induced fit recognition mechanism. On the other hand, the interdomain movement is positively correlated in the case of the trans conformation. This is consistent with a possible evolution of this conformer initially toward the open and subsequently to the cis form in order to be activated for DNA binding. This view is supported by the analysis based on elastic networks, which indicated that the internal dynamics of *Hp*NikR conformers in the open and cis forms favor their mutual interconversion. On the other hand, the internal dynamics of trans conformers were found less favorable for their interconversion toward the other forms compatibly with the lower DBDs mobility found in MD simulations for these conformers compared to the case of cis conformers.

In the presence of 2 or 4 Ni^{2+} ions these correlated motions decrease, suggesting that the presence of Ni^{2+} ions unlocks the reciprocal orientations of the DBDs vs the MBD. The dependence of the correlation pattern on the number of bound metal ions was systematically observed across all types of conformations and hence points to a robust mechanism through which the Ni^{2+} ions binding in a rigid part of the molecule can influence the dynamics of the peripheral domains.

As mentioned above, our MD calculations suggest that the presence of Ni^{2+} ions affects the conformational fluctuations. However, it does not significantly change the average structures in the time scales investigated by our MD simulations. The high CPU cost of the latter prevented us from further prolonging the simulations. Then, to address this issue, we performed high-resolution NMR spectroscopy on *Hp*NikR in solution at the same ionic strength as in the simulations. These experiments reveal that the presence of Ni^{2+} does affect the structure of the protein but by itself the binding of these ions is not sufficient to attain the sole cis form in solution. Hence, additional effects, such as the presence of DNA causing a blockage of a conformational fluxional behavior of the protein, must operate in solution to yield the NikR–DNA complex.

The overall picture emerging from our study is consistently interpreted within the most recent and general interpretative frameworks for protein–macromolecule interaction.⁵⁴ NMR studies here indicate that NikR is present in solution either in a single symmetric conformation or as an average ensemble of conformers interconverting on the submillisecond time scale. However, it is the asymmetric cis conformer (suggested to be stable in solution by our simulations) that binds selectively to DNA, as shown by X-ray crystallography.⁶ The open conformation, also suggested to be stable in solution by MD calculations, is symmetric and fully consistent with NMR data. Other conformations, such as the crystallographically established trans conformation, found to be stable in aqueous solution by MD simulations, could also be present. Hence, here we speculate that the cis conformation (and maybe other conformations such as the trans) are in equilibrium with the symmetric open conformation in the

submilliseconds time scale or less. The proposed picture is fully consistent with NMR data. It needs to be further validated against calculations and/or experiments.

The conformational change from the open (and perhaps other conformers such as the trans) to the cis form must occur upon DNA binding. The observed effect of the bound Ni²⁺ ions in unlocking the relative movement of the DBDs vs the MBD could be instrumental to facilitate protein–DNA molecular recognition. Of course, it cannot be ruled out that the molecular-recognition step is aided by a concurrent induced fit step. Further investigations of these aspects, by means of computational techniques such as advanced thermodynamic sampling as well as simulations using the structure of the *PhNikR*, ought to provide further elements to pinpoint the key steps governing the Ni-regulated interaction of *NikR* and DNA.

Acknowledgment. Computational resources were granted by CINECA (CNR-INFM grant) and by the German Research School. B.B. was a recipient of a grant based on the project “New Antitumoral Technologies”; F.M. was a recipient of a postdoctoral fellowship from UniBO and of a fellowship from CERM-CIRMMMP. Work was supported by Italian PRIN2007.

Supporting Information Available: Details of the results of the MD simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Mulrooney, S. B.; Hausinger, R. P. *FEMS Microbiol. Rev.* **2003**, *27*, 239–261.
- Ragsdale, S. W. *J. Biol. Chem.* **2009**, *284*, 18571–18575.
- Li, Y.; Zamble, D. B. *Chem. Rev.* **2009**, *109*, 4617–4643.
- Schreiter, E. R.; Sintchak, M. D.; Guo, Y.; Chivers, P. T.; Sauer, R. T.; Drennan, C. L. *Nat. Struct. Biol.* **2003**, *10*, 794–799.
- Chivers, P. T.; Tahirov, T. H. *J. Mol. Biol.* **2005**, *348*, 597–607.
- Schreiter, E. R.; Wang, S. C.; Zamble, D. B.; Drennan, C. L. *Proc. Natl. Acad. Sci., U.S.A.* **2006**, *103*, 13676–13681.
- Dian, C.; Schauer, K.; Kapp, U.; McSweeney, S. M.; Labigne, A.; Terradot, L. *J. Mol. Biol.* **2006**, *361*, 715–730.
- Phillips, C. M.; Schreiter, E. R.; Guo, Y.; Wang, S. C.; Zamble, D. B.; Drennan, C. L. *Biochemistry* **2008**, *47*, 1938–1946.
- Chivers, P. T.; Sauer, R. T. *Protein Sci.* **1999**, *8*, 2494–2500.
- Chivers, P. T.; Sauer, R. T. *J. Biol. Chem.* **2000**, *275*, 19735–19741.
- Chivers, P. T.; Sauer, R. T. *Chem. Biol.* **2002**, *9*, 1141–1148.
- Dosanjh, N. S.; Michel, S. L. *Curr. Opin. Chem. Biol.* **2006**, *10*, 123–130.
- Zambelli, B.; Danielli, A.; Romagnoli, S.; Neyroz, P.; Ciurli, S.; Scarlato, V. *J. Mol. Biol.* **2008**, *383*, 1129–1143.
- Zambelli, B.; Bellucci, M.; Danielli, A.; Scarlato, V.; Ciurli, S. *Chem. Commun.* **2007**, 3649–3651.
- Berrera, M.; Pantano, S.; Carloni, P. *J. Phys. Chem. B* **2007**, *111*, 1496–1501.
- Cui, G.; Merz, K. M., Jr. *Biophys. J.* **2008**, *94*, 3769–3778.
- Sindhikara, D. J.; Roitberg, A. E.; Merz, K. M. *Biochemistry* **2009**, *48*, 12024–12033.
- Bradley, M. J.; Chivers, P. T.; Baker, N. A. *J. Mol. Biol.* **2008**, *378*, 1155–1173.
- Phillips, C. M.; Nerenberg, P. S.; Drennan, C. L.; Stultz, C. M. *J. Am. Chem. Soc.* **2009**, *131*, 10220–10228.
- Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acid. Res.* **1994**, *22*, 4673–4680.
- Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.
- Shen, M.; Sali, A. *Protein Sci.* **2006**, *15*, 2507–2524.
- Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
- Wiederstein, M.; Sippl, M. J. *Nucleic Acids Res.* **2007**, *35*, W407–W410.
- Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52–56.
- Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Matthews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER10*; University of California: San Francisco, 2008.
- Auffinger, P.; Cheatham, T., III. *J. Chem. Theory Comput.* **2007**, *3*, 1851–1859.
- 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.; AlLaham, 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. *Gaussian09, revision B.01*; Gaussian, Inc.: Pittsburgh, PA, 2004.
- Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
- Estiu, G.; Merz, K. M., Jr. *Biochemistry* **2006**, *45*, 4429–4443.
- Adam, K. R.; Antolovich, M.; Baldwin, D. S.; Brigden, L. G.; Duckworth, P. A.; Lindoy, L. F.; Bashall, A.; Mcpartlin, M.; Tasker, P. A. *J. Chem. Soc., Dalton Trans.* **1992**, *12*, 1869–1876.
- Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct. Funct. Bioinform.* **2006**, *65*, 712–725.

- (33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (34) Aqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
- (35) Smith, D.; Dang, L. *J. Chem. Phys.* **1994**, *100*, 3757–3766.
- (36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (37) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (38) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–1612.
- (39) Garcia, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.
- (40) Aleksiev, T.; Potestio, R.; Pontiggia, F.; Cozzini, S.; Micheletti, C. *Bioinformatics* **2009**, *25*, 2743–2744.
- (41) Micheletti, C.; Carloni, P.; Maritan, A. *Proteins* **2004**, *55*, 635–645.
- (42) Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12366–12371.
- (43) Bertini, I.; Felli, I. C.; Kummerle, R.; Moskau, D.; Pierattelli, R. *J. Am. Chem. Soc.* **2004**, *126*, 464–465.
- (44) Silva, J. R. A.; Lameira, J.; Santana, P. P. B.; Silva, A.; Schneider, M. P. C.; Alves, C. N. *Int. J. Quantum Chem.* **2010**, *110*, 2067–2075.
- (45) Yarnitzky, T.; Levit, A.; Niv, M. Y. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 317–325.
- (46) Yi, M.; Tjong, H.; Zhou, H. X. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 8280–8285.
- (47) Zhang, Y.; Sham, Y. Y.; Rajamani, R.; Gao, J.; Portoghese, P. S. *ChemBioChem* **2005**, *6*, 853–859.
- (48) Law, R. J.; Sansom, M. S. *Eur. Biophys. J.* **2004**, *33*, 477–489.
- (49) Capener, C. E.; Shrivastava, I. H.; Ranatunga, K. M.; Forrest, L. R.; Smith, G. R.; Sansom, M. S. *Biophys. J.* **2000**, *78*, 2929–2942.
- (50) Potestio, R.; Pontiggia, F.; Micheletti, C. *Biophys. J.* **2009**, *96*, 4993–5002.
- (51) Cascella, M.; Micheletti, C.; Rothlisberger, U.; Carloni, P. *J. Am. Chem. Soc.* **2005**, *127*, 3734–3742.
- (52) Carnevale, V.; Raugei, S.; Micheletti, C.; Carloni, P. *J. Am. Chem. Soc.* **2006**, *128*, 9766–9772.
- (53) Pontiggia, F.; Colombo, G.; Micheletti, C.; Orland, H. *Phys. Rev. Lett.* **2007**, *98*, 048102.
- (54) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796.

CT900635Z

Electric Field Effects on Short Fibrils of A β Amyloid Peptides

Francesca Lugli,^{*,†} Francesca Toschi,[†] Fabio Biscarini,[‡] and Francesco Zerbetto[†]

Dipartimento di Chimica "G. Ciamician", Università di Bologna, via F. Selmi 2, 40126 Bologna, Italy and ISMN CNR, V. Gobetti, Bologna, Italy

Received December 2, 2009

Abstract: Amyloid fibrils are highly ordered protein aggregates, which are associated with many neurodegenerative diseases. The assembling dynamics of monomeric beta-amyloid peptides, A β , into small aggregates (and then into long fibrils) is still debated and has become a hot topic. In this study, we conducted molecular dynamics simulations in explicit water of small A β protofibrils (from monomer to pentamer) under the perturbation of an externally applied electric field with the aim of investigating the fundamental molecular interactions involved in the aggregation mechanism. Dynamics of small adducts of A β (16–42) in the presence of an electric field, which was shown before to accelerate the conformational change of a single molecule, indicate that the structural resilience increases with the number of molecules in the aggregate. In particular, for 50 ns, the pentamer shows an enhanced stability in secondary structure, number of hydrogen bonds, and number of salt bridges, even in the presence of the field perturbation. The resilience to the field perturbation is linked to the variation of the induced dipole moment of the aggregates that tends to level off very rapidly with the growing number of molecules, thereby reducing the energy available per molecule to produce structural changes. The results also show that in the presence of the field the stability of the hydrophobic second β -sheet (β 2, residues 31–42) is higher than that of the first one (β 1, residues 18–26). In particular, we identify Gly33, Gly37, and Met35 as the most important residues that stabilize the intermolecular packing and may act as nucleation sites for fibrillization. Furthermore, dynamics of the full-length A β (1–42) pentameric aggregate, which include the highly charged random coil residues 1–15, confirmed the key role of the second hydrophobic core in the protofibril structure.

1. Introduction

The interest in amyloid fibrils arises because of their association with neurodegenerative diseases, including Alzheimer's disease (AD), type 2 diabetes, prion diseases, Parkinson's disease, senile systematic amyloidosis, and Huntington's disease.^{1–3} Amyloid diseases are characterized by extracellular deposition of insoluble plaques of misfolded polypeptides aggregated into ordered fibril structures. X-ray diffraction from amyloid fibrils shows that, despite differ-

ences in the amino acid sequence, they share a common cross- β structural motif, where β -strands run perpendicular to the long axis of the fibrils while the hydrogen bonds between β -strands are parallel to the axis.⁴

The major component of the amyloid plaques of AD is the 39–42-residue-long amyloid- β (A β) peptide, which is generated from the amyloid precursor protein by the proteolytic activities of β - and γ -secretase.⁵ It has been shown that before fibril formation, natively unstructured monomeric A β peptide undergoes large conformational changes from a high α -helix content peptide into a spherical amyloid intermediate, which contains predominantly β -sheet structures.⁶ Small A β oligomers shows higher toxicity than the fibril, indicating that the β -sheet formation may trigger

* To whom correspondence should be addressed. Tel.: +39 051 2099915. Fax: +39 051 2099456. E-mail: francesca.lugli6@unibo.it.

[†] Università di Bologna.

[‡] ISMN CNR.

neurotoxicity.^{7–10} It is also known that A β 42 is more prone to aggregation than the shorter A β 40.^{11–13}

Understanding the molecular mechanisms of A β amyloid fibril formation and its early detection is important in the elucidation of the role of endogenous factors that may inhibit amyloidogenesis in the normal brain and/or promote it in the AD brain and could also play a role in the design and development of new types of ultrastrong materials because of the high mechanical toughness of the fibrils.¹⁴ Prerequisite to the elucidation of the pathway of aggregation is the understanding of the thermodynamics of the aggregation process and in particular of the conformational rearrangements that carry the starting proteins to the product A β amyloid fibril. Perturbations of the thermodynamics properties of a protein in its soluble or its fibrillar state are usually measured by solid-state NMR spectroscopy, Fourier-transform IR, circular dichroism, gel filtration, electron microscopy, etc., but do not necessarily correlate with the aggregation pathway.¹⁵ Structural information about the fibrils comes primarily from electron microscopy images, X-ray diffraction, solid-state NMR, and biochemical studies.^{16–18} On the basis of such experimental observations, several atomic models for amyloid fibril structure have been proposed.^{19–22} They show that the N-terminal segment, where polar residues are located, exhibits a disordered structure, whereas the C-terminal segment is more ordered and characterized by two hydrophobic β -strands connected by a turn. These two β -sheets are constructed by residues 18–26 (β 1) and residues 31–42 (β 2), while the loop comprises residues 27–30 (see Figure 1).

Experimental studies, supported by computer simulations, have shown that also fragments of A β can form fibrils. In particular, simulations performed on β -strand segments such as A β _{10–35}, A β _{16–35}, and A β _{29–42} peptides^{23,24} form in register parallel β -sheets, while A β _{34–42} and A β _{16–22} prefer an antiparallel organization.^{18,25,26} In the model proposed by Ma and Nussinov²³ the key structural feature of the salt bridge between Asp23–Lys28 is illustrated, along with the intramolecular hydrophobic cluster between Leu17/Phe19 and Ile32/Leu34.

Nussinov and co-workers performed several MD simulations investigating the protofibril structure and stability of A β _{17–42} in linear-like²⁷ and annular-like structures.²⁸ It has also been shown that while fibrillar A β amyloids are always organized in a parallel β -sheet conformation, small oligomeric A β display an antiparallel β -sheet structure.²⁹ Here, we investigate only a fibrillar-like arrangement.

Molecular dynamics simulations of long protofilaments, not considered here, confirmed the role of strong hydrophobic interactions between the β -strands for the stability of the fibril structure.³⁰ Recently, the aggregation behavior of small protofibrillar aggregates (from dimer to pentamer) of A β _{17–42} was studied by MD simulations, suggesting that the second hydrophobic β -strand may act as the nucleation site for aggregation.³¹ The role of the β 2 portion in stabilizing small aggregates and possibly inducing the growth of fibrils was also confirmed by MD simulations performed on the full-length A β (1–42).³²

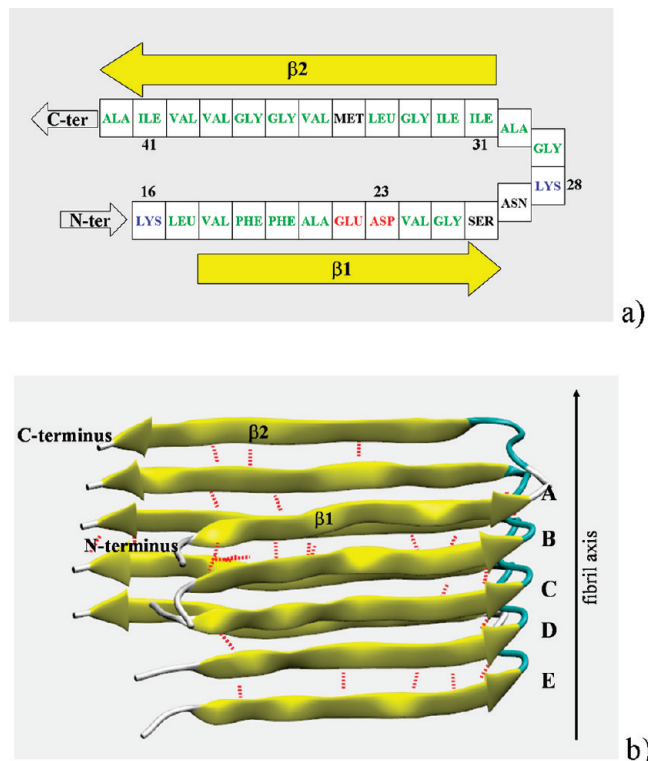


Figure 1. (a) Amino acid sequence of A β (16–42). The hydrophobic residues are in green, positively and negatively charged residues are in blue and red, respectively, while polar residues are in black. β -Sheets, β 1 and β 2, are also indicated. (b) The 3D structure of A β pentamer. Ribbon diagram of the core structure of residues 16–42 showing the intermolecular nature of the inter- β -strands interactions. Red dotted lines indicate the intermolecular H bonds. Individual molecules are labeled as A, B, C, D, and E.

The early detection of the amyloid aggregates could also be important for an early diagnosis of the disease and could be obtained by transistor-based sensors.^{33–35} Structural and functional modifications of biosystems by electromagnetic field exposure are also a growing concern for society. For instance, a recent study showed evidence of nonthermal microwave damage of brain tissue in exposed rats.³⁶ Microwave radiation was shown to induce changes in protein conformations and to promote formation of amyloid fibrils in vitro under nonphysiological conditions.³⁷ Electric fields and electric potentials also have therapeutical effects in deep brain stimulation techniques, where the application of relatively small voltages (\sim 1–3 V) by an electrode implanted in the brain is used to treat Parkinson's disease symptoms.³⁸ Changes in electrical fields associated with membranes can play a role in diseases such as Alzheimer's.^{39,40}

To the best of our knowledge, the possible use of an electric field to influence or modify the strong interactions that exist in amyloid fibrils has not yet been considered. It therefore appears important to investigate the behavior of biosystems under an external electric field (EF). In particular, analysis of the modifications of the structure of individual proteins, their aggregates, and the aggregation mechanism are of practical interest. Computational approaches are rapidly becoming inexpensive and accurate tools to perform such investigation. We recently studied the effect of an

electric field on the coil-to- β sheet conversion of the amyloid peptide,⁴¹ with an approach similar to that of Budi et al.^{42–44} We also investigated the effect of an electric field on a nanosize water drop deposited on a surface⁴⁵ and the effect of water on the response of an organic field effect transistor.⁴⁶ In this work, we examine the early response of small amyloid protofibrils to an externally applied electric field. The external electric field is set to 0.5 V/nm (and in one case to 0.25 V/nm) since this magnitude was shown to be close to the limit of instability of water and strongly modified the conformation of a single amyloid peptide in water.^{41,47}

2. Methods

2.1. Structural Model. The starting structures were extracted from the model, deposited in the Protein Data Bank, whose PDB code is 2BEG.²⁰ The 3D structure of the Alzheimer's $A\beta(1-42)$ protofilament was obtained using hydrogen-bonding constraints from quenched hydrogen/deuterium-exchange NMR, side-chain packing constraints from pairwise mutagenesis studies, and parallel, in-register β -sheet arrangement from previous solid-state NMR studies. The model contains 10 structures of a pentamer of the β -amyloid (17–42) peptides. The first 16 residues are disordered, and the 3D structure contains coordinates of the core structure of residues 17–42.

For the calculations, the first structure of the pentamer was extracted. At each molecule the residue Lys16 was added in order to neutralize the system. In this way, each molecule contains two negatively charged residues (Glu22 and Asp23) and two positively charged residues (Lys16 and Lys28).

The peptides $A\beta(16-42)$ were terminated by an acetyl group at the N-terminus and by N-methyl group at the C-terminus. Figure 1 shows the amino acid sequence and the starting structure of the β -amyloid protofilament.

The core of the amyloid fibril is composed of two β -sheets (β_1 and β_2) whose strands are perpendicular to the fibril axis and hydrogen bonds between peptides that run roughly parallel to the axis. The first, β_1 , comprises residues Val18-Ser26, residues Gly27-Ala30 form a turn, residues Ile31-Ala42 form the second strand, β_2 . Asp23 and Lys28 form a salt bridge, which has been shown in previous computational studies^{2,3} to be fundamental for the fibril structure and in the stabilization the U-shape of the peptide.

The systems simulated were the monomer (molecule A), the dimer adduct (molecules A and B), the trimer adduct (molecules A, B, and C), the tetramer adduct (molecules A, B, C, and D), and pentamer adduct (molecules A, B, C, D, and E) in water boxes of 64 Å side lengths containing about 8000 water molecules.

2.2. Molecular Dynamics. The calculations were performed with the NAMD 2.6 Molecular Dynamics program⁴⁸ using periodic boundary conditions and cubic simulation cells. Electrostatic interactions were calculated with the particle mesh Ewald (PME) method^{49–51} with a grid spacing of 1.2 Å and a spline interpolation of order 4. We used a cutoff of 12 Å for the real-space direct sum part of the Ewald sum and for the van der Waals interactions. The time integration step was set to 2 fs. Simulations at 300 K were

carried out for all systems. Rigid bonds involving hydrogen atoms were constrained using the SHAKE algorithm.⁵² The force field adopted is CHARMM27⁵³ for the peptide, whereas water is modeled by the TIP3P force field.⁵⁴ The calculations were carried out in steps: (a) minimization of the lateral chains, with the backbone frozen; (b) minimization of the water molecules, with the whole protein frozen; (c) minimization of the entire system, with the backbone frozen; (d) MD equilibration in the NVT ensemble for 250 ps with the backbone frozen; (e) MD equilibration in the NPT for 250 ps with the backbone frozen, pressure set to 1 atm, and by using the Langevin dynamics;⁵⁵ (f) MD simulation in the NPT ensemble, where systems were coupled to a barostat with a constant pressure set to 1 atm and a relaxation time of 0.1 ps and to a Berendsen thermostat with temperature set to 300 K.⁵⁶ The validity of this procedure has been confirmed in previous studies.^{41,43} For each of the smallest adducts (dimer, trimer, and tetramer), two 10 ns long molecular dynamics, MD, simulations were performed in the absence and in the presence of a constant external electric field (EF) of 0.5 V/nm, while for the pentamer adduct simulations up to 50 ns were carried out. In all systems, the electric field was applied along the direction of the fibril axis. In order to verify whether the direction of the EF influences the energy and stability of the structures, an additional MD run was performed on the pentamer adduct with the EF perpendicular to the fibril axis. As a reference, two 5 ns long MD simulations (with and without EF) were run for the monomer.

In order to check whether the highly charged and polar N-terminal random coil segment (1–15) influences the structural stability of the core β -sheet domain, three 50 ns MD simulations of the full-length $A\beta(1-42)$ pentamer were carried out: with electric fields of 0.25 and 0.5 V/nm and without electric field. This system was built adding the random coil segment (1–15) to the five peptides of the equilibrated $A\beta(16-42)$ pentamer structure.

2.3. Analysis Tools. The structural stability of the systems under the influence of an external applied field was investigated by determining, as a function of time, the root-mean-square deviations (rmsd) from the structure at the start of the data collection. Also, the radius of gyration (R_g) was examined. The rmsd was calculated only for the C_α atoms, after their least-squared superposition with the initial positions. We also analyzed the time evolution of secondary, tertiary, and quaternary structures. The site-specific propensity for secondary structure formation is determined by the STRIDE program⁵⁷ in the VMD software package.⁵⁸ STRIDE contains a knowledge-based algorithm that uses hydrogen-bond energy and statistically derived backbone torsional angle information to return the secondary structure assignments in maximal agreement with crystallographic designations. Information about tertiary and quaternary structures were obtained by analyzing the intramolecular and intermolecular contact maps. Contacts between pairs of amino acids can be visualized by using contact maps, 2D representations of all amino acid pairs in contact and where contact frequencies are represented by different colors. The intramolecular contact map only takes into account the contacts between pairs of amino acids in the same peptide and

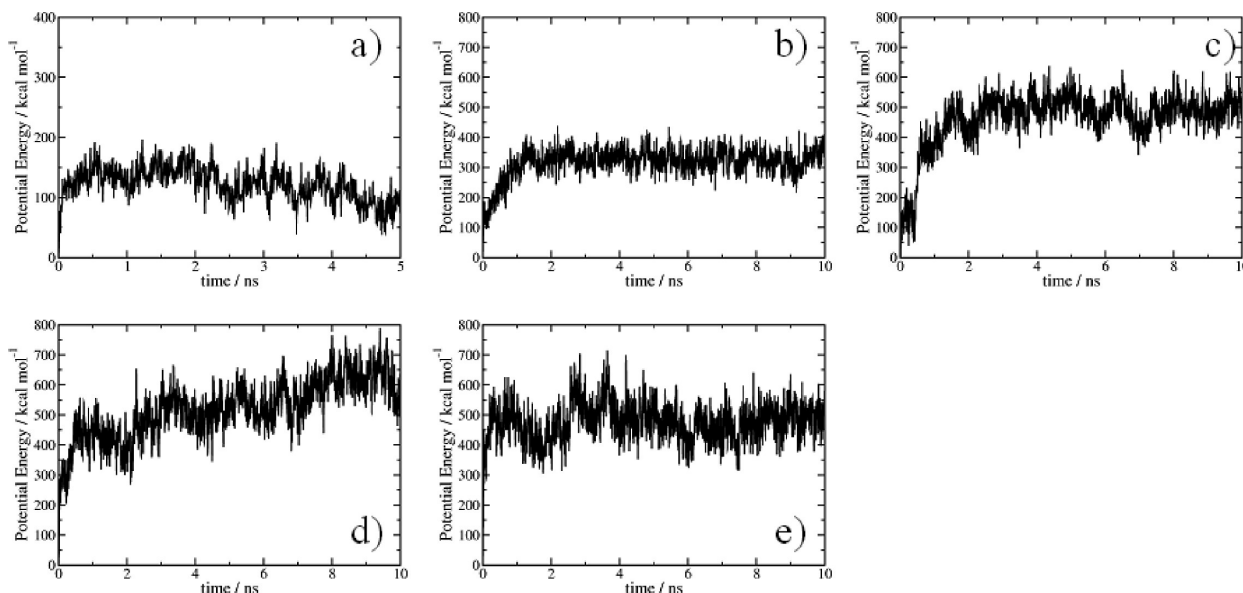


Figure 2. Evolution in time of the energy: (a) monomer, (b) dimer, (c) trimer, (d) tetramer, and (e) pentamer aggregates after an external EF of 0.5 V/nm is switched on. For simplicity, the initial energy is taken as a reference and set to zero.

contains information about the tertiary structure of peptides within the aggregates. The intermolecular contact map only takes into account the contacts between pairs of amino acids that belong to different peptides and yields information on the way different peptides assemble into the aggregate (the quaternary structure). The salt bridges distances were monitored in time by using the VMD Salt Bridges plugin. A salt bridge is considered formed if the distance between the oxygen atom of an acidic residue and the nitrogen atom of a basic residue is within the cutoff distance (default 3.2 Å). Hydrogen bonds are quantified by counting acceptor atom–donor atom pairs that are no further than 3.0 Å and within a cutoff angle of 20°. Total dipole moments of the aggregates were also calculated in time.

3. Results and Discussion

3.1. Electric Field Effect on A β (16–42) Protofibrils.

The early response of small A β (16–42) protofibrils to a strong electric field (EF = 0.5 V/nm) was first analyzed in terms of potential energy profiles in a time range of 10 ns, see Figure 2. The change of potential energy of the A β (16–42) adducts is related to the structural destabilization of the systems due to the field perturbation. The initial energy, at the time when the field is switched on, is taken as a reference and set to zero. The field changes the potential energy of systems and quickly destabilizes them by ~ 142 kcal mol $^{-1}$ for the monomer, ~ 326 kcal mol $^{-1}$ for the dimer, ~ 489 kcal mol $^{-1}$ for the trimer, ~ 581 kcal mol $^{-1}$ for the tetramer, and ~ 470 kcal mol $^{-1}$ for the pentamer. It is instructive to notice that the increase of potential energy destabilization is linear with the number of peptides in the adducts up to the tetramer while the destabilization of the pentamer is less than that calculated for the tetramer. In the absence of the field, the average value of the potential energy is $0.5k_B T$ per degree of freedom. The activation of the field introduces forces that act on the charges and the dipoles present in the simulation box. It is reasonable to expect that

at equilibrium, the effect of the forces is to distort the molecules and increase the total potential energy, proportionally to the number of atoms. The pentamer “stabilization” in the presence of the EF with respect to the other aggregates means that its structural conformation is less affected by the external perturbation.

The root-mean-square deviations (rmsds) of positions of the C $_{\alpha}$ atoms relative to the initial structure of each molecule labeled as A, B, C, D, and E (see Figure 1b) in the dimeric, trimeric, tetrameric, and pentameric aggregates are shown in Figure 3. For each molecule, comparison of the behavior in the presence and in the absence of an electric field, EF, of 0.5 V/nm is possible. The size of the electric field was selected on the basis of previous molecular dynamics simulations performed on the monomer, where 0.5 V/nm showed an acceleration of the denaturation of the peptide in water in 10 ns.⁴¹

In general, application of the electric field affects the smaller aggregates. The effect of the electric field is quite small in the pentamer. The average rmsd values for each of the adducts averaged after equilibration are shown in Table 1 and concur with the more detailed profiles shown in Figure 3. Furthermore, one can notice that the dimer is not stable even without perturbation, and the main effect of the application of the external field is to accelerate the conformational destabilization.

The average radius of gyration (R_g) of each system is also given in Table 1. Only the dimeric and trimeric adducts modify their shape in the presence of the field while tetramer and pentamer tend to retain it.

Figure 4 compares the maps of the evolution of the secondary structure as a function of time both in the absence and in the presence of the field. Noticeable differences appear for the smaller systems when the field is present. In general, also for the larger aggregates, the EF increases the number of residues with a turn conformation. In the pentamer, only the top molecule A (see Figure 1b) is affected, an effect due

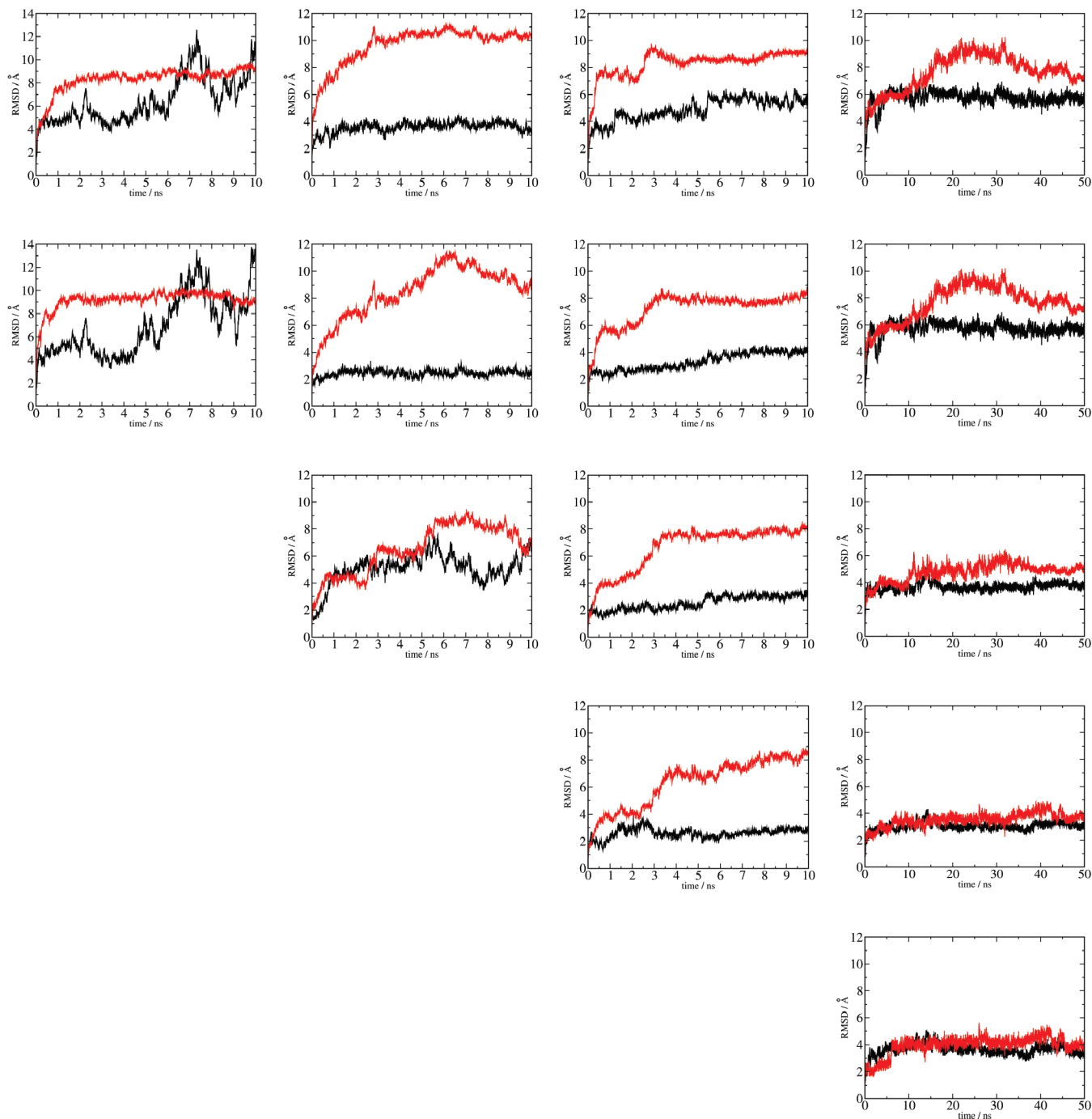


Figure 3. Evolution of the rmsd of C_{α} atoms. From top to bottom, A, B, C, D, and E molecules in the aggregates. From left to right, dimeric, trimeric, tetrameric, and pentamer aggregates. The molecules in each aggregate are labeled from A to E (see Figure 1): (black line) simulation in the absence of the EF, (red line) simulation in the presence of the EF of 0.5 V/nm.

Table 1. RMSDs (\AA) and Radius of Gyration (\AA) Averaged after Equilibration

system	rmsd (\AA), no EF	rmsd (\AA), EF = 0.5 V/nm	R_g (\AA), no EF	R_g (\AA), EF = 0.5 V/nm
monomer	8.5 ± 0.7	10.6 ± 0.4	11.6 ± 0.5	11.7 ± 0.6
dimer aggregate	8.9 ± 1.6	9.1 ± 0.2	16.1 ± 1.3	11.3 ± 0.1
trimer aggregate	4.1 ± 0.4	10.4 ± 0.9	14.4 ± 0.1	13.2 ± 0.4
tetramer aggregate	3.9 ± 0.2	8.1 ± 0.3	14.3 ± 0.1	14.1 ± 0.2
pentamer aggregate	6.0 ± 0.2	6.9 ± 0.3	14.9 ± 0.1	15.1 ± 0.3

to the lack of stabilizing interactions (such as intermolecular salt bridges) with monomers above it.

A good hydrophobic side-chain packing pattern is essential for maintaining the stable β -sheet conformation. Contact

maps can help to rationalize this aspect. Stable and strong hydrophobic lateral chain contacts contribute to stabilize the β -sheet secondary structure, maintaining the hydrogen-bond network between different molecules in the adducts.

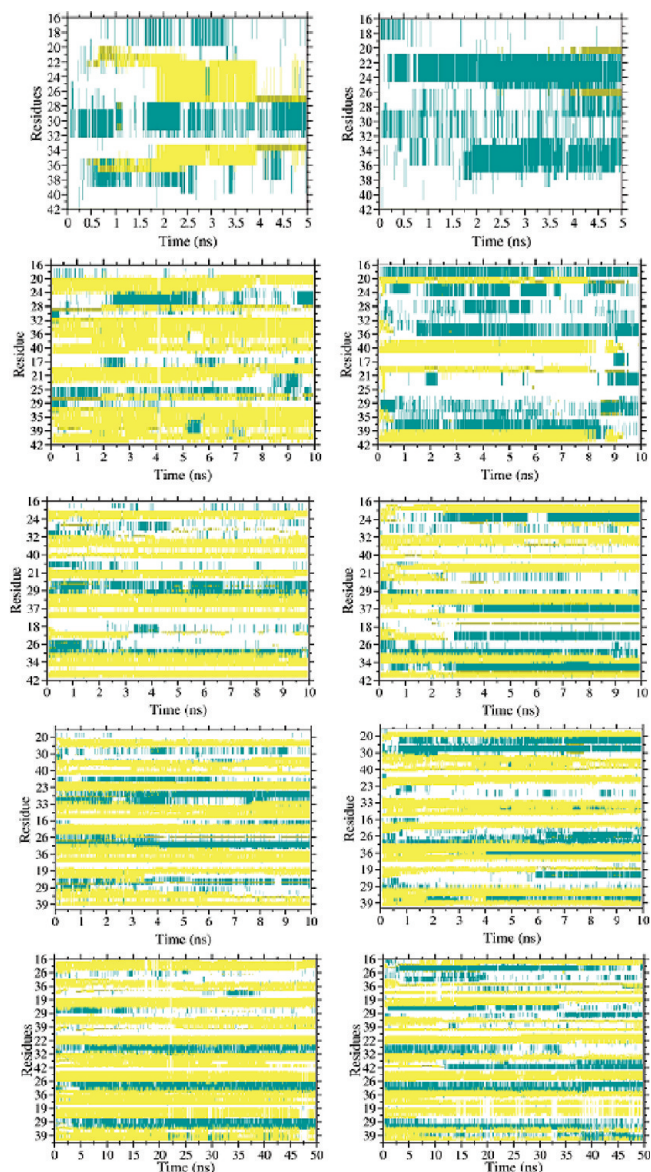


Figure 4. Maps of the evolution of the secondary structure as a function of time. Top to bottom: monomer, dimer, trimer, tetramer, and pentamer. (Left) No electric field; (right) $EF = 0.5$ V/nm. The vertical axis contains the residue number of the 16–42 $A\beta$ sequence. Yellow for β -sheet, ochre and green for turn, and white for coil.

Figure 5 displays the average intramolecular (tertiary structure) and intermolecular (quaternary structure) contact maps for each system. Each pixel in the maps represents the contact strength between two residues. In the intramolecular maps (first and second vertical entries, labeled “Intramolecular”), vertical and horizontal axes represent residues of the same molecule (the dark diagonals are the contact of each residue with itself). The maps are symmetric with respect to the diagonal. The extradiagonal dark pixels represent the intersheet contacts between $\beta 1$ and $\beta 2$ of the same molecule, which are slightly influenced by the field. In particular, the tertiary structure is stabilized by intersheet side-chain interactions between residues Phe19/Gly38 and Ala21/Val36 and by intramolecular salt bridges between residues Asp23 and Lys28. The intermolecular maps (third and fourth vertical entries, labeled “Intermolecular”) repre-

sent contacts between residues of the n th molecule (vertical axis) and the $(n-1)$ th one (horizontal axis). The diagonals represent the distance between the same residues of vicinal molecules, and the maps are not necessarily symmetric. The pentameric protofilament is stabilized by intermolecular side-chain interactions between the odd-numbered residues of sheet $\beta 1$ of the n th molecule and the even-numbered residues of the sheet $\beta 2$ of the $(n-1)$ th molecule (see the dark pixels in the bottom-right region of the contact map) and, more importantly, by salt bridges between residues Asp23 and Lys28. These results are in agreement with the experimental model proposed by Lührs et al.²⁰ The electric field effect will necessarily be smaller in the case of hydrophobic contacts.

The intermolecular salt bridges have a preferential orientation. They form between Asp23 of the n th molecule and Lys28 of the $(n-1)$ th one. The direction of the EF coincides with the fibril axis and should stabilize the intermolecular salt bridges.

The intramolecular salt bridges are affected by the field and tend to be disrupted by the field (see Table 1 in the Supporting Information). The noticeable exception is the pentameric adduct, where, in 50 ns, the bridges of the inner molecules remain stable in the presence of the field. These resilient bridges hold the shape of the molecule in place so that the gyration radius (and the rmsd) of Table 1 and Figure 3 do not vary in the presence of the field.

The intermolecular salt bridges are disrupted by the field in the dimeric and trimeric aggregates. However, they are quite stable in the two higher adducts. Peculiarly, a salt bridge that opposes the field is formed in the tetramer. The simple explanation for its presence is that the local electrostatic effect wins over the nonlocal perturbation of the field, which orients the adduct along its direction in less than 100 ps, see Table 2, and tends to increase the overall dipole moment. Table 2 shows that the dipole moment of the adducts, in the presence of the field, reaches a limiting value already for the trimer and that alignment with the field is very rapid.

The higher resilience of the intermolecular salt bridges, with respect to the intramolecular ones, can be explained by considering the orientation of the dipole moment, which is aligned along the direction of the intermolecular salt bridges and creates a stabilizing interaction when the field is switched on.

Applying the field along the perpendicular direction of the fibril axis effects only the time, τ , required to align the dipole moment of the system to the field but effects neither the potential energy nor the structural features of the aggregates.

The convergence of the dipole moment properties offers a simple explanation for the stability of the aggregates. The interaction energy between the dipole moment and the electric field, $\mu \cdot E$, reaches a plateau already for the trimer adduct. This is the energy available to modify the shape of the adduct, either at the molecular level or its overall shape. Increasing the size of the adduct, the energy available from the field becomes negligible once smeared out over the individual molecules.

The rigidity of the systems is evident if one considers the number of hydrogen bonds present during the dynam-

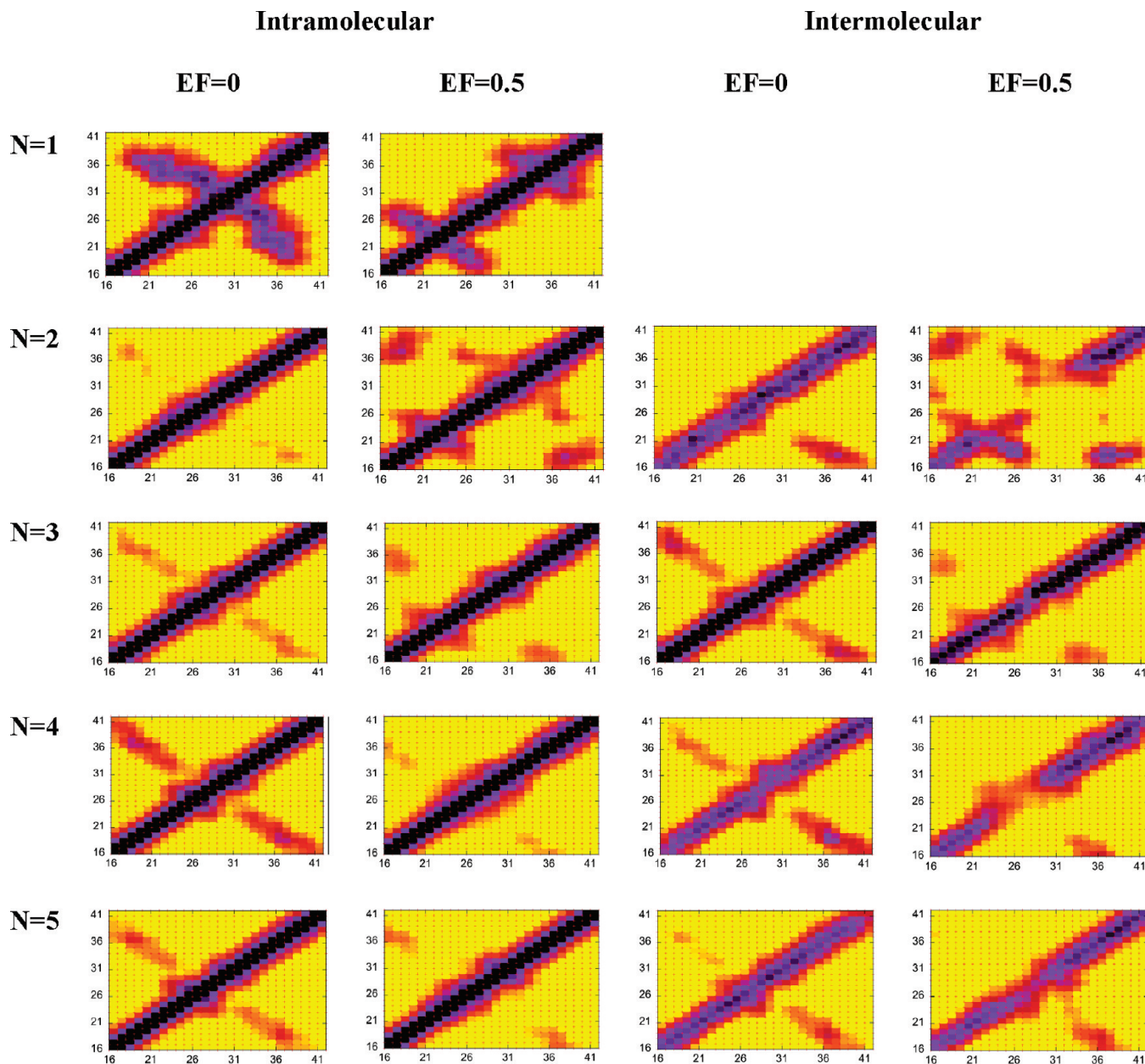


Figure 5. Intramolecular and intermolecular contact maps. Both the x and y axes have the residues number in the peptide $A\beta_{16-42}$ sequence. Top to bottom: monomer ($N = 1$), dimer adduct ($N = 2$), trimer adduct ($N = 3$), tetramer adduct ($N = 4$), pentamer adduct ($N = 5$). Left to right, first entry intramolecular contact maps with no electric field, second entry intramolecular contact maps with $EF = 0.5$ V/nm, third entry intermolecular contact maps with no electric field, and fourth entry intermolecular contact maps with $EF = 0.5$ V/nm. Each pixel in the maps represents the average contact strength between two residues and is color coded: yellow, no contact; orange, red, magenta, to violet, strong contacts; black, very strong contact.

Table 2. Summary of the Dipole Moments, in Debyes, Averaged after Equilibration, and Time, τ in ps, Required To Align the Adduct with the Field

system	μ , no EF	μ , EF = 0.5 V/nm	μ_z , no EF off	μ_z , EF = 0.5 V/nm	τ
monomer	100 ± 30	296 ± 7	18 ± 23	250 ± 6	30
dimer adduct	138 ± 36	539 ± 9	-105 ± 57	536 ± 10	50
trimer adduct	149 ± 28	780 ± 12	-22 ± 75	778 ± 12	50
tetramer adduct	288 ± 19	850 ± 20	230 ± 43	821 ± 17	90
pentamer adduct	470 ± 31	924 ± 19	-343 ± 96	915 ± 19	70

ics, see Table 3. The electric field halves their number in the dimeric aggregate but barely affects it in the higher adducts.

The stability of intermolecular hydrogen bonds in larger adducts is also responsible for the stability of secondary

structure. Previous DFT calculations showed that intermolecular hydrogen bonds formed between the β -sheets of stacked amyloid peptides in the fibril are strongly cooperative and probably may contribute to the nucleation growth kinetics observed experimentally.⁵⁹

Table 3. Number of Hydrogen Bonds Averaged after Equilibration

system	no EF			EF = 0.5 V/nm		
	total	intra	inter	total	intra	inter
monomer	1.3 ± 0.9			0.7 ± 0.7		
dimer adduct	7.4 ± 2.3	0.3 ± 0.5	7.1 ± 2.3	4.1 ± 2.2	0.3 ± 0.6	3.8 ± 2.2
trimer adduct	12.1 ± 2.9	0.2 ± 0.4	11.9 ± 2.8	10.9 ± 2.6	1.1 ± 0.9	9.8 ± 2.5
tetramer adduct	21.0 ± 3.4	2.3 ± 1.0	18.7 ± 3.4	18.1 ± 3.6	1.2 ± 0.8	16.9 ± 3.6
pentamer adduct	30.9 ± 6.1	3.0 ± 1.3	27.9 ± 6.1	28.1 ± 4.9	3.4 ± 1.4	24.7 ± 5.9

The β -sheet content per residue, calculated along the 10 ns MD runs and averaged over all the molecules, is shown in Figure 6. In the absence of the field the β -sheet content increases with the number of monomers. The extension of β_1 increases upon going from trimer to tetramer to pentamer, while β_2 is almost stable for all the adducts, and the β -strand segment, 39–41, is slightly lower for the pentamer. The effect of the EF is to decrease the extension of β_1 , while β_2 exhibits higher secondary structural stability; residues Val40 and Ile41 increase their β -sheet content.

The results are consistent with a recent study, where the stability of $A\beta(17-42)$ aggregates was investigated by MD simulations at high temperature,³¹ and also with experimental and theoretical studies which indicated the importance of Ile41 for paranucleus formation in $A\beta 42$.^{60,61}

The segment β_2 has a higher hydrophobic content than β_1 . Hydrophobic residues such as Ile, Leu, and Met are known to exhibit a positive contribution to protein stability. A good side-chain packing is fundamental for maintaining a stable secondary β -sheet structure because it protects the hydrogen-bonding network from water solvation.⁶² By analyzing inter- and intramolecular contact maps, one can notice that the strongest contacts are between the hydrophobic residues, which are perturbed the least by EF.

These findings imply that hydrophobic interactions play an important role in maintaining the secondary structural stability of the $A\beta$ fibrils, also in presence of external perturbation. The stable β -sheet structure also explains the stabilization of the number of intermolecular hydrogen bonds.

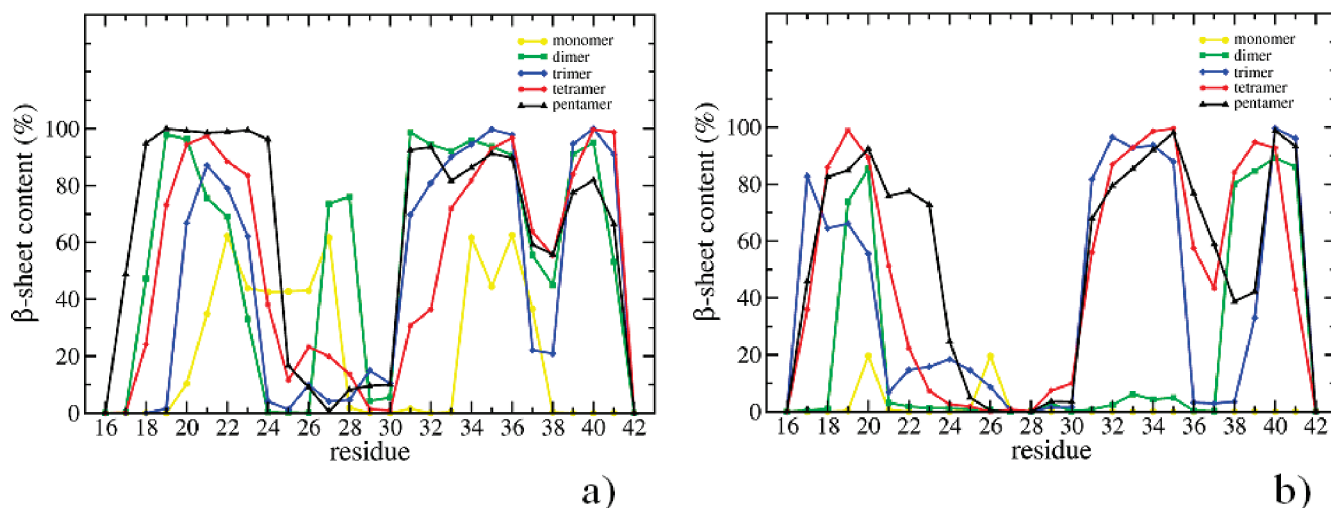
3.2. Simulations of the Full-Length $A\beta(1-42)$ Pentamer. The application of an external perturbation confirms that the strong hydrophobic interactions between the β -strands

are the critical element in maintaining the amyloid fibril stability. This holds also when a full-length $A\beta(1-42)$ protofibril, which contains the highly charged N-terminal residues, is affected by the electric field perturbation.

The effect of an externally applied electric field of 0.5 V/nm on the pentameric $A\beta(1-42)$ was analyzed in terms of root-mean-square displacements of different peptide regions, namely, the random coil, RC (residues 1–17), the first β -sheet, β_1 (residues 18–26), the turn (residues 27–30), and the second β -sheet, β_2 (residues 31–42). The results, averaged over the five molecules of the protofibril, are displayed in Figure 7. The field mainly affects the highly charged random coil portion, causing the consequent structural deformation also of the β_1 region, which is increasingly disrupted. On the contrary, the hydrophobic β_2 region remains stable during the simulation.

Figure 8 shows the comparison of the time evolution of the secondary structure between $A\beta(16-42)$ and $A\beta(1-42)$. The main effect of the EF is to decrease the β -sheet content in the β_1 segment of the full-length $A\beta(1-42)$. After ~ 20 ns, such β -sheet content goes nearly to zero while the β_2 segment shows resistance to the field. In particular, residues Leu34-Met35 and Val40-Ile41 are the most resilient in both pentamers.

The intermolecular contact map averaged over the last 20 ns of simulation (Figure 9) confirms that the most resilient hydrophobic contacts are located in the β_2 segment. In particular, contact pairs Gly33-Gly33, Gly37-Gly37, and Met35-Met35 remain relatively stable. These results agree with previous studies³¹ that suggested that these contacts trigger the β -sheet packing during the early stages of aggregation.

**Figure 6.** β -Sheet content per residue (a) without electric field and (b) with an applied electric field of 0.5 V/nm.

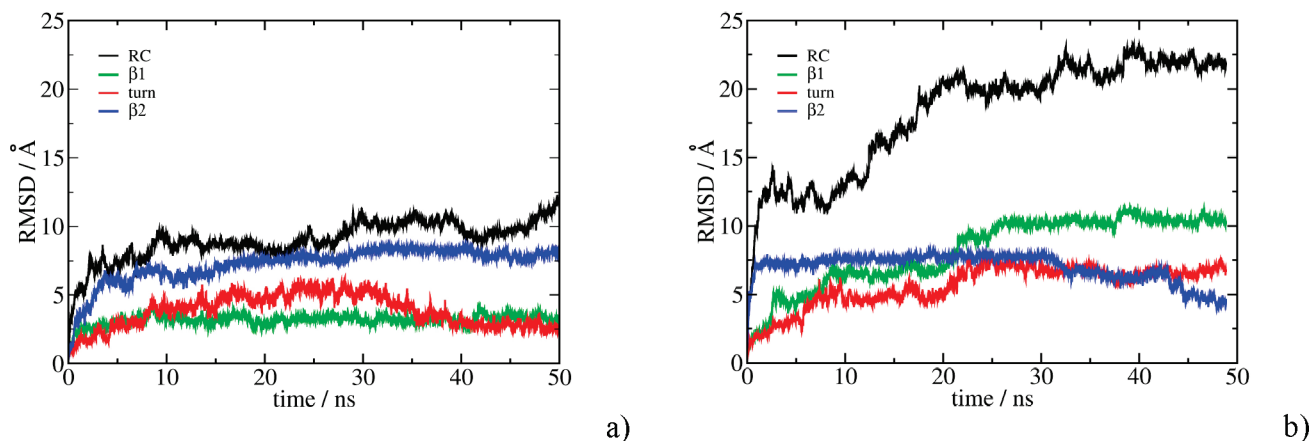


Figure 7. rmsd as a function of time, calculated for the $A\beta(1-42)$ pentamer (a) without EF and (b) with $EF = 0.5$ V/nm.

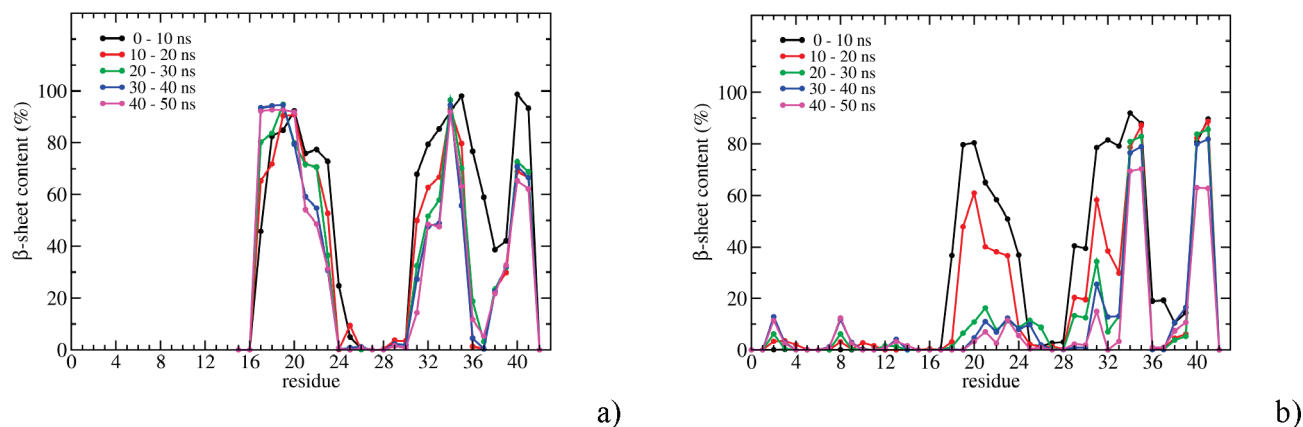


Figure 8. Time evolution of the β -sheet content for all residues (a) in the pentameric adduct of $A\beta(16-42)$ and (b) in the pentameric adduct $A\beta(1-42)$, both under $EF = 0.5$ V/nm.

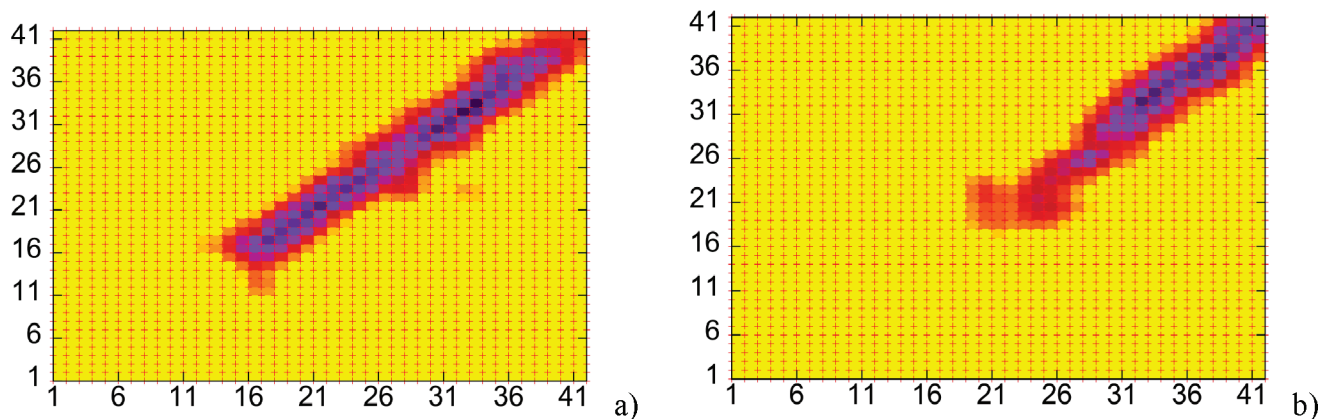


Figure 9. Intermolecular contact maps of the pentameric $A\beta(1-42)$ averaged over the five molecules and the last 20 ns of simulation (a) in the absence and (b) in the presence of EF. The x and y axes contain the residue number in the peptide $A\beta_{1-42}$ sequence. Each pixel in the map represents the contact strength between two residues and is color coded: yellow, no contact; orange, red, magenta, to violet, strong contacts; black, very strong contact.

The electric field mainly acts on the charged residues. Since the charged residues are located in the region of random coil and $\beta 1$, the effect is to “stretch” the molecules. Only two internal salt bridges are stable for the 50 ns of the simulation, namely, those between molecules B and C and between molecules C and D. The intermolecular hydrogen-bond network of $\beta 1$ is slowly reduced, while the H bonds of the $\beta 2$ region are preserved. Representative pictures of

the pentameric $A\beta(16-42)$ and $A\beta(1-42)$ during the simulations are provided in the Supporting Information.

The effect of a different field strength on the full-length $A\beta(1-42)$ pentameric aggregate was also investigated. Since electric fields larger than 0.5 V/nm encounter water instability,^{41,42} which is manifested as an explosion of the water box, we performed a further 50 ns long MD simulation of the pentameric adduct $A\beta_{42}$ under a field of 0.25 V/nm.

Also, in this case, the most internal monomers (molecules B and C) are the most stable. The major difference with the previous simulation, see plots similar to Figures 8 in the Supporting Information, is a partial resilience of the β 1 segment, where residues 18–26 conserve their β -sheet conformation for a longer time.

4. Conclusion

Amyloid proteins may undergo a structural transition from the native soluble monomeric conformation to fibrillar assemblies that have predominantly β -sheet structure.

The study of the principal driving forces that maintain the aggregate state of $A\beta$ peptides is a key issue not only for the understanding of pathological mechanisms involved in many neurodegenerative diseases but also for the development of new nanomaterials that can resist external mechanical perturbation.

The use of modified physiological conditions, such as temperature, pressure, pH, or solvents, has been shown to be a useful tool to study protein aggregation stability. The application of external fields can also be used to rationalize the most important interactions in the formation of amyloid aggregates.

In this study, we conducted numerical simulations of the smallest aggregates of $A\beta(16-42)$ peptides under the perturbation of an electric field. We show that the stability of the aggregates increases with the number of monomers and that larger protofibrils tend to resist to electric field perturbations for at least 50 ns. The number of H bonds and salt bridges is not affected by a field large enough to accelerate the coil-to- β -sheet conformational conversion of the monomer.⁴¹ The overall shape and position of the individual atoms of the adducts are also weakly affected by the electric field. Only some turns are introduced.

In the case of the full-length $A\beta(1-42)$ pentameric adduct, which carries a net negative charge, the main effect of the electric field is to decrease the amount of the β -sheet content in the β 1 segment, leading to a destabilization of intermolecular contact pairs in the N-terminal region (residues 1–21). The hydrophobic β 2 segment is less affected by the strong field, and residues 33–35 and 40–42 tend to maintain their native intermolecular contact pairs.

The present results concur with the hypothesis that the hydrophobic β 2 segment may act as a nucleation site during early stages of $A\beta$ aggregation.

Acknowledgment. We are grateful to Pablo Stoliar, Dago de Leeuw, and Arian Shehu for stimulating discussions. This work was supported by the project EU-STRP 0032652 BIODOT.

Supporting Information Available: Dynamics observed in the simulations for the salt bridges, both intramolecular and intermolecular; results of the 50 ns long MD simulations performed on the pentameric $A\beta(1-42)$ adduct with an electric field of 0.25 V/nm; representative pictures of $A\beta(16-42)$ and $A\beta(1-42)$ pentamers during the simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Rochet, J.-C.; Lansbury, P. T. *Curr. Opin. Struct. Biol.* **2000**, *10*, 60–68.
- (2) Selkoe, D. J. *Nature* **1999**, *399*, A23–A31.
- (3) Selkoe, D. J. *Neuron* **1991**, *6*, 487–498.
- (4) Sunde, M.; Blake, C. *Adv. Protein Chem.* **1997**, *50*, 123–159.
- (5) Yankner, B. A. *Neuron* **1996**, *16*, 921–932.
- (6) Chimon, S.; Shaibat, M. A.; Jones, C. R.; Calero, D. C.; Aizezi, B.; Ishii, Y. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1157–1164.
- (7) Bucciantini, M.; Giannoni, E.; Chiti, F.; Baroni, F.; Formigli, L.; Zurdo, J.; Taddei, N.; Ramponi, G.; Dobson, C. M.; Stefani, M. *Nature* **2002**, *416*, 507–511.
- (8) Bucciantini, M.; Calloni, G.; Chiti, F.; Formigli, L.; Nosi, D.; Dobson, C. M.; Stefani, M. *J. Biol. Chem.* **2004**, *279*, 31374–31382.
- (9) Hardy, J.; Selkoe, D. J. *Science* **2002**, *297*, 353–356.
- (10) Walsh, D. M.; Klyubin, I.; Fadeeva, J. V.; Cullen, W. K.; Anwyl, R.; Wolfe, M. S.; Rowan, M. J.; Selkoe, D. J. *Nature* **2002**, *416*, 535–539.
- (11) Jarrett, J. T.; Berger, E. P.; Lansbury, P. T. *Biochemistry* **1993**, *32*, 4693–4697.
- (12) Bitan, G.; Kirkitadze, M. D.; Lomakin, A.; Vollers, S. S.; Benedek, G. B.; Teplow, D. B. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 330–335.
- (13) Harper, J. D.; Wong, S. S.; Lieber, C. M.; Lansbury, P. T. *Chem. Biol.* **1997**, *4*, 119–125.
- (14) Cherny, I.; Gazit, E. *Angew. Chem., Int. Ed.* **2008**, *47*, 4062–4069.
- (15) Kim, Y. S.; Liu, L.; Axelsen, P. H.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 7720–7725.
- (16) Balbach, J. J.; Petkova, A. T.; Oyler, N. A.; Antzutkin, O. N.; Gordon, D. J.; Meredith, S. C.; Tycko, R. *Biophys. J.* **2002**, *83*, 1205–1216.
- (17) Sachse, C.; Fändrich, M.; Grigorieff, N. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 7462–7466.
- (18) Balbach, J. J.; Ishii, Y.; Antzutkin, O. N.; Leapman, R. D.; Rizzo, N. W.; Dyda, F.; Reed, J.; Tycko, R. *Biochemistry* **2000**, *39*, 13748–13759.
- (19) Petkova, A. T.; Ishii, Y.; Balbach, J. J.; Antzutkin, O. N.; Leapman, R. D.; Delaglio, F.; Tycko, R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16742–16747.
- (20) Lührs, T.; Ritter, C.; Adrian, M.; Riek-Loher, D.; Bohmann, B.; Döbeli, H.; Schubert, D.; Riek, R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17342–17347.
- (21) Nelson, R.; Eisenberg, D. *Curr. Opin. Struct. Biol.* **2006**, *16*, 260–265.
- (22) Tycko, R. *Curr. Opin. Struct. Biol.* **2004**, *14*, 96–103.
- (23) Ma, B.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14126–14131.
- (24) Itoh, S. G.; Okamoto, Y. *J. Phys. Chem. B* **2008**, *112*, 2767–2770.
- (25) Klimov, D. K.; Thirumalai, D. *Structure* **2003**, *11*, 295–307.
- (26) Favrin, G.; Irbäck, A.; Mohanty, S. *Biophys. J.* **2004**, *87*, 3657–3664.

- (27) Zheng, J.; Jang, H.; Ma, B.; Tsai, C.-J.; Nussinov, R. *Biophys. J.* **2007**, *93*, 3046–3057.
- (28) Zheng, J.; Jang, H.; Ma, B.; Nussinov, R. *J. Phys. Chem. B* **2008**, *112*, 6856–6865.
- (29) Cerf, E.; Sarroukh, R.; Tamamizu-Kato, S.; Breydo, L.; Derclaye, S.; Dufrière, Y. F.; Narayanaswami, V.; Goormaghtigh, E.; Ruyschaert, J.-M.; Raussens, V. *Biochem. J.* **2009**, *421*, 415–423.
- (30) Buchete, N.-V.; Tycko, R.; Hummer, G. *J. Mol. Biol.* **2005**, *353*, 804–821.
- (31) Zhao, J.-H.; Liu, H.-L.; Liu, Y.-F.; Lin, H.-Y.; Fang, H.-W.; Ho, Y.; Tsai, W.-B. *J. Biomol. Struct. Dyn.* **2009**, *26*, 481–490.
- (32) Masman, M. F.; Eisel, U. L. M.; Csizmadia, I. G.; Penke, B.; Enriz, R. D.; Marrink, S. J.; Luiten, P. G. M. *J. Phys. Chem. B* **2009**, *113*, 11710–11719.
- (33) Katz, E.; Willner, I. *Electroanalysis* **2003**, *15*, 913–947.
- (34) Heller, I.; Janssens, A. M.; Mannik, J.; Minot, E. D.; Lemay, S. G.; Dekker, C. *Nano Lett.* **2007**, *8*, 591–595.
- (35) Zhou, X.; Moran-Mirabal, J. M.; Craighead, H. G.; McEuen, P. L. *Nat. Nanotechnol.* **2007**, *2*, 185–190.
- (36) Salford, L. G. B. A. E.; Eberhardt, J. L.; Malmgren, L.; Persson, B. R. R. *Environ. Health Perspect.* **2003**, *111*, 881.
- (37) de Pomerai, D. I.; Smith, B.; Dawe, A.; North, K.; Smith, T.; Archer, D. B.; Duce, I. R.; Jones, D.; Candido, E. P. M. *FEBS Lett.* **2003**, *543*, 93–97.
- (38) Benabid, A. L. *Curr. Opin. Neurobiol.* **2003**, *13*, 696–706.
- (39) Blanchard, B. J.; Thomas, V. L.; Ingram, V. M. *Biochem. Biophys. Res. Commun.* **2002**, *293*, 1197–1203.
- (40) Tyner, K. M.; Kopelman, R.; Philbert, M. A. *Biophys. J.* **2007**, *93*, 1163–1174.
- (41) Toschi, F.; Lugli, F.; Biscarini, F.; Zerbetto, F. *J. Phys. Chem. B* **2009**, *113*, 369–376.
- (42) Budi, A.; Legge, F. S.; Treutlein, H.; Yarovsky, I. *J. Phys. Chem. B* **2005**, *109*, 22641–22648.
- (43) Budi, A.; Legge, F. S.; Treutlein, H.; Yarovsky, I. *J. Phys. Chem. B* **2007**, *111*, 5748–5756.
- (44) Budi, A.; Legge, F. S.; Treutlein, H.; Yarovsky, I. *J. Phys. Chem. B* **2008**, *112*, 7916–7924.
- (45) Cramer, T.; Zerbetto, F.; García, R. *Langmuir* **2008**, *24*, 6116–6120.
- (46) Cramer, T.; Steinbrecher, T.; Koslowski, T.; Case, D. A.; Biscarini, F.; Zerbetto, F. *Phys. Rev. B* **2009**, *79*, 155316.
- (47) Xu, D.; Phillips, J. C.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 12108–12121.
- (48) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (50) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (51) Petersen, H. G. *J. Chem. Phys.* **1995**, *103*, 3668–3679.
- (52) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Chem.* **1977**, *23*, 327–341.
- (53) MacKerell, A. D.; Bashford, D.; Bellott, Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (54) Jorgensen, W. L.; Chandrasekhar, J. r.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (55) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- (56) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (57) Frishman, D. A. P. *Proteins: Struct. Funct. Genet.* **1995**, *23*, 566–579.
- (58) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (59) Tsemekhman, K.; Goldschmidt, L.; Eisenberg, D.; Baker, D. *Protein Sci.* **2007**, *16*, 761–764.
- (60) Bitan, G.; Vollers, S. S.; Teplow, D. B. *J. Biol. Chem.* **2003**, *278*, 34882–34889.
- (61) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17345–17350.
- (62) Röhrig, U. F.; Laio, A.; Tantalò, N.; Parrinello, M.; Petronzio, R. *Biophys. J.* **2006**, *91*, 3217–3229.

CT1001335

Heterogeneous Path Ensembles for Conformational Transitions in Semiatomistic Models of Adenylate Kinase

Divesh Bhatt and Daniel M. Zuckerman*

*Department of Computational and Systems Biology, University of Pittsburgh,
Pittsburgh, Pennsylvania 15213*

Received April 27, 2010

Abstract: We performed “weighted ensemble” path-sampling simulations of adenylate kinase, using several semiatomistic protein models. The models have an all-atom backbone with various levels of residue interactions. The primary result is that full statistically rigorous path sampling required only a few weeks of single-processor computing time with these models, indicating the addition of further chemical detail should be readily feasible. Our semiatomistic path ensembles are consistent with previous biophysical findings: the presence of two distinct pathways, identification of intermediates, and symmetry of forward and reverse pathways.

1. Introduction

Fluctuations and conformational changes are of extreme importance in biomolecules.¹ For example, most enzymes show distinctly different conformations in the apo and the holo forms.² Conformational transitions are also typical in nonenzymatic binding proteins,³ and of course are intrinsic to motor proteins.

The fundamental biophysics of conformational transitions in biomolecules is contained in the ensemble of paths – i.e., trajectories in configurational space – defining the transitions. Such path ensembles contain the information about the relevant “mechanisms” for transitions, including possible intermediates (see, for example, discussions by E and vanden Eijnden⁴ and by Hummer⁵). In addition, the transition rates can only be calculated accurately from a path ensemble, which implicitly accounts for all barriers and recrossings.⁶ From a computational point of view, such path ensembles are difficult to obtain due to rugged energy landscapes and the time scales involved.^{6–15} Multiple local minima and/or channels dramatically increase the computational effort required. To put the difficulty of path sampling in perspective, note that equilibrium sampling of fully atomistic models of large biomolecules is not typically feasible.¹⁶ Thus, path sampling using detailed atomistic models for all but the smallest systems is difficult, even with potentially efficient methods developed specifically for path sampling. A number of groups have reported atomistic path sampling studies for

small systems.^{17–19} Radhakrishnan and Schlick²⁰ reported path sampling studies of atomistic DNA polymerase using initial paths generated by targeted molecular dynamics.^{21–23}

Less computationally expensive approaches to determining atomistic paths are available, including targeted and steered molecular dynamics,^{21–23} and more rigorous “nudged elastic band”,^{24–27} and related approaches.^{28–31} However, all these methods yield only a single path or a handful, and not the ensemble required for a correct thermal/statistical description. Specifically, fluctuations in pathways, the possibility of multiple pathways (path heterogeneity), and possible recrossings typically are not accounted for in these approaches.

Coarse-grained (CG) models, on the other hand, permit an alternative strategy for statistical path sampling.^{32–34} Although CG models omit chemical detail, they can be sampled significantly faster than fully atomistic models, and, thus, such models are quite attractive for path sampling studies. For example, Zhang et al.³⁵ showed that a simple α -carbon model of calmodulin can be fully path sampled using the weighted ensemble path sampling method.³⁶ Because full path sampling in this model required only a few weeks of single-processor computing, it is evident that better models and/or larger systems could be studied. Network models have also been used to study conformational transitions.^{37,38}

In this Article, we report weighted-ensemble (WE)^{35,36} path sampling studies of adenylate kinase, which represent improvements over previous work³⁵ in several ways. (i) At 214 residues, adenylate kinase is triple the size of the

* Corresponding author e-mail: ddmmzz@pitt.edu.

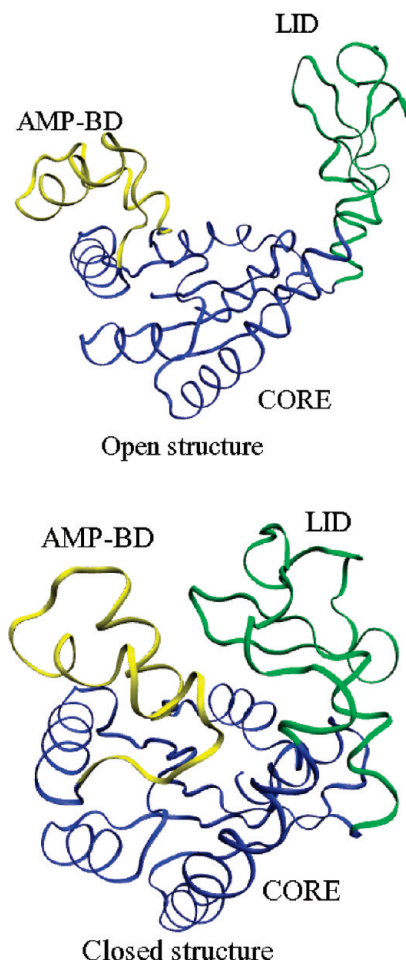
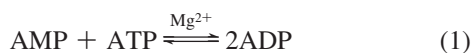


Figure 1. Crystal structures of adenylate kinase. The blue segment represents the core of the protein (CORE), the yellow segment is the AMP binding domain (BD), and the green segment is the flexible lid (LID). The left figure is the Apo (or Open) form, and the right figure is the holo (Closed) form.

calmodulin domain previously path sampled using WE.³⁵ (ii) Our models now include significant atomic details, as explained below. (iii) We examine a series of models to test the sensitivity of the path ensemble to the chosen interactions and parameters. (iv) We investigate symmetry, based on our recent formal derivation,³⁹ between forward and reverse transitions.

Adenylate kinase (Adk) is an enzyme that catalyzes phosphate transfer between AMP and ATP via



and thus helps to regulate the relative amounts of cellular energetic units.^{40–42} The crystal structure of Adk for *E. coli* is available in several conformations. Its native apo form (Protein Data Bank code 4AKE⁴³) is shown in Figure 1a. In the figure, the blue segments represent the core (CORE), the yellow segment represents the AMP binding domain (BD), and the green segment represents the flexible lid (LID). Upon ligand binding, the enzyme closes over the ligands. The crystal structure (1AKE)⁴⁴ of the holo form of the enzyme obtained in complex with a ligand that mimics both AMP

and ATP is shown in Figure 1b. Clearly, in the apo form, the enzyme shows an open structure (that we denote as Open in this Article), and in the holo form, it is closed (denoted by Closed throughout).

Adk has been studied previously via computational methods using both coarse-grained models and fully atomistic simulations. Coarse-grained models used to study transition pathways for Adk have, primarily, utilized network models.^{37,38,45,46} In these methods, the fluctuations in proteins are represented by harmonic potentials, and the deformations due to these fluctuations are used to estimate the free energy in the basins (end states and/or multiple basins). Subsequently, a minimum energy path is calculated to characterize the transition. Coarse-grained models have also been used with molecular dynamics simulations to determine transition paths.^{47,48}

A few groups have also studied conformational fluctuations in Adk using atomistic models. Arora and Brooks⁴⁹ performed atomistic (with implicit solvent) umbrella sampling molecular dynamics (MD) simulations along an initial minimum energy path suggested by “nudged elastic band”. Kubitzki and de Groot⁵⁰ performed replica exchange MD for atomistic Adk to increase conformational sampling of adenylate kinase and observed both Open and Closed conformers; however, a true path ensemble is not obtained from replica exchange. In other work, fully atomistic MD on the two end structures has been performed to observe fluctuations in the two ensembles,^{40,51} but direct conformational transitions were not observed. Beckstein et al.⁵² performed atomistic path sampling studies for conformational transitions in Adk using dynamic importance sampling⁵³ with an approximate biasing scheme.⁵⁴

In the present study, we use semiatomistic models to improve chemical accuracy as compared to typical coarse-grained models while still performing high-quality path sampling. In our models, the backbone is fully atomistic to provide chemically realistic geometry, which is often absent from bead-based models. Inter-residue interactions are modeled at a coarse-grained level via the commonly used double-Gō potentials^{34,46,55,56} that (meta)stabilize two crystal structures. Additionally, one of the models uses residue-specific interactions to probe the effect of such interactions. We use a library-based Monte Carlo (LBMC) scheme to perform sampling.⁵⁷ LBMC was previously developed in our group and shown to facilitate the use of semiatomistic models of the type used here.⁵⁷

Transitions between the Open and the Closed states (both directions) are studied with the weighted-ensemble (WE) path-sampling method³⁶ that has been previously been studied to study folding of proteins,⁵⁸ protein dimerization,⁵⁹ and conformational transitions in an α -carbon model of calmodulin.³⁵ WE was shown to promote efficient path sampling of conformational transitions in the purely α -carbon model of calmodulin.³⁵ Additionally, WE is statistically exact: it preserves natural system dynamics, resulting in an unbiased path ensemble.⁶⁰ Additional strengths of WE include its ability to find multiple pathways,⁶¹ the simultaneous determination of the path ensemble and transition rate,^{35,62} the ease of implementing it at a scripting level, and its natural parallel structure.

Biophysically, we focus on heterogeneity of the path ensemble (multiple pathways) and the forward-reverse “symmetry” of the ensemble. It is possible that evolution has favored the fine-tuned precision of a single pathway in some systems, but the “robustness” of alternative pathways in other cases. Although our semiatomistic models preclude biochemically precise conclusions, high-quality path sampling permits a complete description of a model system.

The goal of this work, in summary, is to explore the application of statistically rigorous path sampling to conformational transitions in proteins using semiatomistic models of a moderately large protein. Methodologically, we want to test whether such models can be fully path sampled. Biophysically, we explore several outstanding issues in computational studies of conformational transitions in Adk. Although two different pathways for conformational transitions in Adk have been identified by models of varying level of details,^{42,47,48} it has variously been suggested that a different pathway is dominant in the forward and reverse directions⁴⁷ versus a more symmetric scenario in which each pathway occurs with same probability in the two directions.⁴⁸ With our path sampling study, we can address this issue. We can also validate our path ensembles by comparison to experimental structures.

This Article is organized as follows. First, in Section 2 we discuss the models we use to depict the protein. Section 3 then describes the method to generate the ensemble of pathways. In Section 4, we present results for transitions in both directions for all the three models we used. We discuss the results, efficiency, and future models in Section 5, with conclusions given in Section 6.

2. Semiatomistic Models

We use three semiatomistic models, expanding on our previous work.⁵⁷ In all the models, the backbone is represented in full atomistic detail, using the three residues alanine, glycine, and proline.⁵⁷ All intraresidue interactions are included explicitly, using the OPLSAA all-atom force field. Both the intraresidue interaction energies and the configurations are stored in libraries as described previously.⁵⁷ In brief, we note that libraries of the three types of residues are pregenerated according to the Boltzmann distribution at 300 K, and alanine is used to represent the backbone of all residues besides glycine and proline (a simplification motivated by the similarity of Ramachandran maps for the residues).⁶³ Ligands are not modeled explicitly in this path sampling study.

The differences in the three models lie in the treatment of inter-residue interactions: two of the models use only double-G \bar{o} interactions at backbone α carbons, whereas one model uses both double-G \bar{o} and residue-specific interactions. Complete information is given below.

All three semiatomistic models employ double-G \bar{o} interactions. Following ref 34 for each of the two crystal structures, residues pairs with α carbons less than 8 Å apart are considered native contacts. In the G \bar{o} energy of an arbitrary configuration, every native contact from the Open form is assigned an energy of $-\varepsilon$, whereas those exclusively found in the Closed form are scaled to be $-e_{\text{scale}}\varepsilon$. G \bar{o} interactions

do not distinguish between different types of residues except in terms of size. This double-G \bar{o} potential between two residues i and j with α carbon distance r_{ij} is given by

$$u^{\text{G}\bar{o}}(r_{ij}) = \begin{cases} \infty, & r_{ij} < r_{ij}^{\text{X}}(1 - \delta) \\ -\varepsilon_{\text{X}}, & r_{ij}^{\text{X}}(1 - \delta) \leq r_{ij} < r_{ij}^{\text{X}}(1 + \delta) \\ 0.3\varepsilon, & r_{ij}^{\text{X}}(1 + \delta) \leq r_{ij} < r_{ij}^{\text{Y}}(1 + \delta) \\ -\varepsilon_{\text{Y}}, & (*)r_{ij}^{\text{Y}}(1 - \delta) \leq r_{ij} < r_{ij}^{\text{Y}}(1 + \delta) \\ 0, & r_{ij} \geq r_{ij}^{\text{Y}}(1 + \delta) \end{cases} \quad (2)$$

where r_{ij}^{X} and r_{ij}^{Y} are the native distances in the two crystal structure ordered such that $r_{ij}^{\text{X}} < r_{ij}^{\text{Y}}$ (X and Y equate to Open or Closed), and δ is a well-width parameter chosen to be 0.05. If X equals Open and Y equals Closed, $\varepsilon_{\text{X}} = \varepsilon$ and $\varepsilon_{\text{Y}} = e_{\text{scale}}\varepsilon$, and vice versa. In the case of overlapping square wells, $r_{ij}^{\text{X}}(1 + \delta) > r_{ij}^{\text{Y}}(1 - \delta)$, the 0.3 ε barrier in the middle does not exist, and the lower limit of the inequality marked with (*) is replaced by $r_{ij}^{\text{X}}(1 + \delta)$.

The total G \bar{o} interactions are therefore

$$U_{\text{G}\bar{o}}(e_{\text{scale}}) = \sum_{i < j} u^{\text{G}\bar{o}}(r_{ij}) \quad (3)$$

For model 1, we have

$$U_1 = U_{\text{G}\bar{o}}(e_{\text{scale}} = 1) \quad (4)$$

The motivation behind using such a double G \bar{o} potential is that the two “end” crystal structures are presumed stable, and the double-G \bar{o} protocol guarantees that bistability. Such double G \bar{o} interactions have been used to probe the biophysics of several systems.^{34,46,55,56}

2.1. Model 1: Pure Double G \bar{o} with Energy Symmetry. Previous path sampling studies of proteins were limited to smaller systems and/or simpler models. We, therefore, first study whether the simplest model within the semiatomistic framework can be fully path sampled. Our Model 1 omits most chemical details and uses only symmetric double-G \bar{o} interactions as given in eq 3. That is, native contacts in the Open and the Closed structure are treated identically ($e_{\text{scale}} = 1$).

Because model 1 is a pure G \bar{o} model, the temperature is specified in units of the well depth of G \bar{o} interactions, ε . We choose the temperature as the highest at which the two experimental crystal structures are stable. We therefore performed a series of Monte Carlo simulations, as described below, at various temperatures. At $T = 0.75\varepsilon/k_{\text{B}}$, both structures melted, but both remained (meta)stable at $T = 0.7\varepsilon/k_{\text{B}}$. Thus, for model 1, all subsequent equilibrium and path sampling simulations were performed at $T = 0.7\varepsilon/k_{\text{B}}$.

2.2. Model 2: Double G \bar{o} with Residue-Specific Interactions. Our second model adds chemical detail, both to improve upon the simplicity of model 1, and to provide a way to check the sensitivity of our results to modeling choices. Model 2 includes atomistic backbone hydrogen bonding, Ramachandran propensities, and residue-specific contact interactions, as detailed below. Because these interactions are implemented as short ranged, Model 2 is

only about 30% slower than model 1, based on wall-clock time per MC step. G \bar{o} interactions are again symmetric, with $e_{\text{scale}} = 1$.

Our semiatomistic LBMC platform makes the inclusion of additional interactions straightforward. Because the backbone is modeled atomistically, backbone–backbone hydrogen bonding is easily incorporated, as described below. However, due to the absence of explicit side chains in the present implementation, residue-specific chemical interactions can only be incorporated at a coarse-grained level. We use residue-specific contact interactions based on the work of Miyazawa and Jernigan (MJ),^{64,65} as discussed below. Specifically, we use the potential energy

$$U_2 = U_{\text{G}\bar{o}}(e_{\text{scale}} = 1) + U_{\text{HB}} + U_{\text{Rama}} + U_{\text{MJ}} \quad (5)$$

where U_{HB} is the hydrogen-bonding potential, U_{Rama} is the potential due to Ramachandran propensities, and U_{MJ} is the residue-specific potential based on MJ interactions. These terms are described below.

Hydrogen bonding for the backbone–backbone interactions is modeled atomistically, but with simplifications appropriate to the otherwise coarse-grained nature of our models. Specifically, we use ordinary Coulomb interactions with OPLSAA charges between the backbone CO and NH groups if the O–H distance less than 2.5 Å. The cutoff was chosen as the distance after which dipole interactions are significantly attenuated. Following previous studies that suggest a dielectric constant of 2–5 inside a protein, we use a value of 3. The use of physical charge and distance units in the hydrogen-bonding interactions allows physical temperature units in the simulation (instead of merely being in relation to the G \bar{o} well depth).

Ramachandran propensities were included via the term U_{Rama} , which is based on a potential of mean force obtained by calculating the distribution of Φ – Ψ dihedral angles in acetaldehyde-alanine-*n*-methylamide using OPLSAA force field. This distribution was tabulated from a Langevin dynamics simulation at 300 K using the GBSA implicit solvent model in Tinker software package.

The construction of MJ-type interactions required some care. Several variants of the original MJ interaction values have been utilized in the literature (such as scaling the MJ interactions energies, as well as shifting),^{66,67} due to the fact that MJ values are based on folded protein data and are not directly applicable for unfolded states. We follow the suggestion of Jernigan and Bahar⁶⁶ to mix MJ values of Table 5 and Table 6 (numbering as in the original MJ paper,⁶⁴ with updated values as in ref 65) so that the residue-specific interactions are modeled as $x \times$ (Table 5) + $(1 - x) \times$ (Table 6). We chose $x = 0.05$ to ensure that the residue-specific interactions are a significant perturbation of the double G \bar{o} interactions.

To make the crystal structures (meta)stable, we “titrated in” double G \bar{o} interactions (ϵ), until bistability was observed at 300 K. Because, as described, hydrogen bonding introduces physical units into model 2, the units of G \bar{o} well depth, ϵ , are also physical. We found that at $\epsilon/k_B = 400$ K, both the structures remained (meta)stable.

2.3. Model 3: Pure Double G \bar{o} without Energy Symmetry. Finally, to facilitate the generation of large path ensembles in both the Open-to-Closed and the Closed-to-Open directions, we also constructed a third model. The new model is designed to overcome the somewhat artifactual overstabilization of the Closed states in models 1 and 2 (see results below in section 4). In brief, our 8 Å cutoff permits significantly more contacts in the Closed state, implicitly but artificially mimicking the presence of ligands in Models 1 and 2. This implicit presence of ligands interferes somewhat with our goal of modeling the ligand-free opening and closing of the enzyme.

In model 3, therefore, we attempt to make the Open and Closed forms of adenylate kinase more comparable in stability. We decrease the strength of G \bar{o} interactions specific to the Closed form to one-half of G \bar{o} interactions (i.e., we set $e_{\text{scale}} = 0.5$). Additionally, to focus on the effect of the reduced stability of the Closed form with respect to the Open form, we use only asymmetric double G \bar{o} interactions (and no H-bonding, Ramachandran, or MJ interactions). That is, we set

$$U_3 = U_{\text{G}\bar{o}}(e_{\text{scale}} = 0.5) \quad (6)$$

3. Methods

3.1. Dynamical Monte Carlo. We follow many precedents^{35,68,69} and use “dynamical” MC for the dynamics of our models. Such an approximation to physical dynamics is consistent with our use of simplified models. Specifically, we use the library-based Monte Carlo (LBMC) algorithm,⁵⁷ to propagate the system in both brute-force simulations for generating equilibrium ensembles and path sampling simulations (discussed below in more detail). For both equilibrium and path sampling, the systems always evolve via “natural” LBMC dynamics, and no artificial forces are used to direct conformational transitions, as explained below.

Our LBMC simulations use the same trial moves described in our earlier work.⁵⁷ One flexible peptide plane in the current configuration is swapped with one stored in the library, and a Ψ angle is also displaced by a small amount.

3.2. Path Sampling. In systems with rugged energy landscapes, such as proteins, regular brute-force simulations are not efficient for studying transitions. For this reason, we use the statistically rigorous weighted-ensemble (WE) path sampling method to generate path ensembles of conformational transitions of adenylate kinase between the Open and the Closed states. This method preserves the natural system dynamics⁶⁰ and was used previously to study protein folding,⁵⁸ protein dimerization,⁵⁹ and conformational transitions of calmodulin using an α -carbon model.³⁵ Weighted ensemble studies the probability evolution of trajectories in the configuration space using any underlying system dynamics.⁶⁰ In this work, we use the WE method to study transitions in more detailed models to evaluate the effect of increasing chemical detail on the transitions, and to study questions of symmetry in forward and reverse directions.

The procedure to use weighted-ensemble path sampling to study conformational transitions is described in detail elsewhere,^{35,36,60} and here we describe our simple imple-

mentation briefly. Prior to beginning the simulations, we divide a one-dimensional projection of the configurational space (i.e., the DRMS from the target structure in the present study) into a number of bins. The DRMS is a “progress coordinate” or “order parameter” and is not necessarily the reaction coordinate. The progress coordinate roughly keeps track of the progress to the target state: the DRMS of structures close to the target state is necessarily small. It is also possible to use multidimensional or adaptively changing progress coordinates,^{35,60} but that was not found necessary here.

In the weighted-ensemble method, an evolving set of trajectories and their probabilities are tracked. Procedurally, several independent trajectories are started in an initial configuration and run for a short time interval τ (consisting of multiple simulation steps) with natural dynamics. At the end of each τ interval, the progress of the trajectories along the progress coordinate is noted (i.e., into which bin along the progress coordinate each trajectory ends). Once bins are tabulated after each τ , trajectories are “split” (replicated with divided probability) and combined. This keeps the same number of trajectories in each occupied bin, prunes low-weight trajectories, and splits trajectories with high probability. This splitting and combining of simulations is performed statistically as discussed elsewhere.^{35,36,60} The probability remains normalized, and all probability flows can be measured.

The full details of our WE simulations are as follows. We employ LBMC to describe the natural system dynamics. We utilize 25 bins between the two states, with 20 simulations (trajectories) in each occupied bin. The end state is defined as being at a DRMS ≤ 1.5 Å from the target crystal structure, a definition used in both directions. Using this definition of the end state, we calculate the probability flux of trajectories entering the target state at the end of each τ .

It should be noted that value of the probability flux into either state, and hence the rate, depends upon the precise definitions of the two states. Although probability flows are good indicators of sampling quality, precise numerical values of the rates are not of great interest in our study of simple models with Monte Carlo dynamics. In this work, we are interested in the path ensembles and not the rates.

4. Results

4.1. Static Analysis of Conformational Differences. For reference, we first analyze the conformational differences between the two end-state static crystal structures to quantify the observed differences in the Open and Closed configurations of Figure 1. Figure 2 shows the α -carbon distance difference map of pairs of residues in the Open and the Closed crystal structures. A large positive value implies that a pair is farther apart in the Open structure than in the Closed structure, whereas a negative value is the opposite. By construction, the figure is symmetric about the diagonal. A few features of the two structures easily emerge from Figure 2. The inter-residue distances for most of the residue pairs are very similar in the two crystal structures. The major differences are that the distances in the Closed structure

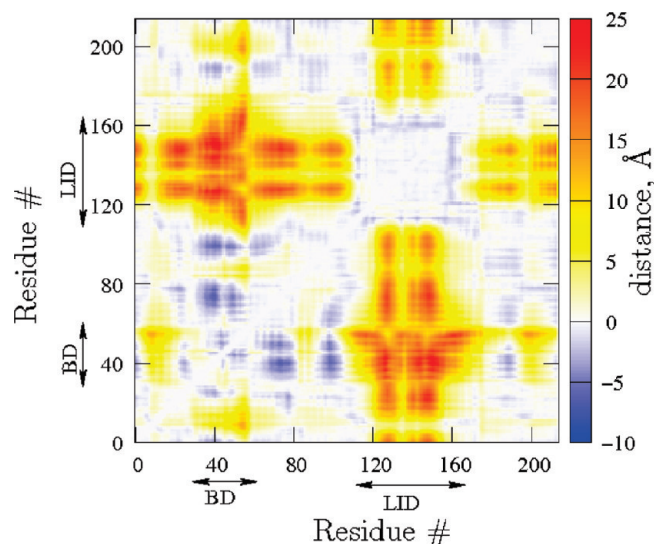


Figure 2. Map of the difference in the inter-residue distances of the Open from the Closed crystal structure. The map is, by definition, symmetric about the diagonal. The white space on the diagonal just implies that we do not calculate distances between residue pairs less than 2 residues apart along the chain. The residues corresponding to the LID and BD are labeled, and the unlabeled residues form the CORE.

between residues labeled LID (114–164) are closer to BD (31–60) and several residues of CORE are smaller than the corresponding distances in the Open structure. Thus, Figure 2 quantifies Figure 1.

From Figures 1 and 2, it is clear that the structural change that characterizes the transition between the Open and the Closed structure is fairly straightforward: the LID and the BD close, and the rest of the protein remains fairly unchanged. Following Figures 1 and 2, for the path sampling studies presented shortly, we monitor inter-residue distances between two pairs of residues: residues 56 (GLY) and 163 (THR), which report on the BD-LID proximity, as well as residues 15 (THR) and 132 (VAL), which report on the CORE-LID proximity. In the Closed structure, $d_{56,163}^c = 4.9$ Å and $d_{15,132}^c = 6.3$ Å. On the other hand, in the Open structure, $d_{56,163}^o = 23.6$ Å and $d_{15,132}^o = 17.8$ Å. Thus, the relation between the CORE and LID is monitored, along with that of the LID and BD.

4.2. Brute-Force Equilibrium Sampling. To demarcate the native basins in our analysis of transitions, we first study equilibrium ensembles for the Open and the Closed states of adenylate kinase. Put another way, we want to quantify the size of native-basin fluctuations in our models. Further, we determine whether transition paths can be obtained without the aid of path sampling.

We quantify fluctuations in the equilibrium ensembles in the two basins by using DRMS from the respective crystal structures. Figure 3a shows two sets of DRMS traces for Model 1 for a simulation started from the Open structure: DRMS-from-Open (black line) and DRMS-from-Closed (blue line). Similarly, Figure 3b shows two sets of DRMS traces for Model 1 for a simulation started from the Closed structure: DRMS-from-Closed (black line) and DRMS-from-Open (blue line). Thus, in each panel, the black line

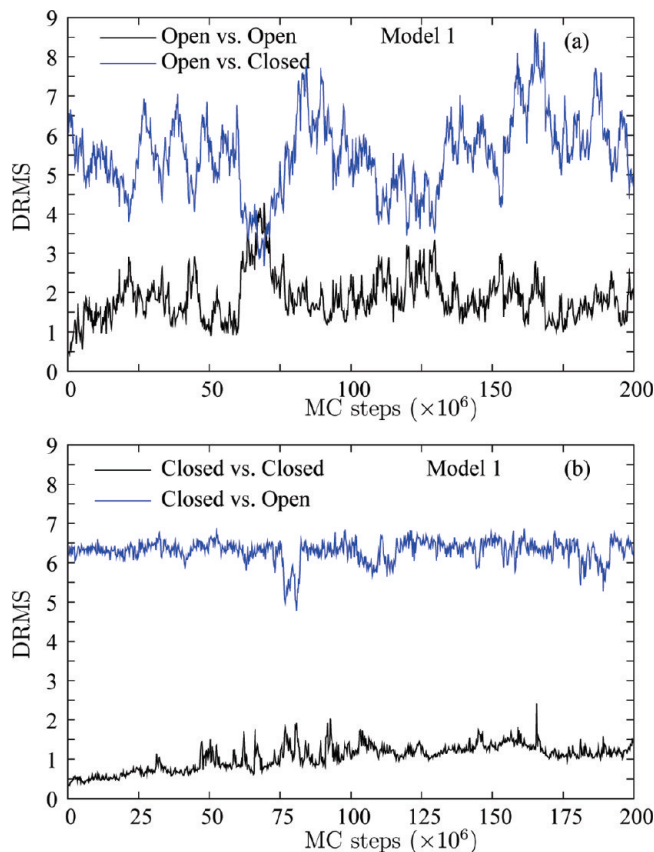


Figure 3. Stability of the native basins in Model 1. The DRMS in Model 1 (pure G \bar{o}) is shown for two simulations (a) starting from the Open structure and (b) starting from the Closed structure. For each simulation, we show the DRMS from the starting structure (black line) and the opposing structure (blue line). Neither simulation shows a transition to the opposing structure, but the scales of fluctuations are very different.

represents DRMS from the starting structure, whereas the blue line represents DRMS from the opposing structure.

A comparison of the two panels of Figure 3 shows that the simulation started from the Open structure shows significantly more fluctuations than the simulation started from the Closed structure. Furthermore, the fluctuations drive the simulation started from the Open structure closer to the Closed structure than vice versa. For example, Figure 3a shows that the simulation started from the Open structure gets to within 3 Å of the Closed structure at approximately 70 million MC steps. On the other hand, the simulation started from the Closed structure (Figure 3b) remains farther from the Open structure.

Most importantly, neither simulation shows a transition to the opposing structure. The DRMS from the opposing structure for a particular simulation is always significantly larger than DRMS values from the starting structure for the other simulation. To elaborate, let us consider the DRMS-versus-Closed structure for the simulation started from the Closed structure (black line in Figure 3b). The fluctuations in DRMS remain less than 1.5 Å in the native basin for the Closed structure. Comparatively, the largest fluctuations in the simulations started from the Open structure bring it only within at most 3 Å of the Closed structure (blue line in Figure 3a). That is, the opposing native basin is never reached.

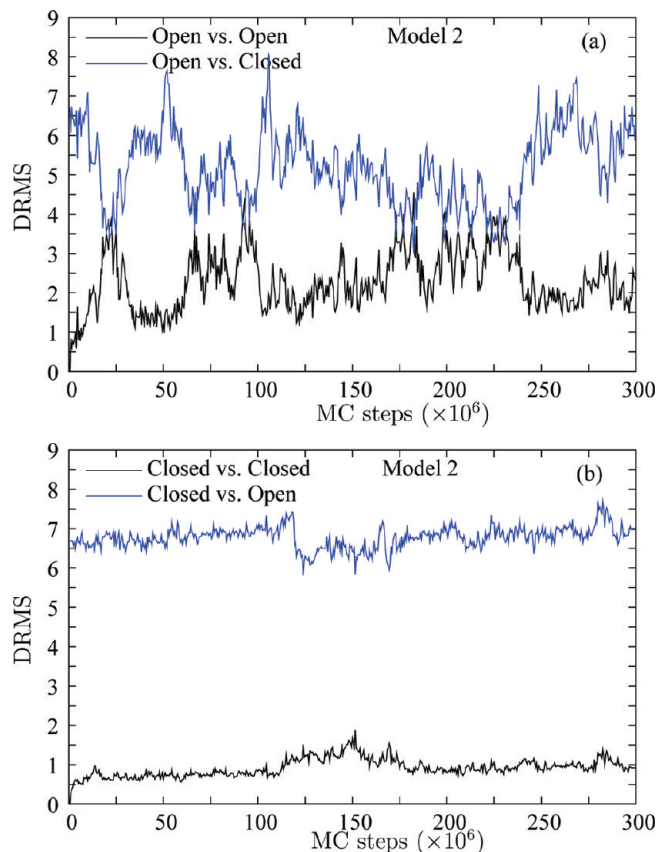


Figure 4. Stability of the native basins in Model 2. The DRMS in Model 2 (G \bar{o} , H-bonding, Ramachandran propensities, and MJ-type residue-specific interactions) is shown for two simulations (a) starting from the Open structure and (b) starting from the Closed structure. For each simulation, we show the DRMS from the starting structure (black line) and the opposing structure (blue line). Neither simulation shows a transition to the opposing structure, but the scales of fluctuations are very different.

We mention that all the DRMS values plotted in Figure 3a and b are based on the first 200 residues. This is because the 14 tail residues, which form a helical segment, are very flexible and the helix unravels in either structure at a much lower temperature than the stable part of the protein. Thus, Figure 3 focuses on the rest of the protein. Additionally, although we show results for $T = 0.7$ here, simulations at lower temperatures also give qualitatively similar results.

We perform an analogous fluctuation analysis for Model 2, which incorporates backbone hydrogen-bonding interactions, Ramachandran propensities, and some residue specificity via MJ-type interactions. Figure 4a shows the DRMS (of the first 200 residues) from the Open (black line) and Closed (blue line) structures for a simulation started from the Open structure. Similarly, Figure 4b shows the DRMS traces for a simulation started from the Closed structure. Again, we observe results similar to those for Model 1: the fluctuations in the Open ensemble are larger than in the Closed ensemble, and no transition to the opposing structure is obtained in either simulation.

4.3. Path Sampling: Models 1 and 2. Because of the inability of brute-force simulations to show transitions, we use weighted-ensemble path sampling to generate an en-

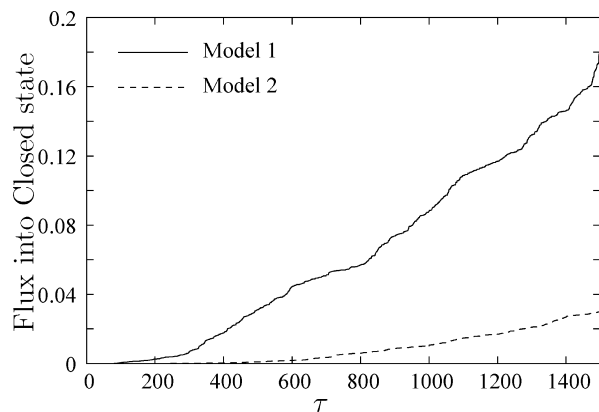


Figure 5. Comparison of the probability fluxes into the Closed state for two models. The fluxes from a pure double-Gō system (Model 1, solid line) and a system with considerable chemical specificity (Model 2, dashed line) are plotted as functions of WE time intervals (each τ interval is 2000 LBMC steps). In both cases, the fluxes reach approximately linear regimes, suggesting transient effects have attenuated.

semble of transition pathways with the aim of assessing path heterogeneity. In particular, we examine transitions in both directions for all three models.

4.3.1. Transition from Open to Closed State. We first check whether our path sampling is sufficient by monitoring the flux into the target state. Figure 5 plots the WE results for probability fluxes obtained into the Closed state for both Models 1 and 2. The “time” axis is merely the number of τ intervals (where one interval contains 2000 LBMC steps). In both models, the fluxes reach linear regimes, indicating that the observed transitions are not merely due to initial fast trajectories and the path ensemble is appropriately sampled.

The sensitivity to the models is also apparent in the fluxes shown in Figure 5: Model 2 (which includes hydrogen-bonding, Ramachandran propensities, and MJ-type residue-specific interactions) has a smaller flux into the Closed state than does Model 1. Residue-specific interactions are expected to roughen the energy landscape, consistent with the observed slowing of transition dynamics. However, the possible change in the Open state basin stability due to addition of these interactions is convoluted with the roughening of the landscape.

We further study the path ensemble by examining individual trajectories. Figure 6 shows, for Model 1, the DRMS from the Closed structure for four typical transitions started in the Open state as a function of time (total number of LBMC steps) obtained via WE path sampling. In contrast with the brute-force simulation in Figure 3a, each trajectory in Figure 6 gets to the Closed state (defined to be within a DRMS of 1.5 Å from the Closed structure). Although the trajectories arrive at the target state with different weights, the ones shown in the figures above are obtained after a simple resampling procedure⁷⁰ and, thus, represent trajectories that arrive with relatively large probabilities. Resampling is a statistically rigorous procedure to prune an ensemble.⁷⁰ In our resampling scheme, a trajectory arriving at the target with weight w is kept with a probability w/w_{\max} .

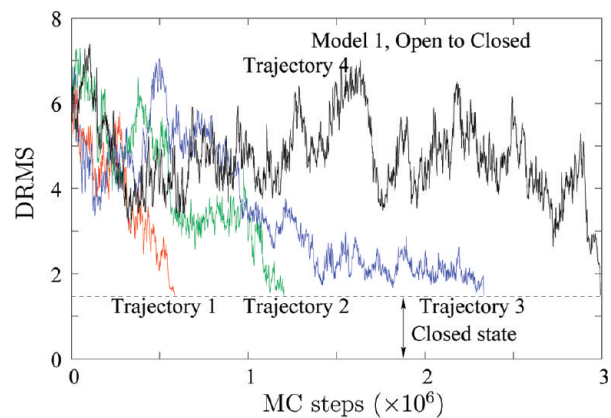


Figure 6. Open-to-Closed transitions observed for Model 1. Four typical DRMS traces are shown, measured from the Closed crystal structure. The Closed basin is delimited by a DRMS of 1.5 Å from the Closed crystal structure.

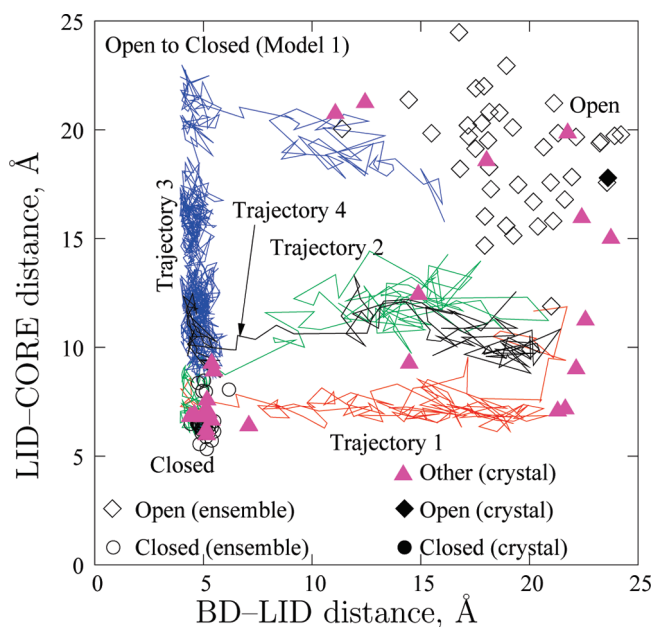


Figure 7. Path heterogeneity in Model 1 transitions. Four typical trajectories for the Open-to-Closed transition (cf., Figure 6) are shown via distances between the CORE and LID (ordinate) and between the BD and the LID (abscissa). Also depicted are experimental configurations, including those from which our double-Gō models were constructed (black filled symbols) and a number of others (pink \blacktriangle), which are tabulated in the Supporting Information. Open symbols depict the equilibrium ensembles obtained from simulation. The “dwell” times for the trajectories in the Open state have been excluded for clarity.

For trajectories that begin transitions at larger times (such as trajectory 4 in Figure 6), a significant amount of time is spent in regions with large DRMS values from the Closed structure. Thus, in Figure 7 and beyond, we do not show the “dwell time” in the Open state.

To analyze the order of domain closing, and, in particular, to study possible heterogeneity in the path ensemble, we study the four Open-to-Closed trajectories of Figure 6 in more detail. Figure 7 plots the projection of the above four trajectories onto the plane of the BD–LID and LID–CORE distances.

The transition paths traced by the four Model 1 trajectories are significantly different. For trajectory 1, the BD shuts after the flexible LID gets close to the CORE. For the following discussion, we call this pathway as Open-LID-BD-Closed (first the LID relaxes, and then the BD shuts close). On the other hand, trajectory 3 shows a dramatically different behavior: the BD snaps shut before the flexible LID gets closer to the CORE (this pathway is labeled as Open-BD-LID-Closed). The other two trajectories are somewhere between the two extremes.

To quantify heterogeneity in the path ensembles, we compare the ratio of trajectories in the two transition pathways. Specifically, we define a trajectory to follow the Open-LID-BD-Closed (lower right) pathway if it first visits the region $d_{56,163} < 10.0$ Å after last leaving the rectangular Open-state region defined by $d_{56,163} > 10.0$ Å and $d_{15,132} > 10.0$ Å. (As in section 4.1, d_{ij} is the distance between α carbons of residues i and j .) On the other hand, a trajectory follows the Open-BD-LID-Closed (upper left) pathway if it first visits the region $d_{56,163} > 10.0$ Å after last leaving the above Open-state rectangular region. We find that for pen-to-Closed transition using Model 1, approximately 60% of the resampled trajectories follow the pen-BD-LID-Closed pathway (akin to trajectory 3 in Figure 7). The remaining 40% follow the Open-LID-BD-Closed pathway.

The model 1 transitions exhibit good correspondence with experimental “intermediates”. Figure 7 also compares the BD–LID and LID–CORE distances of several experimental crystal structures using triangles. Several experimental crystal structures (1DVR, 2AK3, 1AK2, 2BBW – see Supporting Information) lie along the two transition paths. In particular, structures along the Open-LID-BD-Closed pathway approach 1DVR and 1AK2 structures, whereas structures along the Open-BD-LID-Closed pathway approach 2AK3 crystal structure. The identities of the intermediate structures differ from those found by Beckstein et al.⁵² There are two possible reasons for this discrepancy: we use different models, and the effect of the ratcheting bias used by Beckstein et al.⁵² on the pathways and free-energy profiles (from observed populations) is not obvious. More details on the additional experimental structures chosen are given in the Supporting Information.

Further, we look at a few intermediate structures for these two pathways. Figure 8 shows four intermediates along trajectory 3 of Figure 7. The BD and LID domains near one another before the LID closes. On the other hand, Figure 9 shows four intermediates along trajectory 1 in Figure 7. The closing of the LID, followed by snapping shut of the BD, is clearly visible in the figure. As both Figures 8 and 9 show, the rest of the protein (i.e., the CORE region) maintains a stable shape during the transformation.

To determine the sensitivity of the path ensemble to the model, we similarly analyzed results from Model 2 (which includes hydrogen bonding, Ramachandran propensities, and a level of residue specificity). A similar qualitative picture is obtained for Model 2. Figure 10 plots three of the resampled trajectories from the Open to the Closed structure using Model 2. The “dwell times” in the Open state have been omitted for clarity. Again, the symbols have the same

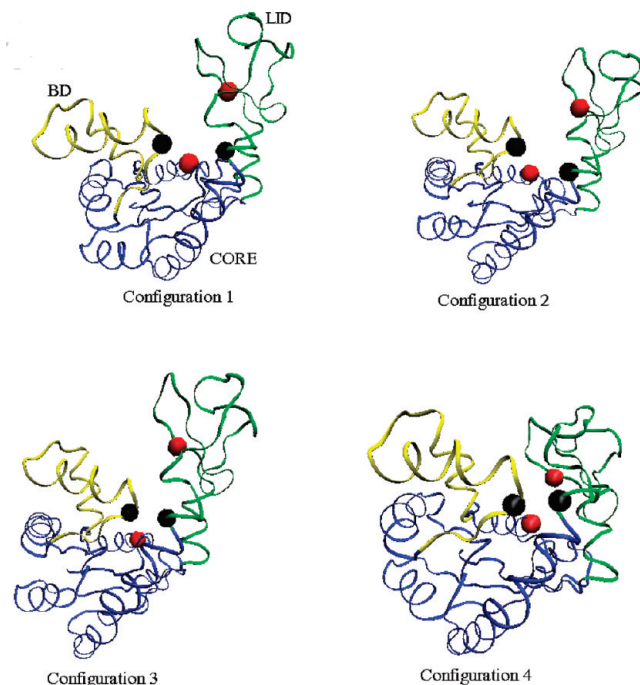


Figure 8. Four time-ordered, representative structures along the Open-BD-LID-Closed path obtained for trajectory 3 in Figure 7 (Model 1). The CORE domain is blue, the BD is yellow, and the LID is green. The configurations correspond to (296 000, 802 000, 822 000, 1 496 000) MC steps, respectively, in trajectory 3 of Figure 6.

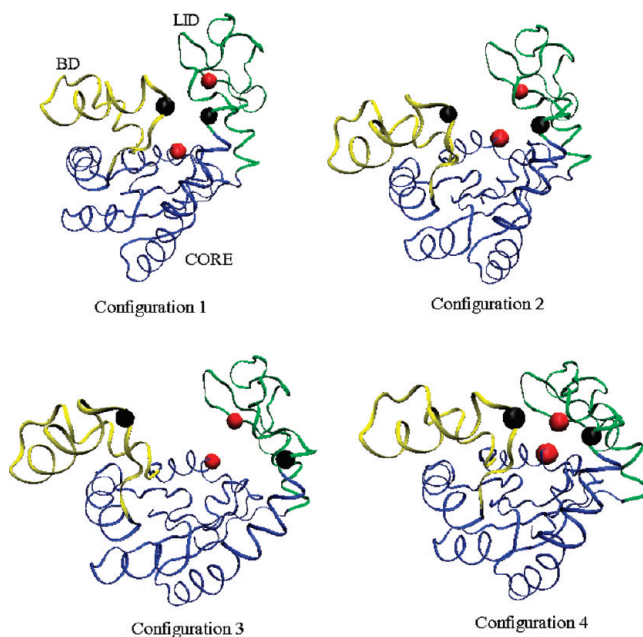


Figure 9. Four time-ordered, representative structures along the Open-LID-BD-Closed path obtained for trajectory 1 in Figure 7 (Model 1). The CORE domain is blue, the BD is yellow, and the LID is green. The configurations correspond to (128 000, 206 000, 306 000, 480 000) MC steps, respectively, in trajectory 1 of Figure 6.

meaning as in Figure 7 (except that Open symbols represent the equilibrium fluctuations obtained using model 2). The transition from the Open-to-Closed structure again takes place by the two pathways: Open-BD-LID-Closed and Open-LID-BD-Closed. The ratio of paths in the two pathways is the

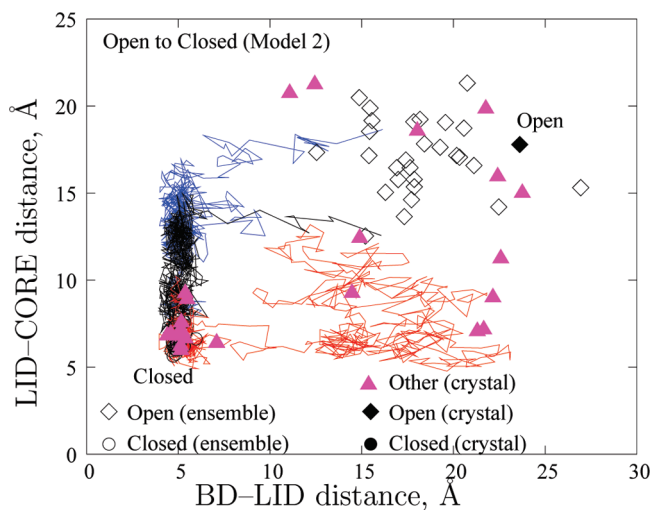


Figure 10. Path heterogeneity in Model 2. Three typical trajectories for the Open-to-Closed transition are shown via the distance between the CORE and LID (ordinate) versus the distance between the BD and the LID (abscissa). Also depicted are experimental configurations, including those from which our double-G \bar{o} models were constructed (black filled symbols) and a number of others (pink \blacktriangle), which are tabulated in the Supporting Information. Open symbols depict the equilibrium ensembles obtained from simulation. The “dwell” times for the trajectories in the Open state have been excluded for clarity.

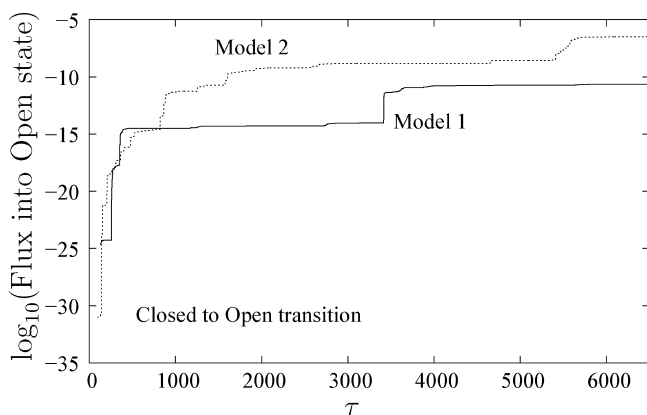


Figure 11. A comparison of probability fluxes into the Open state for two models as a function of WE time increment (each τ is 2000 LBMC steps). The solid line is for Model 1, whereas the dashed line is for Model 2. The log scale emphasizes the small amount of flux in Closed-to-Open direction for both the models, as compared to fluxes in the reverse direction.

same as that for Model 1, and several additional experimental structures show BD–LID and LID–CORE distances similar to structures along the two types of transition pathways.

4.3.2. Transition from Closed to Open State. We also studied “reverse” transitions, from the Closed to the Open state. Figure 11 shows the flux into C as a function of “time” for both Models 1 and 2. As compared to Figure 5 for the transition from the Open to the Closed state, the flux into the Open state is several orders of magnitude lower. This observation mirrors the previously described larger fluctuations in the Open state ensemble. Flux into the Open state for Model 2 with residue-specific chemistry is higher than for Model 1, despite the expected roughening of the energy

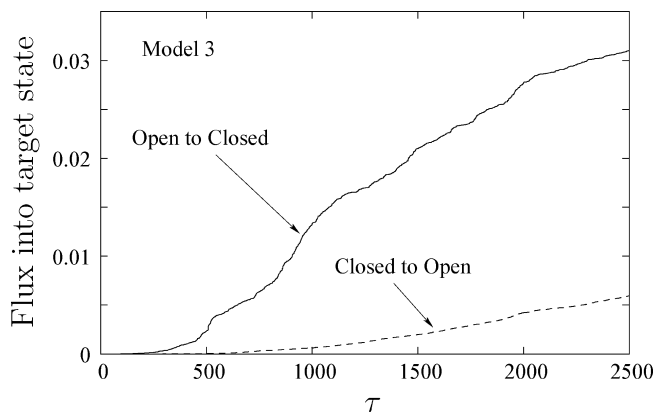


Figure 12. Probability flux in either direction for Model 3 as a function of WE time increment (each τ is 2000 LBMC steps). For this model, the Closed-to-Open flux is of the same order of magnitude as in the reverse direction.

landscape. This necessarily reflects a free energy shift, suggesting MJ interactions destabilize the Closed state as compared to a pure double-G \bar{o} model. Such a shift seems appropriate given that we do not model ligands that implicitly lead to more contacts in the Closed state and consequent overstabilization in the G \bar{o} model.

For the Closed-to-Open transition using either model, we obtain pathways that mirror the Open-to-Closed transition: the LID fluctuates in the Closed state, and this is followed by the BD snapping Open on a relatively fast time scale. For both models 1 and 2, successful trajectories appear to follow only the Closed-LID-BD-Open pathway for Closed-to-Open transition (reverse order of the Open-BD-LID-Closed pathway in the Open-to-Closed transition direction). The absence of symmetry is surprising given our recent formal demonstration,³⁹ and there seem to be two possible reasons. First, the transients for the Closed-BD-LID-Open pathway are long-lived. Lengthy transients are consistent with the low reverse reaction rates, shown in Figure 11, for both Models 1 and 2. Second, our state definitions may be flawed as discussed in section 5.3.

To clarify the issue of the symmetry of path ensembles between forward and reverse directions, we constructed and path sampled model 3.

4.4. Path Ensemble Symmetry Analysis in Model 3.

The slow Closed-to-Open transition indicates that, for Models 1 and 2, the free energy of the Closed structure is significantly lower than that of the Open structure. As discussed in section 2.3, this suggested the use of Model 3, which decreases in magnitude the favorable energy for contacts present only in the Closed state. That is, Model 3 reduces the free energy asymmetry between the Open/Closed states.

Model 3 thereby facilitates study of the symmetry between forward and reverse transitions. As shown in Figure 12, although the flux in the Open-to-Closed direction in Model 3 is higher than that in the Closed-to-Open direction, the difference between the fluxes in the two directions is much less than that for Models 1 and 2. The increased Closed-to-Open rate implies that the relative stability of the Closed state is reduced as compared to Models 1 and 2. Importantly, the relatively linear behavior of fluxes in both directions implies our path sampling is sufficient, well beyond transients.

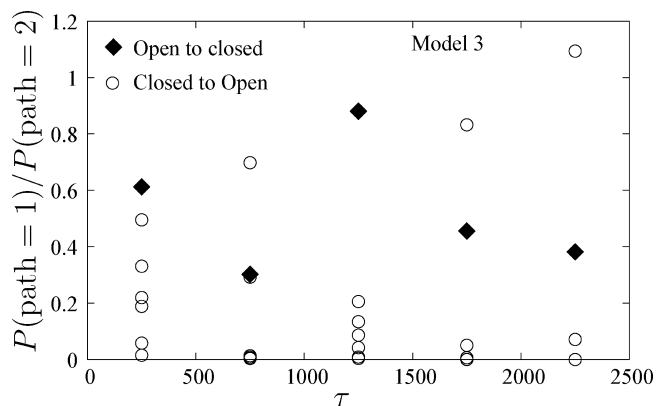


Figure 13. Ratio of probabilities of the two paths, 1 and 2, in the Open-to-Closed (\blacklozenge) and Closed-to-Open directions (\circ). Results from six independent WE simulations are shown in the Closed-to-Open direction to highlight the fluctuations. All data points are window averaged for 500 τ .

For model 3, we examine the same classification of pathways as above. Both paths are frequently observed in both directions. In Figure 13, we show the ratio of probabilities of the two paths as a function of simulation time in the two directions. Values in each window are averaged over 500 τ increments. The results for the Open-to-Closed direction (\blacklozenge) are shown for a single simulation, whereas the Closed-to-Open transitions (\circ) are shown for six independent simulations. Despite large fluctuations, the ratios of paths in the two directions are similar. We discuss the issue of path symmetry further in Section 5.

4.5. Intermediate Detection. The path ensembles obtained can also be used to detect intermediates during the transitions in the two directions. A framework for analyzing the transition path ensemble was recently reviewed by E and vanden-Eijnden.⁴ In particular, transition, or reactive, trajectories can be traced along the configuration space (or a projection of it) to compute populations of regions of the configuration space: regions with high populations correspond to intermediate structures that will be observed more frequently during transitions. Note also the analyses proposed by Hummer.⁵

Figure 14 shows fractional populations in regions of configuration space obtained for Model 3 during transitions in the two directions. Both panels show significant similarity, two distinct channels (or, transition tubes) are visible, corresponding to the two paths. Regions with higher populations suggest metastability, that is, intermediates.

In both figures, the region of the transition channels with LID–CORE distance in the 10–15 Å range and with BD–LID distance similar to that in the Closed state is significantly more populated than other transition regions in the two directions. Such a structure is illustrated by configuration 3 of Figure 8. This not only reiterates the observation in Figure 13 that the Open-BD-LID-Closed pathway is the more dominant transition pathway, but also suggests that the region noted above is a dominant intermediate in both directions. A qualitatively similar result is obtained for the Open-to-Closed direction for Models 1 and 2 (as noted above, the Closed-BD-LID-Open pathway was not observed for Models 1 and 2).

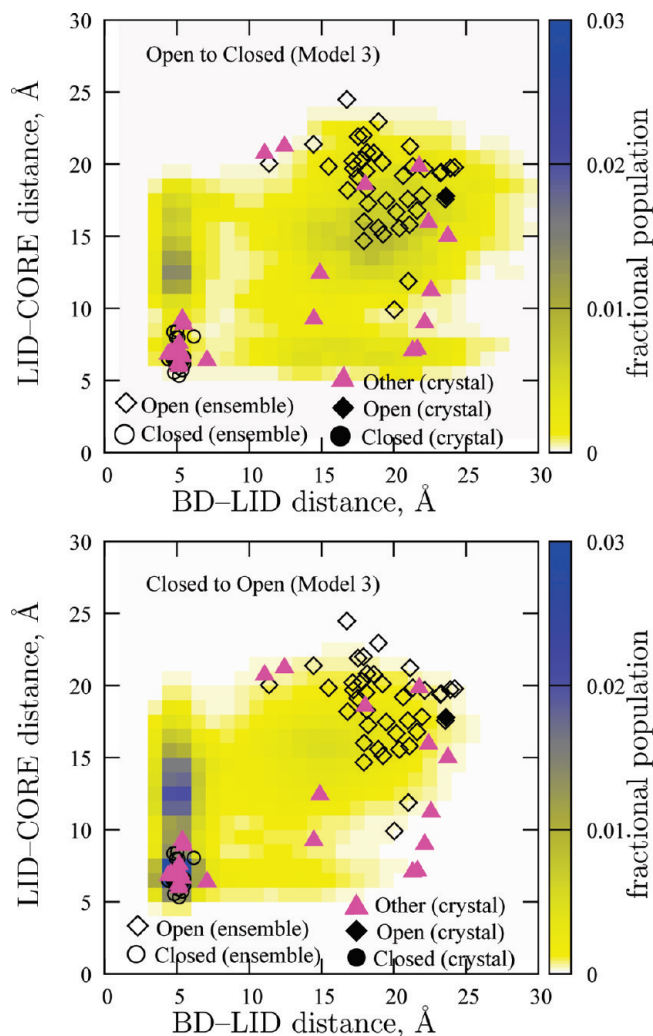


Figure 14. Fractional populations observed in the transition path ensemble. The top panel shows fractional populations for the Open-to-Closed direction, whereas the bottom panel is for the Closed-to-Open direction. Also depicted are experimental configurations, including those from which our double-Gō models were constructed (black filled symbols) and a number of others (pink \blacktriangle), which are tabulated in the Supporting Information. Open symbols depict the equilibrium ensembles obtained from simulation.

5. Discussion

5.1. Models. An important issue in any coarse-grained study is the sensitivity of the results to the particular model(s) used. To address this point, we used three different semi-atomistic models of adenylate kinase. For the models used, we find that the transition pathways are not significantly affected by the models we used. In particular, we find two dominant pathways (Open-LID-BD-Closed and Open-BD-LID-Closed) that occur in all the models. Although the rates vary considerably among models, we do not expect realistic kinetics in simplified models.

Our choice of models was governed by the basic requirement of obtaining full path sampling of conformational transitions, to study path ensembles, heterogeneity, and symmetry. Two of the models are based purely on structure (Gō model), and the other (Model 2) includes some level of residue specificity via Miyajawa–Jernigan interactions, as

well as hydrogen-bonding energies and Ramachandran propensities. In Model 2, the chemical energy terms are significant perturbations to the $G\bar{o}$ interactions (as quantified by MJ interactions between residues). This model is designed to be able to capture a minimum level of biochemistry. However, Model 2 still requires significant $G\bar{o}$ -type interactions to stabilize the two physical states. In the future, we plan to utilize more detailed and explicit side chain–side chain and side chain–backbone interactions to reduce the dependence on $G\bar{o}$ -type interactions.

Another limitation is that we did not consider the ligand in our path sampling simulations. The inclusion of ligand could influence the observed pathways significantly. We have plans for modeling ligands via “mixed models” that include all-atom ligands and binding sites, with a coarse-grained picture for the rest of the protein. Such an explicit inclusion of ligands, with the corresponding degrees of freedom of the unbound ligands in the Open form, should reduce the dependence on arbitrary $G\bar{o}$ interactions. A study with explicit ligands could require higher dimensional progress coordinates to use in weighted ensemble simulations: one coordinate for protein structure (as is done in this work), and a second (or further) coordinate for the distance between ligands and the protein. Note that weighted ensemble simulations can mix real- and configurational-space coordinates: it was originally designed for binding studies.³⁶

5.2. Multiple Paths. The two dominant pathways, Open-LID-BD-Closed and Open-BD-LID-Closed (and reverse on Opening), have also been observed by other groups using coarse-grained models with network analysis⁴⁶ and molecular dynamics.^{47,48} These works utilized double-well potentials. We showed that adding approximate chemical details to the double-well potential preserves the two types of pathways.

Further, Whitford et al.⁴⁷ observed that different pathways were dominant in the two directions. On the other hand, results of Lu and Wang⁴⁸ suggest that any path is similarly favored in both directions. We discuss this symmetry of pathways below.

5.3. Path Symmetry. Recently, we investigated the conditions when there should be symmetry, that is, when pathways in the forward and the reverse directions occur with the same ratio.³⁹ We show that exact symmetry will hold when a specific (equilibrium-based) steady state is enforced. Approximate symmetry is expected if the initial and final states are well-defined physical basins lacking slow internal time scales, so that trajectories emerging from a state “forget” the path by which they entered. Figure 13 suggests that the ratio of the two different pathways in the two directions is very similar for Model 3, which was fully path sampled in both directions.

Such a symmetry is clearly absent from our results (even after accounting for statistical fluctuations) in Models 1 and 2. Although we observed transitions in both directions for all the models, Closed-to-Open transitions in all the models (especially in Models 1 and 2) are harder to obtain. In particular, the Closed-BD-LID-Open pathway is not observed in our simulations for Models 1 and 2. This indicates that the conditions required for symmetry³⁹ are not met in one or more ways: transitions in the Closed-to-Open direction

may be dominated by transients and/or there may be internal barriers within the Closed state. We are currently working on developing WE path sampling methods that allow steady states to be sampled directly and efficiently,⁶² and we hope to investigate the issue of symmetry further. Related steady-state methods are already available.^{71–73}

The above discussion suggests possible reasons behind the lack of path symmetry observed by Whitford et al.⁴⁷ and path symmetry observed by Lu and Wang.⁴⁸ The lack of symmetry observed by Whitford et al.⁴⁷ is, perhaps, due to transitions being studied from one crystal structure conformation (instead of from an ensemble of conformations relevant to that state) to an ensemble of conformations of the opposing state: possible internal barriers within a state may not allow for symmetry to be observed. On the other hand, Lu and Wang⁴⁸ study transitions from an ensemble of one state to the ensemble of the opposing state.

5.4. CPU Time and Efficiency. One of the basic goals of this work was to determine the level of detail we can include in a model, while still allowing for full sampling of the path ensemble. Thus, we now discuss the computational effort that was required. All simulations were performed on single 3 GHz Intel processors. The results shown for Model 1 in the Open-to-Closed direction took approximately 1 week of single CPU time. More simulation was performed in the Closed-to-Open direction, requiring 3–4 weeks of single CPU time. The results for Model 2 were obtained using approximately the same time as Model 1. For Model 3, the Closed-to-Open transition was not much harder to obtain than the Open-to-Closed transition, and a simulation in each direction required approximately 2 weeks of single CPU time. Because of the low CPU usage for obtaining path ensembles for the models used here, obtaining path ensembles of better models using WE is possible. See section 5.1.

It is not hard to estimate the efficiency of WE simulation as compared to brute-force. The transition rates determined from WE simulations indicate the time required for brute-force simulations to achieve transitions and hence permit estimates of efficiency. For example, the rate obtained for Closed-to-Open transition for Model 3 is $2.5 \times 10^{-6}/\tau$. Thus, one brute-force transition can be estimated to require the reciprocal amount of time. Because 2000τ requires approximately 1 week of computing, BF is estimated to take approximately 4 years for a single transition. In contrast, WE yielded 50 transitions after resampling (i.e., 50 transitions with equal weights), in about two weeks of single-processor computing. (Before resampling, there were about 3000 WE transitions for each simulation.) WE is thus significantly more efficient than BF. For transitions in the other direction and/or other models, a qualitatively similar picture for efficiency emerges.

6. Conclusions

We applied weighted ensemble (WE) path sampling to generate ensembles for conformational transitions between Open (apo) and Closed (holo) forms of adenylate kinase using semiatomistic models of the protein. The models have an all-atom backbone including β -carbons, along with varying levels of chemical specificity, and represent a

significant jump in complexity as compared to previous models studied with WE. No additional driving force was used to enable the transitions. We showed that conformational transitions in both directions are possible for such models via WE, while brute-force simulations are less efficient. Given the relatively small computational effort required for observing transitions using WE (weeks of single-CPU time), more detailed models can be used for full path sampling. Models with further reduction in Gō-type interactions are needed, along with ligand modeling, to study the specific enzyme biochemistry.

All the models show significant hereteogeneity in the transition pathways, consistent with previous work and experimental structures. Two dominant pathways are observed, characterized by the order in which the flexible lid and the AMP binding domains close. Although the transition rates (in terms of Monte Carlo steps) varied significantly depending upon the model used, similar dominant pathways are obtained across the models. The model that allows significant transitions in both forward and reverse directions shows an approximate symmetry of pathway populations, consistent with a recently derived symmetry rule.

Acknowledgment. We thank Dr. Bin Zhang and Prof. David Jasnow for helpful discussions. This work was supported by the NIH (grants GM070987 and GM076569) and the NSF (grant MCB-0643456).

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

References

- Berg, J. M.; Stryer, L.; Tymoczko, J. L. *Biochemistry*; Freeman: New York, 2002; pp 227–258.
- Hammes, G. G. *Biochemistry* **2002**, *41*, 8221–8228.
- Berg, J. M.; Stryer, L.; Tymoczko, J. L. *Biochemistry*; Freeman: New York, 2002; pp 269–272.
- vanden, E., W.; Eijnden, E. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420.
- Hummer, G. *J. Chem. Phys.* **2004**, *120*, 516–523.
- Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964–1977.
- Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
- van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762.
- van Erp, T. S.; Bolhuis, P. G. *J. Comput. Phys.* **2005**, *205*, 157–181.
- Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880–10889.
- West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.
- vanden Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.
- Allen, R. J.; Warren, P. B.; ten Wolde, P. R. *Phys. Rev. Lett.* **2005**, *94*, 018104.
- Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 024102.
- Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2007**, *127*, 164101.
- Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876–12882.
- Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246–4257.
- Elber, R. *Biophys. J.* **2007**, *92*, L85–L87.
- Hu, J.; Ma, A.; Dinner, A. R. *J. Chem. Phys.* **2006**, *125*, 114101.
- Radhakrishnan, R.; Schlick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 5970–5975.
- Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. *Mol. Simul.* **1993**, *10*, 291–309.
- Ma, J.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11905–11910.
- Apostolakis, J.; Ferrara, P.; Caffisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.
- Elber, R. *Chem. Phys. Lett.* **1987**, *139*, 375.
- Ulitsky, A. *J. Chem. Phys.* **1990**, *92*, 1519.
- Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252.
- Sevick, E. M.; Bell, A. T.; Theodorou, D. N. *J. Chem. Phys.* **1993**, *98*, 3196.
- Gillilan, R. E.; Wilson, K. R. *J. Chem. Phys.* **1992**, *97*, 1757.
- Elber, R.; Ghosh, A.; Cardenas, A.; Stern, H. *Adv. Chem. Phys.* **2004**, *126*, 123.
- Passerone, D.; Ceccarelli, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 2025.
- Olender, R.; Elber, R. *J. Chem. Phys.* **1996**, *105*, 9299.
- Dill, K. A.; Bromberg, S.; Yue, K.; Fichig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561–602.
- Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.
- Zuckerman, D. M. *J. Phys. Chem. B* **2004**, *108*, 5127–5137.
- Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18043–18048.
- Huber, G. A.; Kim, S. *Biophys. J.* **1996**, *70*, 97–110.
- Maragakis, P.; Karplus, M. *J. Mol. Biol.* **2005**, *352*, 807–822.
- Chennubhotla, C.; Bahar, I. *PLoS Comput. Biol.* **2007**, *3*, 1716–1726.
- Bhatt, D.; Zuckerman, D. arXiv. <http://arxiv.org/abs/1002.2402> (accessed February 11, 2010).
- Pontiggia, F.; Zen, A.; Micheletti, C. *Biophys. J.* **2008**, *95*, 5901–5912.
- Shapiro, Y. E.; Sinev, M. A.; Sineva, E. V.; Tugarinov, V.; Meirovitch, E. *Biochemistry* **2000**, *39*, 6634–6644.
- Hanson, J. A.; Duderstadt, K.; Watkins, L. P.; Bhattacharyya, S.; Brokaw, J.; Chu, J.-W. *Proc. Natl. Acad. Sci. U.S.A.* **104**, *46*, 18055–18060.
- Muller, C. W.; Schlauderer, G.; Reinstein, J.; Schultz, G. E. *Structure* **1996**, *4*, 147–156.
- Muller, C. W.; Schultz, G. E. *J. Mol. Biol.* **1992**, *224*, 159–177.
- Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. *J. Mol. Biol.* **2007**, *366*, 1661–1671.

- (46) Chu, J. W.; Voth, G. A. *Biophys. J.* **2007**, *93*, 3860–3871.
- (47) Whitford, P. C.; Gosavi, S.; Onuchic, J. N. *J. Biol. Chem.* **2008**, *283*, 2042–2048.
- (48) Lu, Q.; Wang, J. *J. Am. Chem. Soc.* **2008**, *130*, 4772–4783.
- (49) Arora, K.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18496–18501.
- (50) Kubitzki, M. B.; de Groot, B. L. *Structure* **2008**, *16*, 1175–1182.
- (51) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Karplus, M.; Hubner, C. G.; Kern, D. *Nature* **2007**, *450*, 838–844.
- (52) Beckstein, O.; Denning, E. J.; Perilla, J. R.; Woolf, T. B. *J. Mol. Biol.* **2009**, *394*, 160–176.
- (53) Zuckerman, D. M.; Woolf, T. B. *J. Chem. Phys.* **1999**, *111*, 9475–9484.
- (54) Zuckerman, D. M.; Woolf, T. B. arXiv. <http://arxiv.org/abs/physics/0209098> (accessed September 27, 2002).
- (55) Best, R. B.; Chen, Y.-G.; Hummer, G. *Structure* **2005**, *13*, 1755–1763.
- (56) Levy, Y.; Cho, S. S.; Shen, T.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2373–2378.
- (57) Mamonov, A. B.; Bhatt, D.; Cashman, D. J.; Ding, Y.; Zuckerman, D. M. *J. Phys. Chem. B* **2009**, *113*, 10891–10904.
- (58) Rojnuckarin, A.; Kim, S.; Subramanian, S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4288–4292.
- (59) Fisher, E. W.; Rojnuckarin, A.; Kim, S. *J. Mol. Struct. (THEOCHEM)* **2000**, *529*, 183–191.
- (60) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*, 054107.
- (61) Zhang, B. W.; Jasnow, D.; Zuckerman, D. arXiv. <http://arxiv.org/abs/0902.2772> (accessed February 16, 2009).
- (62) Bhatt, D.; Zhang, B. W.; Zuckerman, D. *J. Chem. Phys.* **2010**, *133*, 014110.
- (63) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 389–408.
- (64) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534–552.
- (65) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–644.
- (66) Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195–209.
- (67) Gan, H. H.; Tropsha, A.; Schlick, T. *J. Chem. Phys.* **2000**, *113*.
- (68) Shimada, J.; Kussell, E. L.; Shakhnovich, E. I. *J. Mol. Biol.* **2001**, *308*, 79–95.
- (69) Shimada, J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11175–11180.
- (70) Liu, J. S. *Monte Carlo Strategies in Scientific Computing*; Springer: New York, 2004.
- (71) Warmflash, A.; Bhimalapuram, P.; Dinner, A. R. *J. Chem. Phys.* **2007**, *127*, 154112.
- (72) Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *130*, 074104.
- (73) vanden Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *131*, 044120.

CT100406T

JCTC

Journal of Chemical Theory and Computation

Modeling Protein–Ligand Binding by Mining Minima

Wei Chen,[†] Michael K. Gilson,^{*,†} Simon P. Webb, and Michael J. Potter^{*}

VeraChem LLC, P.O. Box 2206, Germantown, Maryland 20875-2206

Received May 10, 2010

Abstract: We present the first application of the mining minima algorithm to protein–small molecule binding. This end-point approach uses an empirical force field and implicit solvent models, treats the protein binding site as fully flexible, and estimates free energies as sums over local energy wells. The calculations are found to yield encouraging agreement with experimental results for three sets of HIV-1 protease inhibitors and a set of phosphodiesterase 10a inhibitors. The contributions of various aspects of the model to its accuracy are examined, and the Poisson–Boltzmann correction is found to be the most critical. Interestingly, the computed changes in configurational entropy upon binding fall roughly along the same entropy–energy correlation previously observed for smaller host–guest systems. Strengths and weaknesses of the method are discussed, as are the prospects for enhancing accuracy and speed.

1. Introduction

The appeal and promise of structure-based drug design has grown in recent decades as computer power has increased and as the three-dimensional structures of proteins have become easier to solve by crystallographic and NMR methods (e.g., refs 1–6). Indeed, advances in protein crystallography are now extending the reach of structure-based approaches into the medically critical yet previously inaccessible realm of G-protein coupled receptors.^{7–11} Computational tools for the selection or design of compounds to bind a targeted protein also have advanced substantially in recent years (e.g., refs 12–17), but there is still a need for more reliable methods of ranking candidate ligands of a targeted protein according to affinity.^{18,19}

Approaches to modeling protein–small molecule interactions may be viewed, very broadly, as falling along a spectrum of complexity. At one end lie the simplest and most efficient docking methods (e.g., refs 20–28), which seek to identify the single most stable conformation of a protein–ligand complex and estimate binding affinity with a scoring function that is often conceptualized as a sum of free energy contributions from, for example, hydrogen bonding, the hydrophobic effect, and Coulomb interactions. (Recent

progress in the use of relatively detailed electronic structure calculations as part of protein–ligand scoring functions also deserves mention (e.g., refs 29–36).) These methods offer the speed required to rapidly screen large compound databases for promising candidate ligands but typically neglect or, arguably, oversimplify significant free energy contributions, such as the energetic and entropic consequences of ligand preorganization or the lack thereof. At the other end of the spectrum lie free energy pathway methods, like free energy perturbation and thermodynamic integration, which employ relatively detailed Monte Carlo or molecular dynamics methods, typically using explicit solvent models, to compute the absolute or relative work of binding for candidate ligands (e.g., refs 37–47). Such approaches have the potential to capture much of the relevant physics but still tend to be too computationally demanding for routine use in drug-design projects. Toward the middle of the spectrum of complexity lie so-called “end-point” free energy methods, such as MMPBSA^{48–51} and LIE.^{52–55} These aim to provide greater physical detail than docking and scoring methods by accounting for both the bound and unbound states of the protein–ligand system, while potentially avoiding the computational costs of free-energy simulations because they do not involve computing the work of a multistep binding process or “alchemical”³⁷ ligand change.

The present study focuses on one of a class of end-point methods,⁵⁶ including mining minima (here M2)^{57–59} and MINTA,⁶⁰ which approximate the thermodynamics of bind-

* To whom correspondence should be addressed. Phone: (240) 686-0565. Fax: (240) 686-0564. E-mail: gilson@verachem.com (M.K.G.); potterm@verachem.com (M.J.P.).

[†] These authors contributed equally to this work.

ing by identifying a manageably small set of conformations (local energy minima) of the free and bound protein and ligand. The free energy associated with each local energy minimum of the ligand, receptor or complex, is evaluated on the basis of its depth and width, and the contributions of the energy wells are combined to yield an estimate of the overall free energy. A potential merit of this approach is that one may employ conformational search approaches that aggressively cross energy barriers to identify the stable conformations, rather than having to wait for barriers to be crossed through thermal motion, as in typical simulation-based methods. On the other hand, in order to avoid a combinatorial explosion of conformations, such methods must limit the number of explicit degrees of freedom by using an implicit solvent model and, for large receptors, treating part of the system as restrained or rigid. Applications of such models to binding in host–guest systems have shown encouraging agreement with experimental results,^{57,59–62} suggesting that this approach, used with current energy models, captures much of the relevant physical chemistry of molecular recognition. The host–guest studies also have yielded unexpected insights into the role of configurational entropy in binding, providing evidence that these entropy contributions are commensurate with more familiar energy terms and may differ enough across ligands to significantly affect their affinity rankings. It is thus of considerable interest to apply this approach to protein–small molecule binding.

We have now developed an implementation of the M2 algorithm, VM2, that is suitable for protein–ligand affinity calculations. The software includes more aggressive search algorithms than those used for the prior host–guest calculations and allows part of the target protein to be held rigid while a user-defined binding site region, including a ligand and any number of additional solvent molecules or cofactors, is treated as flexible. This paper describes the algorithm and characterizes its performance for three groups of compounds binding two very different proteins, the human immunodeficiency virus 1 protease (HIVP) and phosphodiesterase 10a (PDE 10a).

2. Methods

2.1. Application of Mining Minima to Protein Ligand Binding. *2.1.1. Overview.* As previously described,^{57,59} the second-generation mining minima method, M2, computes the standard free energy (or, more properly, the standard chemical potential) of a bound complex, AB, and the free molecules, A and B, and obtains the standard binding free energy as the difference:

$$\Delta G^\circ = G_{AB}^\circ - G_A^\circ - G_B^\circ \quad (1)$$

M2 is thus an “end-point” method.¹² In brief, the free energy, G_X , of a molecule or complex $X = A, B,$ or AB is estimated by summing local configuration integral contributions z_i from distinct energy wells (conformations) i :

$$G_X^\circ \approx -RT \ln \sum_i z_i \quad (2)$$

Each local integral, z_i , is obtained by computing the Hessian matrix with respect to bond-angle-torsion coordinates at the

corresponding local energy minimum and applying an enhanced harmonic approximation which accounts for anharmonicity along those eigenvectors of the Hessian having eigenvalues up to 2 kcal/mol/xy, where x and y are either Angstroms or radians, depending on the coordinates involved. Local energy minima are found with an aggressive conformational search which, unlike typical molecular dynamics simulations, can readily cross energy barriers (see below). The energy model comprises an empirical force field and an implicit solvent model.

The present study uses an implementation of M2 that is suitable for protein–ligand modeling. In particular, it allows part of the receptor (the “real set”) to be held rigid while treating only a user-defined binding-site region (the “live set”) as flexible. The rigid part of the receptor acts as a framework that holds the flexible binding site region in the right overall conformation, while the flexibility of the binding site allows its conformation to adapt to different ligands. The present implementation also affords additional, more aggressive, conformational search methods and permits the complex to include more than one bound molecule; for example, it can include a drug-like ligand and one or more explicit water molecules. A ligand or solvent molecule can, optionally, be restrained in a spherical region near a desired location by applying a flat-bottomed energy well to one of its atoms:

$$E_{\text{restraint}} = k \frac{(r - r_0)^p}{R_{\text{restraint}}^p} \quad (3)$$

Here, k is the force constant of the restraint, r is the location of the restrained atom, r_0 is the center of the flat-bottomed well, in coordinates rooted in the rigid part of the receptor, $R_{\text{restraint}}$ is the radius of the restraint region, and p is a parameter controlling the “hardness” of the energy wall around the energy well. A value of $p = 12$ was used in all calculations presented here.

2.1.2. Conformational Search. The conformational search method is a variant of the previously described Tork algorithm,⁶³ which in turn is inspired by the low mode search method.^{64,65} The search typically starts with a single initial conformation, although more than one can be used. The conformation is energy-minimized to a local energy minimum, and the second-derivative (Hessian) matrix of the energy in Cartesian coordinates is computed, transformed into the Hessian for bond-angle-torsion coordinates, as previously described,⁶³ reduced by elimination of rows and columns corresponding to bond-stretch and bond-angle coordinates, and diagonalized. A “distortion mode” is then defined by choosing a normalized eigenvector e_i with a low eigenvalue v_i , or a linear combination of two normalized eigenvectors e_i and e_j with low eigenvalues v_i and v_j . The mode is then modified by zeroing any eigenvector components less than 0.1. The system is distorted along this distortion mode, and then 10 steps of quasi-Newton energy minimization are carried out with the distorted atoms held fixed and the rest of the system free. If the final energy is less than a threshold of E_t and no torsional change has exceeded a user-defined threshold of ϕ_t , then the process is

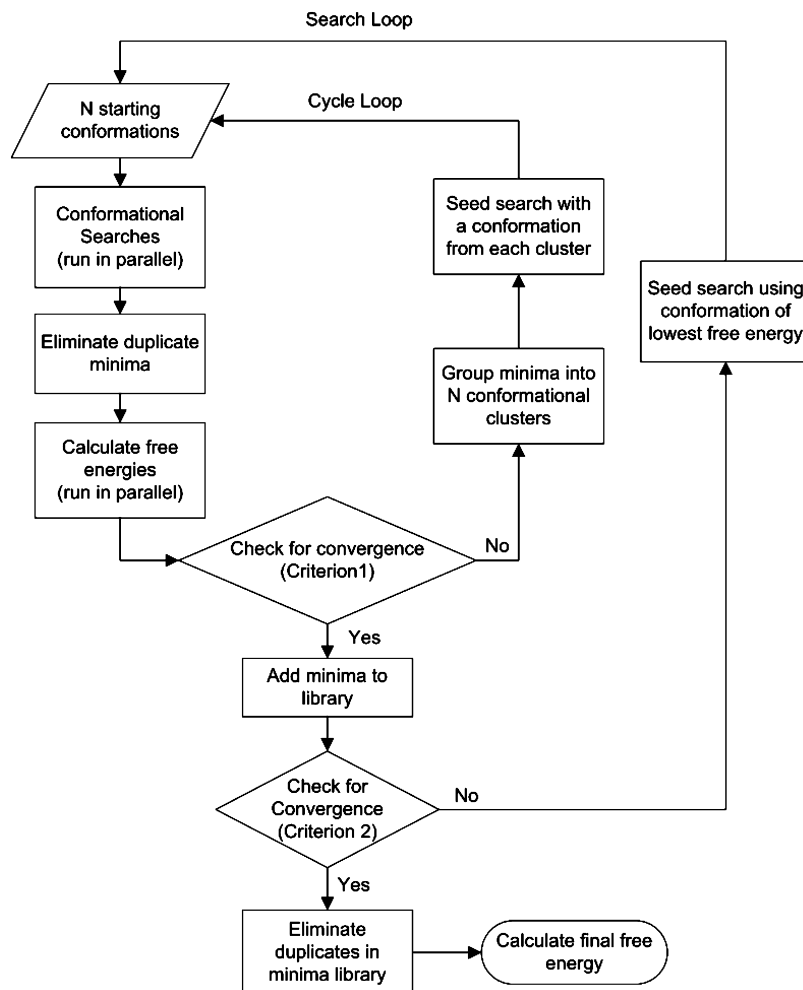


Figure 1. High-level flow-chart of the mining minima procedure, showing nested iterative loops over cycles and searches, each with its own convergence criterion, as detailed in the text.

repeated. When this iterative process crosses either the energy threshold or the torsional threshold, the distorted conformation is fully energy-minimized to a local energy minimum. Here, the 100 eigenvectors with the lowest eigenvalues were considered “low”, the energy threshold, E_t , was set to 2000 kcal/mol, and the torsional threshold, ϕ_t , was set to 180° . For the pairwise distortions, the contribution of each eigenvector to the combined distortion vector, e_{combo} , was weighted according to the reciprocal of its eigenvalue:

$$e_{\text{combo}} = \frac{e_i/v_i + e_j/v_j}{1/v_i + 1/v_j} \quad (4)$$

The present software implementation also includes two aggressive search methods. In one, only the ligand is moved stepwise along its soft modes, and after each step, the live atoms of the protein are energy-minimized for several steps so that they move out of the way of the moving ligand. This process is iterated until an energy or displacement is reached, at which point the entire system is energy-minimized. The second aggressive search method is the same, except that only the protein is moved stepwise along its soft modes, while the ligand is relaxed in response. These more aggressive methods are most appropriate for cases where the binding modes are unknown and were not used here because

we wished to leave the ligands’ scaffolds close to their positions in the available crystallographic structures.

2.1.3. Monitoring Free Energy Convergence. The free energy calculation for the protein, ligand, or complex begins with a single energy-minimized conformation. As diagrammed in the flowchart in Figure 1, the algorithm then follows a convergent, iterative procedure involving a series of “searches”, each composed of “cycles”, where each search is run to convergence (criterion 1 in the flowchart) and the overall set of searches is also run to convergence (criterion 2 in the flowchart). The first search is carried out as follows. The starting conformation is used to seed a Tork conformational search (see prior section) and thereby generate a set of new conformations. Any duplicate conformations are removed from the list, and the configuration integral z_i is computed for each of the N distinct local energy minima. The free energy is obtained from eq 2, where the sum ranges over the energy minima i found to date.

This completes the first cycle of the first search. All conformations from the current search iteration are then superimposed and grouped into N clusters based upon the atomic root-mean-square distances between conformations, and the lowest free energy conformation of each cluster is used to start a new Tork conformational search. Thus, each cycle after the initial one starts with N , rather than 1, initial

conformations, but the algorithm is otherwise the same. This procedure aims to generate a diversity of starting structures for the new cycle. Duplicates are again removed, with reference to all conformations generated in the current search. Configuration integrals are computed for the remaining distinct conformations, and the free energy is updated to reflect all distinct conformations found in the current cycle. Additional cycles are run in this manner until the last cycle meets criterion 1: it must lower the free energy by less than 0.05 kcal/mol, and the last three cycles must cumulatively lower G by less than 0.5 kcal/mol. At this point, the first search is considered to be converged. A new search then begins from the lowest free energy conformation found in the prior search. Additional searches are run until criterion 2 is met; i.e., the last search must lower the free energy by less than 0.05 kcal/mol. In the present calculations, the number of clusters, N , is set to 4.

2.1.4. Energy Model. The protein's potential energy as a function of conformation was approximated with the CHARMM⁶⁶ united-atom force field. For the ligands, we used VeraChem's enhanced version of the Dreiding⁶⁷ force field for bond-stretch, bond-angle, and torsional terms. The Lennard-Jones terms were automatically assigned to those of chemically related atoms in the CHARMM force field, and partial charges were assigned with the electronegativity equalization method Vcharge.⁶⁸ The topology files for the ligands are provided in the Supporting Information. During the conformational search procedure, solvation is modeled with a generalized Born model (GB).^{69,70} The free energy of each energy minimum is subsequently corrected, as previously described,⁵⁷ by subtracting its GB energy and adding in the Poisson–Boltzmann/surface area (PB/SA) solvation energy, where the Poisson–Boltzmann equation is solved with a fast finite difference code.^{71,72} It is this PB-corrected free energy that is checked to determine whether the new conformation is more stable than prior conformations. Using the corrected free energy helps prevent the search algorithm from finding conformations that are stable according to the GB model but not according to the PB/SA model. The dielectric cavity radius of each atom is set to the value of its Lennard-Jones parameters, i.e., to the distance at which the Lennard-Jones energy of two atoms of this type has its local energy minimum.

2.2. Protein–Ligand Systems and Setups. We applied the method described above to 24 HIV-1 protease inhibitors and 20 phosphodiesterase 10a (PDE 10a) inhibitors. The protease inhibitors are divided into two groups. Group 1 is a heterogeneous collection of drugs and nonclinical compounds with measured binding free energies of -19.2 to -6.6 kcal/mol, as determined by calorimetric and enzymatic K_i measurements cited in the caption of Table 2. Although these data derive from six different research groups, their wide range of binding free energies reduces the quantitative significance of methodologic variations among the groups. Group 1 is subdivided into group 1a compounds (Figure 2), which are known or presumed to bind with a bridging “flap water”, and group 1b compounds (Figure 3), cyclic and azacyclic ureas which bind without a flap water. Group 2 (Figure 4) is a relatively homogeneous set of 10 compounds⁷³

from a project aimed at discovering new protease inhibitors that would be robust to mutational resistance, with binding free energies of -16.7 to -9.3 kcal/mol, on the basis of enzymatic K_i data obtained by a single research group. The PDE 10a inhibitors⁷⁴ possess a common chemical scaffold (Figure 5) and range from -11.5 to -9.2 kcal/mol in binding free energy, also on the basis of K_i data obtained by a single research group.

The evaluation of a method of ranking ligand affinities should include ligand series for which affinity is not strongly correlated with ligand size, so that the arguably trivial tendency of many scoring functions to correlate with ligand size⁷⁵ does not lead to an overly optimistic assessment of the method. Here, the correlation coefficients (R values) of binding free energy with the number of non-hydrogen atoms are 0.75, 0.63, 0.00, and 0.44 for the group 1a, group 1b, and group 2 protease inhibitors and PDE 10a, respectively. We thus have a mix of correlated and uncorrelated series, much as previously observed for a variety of experimental data sets.⁷⁶

Calculations for the group 1 HIV protease calculations were based on PDB structure 1HVR,⁷⁷ which was solved with a bound cyclic urea inhibitor, while group 2 calculations used 2IOD,⁷³ which was solved with ligand AD81 of this group. Except as otherwise noted, all crystallographic solvent molecules were omitted. For group 1a ligands, an explicit flap water was included with a flat-bottomed restraint on the water's oxygen atom, with $R = 0.5$ Å and $k = 1.0$ kcal/mol/Å.¹² The location of the oxygen was obtained by superimposing the protein parts of structures 1HVR and 2FDE⁷⁸ (HIV protease with GW0385) and then using the water coordinates from the superimposed 2FDE structure. On the basis of prior analyses, Asp 25 was treated as protonated on oxygen, and Asp 25' was treated as unprotonated for the group 1a compounds,⁷⁹ while Asp 25 and 125 were both protonated for the group 1b compounds.⁸⁰ The initial bound conformation of each HIVP ligand was generated by superimposing the common parts of its structures with the highest affinity compound in its group.

Calculations for the PDE 10a inhibitors were based on PDB structure 2OVV,⁷⁴ which was solved with inhibitor **21** and is very similar to structure 2OVY,⁷⁴ which was solved with inhibitor **29**. Examination of these structures revealed that water 85 in 2OVV and 2OVY appears to play a particularly important bridging role, as it is the only water within 3 Å of both the ligand and the protein. Moreover, initial conformational searches with no crystallographic waters generated conformations in which the ligand wandered from its crystallographic location, whereas including water 85 with a flat-bottomed restraint ($R = 0.5$ Å, $k = 1.0$ kcal/mol/Å¹²) centered on its crystallographic site caused the ligands to remain close to their crystallographic locations. This water was therefore included in all subsequent calculations, with the same restraining potential. Initial conformations of the bound PDE 10a ligands were generated by superimposing them on the crystal conformation of the tightest-binding ligand, **29** in 2OVY. The bound Zn and Mg atoms were part of the real set and therefore treated as rigidly fixed. Both metals were treated as electrically neutral, as were

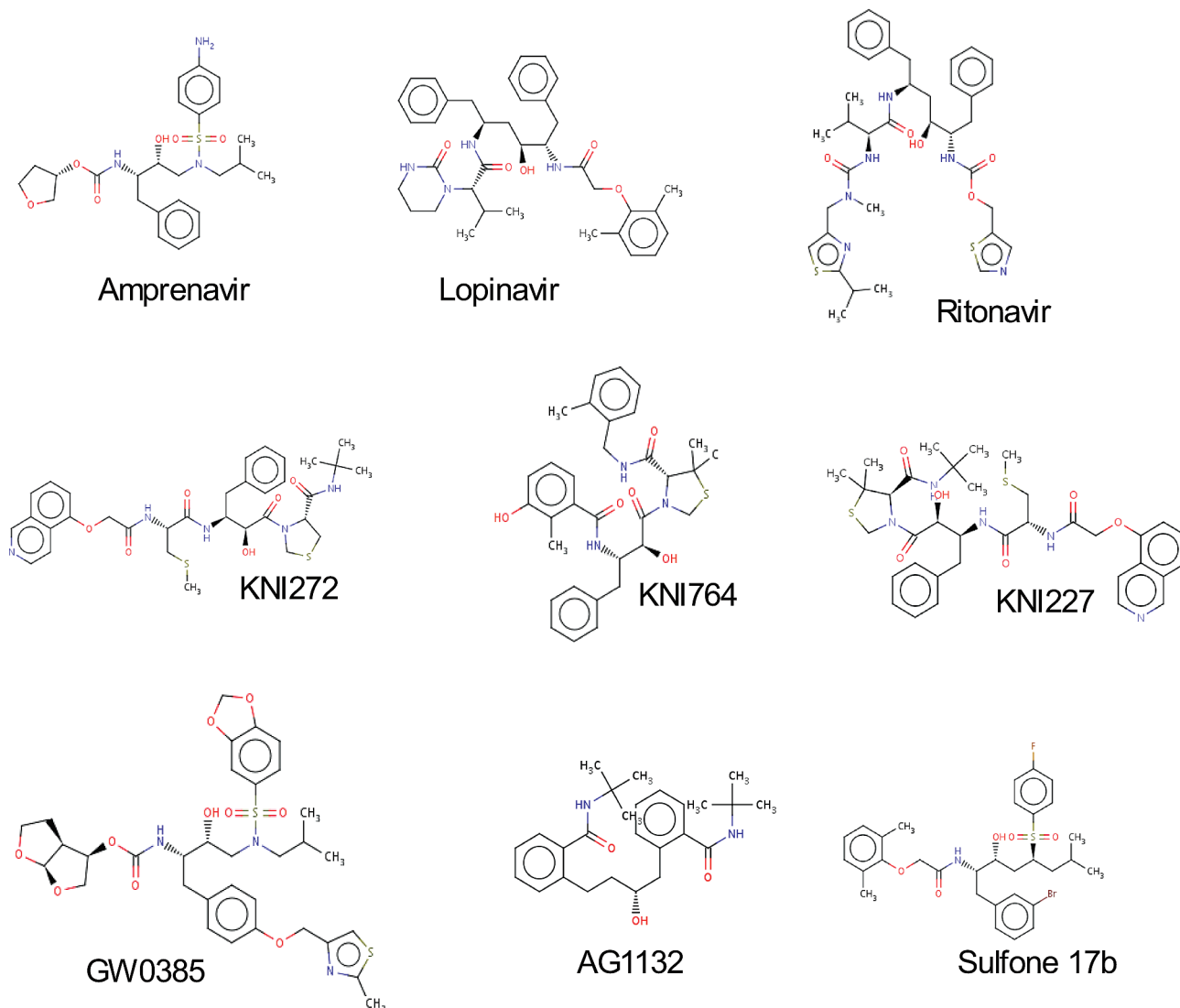


Figure 2. Group 1a HIV-1 protease inhibitors.

their first-shell histidine ligands. We do not know the actual protonation states of the first-shell ligands, but we observed that adding +2 charge to both metals (while keeping the first-shell histidines neutral) caused the conformational searches to yield grossly incorrect ligand poses. Fortunately, although the scaffolds of these compounds lie about 7–8 Å from the metals, the closest atoms of variable substituents (R in Figure 5) lie at least 12.7 Å away from the metals, on the basis of the available crystal structures, and the substituents are uncharged and mainly aromatic in nature. As a consequence, a detailed treatment of the metal cluster is not expected to be critical in this series.

For both HIVP and PDE 10a, the live set—the set of binding-site atoms treated as mobile—was defined as all atoms within 7 Å of any atom of any ligand. The real set—the set of protein atoms treated as rigid and thus supporting the live set—comprised all protein residues having any atom within 5 Å of any live-set atom. For the HIVP structures, the initial live and real sets based on these criteria were symmetrized, so that both protein monomers would have the same live and real sets. This was done by expanding, rather than contracting, the live and real sets. All other protein atoms were deleted, in order to reduce the size of the

nonbonded pairlist and thus speed the calculations. In order to diminish any initial stress in the starting conformations that might artifactually drive the binding site away from its crystallographic starting conformation, the protein models were subjected to an initial relaxation step in which both the live and real sets were temporarily treated as live, except for the two metals, which were still held fixed. All protein residues having an atom within 5 Å of any live atom were temporarily treated as real, and the entire temporary live set was energy-minimized. Figures 6–8 highlight the live sets used in the present calculations. The program VMD⁸¹ was used for all 3D molecular graphics in this paper.

3. Results

3.1. Free Energy Calculations. *3.1.1. Group 1 HIVP Inhibitors.* The computed binding free energies for the group 1a and group 1b protease inhibitors are shifted relative to experimental results but correlate well (correlation coefficient $r \sim 0.8$), as shown by the scatter plots in Figure 9 and the linear regression parameters and mean errors in Table 1. Interestingly, the computed free energies of group 1b run lower than those of group 1a by about 10 kcal/mol (Figure

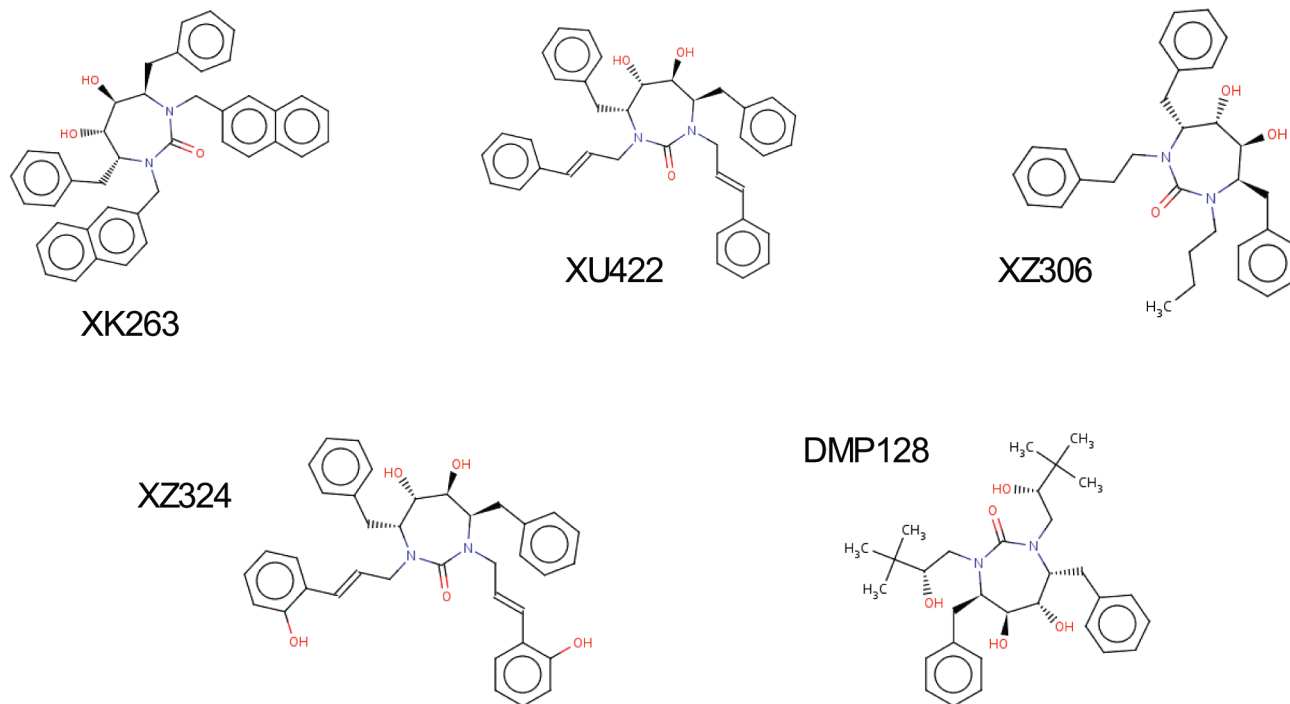


Figure 3. Group 1b HIV-1 protease inhibitors.

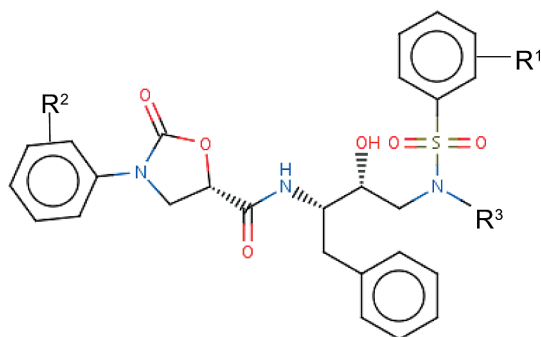


Figure 4. Scaffold of group 2 HIV-1 protease inhibitors.

9). We surmise that this offset stems primarily from the difference in the setup of the two protease structures. Thus, as described in the Methods section, the group 1a calculations include an explicit “flap” water molecule and treat one of the catalytic aspartyls as ionized, whereas the group 1b calculations omit the flap water and treat both aspartyls as neutral. Table 2 provides information on the root-mean-square deviation (RMSD) of the most stable computed conformations relative to crystallographic structures, where available. Results are presented for the ligand alone, and for the entire mobile region. The results are rather mixed, with ligand-only RMSDs ranging from 1.4 to 3.8. Examination of these structures indicates that the structural deviations are primarily associated with extended substituents, rather than with core parts of the ligand that interact with the central Asp groups and the flap Ile residue. Note that all group 1 calculations were based upon the same protein structure, 1HVR,⁷⁷ which was solved with bound XK263. It is also of interest to examine the conformations found in the calculations. Figures 10 and 11 give a sense for this range by overlaying representative conformations of the four most stable conformational clusters found for the bound states of

amprenavir (group 1a) and XK263 (group 1b). In both cases, the representative conformations may be viewed as minor modulations of a basic binding mode.

3.1.2. Group 2 HIVP Inhibitors. The group 2 inhibitors were all modeled on the basis of PDB structure 2I0D,⁷³ which was solved with bound AD81 from this series. The scatter plot of calculation versus experiment, Figure 12, shows nine nearly collinear points and one gross outlier, which corresponds to inhibitor KB19. Corresponding linear regression parameters, omitting KB19, are provided in Table 1. The mean deviation of calculation from experimental results for the group 2 inhibitors, -7.1 kcal/mol, is similar to that for the group 1a inhibitors, -5.1 kcal/mol, while the mean deviation of the group 1b inhibitors is considerably greater, at -16.5 kcal/mol. This similarity of the group 2 results to the group 1a results may stem from the fact that the group 1a and group 2 calculations both included an explicit flap water and treated the two catalytic aspartyls as having a net charge of -1 . In contrast, the group 1b calculations did not include a flap water and treated the aspartyls as neutral. For those compounds in group 2 for which crystal structures are available, the RMSDs range from 1.9 to 2.4 for the ligand only, and from 2.0 to 2.8 for the entire mobile region. Compared with the RMSDs for group 1, these values are relatively moderate and uniform. This is not unexpected, given the comparative uniformity of the compound series. The variation of the computed structures across ligands is examined in Figure 13, which overlays the most stable bound conformation computed for each of the 10. The conformations are all quite similar, with most of the conformational differences localized in the variable substituents.

Examination of these bound conformations led to a potential explanation of why inhibitor KB19 is an outlier in the scatter plot (Figure 12). When the conformation of AD32,

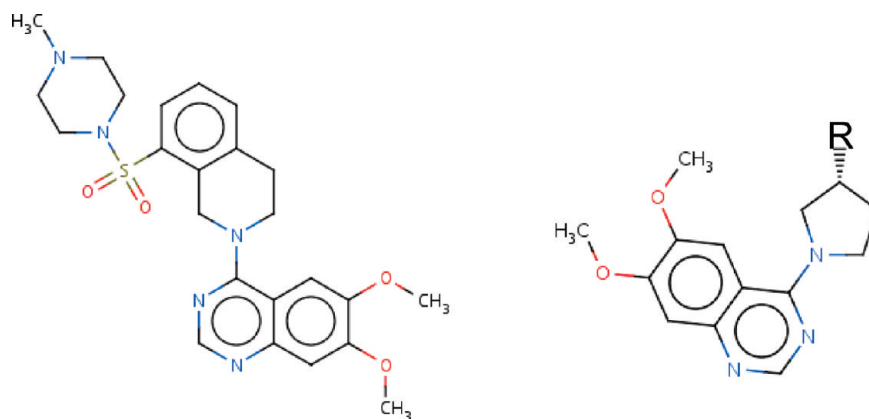


Figure 5. PDE 10a inhibitor **1** (left) and scaffold for inhibitors **11–29**.

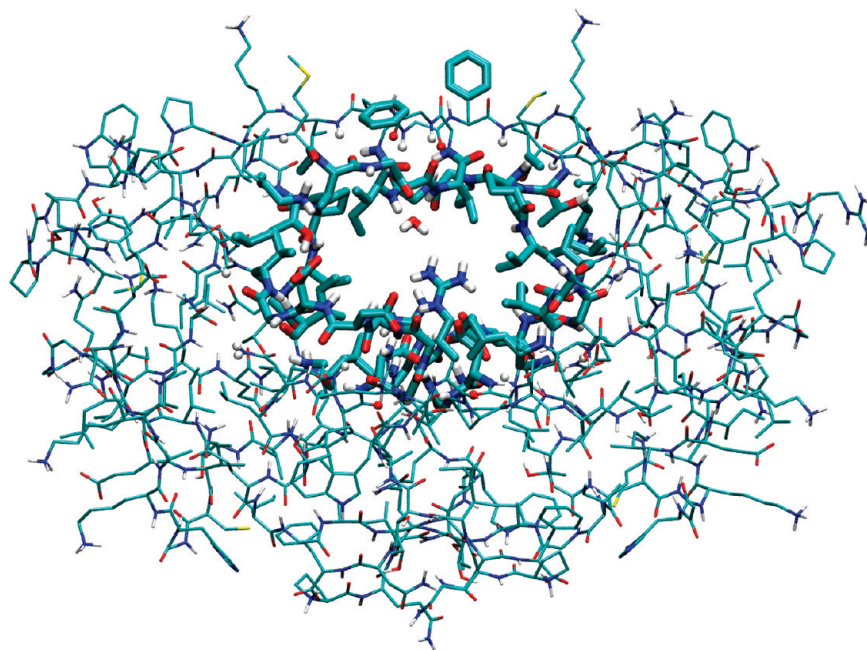


Figure 6. HIV-1 protease structures used for group 1 inhibitors, with mobile (“live”) set highlighted. (The inclusion in the live set of two apparently isolated phenyl rings results from the proximity of an inhibitor to one of the rings, combined with the symmetrization of the live set across the two protease monomers.) This and other 3D molecular graphics were generated with the program VMD.⁸¹

which also has a 4-acetyl substituent at the R^2 location, was compared, we noticed that this substituent lies in opposite orientations in the two compounds. That is, in the most stable conformations computed for these two compounds, the oxygen of each lay on the methyl carbon of the other. Because AD32 is not an outlier, we restarted the KB19 conformational search from a conformation modeled on the most stable conformation of AD32. This led to a conformation of the KB19 complex in which the 4-acetyl group is oriented as in AD32, and whose energy is about 8 kcal/mol more stable than any found before. (See second KB19 line (*italic*) in Table 2.) This lower energy brings KB19 largely into line with the other group 2 compounds (Figure 12) and lowers its RMSD somewhat (Table 2).

3.1.3. PDE 10a Inhibitors. Computed and measured binding free energies for the 20 congeneric PDE 10a inhibitors are compared in Table 3 and Figure 14, and the corresponding linear regression parameters and mean error are presented in Table 1. A significant correlation is observed,

and the correlation coefficient of 0.84 is similar to those obtained for the HIV protease inhibitors (above). However, the slope of the linear fit is about half as great as that obtained for the protease inhibitors. Figure 15 gives a sense of the conformational variation among the 20 bound ligands via an overlay of their most stable computed conformations, along with mobile residues in the PDE 10a binding site. The common scaffold occupies a uniform pose, while, not surprisingly, the variable substituent shows a wider range of positions. The side chains showing greatest conformational flexibility appear to be Met 251 and Met 252, which lie near the variable ligand substituent. The higher RMSD (Table 3) for compound **1**, relative to the other ligands with available crystal structures, **21** and **29**, may result from the fact that this compound is not based on the same chemical scaffold as any of the other compounds in the series (Figure 5) and that the protein structure used in the calculations was solved with one of the other compounds.

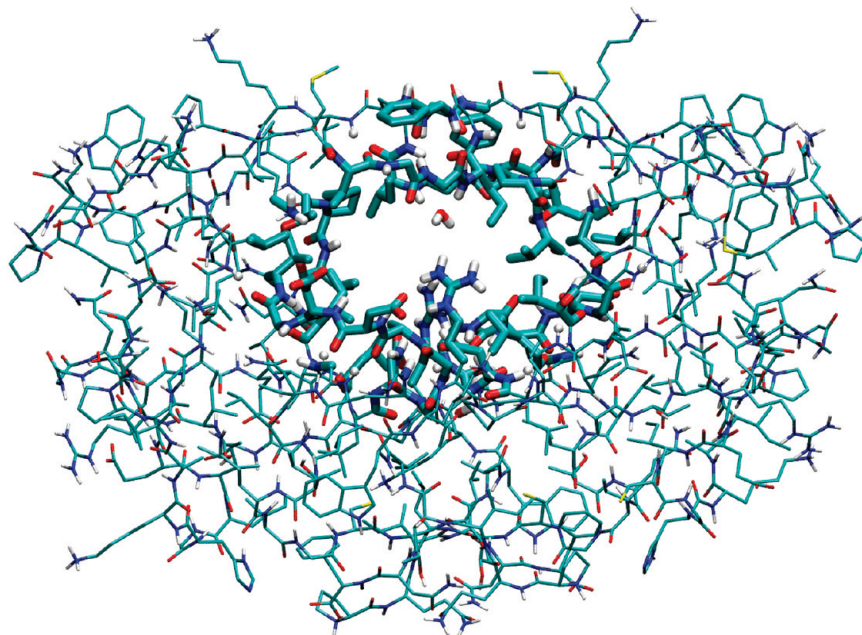


Figure 7. HIV-1 protease structure used for the group 2 inhibitors, with mobile (“live”) set highlighted.

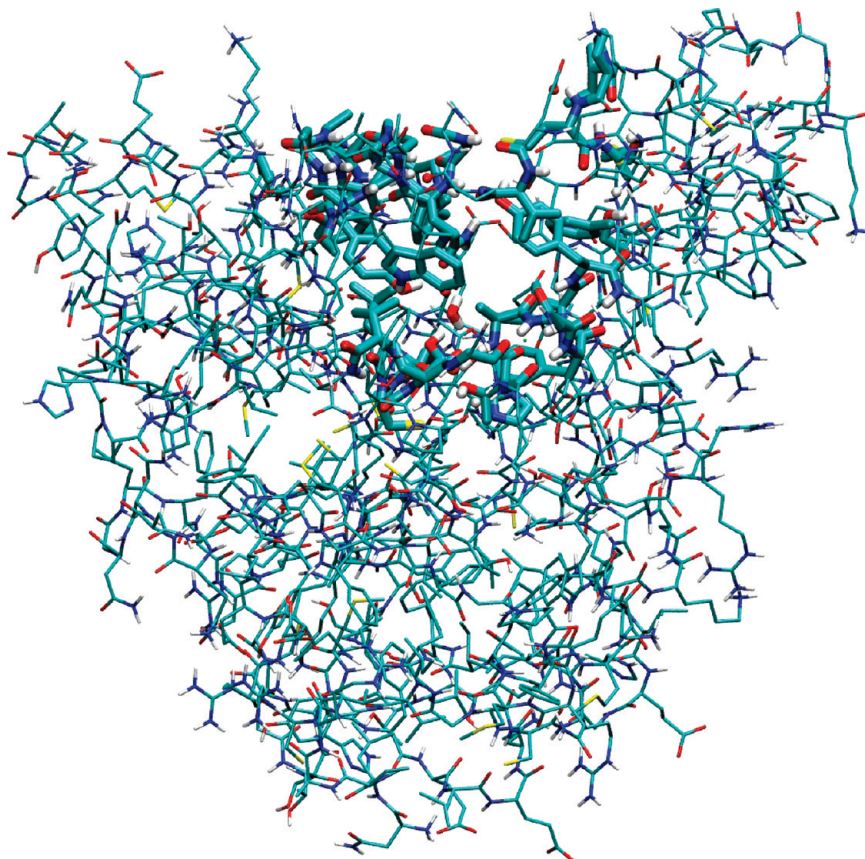


Figure 8. Phosphodiesterase 10a (PDE 10a) structure used for all the PDE 10a inhibitors, with mobile (“live”) set highlighted.

3.2. *Entropy and Energy.* The VM2 method yields not only overall binding free energies but also changes in mean energy components and in the configurational entropy. These breakdowns are provided in Tables 2 and 3 for the HIV protease and PDE 10a inhibitors, respectively. As in prior mining minima calculations on simpler host–guest systems, the net free energy changes are balances of considerably

larger energy and entropy contributions. In particular, the change in the Boltzmann averaged energy, $\Delta\langle E \rangle$, is considerably more negative than the binding free energy, ΔG° , while the configurational entropy makes a markedly positive contribution, $-T\Delta S^\circ$, and thus strongly opposes binding. The binding free energy can thus be viewed as a small difference between large numbers and is hence potentially subject to

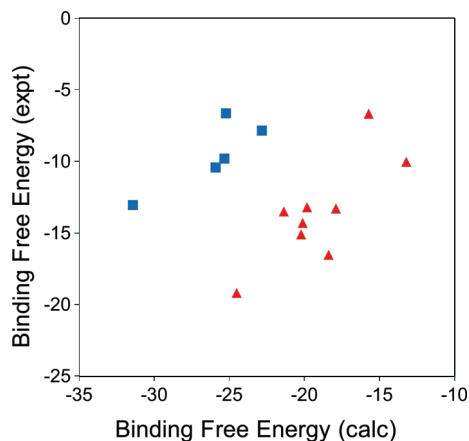


Figure 9. Experimental vs calculated binding free energies (kcal/mol) for group 1a (red triangles) and group 1b (blue squares) HIV-1 protease inhibitors.

Table 1. Linear Regressions and Mean Errors of Computed Binding Free Energies (ΔG°) versus Measured Binding Free Energies, Showing Correlation Coefficient (R), Slope and Y Intercept (kcal/mol) of Linear Regression Fits, and Mean Deviation (kcal/mol) of Calculation Relative to Experimental Results^a

	R	slope	Y intercept	mean error
HIVP group 1a	0.79	0.88	3.2	-5.5
HIVP group 1b	0.84	0.65	7.5	-16.6
HIVP group 2 ^b	0.98	0.56	-1.1	-7.1
PDE 10a	0.84	0.38	-8.3	5.2

^aIn all cases, the experimental binding free energy is considered to be on the ordinate (y -axis). ^bGroup 2 regressions omit the outlier, KB-19.

large uncertainties on this account. It is thus worth emphasizing that the mining minima algorithm does not generate the free energy by computing energy and entropy separately, and then adding them. Rather, it yields the free energy directly, and this can later be decomposed into entropy and energy. This integrated approach represents a potential advantage over methods which compute the two terms separately. It is also important to recognize that, although $\Delta\langle E\rangle$ omits the configurational entropy, it implicitly includes the change in solvent entropy via the implicit solvent model. As a consequence, the entropy values reported here should not be directly compared with experimental entropy changes of binding for these systems.

The computed entropic penalties range from 25 to 37 kcal/mol and average 30 kcal/mol. The energy changes are of greater magnitude and are negative in sign (average -44 kcal/mol), as indeed they must be if the computed binding free energies are to be negative, since $\Delta G^\circ = \Delta\langle E\rangle - T\Delta S^\circ$. In addition, the range of energy changes ($\Delta\langle E\rangle$), -63 to -27 kcal/mol, is about 3 times broader than the range of entropic contributions. Not surprisingly, both the entropy and energy contributions for these protein-ligand systems are considerably larger, by factors of ~ 2.8 , than those computed previously for smaller host-guest systems,^{57,59,61,62} where $-T\Delta S^\circ$ and $\Delta\langle E\rangle$ averaged about 11 and -16 kcal/mol.

The importance of the configurational entropy to the correlations between calculation and experimentation re-

ported above may be examined by comparing the measured binding free energies with the computed values of $\Delta\langle E\rangle$ instead of ΔG° . Figures 16 and 17 provide such comparisons in graphical format for the HIV protease inhibitors, and Figure 18 provides the comparison for the PDE 10a inhibitors. The corresponding linear regression parameters and mean errors are provided in Table 4 (first 4 rows of data). (The group 2 regression again omits the outlier KB19.) Overall, the new correlation coefficients tend to be somewhat lower than those found for the full computed free energy, with the mean R lower by 15%, especially for PDE 10a. Even more striking are the large reductions in the regression slopes by a mean factor of 0.59, and the markedly more negative mean errors by about -30 kcal/mol. These systematic changes highlight the fact that the computed entropies strongly oppose binding and that the entropic penalty correlates with the binding energy, as further analyzed later in this section.

The values of $\Delta\langle E\rangle$ considered in the prior paragraph are Boltzmann-weighted averages, where the Boltzmann weighting is based upon the computed free energy of each energy well. Computing such average energies thus requires the full mining minima algorithm. It is also of interest to examine the even simpler energy-based approach of simply subtracting the global energy minima of the free protein and ligands from those of the protein-ligand complexes. The resulting energy changes, ΔE_{\min} , are not averages, and they entirely exclude any consideration of configurational entropy. However, they do require a conformational search and do include the PB/SA correction to the GB solvation energy (see the Methods section). These energy differences are compared with experimental binding free energies in the last four rows of Table 4. The results are much the same as for $\Delta\langle E\rangle$.

Prior mining minima calculations for host-guest systems yielded a striking correlation between configurational entropy and energy.^{57,59,61} It is thus of interest to put the energy and entropy results for the present protein-ligand calculations into the context of these prior host-guest results. As shown in Figure 19, the protein ligand results may be viewed as broadly extending the approximately linear entropy-energy relationship found for the host-guest systems. This observation is consistent with the observation that linear regressions of measured binding free energy versus computed binding energy yield lower slopes and more positive y intercepts than linear regressions of measured binding free energy versus computed binding *free* energy (Tables 2 and 4). This is because including the entropic contribution reduces the range of the computed values and makes the computed results more positive (opposing binding).

A more careful look at Figure 19 indicates that the PDE 10a data follow the host-guest trend particularly closely. The combined host-guest and PDE 10a data yield the following regression fit: $-T\Delta S^\circ = -0.88\Delta E - 1.56$. Since $\Delta G^\circ = \Delta E - T\Delta S^\circ$, this regression corresponds to the energy-free energy relationship $\Delta G^\circ = 0.12\Delta E - 1.56$, which indicates that most of the energetic driving force for binding is predicted to be canceled by a proportionate entropic penalty. The entropic penalties for the HIVP systems fall below the trend established by the host-guest and PDE 10a

Table 2. Detailed Energy (kcal/mol) Breakdowns of Computed HIVP-Inhibitor Binding Free Energies, $\Delta G^\circ(\text{calc})$, along with Experimental Binding Free Energies, $\Delta G^\circ(\text{expt})$; Change in Mean Energy Associated with Force-Field Bond-Stretch, Angle-Bend, and Dihedral Terms (Valence); Change in Mean Force-Field Coulombic Energy (Coulomb); Change in Mean Poisson–Boltzmann Solvation Energy (PB); Change in Mean Nonpolar Surface Energy (NP); Change in Mean Force-Field Lennard-Jones Energy (VDW); Change in Mean Total Energy (ΔE , the Sum of the Prior Five Terms); and Change in Configurational Entropy Contribution to the Free Energy ($-T\Delta S^\circ$)^a

inhibitor	Boltzmann-averaged energy changes									RMSD (Å)
	$\Delta G^\circ(\text{expt})$	$\Delta G^\circ(\text{calc})$	valence	Coulomb	PB	NP	VDW	ΔE	$-T\Delta S^\circ$	
group 1a										
Amprenavir	-13.2 ^b	-19.82	-0.15	-33.28	61.00	-6.60	-69.40	-48.43	28.61	1.4/3.7 ^k
Lopinavir	-15.1 ^b	-20.22	0.87	-49.32	88.00	-7.73	-88.33	-56.51	36.30	1.4/1.1 ^l
Ritonavir	-13.5 ^c	-21.37	14.26	-78.40	101.30	-7.61	-87.99	-58.45	37.08	3.8/1.6 ^m
KNI272	-13.3 ^b	-17.90	9.07	-53.65	90.00	-7.38	-90.88	-52.85	34.95	3.3/1.6 ⁿ
KNI764	-14.3 ^b	-20.12	1.72	15.08	29.62	-6.80	-85.03	-45.41	25.29	2.5/2.3 ^o
KNI227	-16.5 ^d	-18.39	11.58	-55.50	91.13	-7.54	-91.80	-52.13	33.74	
GW0385	-19.2 ^e	-24.51	8.23	-5.45	38.77	-7.24	-88.13	-53.83	29.31	2.2/1.3 ^p
AG1132	-6.7 ^f	-15.72	12.12	-61.17	71.65	-6.11	-60.81	-44.32	28.60	
Sulfone 17b	-10.0 ^g	-13.21	6.60	-37.88	81.15	-7.39	-84.94	-42.46	29.25	
group 1b										
XK263	-13.1 ^h	-31.43	-10.03	-35.08	46.10	-7.44	-55.99	-62.43	31.01	1.5/1.3 ^q
XU422	-7.9 ⁱ	-22.84	-10.73	3.80	22.84	-7.54	-60.85	-52.49	29.65	
XZ306	-9.8 ⁱ	-25.33	-7.22	-28.51	36.26	-6.65	-47.49	-53.61	28.28	
XZ324	-10.4 ⁱ	-25.92	-9.15	-51.16	56.29	-7.60	-49.51	-61.12	35.21	
DMP128	-6.6 ⁱ	-25.23	-4.40	24.82	7.64	-6.64	-74.31	-52.88	27.66	–
group 2 ^j										
AD23 (29a)	-11.6	-19.88	2.92	-69.42	86.32	-6.98	-67.43	-54.59	34.71	
AD32 (37f)	-10.4	-16.74	-0.34	-71.59	90.86	-7.61	-62.33	-51.00	34.25	
AD67 (27b)	-14.0	-22.18	4.57	-74.23	94.32	-7.47	-74.40	-57.22	35.04	
AD74 (25c)	-13.3	-22.64	2.37	-85.13	94.82	-7.02	-62.91	-57.85	35.22	
AD81 (21e)	-16.7	-27.51	7.52	-75.63	90.62	-7.49	-76.81	-61.78	34.27	2.1/2.0 ^r
KB02 (21a)	-13.8	-22.27	0.85	-69.60	84.38	-7.06	-65.50	-56.92	34.65	
KB19 (21f)	-15.7	-12.99	-1.03	-62.36	81.58	-7.28	-60.32	-49.41	36.41	2.4/2.2 ^s
KB19 (21f)	-15.7	-20.83	-2.18	-57.59	73.80	-7.30	-60.50	-53.79	32.95	2.1/2.0
KB56 (36c)	-12.7	-19.63	-1.26	-74.88	102.96	-7.33	-70.37	-50.88	31.25	
KB60 (26d)	-14.9	-24.69	4.98	-75.40	91.24	-7.45	-74.57	-61.20	36.50	1.9/2.8 ^t
KK98 (37b)	-9.3	-14.38	3.10	-63.73	88.10	-7.49	-70.78	-50.80	36.42	

^a RMSD: root-mean-square deviation of most stable computed conformation of the bound complex for ligand alone/all mobile atoms (Angstroms). Experimental free energies sourced as follows: ^b Ref 96. ^c Ref 97. ^d Ref 98. ^e Ref 99. ^f Ref 100. ^g Ref 101. ^h Refs 77, 77, 102, and 103. ⁱ BindingDB Entry 285. ^j Ref 73. Crystal structures as follows: ^k 1HPV. ^l 1MUJ. ^m 1HXW. ⁿ 1HPX. ^o 1KZK. ^p 2FDE. ^q 1HVR. ^r 2IOD. ^s 2IOA. ^t 3GI4. ¹⁰⁹

systems, indicating that a smaller proportion of binding energy is canceled by entropy losses. Although it is tempting to interpret this result as implying that the HIVP systems tend to overcome entropic compensation, it might equally well indicate that keeping much of the protein rigid, as done in these calculations, has led to a disproportionate underestimate of entropic losses for HIVP, relative to PDE 10a. The HIVP points in the entropy-energy scatter plot also show considerably greater scatter than the PDE or host–guest results; the reasons for this are unclear.

3.3. Mode Scanning and Poisson–Boltzmann Corrections. The PB/SA solvation energy correction and the mode-scanning correction to the harmonic approximation in free energy in the energy well (see the Methods section) make the calculations somewhat more complicated and time-consuming. It is therefore of interest to examine the extent to which they contribute to accuracy. Accordingly, Table 5 presents linear regression analyses of measured binding free energies vs binding free energies computed without the mode scanning and PB/SA corrections, respectively. Omitting mode scanning has minimal effect on the results (Table 5, top), but omitting the PB/SA correction seriously damages the correlations (Table 5, bottom). This is clearly due to the

PB, rather than the SA part, given the comparatively small and uniform values of the latter (Tables 2, 3).

3.4. Convergence and Timings. The progress of the computed free energies of the various protein–ligand complexes as a function of the number of conformational search cycles (see the Methods section) is displayed in Figures 20–22 for the group 1 and group 2 HIV protease inhibitors and PDE 10a inhibitors, respectively. The graphs all trend downward because accounting for additional conformations in the mining minima method can only lower the overall free energy. Thus, the sudden downward jumps in the graphs result from the discovery of new conformations of markedly lower free energy as the search proceeds. Note that the absolute free energies displayed in these graphs are not physically interpretable on their own, because the free energies of the free proteins and their respective ligands have not been subtracted out.

The free energy calculations reported here required an average of about 4.5 h per search cycle on a 2.6 GHz dual-core AMD Opteron 2218 CPU with 4 GB of RAM. Roughly 80% of this time is spent on conformational search and 20% on calculating the free energies of individual energy wells. The CPU time for computing the

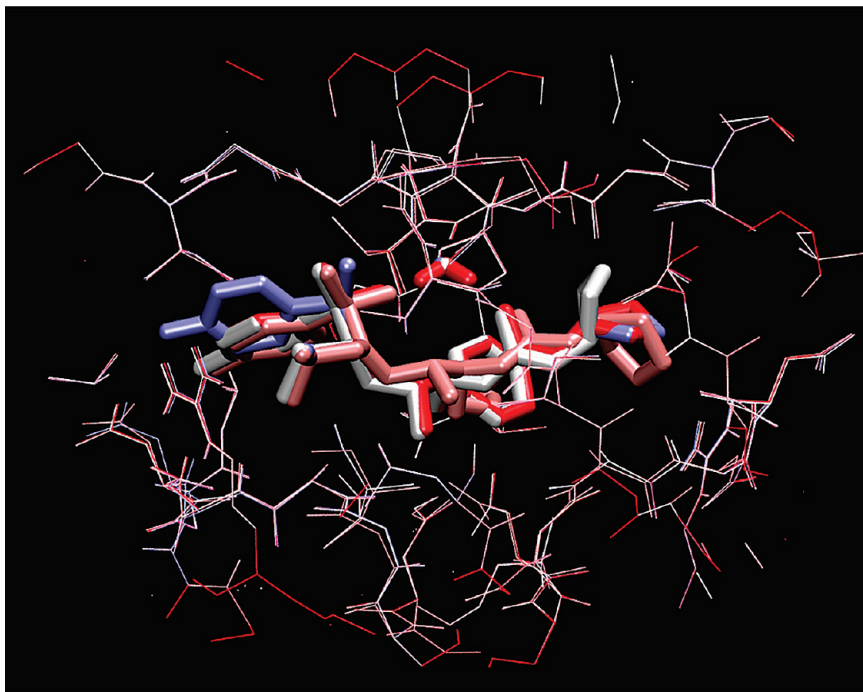


Figure 10. Lowest-energy conformers from each of the four most stable conformational clusters found for the complex of amprenavir with HIV-1 protease. Color indicates computed stability in the order red > pink > white > blue, from highest to lowest stability. Ligands and flap water shown as tubes, protein as lines.

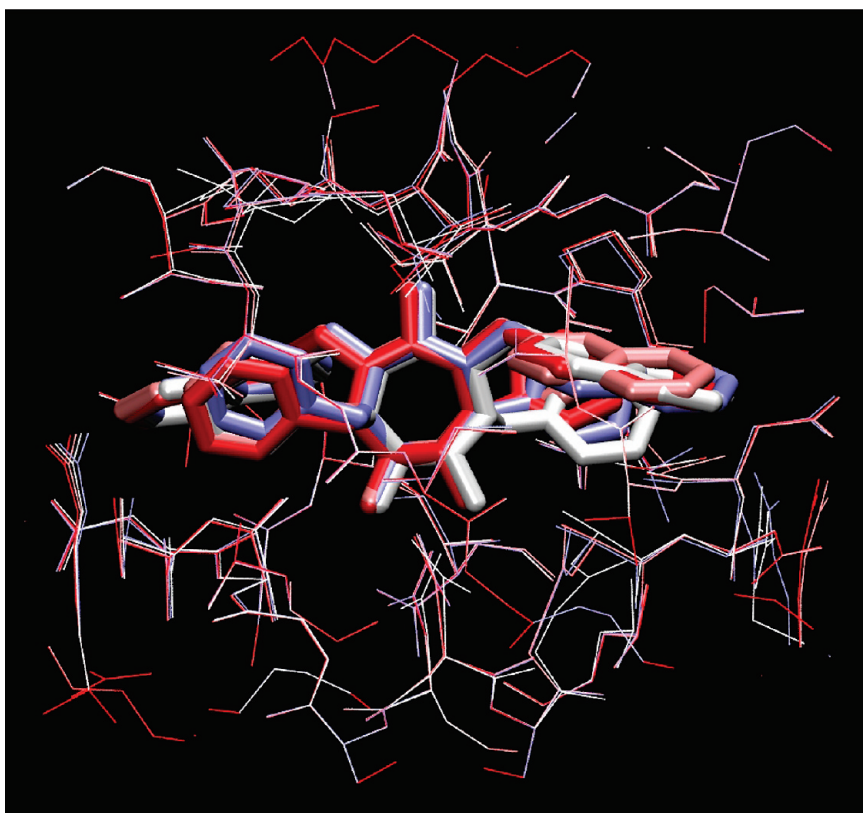


Figure 11. Lowest-energy conformers from each of the four most stable conformational clusters found for the complex of XK263 with HIV-1 protease. See caption of prior figure for details.

free energies of the individual energy wells partitions approximately as follows: 20% harmonic free energy approximation, 20% mode scanning, and 60% PB/SA solvation correction.

4. Discussion

These first applications of the mining minima approach to protein–ligand binding yield encouraging correlations be-

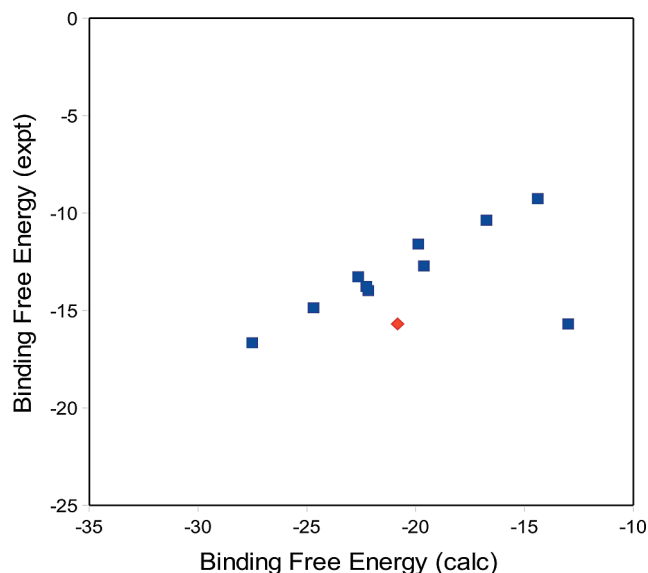


Figure 12. Measured vs computed binding free energies (kcal/mol) for group 2 HIV-1 protease inhibitors. Blue squares: initial calculations for all inhibitors. Red diamond: KB19 after recalculating the free energy of its complex based on the most stable conformation of AD32. (See text.)

tween calculated and measured binding free energies for inhibitors of two enzymes, HIV-1 protease and PDE 10a. There are several features of the present approach that may help it achieve these correlations. First, it is based upon a physics-based energy model comprising both an empirical force field and an implicit solvent model. Second, it accounts for the conformational flexibility of not only the ligand but also the side-chain and backbone atoms in the protein binding site. Third, the free energy of the ligand is computed and subtracted from the free energy of the complex, so that the computed binding free energies can account for the degree of ligand preorganization. Finally, the method accounts for changes in configurational entropy, not just energy.

The significance of the entropic contribution was tested by artificially omitting it. In the present calculations, neglecting the computed change in configurational entropy tends to weaken the correlation with experimental results, although this is not seen uniformly. In particular, omitting entropy for the group 1b HIVP inhibitors if anything slightly strengthens the correlation with experimental results. It is

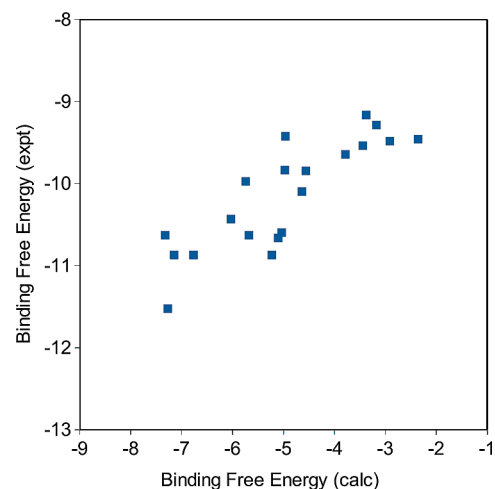


Figure 14. Experimental vs calculated binding free energies (kcal/mol) for 20 PDE 10a inhibitors.

interesting to speculate that this might be related to the greater rigidity and structural uniformity of these compounds. It is also of scientific interest that the calculations yield correlations between the change in configurational entropy on binding and the change in energy, very much like those seen in prior mining minima calculations for host–guest systems, and highly analogous to the entropy–enthalpy compensation seen in many experimental studies.^{82,83} This observation is consistent with the tendency of purely energetic physics-based models to overestimate the scale of protein–ligand binding affinities⁵⁹ since such models omit a systematically varying free energy penalty.

The limitations of this study also deserve comment. One is that we have simplified the challenge of achieving correlation with experimental results by focusing primarily on congeneric compound series and taking advantage of existing knowledge about each series, such as by including the flap water in two series of HIVP inhibitors. Also, the rather mixed structural RMSDs found for the group 1a HIVP inhibitors are a concern. On the other hand, highly accurate conformations may not be necessary in order to generate at least an approximate ranking of binding energies over the ~ 15 kcal/mol range spanned by this set. (The experimental affinity ranges are markedly smaller for group 2 and especially for the PDE 10a system.) It is also worth noting

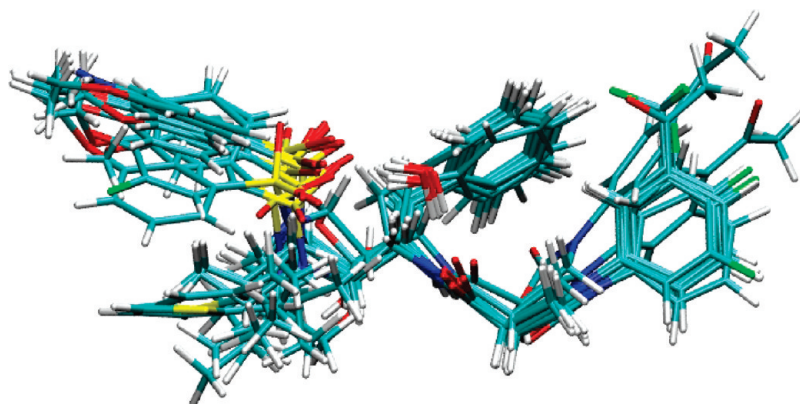


Figure 13. Most stable computed bound conformations of all 10 group 2 HIV-1 protease inhibitors, along with the flap water. Cyan: carbon. White: hydrogen. Red: oxygen. Blue: nitrogen. Green: fluorine. Yellow: sulfur.

Table 3. Detailed Energy Breakdowns for PDE 10a Inhibitors (see Table 2 for details)

inhibitor	$\Delta G^\circ(\text{expt})^a$	$\Delta G^\circ(\text{calc})$	Boltzmann-averaged energy changes					ΔE	$-\Delta\Delta S^\circ$	RMSD (Å)
			valence	Coulomb	PB	NP	VDW			
1	-10.43	-6.03	4.52	-47.90	59.50	-5.03	-46.13	-35.04	29.01	3.0/1.6 ^b
11	-9.97	-5.74	-5.42	-31.69	41.45	-5.11	-33.23	-34.00	28.26	
12	-9.29	-3.17	0.50	-18.76	30.28	-4.78	-36.91	-29.68	26.50	
13	-9.42	-4.96	-2.73	-39.91	51.37	-5.31	-36.60	-33.17	28.21	
14	-9.64	-3.78	-2.71	-38.51	51.92	-4.84	-39.00	-33.15	29.36	
15	-9.84	-4.97	-2.92	-35.10	46.50	-4.71	-36.43	-32.66	27.69	
16	-9.85	-4.56	-2.34	-39.82	54.45	-5.02	-38.81	-31.54	26.98	
17	-10.10	-4.64	-6.17	-39.27	53.23	-4.60	-33.12	-29.93	25.29	
18	-9.48	-2.91	-6.70	-34.35	49.75	-4.54	-32.35	-28.19	25.28	
19	-9.46	-2.36	-2.70	-40.66	51.79	-4.35	-31.30	-27.22	24.86	
20	-9.54	-3.44	-4.32	-32.57	38.56	-4.34	-25.40	-28.08	24.63	
21	-10.87	-7.15	-3.68	-30.23	38.97	-4.92	-33.07	-32.94	25.79	1.2/1.2 ^c
22	-9.16	-3.38	-1.87	-28.31	35.93	-4.73	-32.33	-31.31	27.94	
23	-10.63	-5.68	-1.87	-42.53	53.22	-5.09	-38.74	-35.02	29.34	
24	-10.66	-5.10	-3.80	-39.06	50.73	-5.05	-38.12	-35.30	30.20	
25	-10.87	-6.77	-0.11	-36.38	42.57	-5.09	-34.23	-33.24	26.48	
26	-10.63	-7.32	-1.65	-33.18	40.46	-5.11	-34.49	-33.97	26.65	
27	-10.87	-5.23	-2.52	-39.64	51.38	-4.83	-35.81	-31.43	26.20	
28	-10.60	-5.03	-6.43	-33.56	49.85	-4.91	-35.60	-30.65	25.62	
29	-11.53	-7.27	-5.62	-39.45	54.14	-4.91	-37.82	-33.67	26.40	0.6/1.2 ^d

^a Ref 74. ^b 2O8H. ^c 2OVV. ^d 1OVY. ⁷⁴

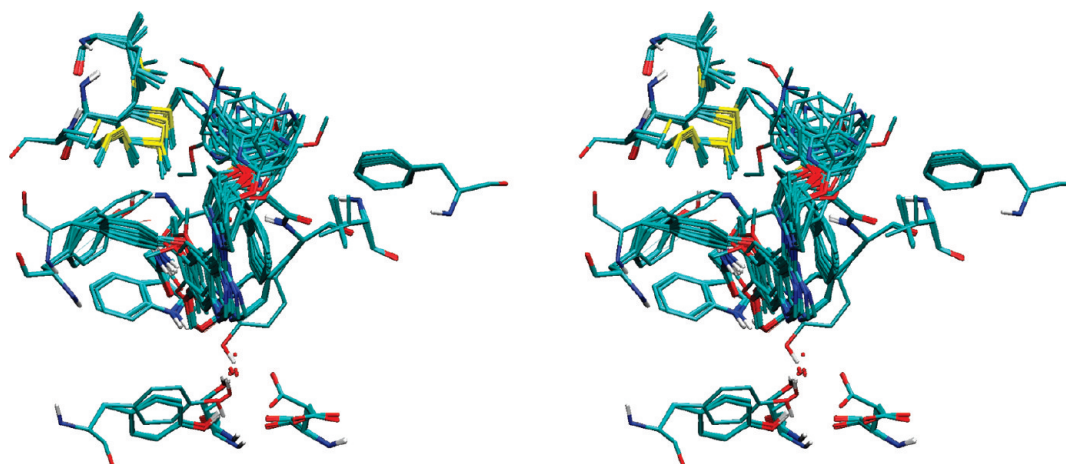


Figure 15. Most stable bound conformations computed for the complexes of the PDE 10a inhibitors, showing the ligand surrounded by mobile residues in cross-eyed stereo.

that we have not simply redocked each ligand into a protein structure solved with that ligand but have instead used a single protein structure for all calculations within the ligand series. This approach provides a more challenging and real-world test of the methodology. Finally, this is a retrospective validation study, and a full evaluation of the method will require testing of prospective predictions. The results presented here suggest that it will be worth making such an effort.

The present calculations are time-consuming, yet, as with many other free energy methods, there are still concerns regarding convergence. The convergence problem is highlighted by the example of the HIVP inhibitor KB19, which appeared to be a gross outlier until we manually adjusted its conformation and restarted the conformational search, upon which its energy fell, and it came into line with the other group 2 inhibitors. It is essential to improve this aspect of the method so that one can be more confident of obtaining converged results. It is worth noting at the outset that the specific problem with KB19 could likely have been prevented

by adjustments to the run parameters of the existing conformational search algorithm, and, indeed, we are still defining “best practices” for the use of this software. At the same time, it will clearly be advantageous to speed up the calculations in order to accommodate more extensive conformational searches. One approach is to eliminate any unneeded, time-consuming steps and thus leave more time for the conformational search. Two possibilities were considered here: mode scanning and the PB/SA solvation correction. Omitting mode scanning, which provides a correction to the anharmonicity of the energy surface near the base of each energy well, made very little difference in the results, so it would seem reasonably safe to omit this part of the procedure. On the other hand, this step uses only about 2% of the total CPU time, and further testing might uncover cases where it is important. Omitting the PB/SA correction, on the other hand, seriously damaged the correlations of calculation with experimental results. Thus, although this step is computationally costly, it is indispensable. An alternative approach is to parallelize the calcula-

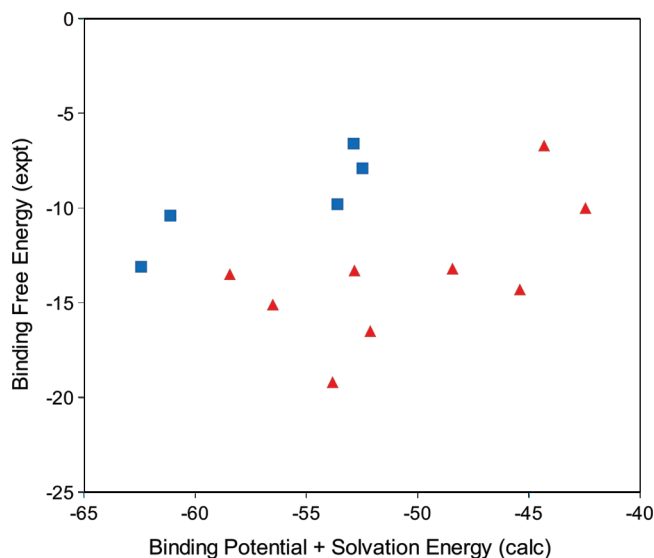


Figure 16. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for groups 1a (red triangles) and 1b (blue squares) HIV-1 protease inhibitors (kcal/mol).

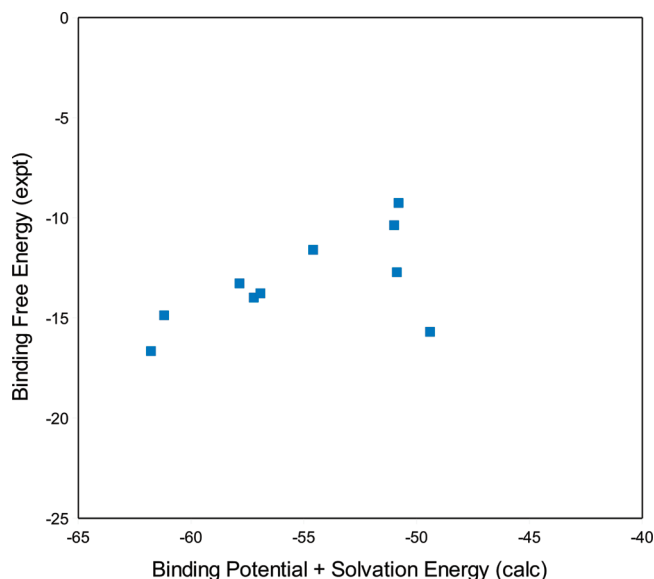


Figure 17. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for group 2 HIV-1 protease inhibitors (kcal/mol).

tions so that more extensive conformational searches can be carried out at low cost. Our preliminary efforts at fine-grained OpenMP parallelization and at porting key loops to graphical processor units (GPUs) indicate that substantial speedups can be achieved by these means. It will also be of interest to determine whether the search algorithm itself can be enhanced, such as by seeding it with candidate low-energy conformations generated with a simpler, faster algorithm.

It is of interest to compare the present mining minima method with the MMPBSA approach, because both are endpoint methods that use a force-field and an implicit solvent model to estimate binding free energies. One of the key differences is that mining minima simplifies the calculations by treating only the protein binding site as flexible. Treating

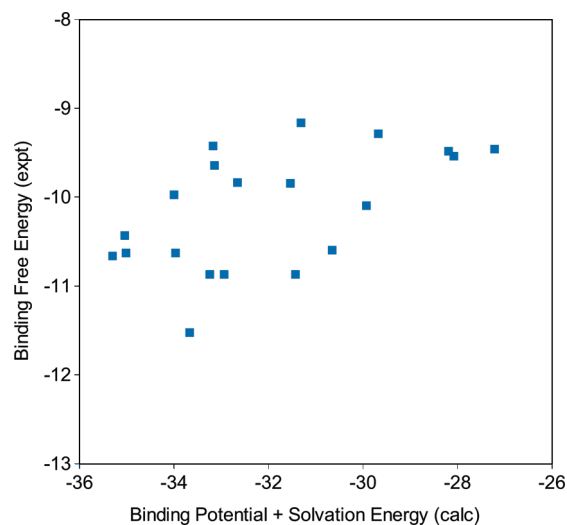


Figure 18. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for PDE 10a inhibitors (kcal/mol).

Table 4. Linear Regressions and Mean Errors of Computed Mean Binding Energies (Boltzmann-Averaged, $\Delta\langle E \rangle$ and Global Minimum, ΔE_{\min}) versus Measured Binding Free Energies, Showing Correlation Coefficient (R), Slope, and Y Intercept (kcal/mol) of Linear Regression Fits, and Mean Deviation (kcal/mol) of Calculation Relative to Experiment^a

	R	slope	Y intercept	mean error
		$\Delta\langle E \rangle$		
HIVP group 1a	0.63	0.40	6.5	-37.0
HIVP group 1b	0.86	0.44	15.4	-47.0
HIVP group 2 ^b	0.89	0.47	13.5	-42.9
PDE 10a	0.56	0.15	-5.2	-21.9
		ΔE_{\min}		
HIVP group 1a	0.67	0.42	7.3	-35.4
HIVP group 1b	0.89	0.48	16.3	-44.8
HIVP group 2 ^b	0.88	0.50	14.3	-40.7
PDE 10a	0.61	0.19	-4.3	-21.1

^a In all cases, the experimental binding free energy is considered to be on the ordinate (y axis). ^b Group 2 regressions omit the outlier, KB-19.

the entire protein as flexible, as is frequently done in MMPBSA calculations, in principle offers the possibility of accounting fully for protein-wide energy changes on binding. However, in practice, the large fluctuations of the energy of the entire protein makes it very difficult to converge the energy difference on binding. As a consequence, this contribution to the binding free energy is often completely discarded by using the single-trajectory MMPBSA approach.^{48,84} Focusing on the binding site allows mining minima to provide a detailed accounting of changes in the binding site energy, which is expected to be the most important part. Treating the protein as having a flexible binding site backed up by a shell of rigid atoms has the further benefit of allowing us to use conformational search algorithms that are much more aggressive than molecular dynamics, which is typically used in MMPBSA, without disrupting the overall protein structure. A second key difference is that MMPBSA calculations typically either neglect configurational entropy or else approximate it as an average vibrational entropy over

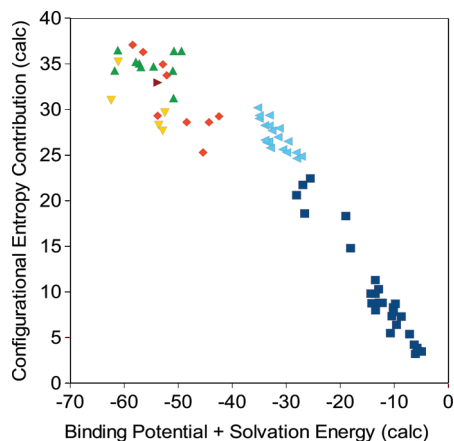


Figure 19. Computed configurational entropy contributions, $-T\Delta S^\circ$ vs energy contributions, ΔE , for the present protein–ligand calculations, along with prior host–guest results. Host–guest: blue squares. Group 1a HIVP inhibitors: red diamonds. Group 1b HIVP inhibitors: point-down yellow triangles. Group 2 HIVP inhibitors: point-up green triangles. PDE 10a: slanted, pale blue triangles. Recalculated KB19 (group 2) HIVP inhibitor: right-pointing brown triangle (kcal/mol).

Table 5. Linear Regressions of Experiment versus Computed Binding Free Energies with Neglect of Either the Mode Scanning Correction for Anharmonicity (Harmonic) or the PB/SA Solvation Correction (GB)

	<i>R</i>	slope	Y intercept
harmonic			
HIVP group 1a	0.81	0.89	2.5
HIVP group 1b	0.84	0.65	7.5
HIVP group 2*	0.97	0.56	−1.5
PDE 10a	0.83	0.37	−8.5
GB solvation			
HIVP group 1a	0.06	0.01	−12.5
HIVP group 1b	0.41	−0.08	−18.8
HIVP group 2*	0.52	0.12	−5.7
PDE 10a	0.21	0.029	−10.1

essentially randomly selected, energy-minimized molecular dynamics snapshots. We are not aware of any systematic effort to characterize or validate the MMPBSA approach. The mining minima approach, in contrast, is clearly tied to theory and has been numerically tested for small systems where reliable entropies can be obtained by brute-force methods.⁸⁶

Potential avenues for improving the accuracy of the method center on the representation of the protein and the choice of energy and solvent models. One issue is that the present calculations were simplified by omitting parts of the protein remote from the binding site. This approximation could become problematic for ionized ligands, and especially when comparing ionized ligands with neutral ones, because then long-ranged electrostatic interactions will become more important. It should be possible to address this omission efficiently and straightforwardly by using pre-computed grids of potentials to account for the influence of remote parts of the protein on the mobile atoms of the ligand and protein in the binding site.⁸⁵ The force-field used here is a Dreiding-based model, with Vcharge partial charges for

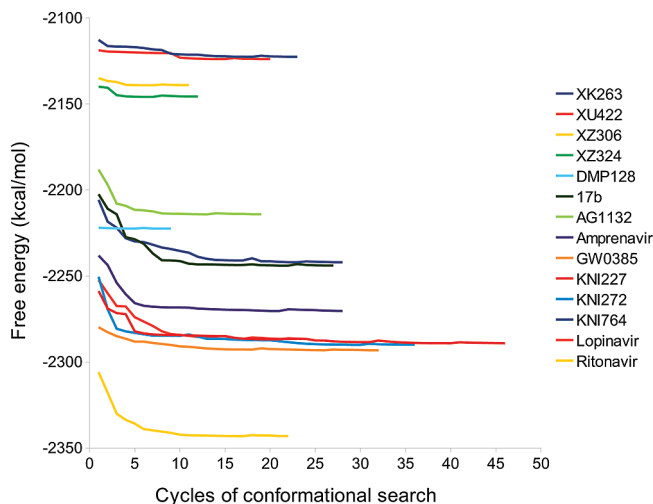


Figure 20. Convergence of computed free energies of protein–ligand complexes for group 1a and 1b HIV protease inhibitors, as a function of the number of search cycles (see Methods).

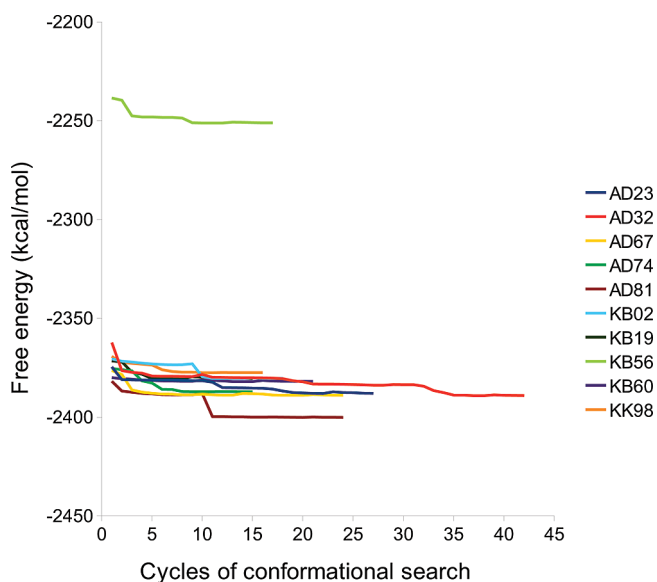


Figure 21. Same as prior figure, for group 2 HIV protease inhibitors. (The high energies of KB56 result from its cyclopropane ring, whose strain energy, evident here, cancels when the free energy of unbound KB56 is subtracted according to eq 1).

the inhibitors and CHARMM for the proteins. It is reasonable to ask whether more advanced energy models, such as ones that account for electronic polarizability,^{87–91} could increase reliability. More sophisticated energy models, potentially including quantum mechanical approaches,⁹² are likely to be important when metal atoms are closely involved in binding. It may be practical to include these as energy corrections as now done for the PB corrections, rather than having to pay the high computational cost of including them during the conformational search. Future calculations should also be able to take advantage of ongoing efforts aimed at assigning reliable parameters to a variety of drug-like ligands.^{93–95} Perhaps more critical is the treatment of the solvent. Here, we have effectively used a hybrid implicit–explicit model, since we have used knowledge of representa-

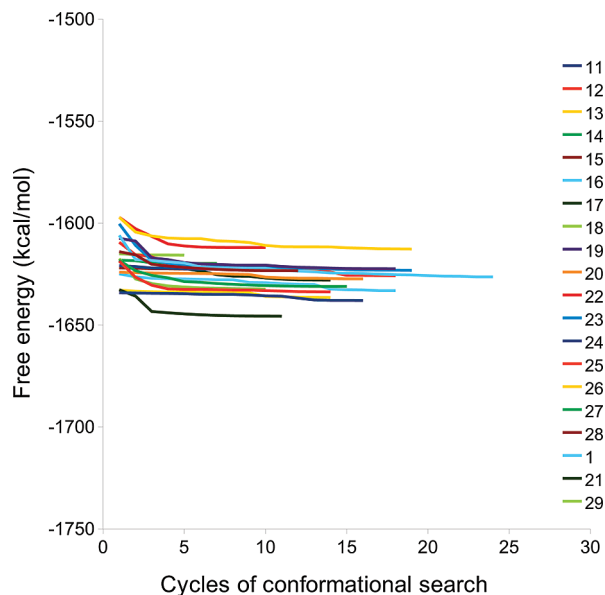


Figure 22. Same as prior two figures, for PDE 10a inhibitors. (The range of the vertical axis is matched to the prior figures to facilitate comparison.)

tive crystal structures to include key bridging waters explicitly. It would clearly be preferable to have an automatic means of accounting for the consequences of water's granular nature in the protein–ligand interface. One approach might be to routinely include one or a few explicit waters as additional “ligands” in the calculations. The present software implementation already allows this, and the added computational cost might remain manageable through parallelization, as discussed above. It should be noted, though, that such a procedure raises the theoretical issue of how to put the affinities of ligands computed with different numbers of explicit waters onto the same free energy scale. For example, part of the shift between the computed affinities of the present group 1a and group 1b HIVP inhibitors presumably results from the fact that group 1a includes one explicit water, while group 1b does not. We are currently working on this issue.

In summary, the mining minima approach has provided encouraging correlations between calculation with experimentation, and a number of avenues for further testing and improvement of the method have been identified. The approach appears to hold significant promise as a predictive method intermediate in complexity between fast, approximate docking and scoring methods and more time-consuming, rigorous free energy integration methods. It may thus be useful as a secondary computational filter for candidate ligands identified as promising by simpler, high-throughput docking methods, and as a tool to support the chemical optimization of lead compounds.

Acknowledgment. This work was made possible by grant R44GM075350 to M.J.P. from the National Institute of General Medical Sciences of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of General Medical Sciences.

Supporting Information Available: A zip file with the molecular topologies files for the drugs and drug-like

molecules studied in our paper, along with PDF document describing the parameters and units used in the topology files. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Chayen, N. *Struct. Genom., Part C* **2009**, *77*, 1–22.
- (2) Musa, T.; Ioerger, T.; Sacchettini, J. *Struct. Genom., Part C* **2009**, *77*, 41–76.
- (3) Joachimiak, A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 573–584.
- (4) Terwilliger, T.; Stuart, D.; Yokoyama, S. *Ann. Rev. Biophys.* **2009**, *38*, 371–383.
- (5) Shapiro, L.; Lima, C. *Structure* **1998**, *6*, 265–267.
- (6) Kim, S. *Curr. Opin. Struct. Biol.* **2000**, *10*, 380–383.
- (7) Congreve, M.; Marshall, F. *Br. J. Pharmacol.* **2010**, *159*, 986–996.
- (8) Rasmussen, S.; Choi, H.; Rosenbaum, D.; Kobilka, T.; Thian, F.; Edwards, P.; Burghammer, M.; Ratnala, V.; Sanishvili, R.; Fischetti, R.; Schertler, G.; Weis, W.; Kobilka, B. *Nature* **2007**, *450*, 383–U4.
- (9) Cherezov, V.; Rosenbaum, D.; Hanson, M.; Rasmussen, S.; Thian, F.; Kobilka, T.; Choi, H.; Kuhn, P.; Weis, W.; Kobilka, B.; Stevens, R. *Science* **2007**, *318*, 1258–1265.
- (10) Jaakola, V.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. *Science* **2008**, *322*, 1211–1217.
- (11) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Le Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M.; Miyano, M. *Science* **2000**, *289*, 739–745.
- (12) Gilson, M.; Zhou, H. *Ann. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (13) Villoutreix, B.; Eudes, R.; Miteva, M. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 1000–1016.
- (14) Andricopulo, A.; Salum, L.; Abraham, D. *Curr. Top. Med. Chem.* **2009**, *9*, 771–790.
- (15) Prathipati, P.; Dixit, A.; Saxena, A. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 133–148.
- (16) Huang, N.; Jacobson, M. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 325–331.
- (17) Lundstrom, K. *J. Cell. Mol. Med.* **2007**, *11*, 224–238.
- (18) Coupeuz, B.; Lewis, R. *Curr. Med. Chem.* **2006**, *13*, 2995–3003.
- (19) Warren, G.; Andrews, C.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I.; Woolven, J.; Peishoff, C.; Head, M. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (20) Shoichet, B.; Bodian, D.; Kuntz, I. *J. Comput. Chem.* **1992**, *13*, 380–397.
- (21) Jones, G.; Willett, P.; Glen, R. *J. Mol. Biol.* **1995**, *245*, 43–53.
- (22) Oshiro, C.; Kuntz, I.; Dixon, J. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 113–130.
- (23) Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–748.

- (24) McMartin, C.; Bohacek, R. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
- (25) Totrov, M.; Abagyan, R. *Prot. Struct. Funct. Genet.* **1997**, 215–220.
- (26) Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R.; Olson, A. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (27) Kramer, B.; Rarey, M.; Lengauer, T. *Prot. Struct. Funct. Genet.* **1999**, *37*, 228–241.
- (28) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (29) Raha, K.; Merz, K. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
- (30) Vasilyev, V.; Bliznyuk, A. *Theor. Chem. Acc.* **2004**, *112*, 313–317.
- (31) Grater, F.; Schwarzl, S.; Dejaegere, A.; Fischer, S.; Smith, J. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.
- (32) Curioni, A.; Mordasini, T.; Andreoni, W. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 773–784.
- (33) Raha, K.; Merz, K. *J. Med. Chem.* **2005**, *48*, 4558–4575.
- (34) Rajamani, R.; Good, A. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 308–315.
- (35) Fong, P.; McNamara, J.; Hillier, I.; Bryce, R. *J. Chem. Inf. Model.* **2009**, *49*, 913–924.
- (36) Cho, A.; Chung, J.; Kim, M.; Park, K. *J. Chem. Phys.* **2009**, 131.
- (37) Straatsma, T.; McCammon, J. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (38) Tembe, B.; McCammon, J. *Comput. Chem.* **1984**, *8*, 281–283.
- (39) Wang, D.; Rizzo, R.; Tirado-Rives, J.; Jorgensen, W. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2799–2802.
- (40) McCarrick, M.; Kollman, P. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 109–121.
- (41) Lau, F. T. K.; Karplus, M. *J. Mol. Biol.* **1994**, *236*, 1049–1066.
- (42) Dolenc, J.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. *Nucleic Acids Res.* **2005**, *33*, 725–733.
- (43) Woo, H.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.
- (44) Mobley, D.; Chodera, J.; Dill, K. *J. Chem. Phys.* **2006**, 125.
- (45) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C.; Shirts, M.; Sorin, E.; Pande, V. *J. Chem. Phys.* **2005**, 123.
- (46) McDonald, J.; Brooks, C., III *J. Am. Chem. Soc.* **1991**, *113*, 2295–2301.
- (47) Reddy, M.; Viswanadhan, V.; Weinstein, J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10287–10291.
- (48) Brown, S.; Muchmore, S. *J. Chem. Inf. Model.* **2006**, *46*, 999–1005.
- (49) Gouda, H.; Kuntz, I.; Case, D.; Kollman, P. *Biopolymers* **2003**, *68*, 16–34.
- (50) Pearlman, D. *J. Med. Chem.* **2005**, *48*, 7796–7807.
- (51) Srinivasan, J.; Cheatham, T.; Cieplak, P.; Kollman, P.; Case, D. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (52) Aqvist, J.; Medina, C.; Samuelsson, J. *Protein Eng.* **1994**, *7*, 385–391.
- (53) Carlsson, J.; Ander, M.; Nervall, M.; Aqvist, J. *J. Phys. Chem. B* **2006**, *110*, 12034–12041.
- (54) JonesHertzog, D.; Jorgensen, W. *J. Med. Chem.* **1997**, *40*, 1539–1549.
- (55) Zhou, R.; Friesner, R.; Ghosh, A.; Rizzo, R.; Jorgensen, W.; Levy, R. *J. Phys. Chem. B* **2001**, *105*, 10388–10397.
- (56) Gilson, M.; Given, J.; Head, M. *Chem. Biol.* **1997**, *4*, 87–92.
- (57) Chang, C.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- (58) Head, M.; Given, J.; Gilson, M. *J. Phys. Chem. A* **1997**, *101*, 1609–1618.
- (59) Chen, W.; Chang, C.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 3035–3049.
- (60) Kolossvary, I. *J. Am. Chem. Soc.* **1997**, *119*, 10233–10234.
- (61) Chen, W.; Chang, C.; Gilson, M. K. *J. Am. Chem. Soc.* **2006**, *128*, 4675–4684.
- (62) Moghaddam, S.; Inoue, Y.; Gilson, M. *J. Am. Chem. Soc.* **2009**, *131*, 4012–4021.
- (63) Chang, C.; Gilson, M. K. *J. Comput. Chem.* **2003**, *24*, 1987–1998.
- (64) Kolossvary, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.
- (65) Keserü, G. M.; Kolossvary, I. *J. Am. Chem. Soc.* **2001**, *123*, 12708–12709.
- (66) Momany, F. A.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888–900.
- (67) Mayo, S.; Olafson, B.; Goddard, W. *J. Phys. Chem.* **1990**, *94*, 8897–8909.
- (68) Gilson, M.; Gilson, H.; Potter, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.
- (69) Gilson, M.; Honig, B. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 5–20.
- (70) Qiu, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.
- (71) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.
- (72) Cai, Q.; Hsieh, M.; Wang, J.; Luo, R. *J. Chem. Theory Comput.* **2010**, *6*, 203–211.
- (73) Ali, A.; Reddy, G. S. K. K.; Cao, H.; Anjum, S. G.; Nalam, M. N. L.; Schiffer, C. A.; Rana, T. M. *J. Med. Chem.* **2006**, *49*, 7342–7356.
- (74) Chappie, T. A.; Humphrey, J. M.; Allen, M. P.; Estep, K. G.; Fox, C. B.; Lebel, L. A.; Liras, S.; Marr, E. S.; Menniti, F. S.; Pandit, J.; Schmidt, C. J.; Tu, M.; Williams, R. D.; Yang, F. V. *J. Med. Chem.* **2007**, *50*, 182–185.
- (75) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
- (76) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III *J. Med. Chem.* **2004**, *47*, 3032–3047.
- (77) Lam, P.; Jadhav, P.; Eyermann, C.; Hodge, C.; Ru, Y.; Bacheler, L.; Meek, J.; Otto, M.; Rayner, M.; Wong, Y. *Science* **1994**, *263*, 380–384.
- (78) Miller, J. F.; Andrews, C. W.; Brieger, M.; Furfine, E. S.; Hale, M. R.; Hanlon, M. H.; Hazen, R. J.; Kaldor, I.

- McLean, E. W.; Reynolds, D.; Sammond, D. M.; Spaltenstein, A.; Tung, R.; Turner, E. M.; Xu, R. X.; Sherrill, R. G. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1788–1794.
- (79) Silva, A. M.; Cachau, R. E.; Sham, H. L.; Erickson, J. W. *J. Mol. Biol.* **1996**, *255*, 321–340.
- (80) Yamazaki, T.; Nicholson, L.; Torchia, D.; Wingfield, P.; Stahl, S.; Kaufman, J.; Eyermann, C.; Hodge, C.; Lam, P.; Ru, Y.; Jadhav, P.; Chang, C.; Weber, P. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.
- (81) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14* (33–38), 27–28.
- (82) Lumry, R.; Rajender, S. *Biopolymers* **1970**, *9*, 1125–1227.
- (83) Sharp, K. *Protein Sci.* **2001**, *10*, 661–667.
- (84) Lee, M. S.; Olson, M. A. *Biophys. J.* **2006**, *90*, 864–877.
- (85) Given, J. A.; Gilson, M. K. *Proteins* **1998**, *33*, 475–495.
- (86) Chang, C.; Potter, M. J.; Gilson, M. K. *J. Chem. Phys. B* **2003**, *107*, 1048–1055.
- (87) Lopes, P.; Roux, B.; MacKerell, A. *Theor. Chem. Acc.* **2009**, *124*, 11–28.
- (88) Maple, J.; Cao, Y.; Damm, W.; Halgren, T.; Kaminski, G.; Zhang, L.; Friesner, R. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.
- (89) Ren, P.; Ponder, J. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- (90) Cho, A.; Guallar, V.; Berne, B.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915–931.
- (91) Kaminski, G.; Stern, H.; Berne, B.; Friesner, R.; Cao, Y.; Murphy, R.; Zhou, R.; Halgren, T. *J. Comput. Chem.* **2002**, *23*, 1515–1531.
- (92) Peters, M.; Raha, K.; Merz, K. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 370–379.
- (93) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (94) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. *J. Comput. Chem.* **2010**, *31*, 671–690.
- (95) Jakalian, A.; Jack, D.; Bayly, C. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (96) Ohtaka, H.; Schon, A.; Freire, E. *Biochemistry* **2003**, *42*, 13659–13666.
- (97) Todd, M. J.; Luque, I.; Velazquez-Campoy, A.; Freire, E. *Biochemistry* **2000**, *39*, 11876–11883.
- (98) Kageyama, S.; Mimoto, T.; Murakawa, Y.; Nomizu, M.; Ford, H.; Shirasaka, T.; Gulnik, S.; Erickson, J.; Takada, K.; Hayashi, H. *Antimicrob. Agents Chemother.* **1993**, *37*, 810–817.
- (99) Hanlon, M. H.; Porter, D. J. T.; Furfine, E. S.; Spaltenstein, A.; Carter, H. L.; Danger, D.; Shu, A. Y. L.; Kaldor, I. W.; Miller, J. F.; Samano, V. A. *Biochemistry* **2004**, *43*, 14500–14507.
- (100) Melnick, M.; Reich, S. H.; Lewis, K. K.; Mitchell, L. J.; Nguyen, D.; Trippe, A. J.; Dawson, H.; Davies, J. F.; Appelt, K.; Wu, B.; Musick, L.; Gehlhaar, D. K.; Webber, S.; Shetty, B.; Kosa, M.; Kahil, D.; Andrada, D. *J. Med. Chem.* **1996**, *39*, 2795–2811.
- (101) Specker, E.; Bottcher, J.; Heine, A.; Sottriffer, C. A.; Lilie, H.; Schoop, A.; Muller, G.; Griebenow, N.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6607–6619.
- (102) Hodge, C. N.; Aldrich, P. E.; Bacheler, L. T.; Chang, C.; Eyermann, C. J.; Garber, S.; Grubb, M.; Jackson, D. A.; Jadhav, P. K.; Korant, B. *Chem. Biol.* **1996**, *3*, 301–314.
- (103) Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLucca, G. V.; Eyermann, C. J.; Chang, C.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Jackson, D. A.; Erickson-Viitanen, S.; Hodge, C. N. *J. Med. Chem.* **1996**, *39*, 3514–3525.
- (104) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. *J. Am. Chem. Soc.* **1995**, *117*, 1181–1182.
- (105) Stoll, V.; Qin, W.; Stewart, K. D.; Jakob, C.; Park, C.; Walter, K.; Simmer, R. L.; Helfrich, R.; Bussiere, D.; Kao, J.; Kempf, D.; Sham, H. L.; Norbeck, D. W. *Bioorg. Med. Chem.* **2002**, *10*, 2803–2806.
- (106) Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P. *Proc. Natl. Acad. Sci. U.S.A* **1995**, *92*, 2484–2488.
- (107) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Liu, B.; Topol, I. A.; Kiso, Y.; Mimoto, T.; Mitsuya, H.; Erickson, J. W. *Structure* **1995**, *3*, 581–590.
- (108) Reiling, K. K.; Endres, N. F.; Dauber, D. S.; Craik, C. S.; Stroud, R. M. *Biochemistry* **2002**, *41*, 4582–4594.
- (109) Nalam, M. N.; Ali, A.; Altman, M. D.; Reddy, G. S. K. K.; Chellappan, S.; Kairys, V.; Ozen, A.; Cao, H.; Gilson, M. K.; Tidor, B.; Rana, T. M.; Schiffer, C. A. *J. Virol.* **2010**, 02531–09.

CT100245N

Nonpolar Solvation Free Energies of Protein–Ligand Complexes

Samuel Genheden,[†] Jacob Kongsted,[‡] Pär Söderhjelm,[§] and Ulf Ryde^{*†}

Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, SE-221 00 Lund, Sweden; Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; and Department of Chemistry and Applied Biosciences—Computational Science, ETH Zürich, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

Received May 24, 2010

Abstract: Recent investigations have indicated that different solvation methods give qualitatively different results for the nonpolar solvation contribution to ligand-binding affinities. Therefore, we have calculated the nonpolar solvation contribution to the free energy of benzene binding to the T4 lysozyme Leu99Ala mutant using thermodynamic integration (TI) and three approximate methods. The total binding free energy was calculated with TI and then decomposed into contributions from the solvent and the solute. The nonpolar contribution from the solute was compared to approximate methods within the framework of the molecular-mechanics and generalized Born with surface area method (MM/GBSA). First, the nonpolar solvation energy was calculated with a linear relation to the solvent-accessible surface area (SASA). Second, a recent approach that divides the nonpolar solvation energy into cavity and dispersion parts was used, and third, the nonpolar solvation energy was calculated with the polarized continuum model (PCM). Surprisingly, the simple SASA estimate reproduces the TI results best. However, the reason for this is that all continuum methods assume that the benzene cavity is filled with water for the free protein, contrary to both experimental and simulation results. We present a method to avoid this assumption and then, PCM provides results that are closest to the results obtained with TI.

Introduction

Solvation plays a crucial role when a small ligand binds to a protein. Therefore, accurate methods to estimate the solvation free energy, ΔG_{solv} , of protein–ligand complexes are of great importance in drug design.^{1,2} The free energy of solvation can in principle be calculated by thermodynamic integration (TI) or free energy perturbation (FEP), but these methods are computationally very expensive and can hardly

be used to calculate the solvation energy of a full protein. Therefore, a large number of less computationally demanding and more approximate methods have been developed, based on a dielectric continuum treatment of the solvent, and they have been successfully used in many applications.^{3–6}

Continuum solvation methods normally split ΔG_{solv} into a polar and a nonpolar contribution.^{7,8} The polar solvation free energy, $\Delta G_{\text{polsolv}}$, can be accurately estimated by various methods, e.g., the polarized continuum model (PCM), the generalized Born methods, or by solving the Poisson–Boltzmann equation.³ The accuracy of such calculations in the context of protein–ligand complexes has been extensively studied elsewhere.^{9,10} In this work, we are mainly interested in the nonpolar contribution to the solvation free energy, $\Delta G_{\text{nonpolsolv}}$.

* To whom correspondence should be addressed. Tel: +46 - 46 2224502. Fax: +46 - 46 2228648. E-mail: Ulf.Ryde@teokem.lu.se.

[†] Department of Theoretical Chemistry, Lund University, Chemical Centre.

[‡] Department of Physics and Chemistry, University of Southern Denmark.

[§] Department of Chemistry and Applied Biosciences—Computational Science, ETH Zürich.

A commonly used method to estimate $\Delta G_{\text{nonpolsolv}}$ is to relate it to the solvent-accessible surface area (SASA) through a linear relation:

$$\Delta G_{\text{nonpolsolv}}^{\text{SASA}} = \gamma_{\text{SASA}} \text{SASA} + b_{\text{SASA}} \quad (1)$$

where γ_{SASA} and b_{SASA} are taken from a linear regression of the solvation free energy of a set of small apolar molecules in water.^{11,12} This is the method used in the MM/GBSA (molecular mechanics with generalized Born and surface-area) approach to calculate ligand-binding free energies. Equation 1 will be referred to as the SASA approach in the following.

More rigorous approaches can be derived by considering the binding process in more detail. At the next level of approximation, the solvent contribution to the nonpolar binding free energy has been divided into two parts.^{13,14} First, a cavity in the solvent is created that can accommodate the solute. Then, the apolar solute is introduced into the cavity. The energy of the first step is usually estimated by multiplying some kind of molecular surface (MS) with the surface tension of water, γ_{wat} ,

$$\Delta G_{\text{cavity}} = \gamma_{\text{wat}} \text{MS} \quad (2)$$

Even though this resembles the SASA approach in eq 1, the two γ values are intrinsically different: γ_{SASA} in eq 1 is obtained from a linear fit, whereas γ_{wat} in eq 2 is a measurable physical constant of water. The energy of the second step can be estimated from the mean-field interaction energy between the solute and the solvent. In most force fields, it is modeled by a Lennard-Jones potential, i.e., by an attractive dispersion term and a repulsive exchange-repulsion term,

$$\Delta G_{\text{insertion}} = \Delta E_{\text{disp}} + \Delta E_{\text{rep}} \quad (3)$$

In a recent approach of Tan et al.,¹⁵ ΔE_{rep} was merged into the cavity term, so that the nonpolar solvation free energy was estimated by the following:

$$\Delta G_{\text{nonpolsolv}}^{\text{CD}} = \gamma_{\text{CD}} \text{MS} + b_{\text{CD}} + \Delta E_{\text{disp}} = \Delta G_{\text{CR}} + \Delta E_{\text{disp}} \quad (4)$$

Several molecular surfaces and volume estimates (in the latter case, γ_{CD} is a pressure term) were examined in ref 15, but it was found that solvation free energies of small drug-like molecules were indifferent to the choice of the surface or volume (probably because these two measures are strongly correlated for small molecules). The dispersion term was estimated by a volume integral. Methods based on eq 4 are referred to as the cavity–dispersion (CD) method.

Another, still more complicated, approach is used in the polarized continuum model (PCM).¹⁶ In this approach, a separate term for the exchange repulsion is also included, giving three terms for the nonpolar solvation free energy:

$$\Delta G_{\text{nonpolsolv}}^{\text{PCM}} = \Delta G_{\text{cavity}} + \Delta E_{\text{disp}} + \Delta E_{\text{rep}} \quad (5)$$

The cavity term is calculated from expressions of the radius of each atom to the power of 0 to 3,¹⁷ i.e., including both

area and volume terms. The two other terms are based on volume integrals.¹⁸

Recently, we have used PCM in an MM/GBSA framework to estimate ligand-binding affinities, i.e., we replaced the generalized Born and the SASA estimate with PCM calculations.¹⁹ In those calculations, we observed a qualitative difference between the SASA and PCM estimates of the nonpolar solvation free energy. For the binding of a series of seven biotin analogs to avidin, SASA predicted a small and negative nonpolar contribution to the binding affinity, whereas the nonpolar part of PCM was three times larger and positive. Interestingly, the 3D-RISM method gives a nonpolar estimate of the solvation free energy similar to that of PCM for the same test case.¹⁰ Of course, it is a major problem if different continuum solvation approaches give qualitatively different estimates of the same contribution. In this work, we use strict TI calculations of the absolute binding affinity to establish which of the SASA and PCM estimates is more reliable.

As a test case, we use the binding of benzene to an engineered apolar cavity in T4 lysozyme. This system has previously been used in theoretical studies because the system is experimentally well characterized²⁰ and the benzene molecule is small so it is possible to compute accurate absolute binding free energies.^{21,22}

Methods

TI Calculations. The binding of a ligand to its receptor can be described with the thermodynamic cycle in Figure 1. This double-decoupling method has been employed extensively.^{21–25} P is the protein and L is the ligand. L' denotes an apolar ligand, i.e., a ligand with all the atomic charges set to zero, and L'' denotes an apolar ligand where also the van der Waals interactions with the surroundings have been turned off. Restraints are applied to the ligand relative to the receptor so that a standard state can be defined. This also improves the convergence of the simulations.^{23,25} The restraints are denoted by “restr” in Figure 1. According to the thermodynamic cycle in the figure, the binding free energy can be written as follows:

$$\Delta G_{\text{bind}} = \Delta G_{\text{ele}}^{\text{free}} + \Delta G_{\text{vdW}}^{\text{free}} - (\Delta G_{\text{ele}}^{\text{bound}} + \Delta G_{\text{vdW}}^{\text{bound}}) + \Delta \Delta G_{\text{restr}} \quad (6)$$

where the terms denote the free energy of removing the charges of the unbound ligand in bulk water, the free energy of turning off the van der Waals interaction of the unbound ligand, the free energy of removing the charges of the bound ligand, the free energy of turning off the van der Waals interaction of the ligand with the surrounding protein and water, and the free energy difference of applying and removing the restraints, $\Delta \Delta G_{\text{restr}} = -(\Delta G_{\text{restr}+} + \Delta G_{\text{restr}-})$. We define the nonpolar contribution to the free energy as $\Delta G_{\text{nonpol}} = \Delta G_{\text{vdW}}^{\text{free}} - \Delta G_{\text{vdW}}^{\text{bound}}$ and similar for the polar contribution, $\Delta G_{\text{pol}} = \Delta G_{\text{ele}}^{\text{free}} - \Delta G_{\text{ele}}^{\text{bound}}$.

The free energies associated with removing the charges or the van der Waals interactions were calculated with TI.²⁶ By TI, the free energy between two states is calculated as follows:

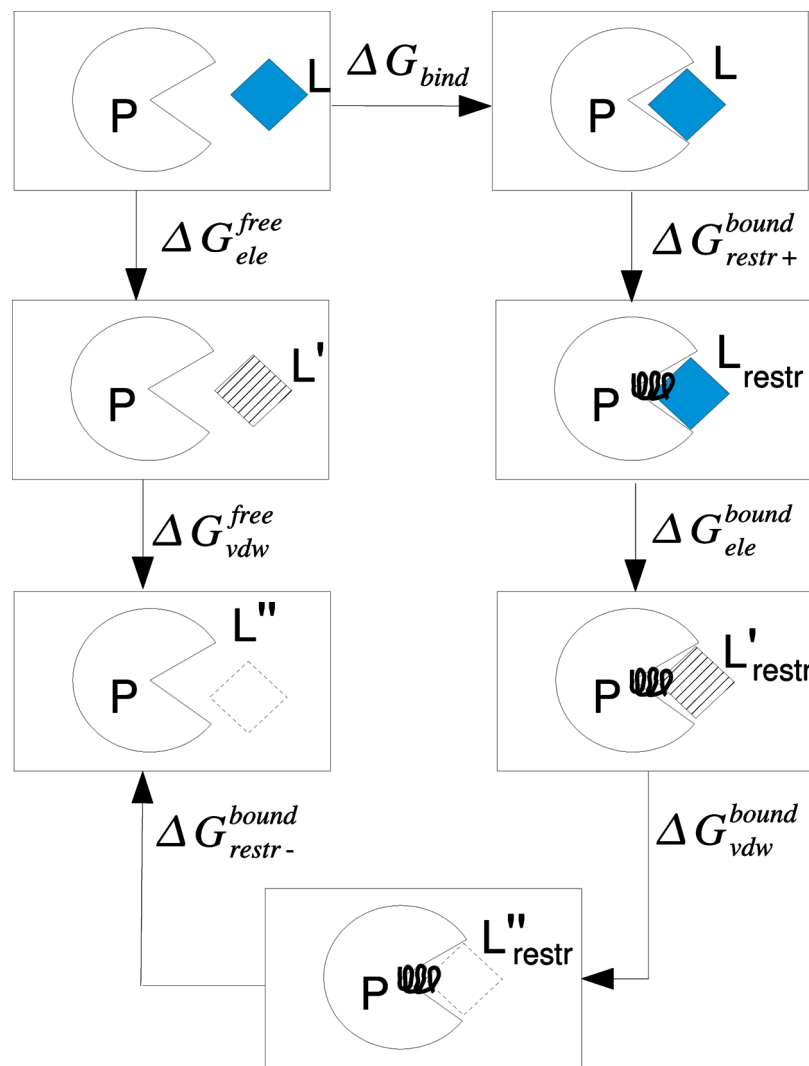


Figure 1. The thermodynamic cycle used to describe the binding of a ligand to its receptor.

$$\Delta G = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (7)$$

where the brackets indicate an ensemble average and V is defined as follows:

$$V(\lambda) = (1 - \lambda)V_0 + \lambda V_1 \quad (8)$$

where V_0 and V_1 are the potential energy of initial and final states, respectively. λ is a coupling parameter, describing the amount of the initial and final states in V . The integral in eq 7 was estimated by simulating the system at 11 distinct values of λ (0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95), followed by a trapezoid integration. The λ values at 0 and 1 were estimated with linear extrapolation from the two nearest points. Additional λ points were added where the formula $\delta V/\delta \lambda$ curve was less smooth or where there was a sharp change in the curvature, but no significant changes were obtained. The overlap between the $\delta V/\delta \lambda$ distribution at the different integration points was also checked. To avoid end-point problems in the van der Waals calculations, soft-core potentials, as implemented in the Amber simulation package, were used.²⁷

The value of $\Delta \Delta G_{\text{restr}}$ can in principle be calculated by TI as well, but we used an approximation applicable in the limit

of strong restraints.^{21,24} The restraints are constructed with respect to three points in the protein (P_1, P_2, P_3) and three points in the ligand (L_1, L_2, L_3). Following previous work on lysozyme,²¹ P_1 should be an atom lying close to the center of mass of the protein and was chosen to be the CB atom of Phe104, P_2 is the N atom in Met1, and P_3 is the C atom of Tyr161. For the benzene, three not connected carbon atoms were selected, C1, C3, and C5. One distance, two angles, and three dihedral angles were then defined based on these points, viz. the distance from P_1 to L_1 (r) the angles P_2 - P_1 - L_1 and P_1 - L_1 - L_2 (Θ and α), and the dihedral angles P_3 - P_2 - P_1 - L_1 , P_2 - P_1 - L_1 - L_2 and P_1 - L_1 - L_2 - L_3 (ϕ , β , and ζ). These geometric parameters were harmonically restrained to their values in the crystal structure with the force constants of 41.84 kJ/mol/Å² for the bond and 836.8 kJ/mol/rad² for the angles and the dihedral angles.²¹ In the limit of strong restraints, the free energy of applying the restraints can be estimated as follows:²⁴

$$\exp\left(-\frac{\Delta \Delta G_{\text{restr}}}{kT}\right) = C^\circ (2\pi)^{(3/2)} r_0^2 \sigma_r \sigma_\Theta \sigma_\phi \sin \theta_0 + \frac{\sigma_\alpha \sigma_\beta \sigma_\zeta \sin \alpha_0}{\sqrt{8\pi}} \quad (9)$$

where k is Boltzmann's constant, T the absolute temperature, C° is the standard concentration ($1/1661 \text{ \AA}^{-3}$), r_0 , θ_0 , and α_0 are the initial values of r , θ , and α , i.e., the values in the crystal structure, and the six σ values correspond to the standard deviations of the corresponding geometric parameter in an unrestrained simulation. These standard deviations were estimated from the same simulations that were used to compute the MM/GBSA estimates. A symmetry number of 12 for benzene was used for the rotational restraints.²¹

MM/GBSA Calculations. The binding free energy, ΔG_{bind} , of the benzene to T4 lysozyme was also estimated with the MM/GBSA method^{28,29} according to the following:

$$\Delta G_{\text{bind}} = \langle G^{\text{PL}} \rangle - \langle G^{\text{P}} \rangle - \langle G^{\text{L}} \rangle \quad (10)$$

where PL is the protein–ligand complex, P is the protein, and L is the ligand. The brackets indicate ensemble averages over a MD simulation. Each free energy is estimated by the following:

$$G = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdW}} + G_{\text{polsolv}} + G_{\text{nonpolsolv}} - TS_{\text{MM}} \quad (11)$$

where E_{int} , E_{ele} , and E_{vdW} are the internal, electrostatic, and van der Waals interactions, calculated with infinite cutoff. G_{polsolv} is the polar solvation energy, estimated by the generalized Born method of Onufriev et al., model I (GB^{OBCL}),³⁰ i.e., with $\alpha = 0.8$, $\beta = 0$, and $\gamma = 2.91$. $G_{\text{nonpolsolv}}$ is the nonpolar solvation energy, estimated from the SASA, using eq 1 with $\gamma_{\text{SASA}} = 0.0227 \text{ kJ/mol/\AA}^2$ and $b_{\text{SASA}} = 3.85 \text{ kJ/mol}$.^{29,31} In addition, we also estimated $G_{\text{nonpolsolv}}$ by the CD method,¹⁵ using the σ decomposition and radii optimized by Tan et al.³² For the cavity term, the solvent-accessible volume was used as the MS,³³ $\gamma_{\text{CD}} = 0.0378 \text{ kJ/mol/\AA}^2$, $b_{\text{CD}} = -0.5692 \text{ kJ/mol}$, and the probe radius was 1.3 \AA . For the dispersion term, a probe radius of 0.557 \AA was used and the water density was set to 1.129 kg/L .^{15,33} S_{MM} is the entropy, estimated from harmonic frequencies, calculated at the MM level²⁹ on a truncated and buffered system, as described previously, to improve the statistical precision of the estimate.³⁴ All the terms in eq 11 were averages over the last snapshot from 20 independent MD simulations. In general, more snapshots are required to obtain a high precision in ΔG_{bind} ,³⁵ but the $\Delta G_{\text{nonpolsolv}}$ term is usually very stable and 20 snapshots is enough to obtain converged results for this term. As is customary in MM/GBSA, the same geometry was used for all three reactants (i.e., only the complex was simulated).²⁹ Then, the E_{int} term cancels in eq 11. All MM/GBSA calculations were performed with the Amber 10 software.³⁶

PCM Calculations. We also calculated $\Delta G_{\text{nonpolsolv}}$ with PCM on the same snapshots that were used to estimate MM/GBSA energies. We used the integral-equation formulation of PCM, IEFPCM,³⁷ which exhibits a better numerical stability than other formulations of PCM and is the default PCM method in the Gaussian software.³⁸ Owing to the large size of the molecular systems, the PCM induced charges were obtained using a direct inversion of the iterative subspace procedure,³⁹ as implemented in the GAMESS software.⁴⁰ Thus, no explicit matrix inversion is needed. We employed

the UAKS radii as implemented in Gaussian³⁸ and a scaling factor for the polar part of 1.15.¹⁹

Explicit Calculations of Ligand–Solvent van der Waals Interaction. The van der Waals interaction energy between the benzene molecule and the surrounding water molecules was estimated using 20 independent simulations of benzene free in bulk water or bound to T4 lysozyme. The interaction energy was calculated every 5 ps using infinite cutoff, but with no correction for long-range interactions. The total van der Waals interaction energy was decomposed into repulsive and attractive contributions using three schemes as described in.¹⁵ These were 12–6, which just separates the r^{-12} and r^{-6} terms in the Lennard-Jones potential, the Weeks–Chandler–Anderson (WCA) approach,⁴¹ which separates the potential at the point where the force changes sign, and the σ approach,¹⁵ which separates the potential at the point where the energy changes sign.

System Preparation. The calculations were based on the crystal structure 181L of the Leu99Ala mutant of T4 lysozyme in complex with benzene⁴² and the protein was prepared as described previously.⁴³ All Asp and Glu residues were assumed to be negatively charged and all Lys and Arg residues were positively charged. The single histidine residue was assumed to be protonated on the ND1 atom, based on the local surroundings and the hydrogen-bond network. The protein was described by the Amber99SB force field⁴⁴ and the benzene molecule with the generalized Amber force field⁴⁵ with charges derived from a RESP calculation,⁴³ using electrostatic potentials calculated at the HF/6-31G* level in points sampled according to the Merz–Kollman scheme.⁴⁶ The protein–ligand system was immersed in a truncated octahedral box of TIP4P-Ewald waters,⁴⁷ extending at least 8 \AA from the protein (7774 water molecules and 33711 atoms in total for the simulations of the complex).

MD Simulations. All MD simulations were run by the sander module of Amber 10.³⁶ The SHAKE algorithm⁴⁸ was used to constrain bonds involving hydrogen atoms so that a time step of 2 fs could be used. The temperature was kept constant at 300 K using a Langevin thermostat⁴⁹ with a collision frequency of 2.0 ps^{-1} . The pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm⁵⁰ with a relaxation time of 1 ps. Particle-mesh Ewald summation,⁵¹ with a fourth-order B spline interpolation and a tolerance of 10^{-5} was used to handle long-range electrostatics. The nonbonded cutoff was 8 \AA and the nonbonded pair list was updated every 50 fs.

Unconstrained protein–ligand simulations, used in the MM/GBSA analysis, were generated in the following way: A single system was first energy minimized for 100 cycles of steepest descent, with all atoms, except water molecules and hydrogen atoms, restrained to their start position with a force constant of 418 kJ/mol/\AA^2 . This was followed by a 20 ps NPT simulation with the same restraints, and a 100 ps unconstrained NPT equilibration. From this equilibrated structure, 20 independent simulations were initiated by assigning random starting velocities. These simulations were further equilibrated for 50 ps in the NPT ensemble and then, a 200 ps production run were performed, also in the NPT

Table 1. ΔG_{nonpol} as Calculated by TI in kJ/mol

	total	solvent	solute
$\Delta G_{\text{vdW}}^{\text{bound}}$	38.7 ± 1.6	2.2 ± 0.0	36.5 ± 1.6
$\Delta G_{\text{vdW}}^{\text{free}}$	-5.0 ± 0.7	-5.0 ± 0.7	0
$\Delta G_{\text{nonpol}}^a$	-43.7 ± 1.7	-7.3 ± 0.7	-36.5 ± 1.6

$$^a \Delta G_{\text{nonpol}} = \Delta G_{\text{vdW}}^{\text{free}} - \Delta G_{\text{vdW}}^{\text{bound}}$$

ensemble. The same protocol was also used to simulate an apolar benzene molecule, both in the bulk and bound to T4 lysozyme.

The TI simulations were performed as follows: 10 independent simulations were initiated from the crystal structure by assigning random starting velocities. These were then simulated for 20 ps in a NPT ensemble with restraints to all atoms, except the hydrogen atoms and water molecules, as described above. This was followed by a 100-ps equilibration run and a 200-ps production run, both in the NPT ensemble. Snapshots were saved every 5 ps for analysis of $\delta V/\delta\lambda$. Restraints were applied in the bound simulations, as described above. The reported TI results are averages over the 10 simulations. A similar protocol has been used before.²¹ For several perturbations, we tested to use 1 ns production simulations, but this did not change the results significantly.

Results

Nonpolar Solvation Energies. As detailed above, we have estimated the binding free energy of benzene to T4 lysozyme using TI and three approximate methods. This binding free energy has then been decomposed into various terms according to the MM/GBSA formalism to allow for the investigation of the nonpolar contribution to the solvation free energy.

We employed the TI method with double decoupling as a computational benchmark. This method estimated ΔG_{bind} to

be -24 ± 2 kJ/mol (see below), which is in good agreement with the experimental value of -22 ± 1 kJ/mol.²⁰ This shows that the TI results are reliable and can be used as the benchmark for more approximate methods. Several other groups have studied the same system with free-energy methods, yielding results of -15 to -38 kJ/mol,⁵² -36 to -38 kJ/mol,⁵³ -19.1 ± 0.8 kJ/mol,²² and -24.9 ± 0.8 kJ/mol.²¹ Thus, our results are among the best available.

All nonbonded terms are pairwise additive, so $\delta V/\delta\lambda$ can be decomposed on a residue-wise basis, making it possible to investigate the contributions from the solvent and the solute to the calculated free energies. Admittedly, these contributions may depend on how the decomposition is performed and therefore are not fully unambiguous. All contributions involving solvent molecules were assigned to the solvent and all the rest to the solute. Strictly, the TI free energies calculated for the bound state depend on the applied restraints, but it is assumed that these restraints have a negligible effect on the solvent molecules. In Table 1, the calculated ΔG_{nonpol} , as well as its components, $\Delta G_{\text{nonpol}} = \Delta G_{\text{vdW}}^{\text{free}} - \Delta G_{\text{vdW}}^{\text{bound}}$ (cf. Figure 1), and the solvent and solute contributions are shown. Overall, $\Delta G_{\text{nonpol}} = -44$ kJ/mol and only -7 kJ/mol of these comes from the solvent molecules. The solvent contribution is favorable for the free ligand, whereas it is unfavorable in the complex (-5 and 2 kJ/mol; cf. Table 1). The $\delta V/\delta\lambda$ curves for the bound and free simulations are shown in Figure 2. It can be seen that the $\delta V/\delta\lambda$ curves are rather smooth, so the trapezoid integration is reliable.

Next, we estimated the binding free energy of benzene to T4 lysozyme within a MM/GBSA framework. As specified in the Methods section, we used three continuum methods to estimate the solvation energy. In the standard MM/GBSA approach, $\Delta G_{\text{polsolv}}$ is calculated by the generalized Born

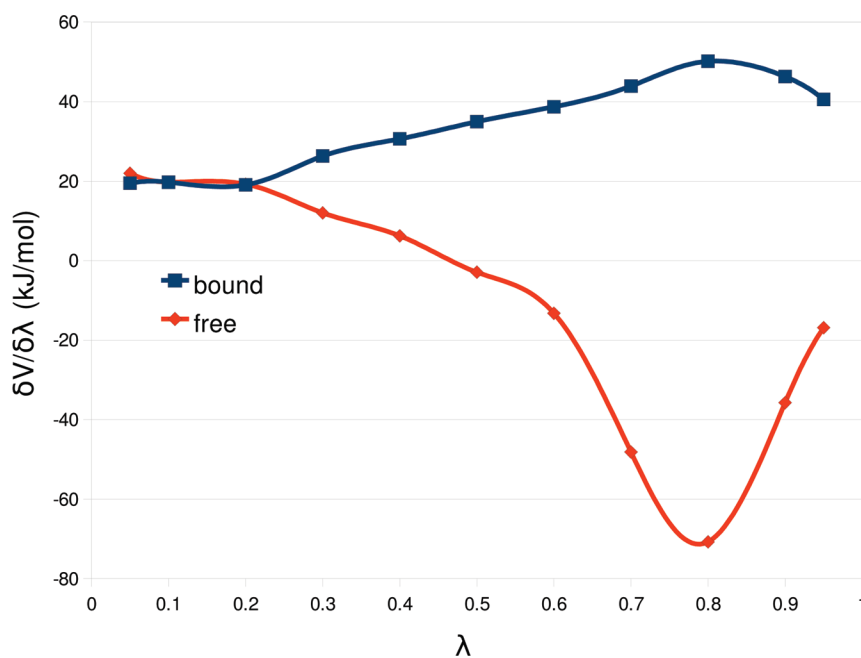


Figure 2. The derivative of the potential with respect to the coupling parameter ($\delta V/\delta\lambda$) in kJ/mol. The curves show $\Delta G_{\text{vdW}}^{\text{bound}}$ and $\Delta G_{\text{vdW}}^{\text{free}}$. The state $\lambda = 0$ contains a benzene molecule without charges, whereas in state $\lambda = 1$, the Lennard-Jones parameters of benzene have been turned off also.

Table 2. $\Delta G_{\text{nonpolsolv}}$ Calculated with Various Methods (kJ/mol)

	SASA	SASA(P0) ^d	CD	CD(P0) ^d	PCM	PCM(P0) ^d	TI
bound ^a	1.3 ± 0.6	0.0 ± 0.0	-27.8 ± 0.4	2.1 ± 0.1	-44.9 ± 1.0	1.1 ± 0.2	2.2 ± 0.0
free ^b	-9.4 ± 0.0	-9.4 ± 0.0	-8.5 ± 0.1	-8.5 ± 0.1	-4.5 ± 0.0	-4.5 ± 0.0	-5.0 ± 0.7
$\Delta G_{\text{nonpolsolv}}^c$	-10.7 ± 0.6	-9.4 ± 0.0	19.2 ± 0.4	-10.7 ± 0.1	40.4 ± 1.0	-5.6 ± 0.2	-7.3 ± 0.7

^a For TI this is $\Delta G_{\text{vdWSolv}}^{\text{bound}}$, for the other method it is the difference $\langle G_{\text{nonpolsolv}}(P) \rangle - \langle G_{\text{nonpolsolv}}(PL) \rangle$. ^b For TI this is $\Delta G_{\text{vdWSolv}}^{\text{free}}$, for the other methods it is $-\langle G_{\text{nonpolsolv}}(L) \rangle$. ^c $\Delta G_{\text{nonpolsolv}} = \Delta G_{\text{vdWSolv}}^{\text{free}} - \Delta G_{\text{vdWSolv}}^{\text{bound}}$. ^d Here, we assume that the ligand cavity in the free-protein calculations is filled with a dummy benzene molecule with zeroed charges, dispersion, and exchange-repulsion parameters.

Table 3. Components of the PCM and SASA Non-Polar Energy, As Well As Area and Volume Estimates for Lysozyme with (PL) and without (P) the Ligand (L)^a

nonpolar energies (kJ/mol)	PL	P	L	P–PL	PL–P–L	P0	P0–PL	PL–P0–L
$\Delta G_{\text{cavity}}^{\text{PCM}}$	9319	9282	53	-38	-16	9319	0	-53
$\Delta G_{\text{disp}}^{\text{PCM}}$	-3325	-3335	-60	-10	70	-3324	1	58
$\Delta G_{\text{rep}}^{\text{PCM}}$	817	820	11	3	-13	817	0	-11
$\Delta G_{\text{nonpolsolv}}^{\text{PCM}}$	6811	6766	4	-45	40	6812	1	-6
$\Delta G_{\text{CR}}^{\text{CD}}$	5453	5452	55	-1	-54	5453	0	-55
$\Delta E_{\text{disp}}^{\text{CD}}$	-3281	-3307	-46	-26	73	-3278	2	44
$\Delta G_{\text{nonpolsolv}}^{\text{CD}}$	2173	2145	9	-28	19	2175	2	-11
$\Delta G_{\text{nonpolsolv}}^{\text{SASA}}$	209	210	9	1	-11	209	0	-9
area (Å ²)								
vdWS	16855	16765	91	-91	0	16855	0	-91
SES	7809	7975	90	166	-255	7809	0	-90
SAS	8590	8665	211	75	-286	8590	0	-211
volume (Å ³)								
vdWS	14056	13985	71	-71	0	14056	0	-71
SES	20627	20453	77	-174	97	20627	0	-77
SAS	31162	31134	276	-28	-248	31162	0	-276

^a P0 is the free protein in the cavity of the complex. The areas were estimated by the molsurf program,³⁶ whereas the volumes were estimated by a local software with a grid spacing of 0.3 Å. In both cases, Bondi radii were used and a probe radius of 1.4 Å.

method and $\Delta G_{\text{nonpolsolv}}$ from the SASA, according to eq 1. Second, $\Delta G_{\text{nonpolsolv}}$ was estimated by the CD method, which divides $\Delta G_{\text{nonpolsolv}}$ into a cavity and dispersion term (eq 4). Third, we replaced the entire solvation free energy estimate with that obtained by PCM as implemented in the GAMESS software.³⁹ The three $\Delta G_{\text{nonpolsolv}}$ estimates are compared to the result of the TI calculations in Table 2. To make the analysis easier, we divided the free energy into contributions from the bound and free states. For TI, this is $\Delta G_{\text{vdWSolv}}^{\text{free}}$ and $\Delta G_{\text{vdWSolv}}^{\text{bound}}$ (cf. Figure 1, divided into solvent and solute contributions), and for the other methods it is $\langle G_{\text{nonpolsolv}}(P) \rangle - \langle G_{\text{nonpolsolv}}(PL) \rangle$ and $-\langle G_{\text{nonpolsolv}}(L) \rangle$, respectively.

From Table 2, it can be seen that SASA gives almost the same result as TI for the nonpolar solvation energy in the bound state, 1 kJ/mol compared to 2 kJ/mol. In contrast, the CD method gives the wrong sign; it estimates a negative free energy of -28 kJ/mol. PCM also gives the wrong sign; it assigns a negative free energy of -45 kJ/mol to the process.

To gain some further understanding of this discrepancy, the PCM nonpolar solvation energy was divided into its three contributions, viz. the cavitation, the dispersion, and exchange-repulsion energies (Table 3). These are -38, -10, and 3 kJ/mol, respectively. Thus, the PCM nonpolar solvation energy is dominated by the cavitation term. This gives us a clue to the cause of the problem: PCM predicts an appreciable difference in the size of the complex and the free protein, i.e., it predicts that the benzene cavity in lysozyme is filled with solvent in the free protein. This is also confirmed by the other two PCM terms: Both the dispersion

and exchange repulsion are larger in the free protein than in the complex (i.e., the dispersion term is more negative and the exchange-repulsion term more positive; cf. Table 3). Considering that the binding site in lysozyme is completely hidden inside the protein (cf. Figure 3), this shows that PCM considers the cavity to be filled with water in the free protein.

On the other hand, the MD simulations do not show any water molecules in the cavity in the unbound state. Likewise, Figure 4 shows that $\delta V/\delta \lambda$ for $\Delta G_{\text{vdWSolv}}^{\text{bound}}$ is flat, indicating that the water molecules do not go through any large changes as we remove the benzene molecule. This is also supported by experiments, which show that the cavity in the protein is empty at normal pressure⁵⁴ and that there is no change in the general shape of the protein when benzene binds.

However, this is not necessarily an erroneous behavior of PCM. On the contrary, large protein cavities are typically filled with water molecules, and also smaller cavities may contain water.⁵⁵ Therefore, it cannot be known beforehand whether a cavity is filled with water molecules or not. It is quite natural to let the continuum models assume that all cavities are water-filled, although it would be better if this could be affected as an input option.

In Table 3, we also provide estimates of the area and volume of lysozyme (from the crystal structure). Unfortunately, there are several definitions of the surface of a molecule. The van der Waals surface (vdWS) is the union of the van der Waals surfaces of all atoms in the molecule. The solvent-excluded surface (SES) is the surface accessible to a solvent probe, i.e., the vdWS, but excluding crevices and cavities that are not large enough to room a solvent molecule. Finally, the solvent-accessible surface (SAS) is

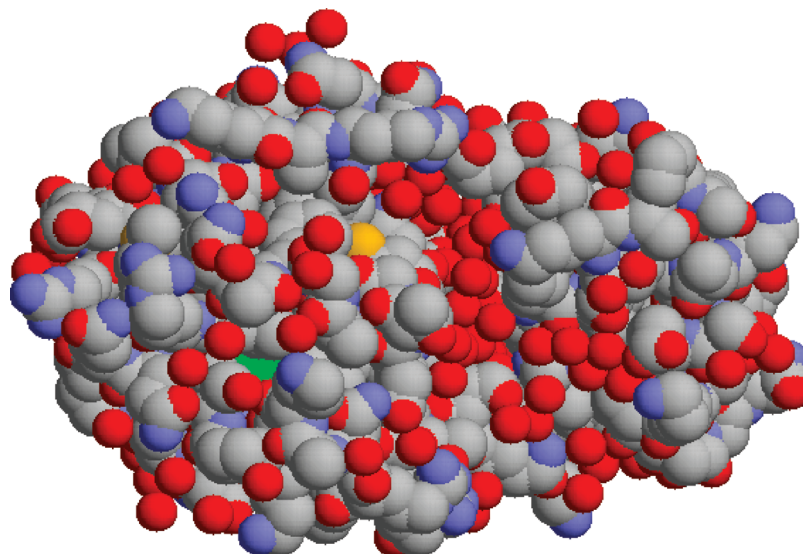


Figure 3. Space-filling model of lysozyme, showing that the benzene molecule (light green) is almost completely buried in the protein.

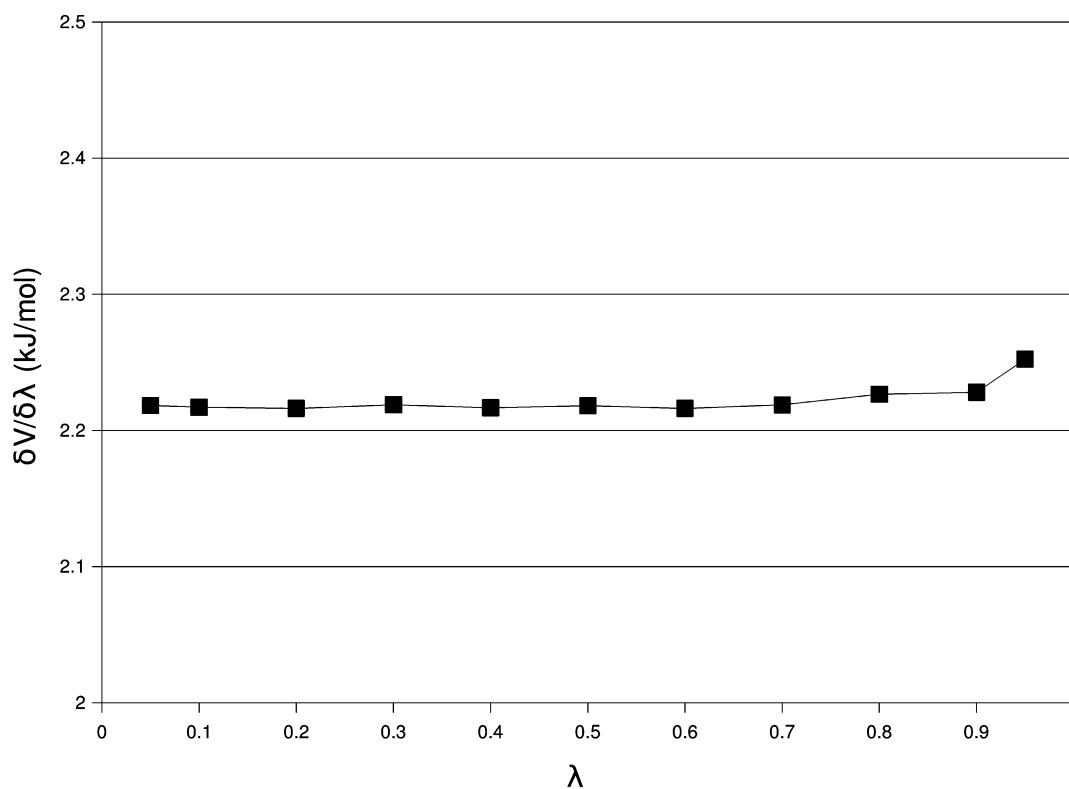


Figure 4. The derivative of the potential with respect to the coupling parameter in kJ/mol for the $\Delta G_{\text{vdWSolv}}^{\text{bound}}$ process.

the surface followed by the center of the solvent probe when rolling on the vdWS. In practice, the SAS is equivalent to a vdWS for which all atomic van der Waals radii have been enhanced by the solvent-probe radius.

The various solvation terms employ different size measures: $\Delta G_{\text{nonpolsolv}}^{\text{SASA}}$ is of course based on the SAS area, as are the PCM dispersion and repulsion terms. However, the polar PCM solvation energy is based on the SES area, whereas the PCM cavity energy is based on the vdWS, but with both area and volume terms.⁵⁶ As an effect, the PCM cavity term shows a different behavior than the other two PCM terms: As mentioned above, the absolute value of the dispersion

and exchange-repulsion terms is larger for the free protein than for the complex, in agreement with the SAS area (cf. Table 3). On the other hand, the cavity term is larger in the complex, in agreement with both the vdWS area term and all volume terms. It is not clear whether the area or volume terms dominate the cavity term, but it seems questionable to use the vdWS in a protein, which will include crevices and cavities that are too small to room a water molecules (in fact, all atoms in lysozyme contributes to the vdWS, whereas only 59% of the atoms contribute to the SAS). As can be seen from Table 3, the vdWS gives the opposite sign for the prediction of the difference between

Table 4. ΔG_{solv} of Benzene (in kJ/mol) Estimated by Four Different Methods^a

Method	ΔG_{solv}
GB/SASA	-8.3 ± 0.0
GB/CD	-9.1 ± 0.2
PCM	-4.6 ± 0.0
TI	-0.1 ± 0.7
experiment ¹²	-3.6

^a For TI, this energy is $-(\Delta G_{\text{vdW}}^{\text{free}} + \Delta G_{\text{elec}}^{\text{free}})$, whereas for the other methods it is $\langle G_{\text{nonpolsolv}}(L) + G_{\text{polsolv}}(L) \rangle$. A small gas-phase contribution has been neglected.

the complex and the protein. PCM was calibrated for small molecules, for which there is little difference between the vdWS and SAS areas, but for large molecules, they are widely different, as can be seen in Table 3.

Thus, we can conclude that the prime reason for the poor result of the PCM solvation model is the fact that it predicts that the lysozyme cavity is full with water in the free protein. A simple way to avoid this problem is to recalculate the PCM solvation energy for the free protein in the cavity of the complex (in practice, this is obtained by using a dummy benzene molecule with all charges, as well as all dispersion and exchange-repulsion coefficients zeroed). Thereby, we prevent the cavity from being filled with water molecules. As a consequence, the cavitation energy will cancel between the complex and the free protein, whereas the other terms will change somewhat, as shown in Table 3 (the polar solvation energy also changes by 3 kJ/mol). The net bound $\Delta G_{\text{nonpolsolv}}$ is shown in Table 2 (column PCM(P0)). It can be seen that now PCM gives a result close to that of TI, 1 kJ/mol, compared to 2 kJ/mol. This supports our suggestion that the main problem with PCM is that it assumes that the cavity in lysozyme is filled with water.

Interestingly, the SASA and CD models also assume that the cavity in lysozyme is filled with solvent. This can for example be seen from the fact that the SASA in Table 3 changes between the free protein (P) and the complex (PL), giving a positive sign of the bound SASA term in Table 2, which shows that the SASA of the free protein is larger than that of the complex. However, the SASA term is much smaller than the corresponding PCM terms, so that it does not cause any problems.

Therefore, we can cure the CD and SASA results in the same way as for PCM. This changes $G_{\text{nonpolsolv}}$ for CD from -28 to 2 kJ/mol, i.e., to a close agreement with the TI result (Table 2, column CD(P0)). Thus, this simple approach also cures the poor results of the CD approach. However, the polar GB solvation energy for the free protein changes by 8 kJ/mol. For SASA, the result hardly changes, because $\Delta G_{\text{nonpolsolv}}$ for the bound state was only 1 kJ/mol (Table 2, column SASA(P0)).

For the free ligand, TI predicts a decrease in $\Delta G_{\text{nonpolsolv}}$ of -5 kJ/mol. This value is closely reproduced by PCM (-4 kJ/mol), whereas the SASA and CD results are slightly more negative (-9 kJ/mol).

These results can be combined with the polar solvation energy of the free ligand to give the solvation free energy of benzene (neglecting a small gas-phase correction), which experimentally is -4 kJ/mol.¹² From Table 4, it can be seen

that PCM obtains a value close to experiments (-5 kJ/mol). The GB/SASA and GB/CD estimates are 4 – 5 kJ/mol too negative, whereas the TI estimate is 3 kJ/mol too positive. Thus all three continuum methods give similar results for the solvation free energy of the free ligand, which is expected, because they were calibrated to give accurate solvation energies for small ligand. Major differences are observed only for large ligands and macromolecules, for which experimental solvation energies are missing, and for which the various area and volume definitions give differing results.

An interesting consequence of the differing surface definitions for the nonpolar terms is that PCM and SASA give the opposite sign for $\Delta G_{\text{nonpolsolv}}$, which was also observed for the binding of seven biotin analogues to avidin.¹⁹ In general, both the vdWS area and volume terms nearly cancel for the net PL–P–L term, as can be seen in Table 3. Therefore, the net PCM cavity term also nearly cancels. Among the remaining PCM terms, the dispersion term is larger than the repulsion term in all cases we have studied, and therefore the net PCM nonpolar solvation energy is positive (the SAS area is larger for P than for PL, and this difference is enhanced by the ligand, giving a negative result for PL–P–L; then the dispersion is always negative, i.e., attractive). However, the SASA term is always positive (repulsive) and consequently, the nonpolar PCM and SASA energies always have the opposite sign for a buried cavity. Alternatively, if the cavity is empty, then all PCM terms are nearly identical for PL and P0, so the net nonpolar solvation energy is close to the negative of that of the ligand, which can have any sign, depending on the relative size of the dispersion and the sum of the repulsion and cavity terms.

Dispersion and Repulsion Terms. The dispersive and repulsive parts of PCM and the dispersion part in the CD method can be further analyzed by relating them, in a mean-field approximation, to the average van der Waals interaction between the apolar benzene molecule and the surrounding water molecules.¹⁵ Therefore, explicit simulations of an apolar benzene molecule both bound to T4 lysozyme and free in bulk water were performed. For such a comparison of explicit simulations with the approximate methods, it is necessary to decompose the Lennard-Jones potential into attractive and repulsive parts.¹⁴ We have used three common schemes for such a decomposition, as described in the Methods section. The explicit simulations are compared to PCM and CD in Table 5. The van der Waals interactions in the bound simulations are small and all three decomposition schemes give the same results (because all water molecules are far away from the benzene molecules). The small value is consistent with the TI results as discussed above. However, the PCM and CD continuum models give much larger values with the opposite sign. The reason is that they assume a water-filled cavity, as discussed above. If we force the cavity to be empty, both the PCM and CD results (PCM(P0) and CD(P0) in Table 5) closely reproduce the results of the explicit simulation.

The van der Waals interactions in the simulation of the free ligand are larger and negative. In this case, different decomposition methods give different contributions for

Table 5. Average Benzene–Water van der Waals Interaction Energies (kJ/mol) in Explicit Simulations (Using Three Different Decomposition Schemes) Compared to the PCM and CD Estimates^a

method	free			bound		
	attractive	repulsive	total	attractive	repulsive	total
explicit (12–6)	−81.8 ± 0.4	38.4 ± 0.4	−43.4 ± 0.2	1.3 ± 0.0	0.0 ± 0.0	1.3 ± 0.0
explicit (WCA)	−52.1 ± 0.1	8.7 ± 0.2	−43.5 ± 0.2	1.3 ± 0.0	0.0 ± 0.0	1.3 ± 0.0
explicit (σ)	−48.8 ± 0.1	5.3 ± 0.1	−43.5 ± 0.2	1.3 ± 0.0	0.0 ± 0.0	1.3 ± 0.0
PCM	−59.5 ± 0.0	10.7 ± 0.0	−48.8 ± 0.0	−10.0 ± 0.5	2.8 ± 0.2	−7.3 ± 0.3
PCM(P0) ^b	−59.5 ± 0.0	10.7 ± 0.0	−48.8 ± 0.0	1.1 ± 0.2	0.0 ± 0.0	1.1 ± 0.2
CD	−46.3 ± 0.1		−46.3 ± 0.1	−26.5 ± 0.4		−26.5 ± 0.4
CD(P0) ^b	−46.3 ± 0.1		−46.3 ± 0.1	2.1 ± 0.1		2.1 ± 0.1

^a The results for the free state are taken from the simulations with the free ligand in water. The results for the bound state are the negative interaction energy between the ligand and the solvent in the explicit simulations, and the difference between simulations of the free protein and of the complex for the continuum methods. ^b Here, we assume that the ligand cavity in the free-protein calculations is filled with a dummy benzene molecule with zeroed charges, dispersion, and exchange-repulsion parameters.

Table 6. Contributions to ΔG_{bind} (kJ/mol) for Benzene Binding to Lysozyme, Calculated with Four Variants of MM/GBSA and with TI^a

Contribution	MM/GBSA		MM/GBCD		MM/PCM		TI	Exp.
	P	P0 ^b	P	P0 ^b	P	P0 ^b		
Polar, solute	−6.6 ± 0.4	−6.6 ± 0.4	−6.6 ± 0.4	−6.6 ± 0.4	−6.6 ± 0.4	−6.6 ± 0.4	0.5 ± 0.3	
Polar, solvent	26.3 ± 0.3	26.3 ± 0.3	26.3 ± 0.3	18.1 ± 0.3	11.8 ± 0.3	8.8 ± 0.3	5.3 ± 0.2	
Nonpolar, solute	−11.0 ± 3.2	−11.0 ± 3.2	−11.0 ± 3.2	−11.0 ± 3.2	−11.0 ± 3.2	−11.0 ± 3.2	−23.0 ± 1.7	
Nonpolar, solvent	−10.7 ± 0.1	−9.4 ± 0.0	19.2 ± 0.5	−10.7 ± 0.1	40.4 ± 1.0	−5.6 ± 0.2	−7.3 ± 0.7	
Sum	−1.9 ± 3.1	−8.9 ± 3.0	27.9 ± 3.0	−10.2 ± 3.0	34.6 ± 3.2	−14.4 ± 3.0	−24.5 ± 2.1	−21.7

^a For the continuum solvation methods, the terms (polar solute, polar solvent, nonpolar solute, and nonpolar solvent) are $\langle \Delta E_{\text{ele}} \rangle$, $\langle \Delta G_{\text{vdWsol}}^{\text{polsolv}} \rangle$, $\langle \Delta E_{\text{vdW}} \rangle - \langle T\Delta S \rangle$, and $\langle \Delta G_{\text{nonpolsolv}} \rangle$, respectively. For TI, the terms are $\Delta G_{\text{ele}}^{\text{free}} - \Delta G_{\text{ele}}^{\text{bound}}$, $\Delta G_{\text{vdWsol}}^{\text{free}} - \Delta G_{\text{vdWsol}}^{\text{bound}}$, $\Delta G_{\text{vdWsol}}^{\text{polsolv}} - \Delta G_{\text{vdWsol}}^{\text{nonpolsolv}} + \Delta \Delta G_{\text{rest}}$, and $\Delta G_{\text{vdWsol}}^{\text{free}} - \Delta G_{\text{vdWsol}}^{\text{bound}}$, respectively. ^b Here, we assume that the ligand cavity in the free-protein calculations is filled with a dummy benzene molecule with zeroed charges, dispersion, and exchange-repulsion parameters.

dispersion and repulsion. The CD method should be comparable to the σ decomposition of the explicit simulations and in this case, the two results agree within 3 kJ/mol. PCM gives results that resemble the WCA decomposition, although the original implementation used a 12–6 decomposition.⁵⁷

Binding Free Energies. Finally, we have collected in Table 6 all contributions to the binding free energy of benzene to T4 lysozyme, divided into polar and nonpolar contributions, as well as solvent and solute contributions. The entropic term in MM/GBSA is added to the nonpolar contribution of the solute and the contribution from releasing the TI restraints (17 kJ/mol) is assigned to the same contribution, because this is mainly an entropic term. We have tested this approximation by calculating $\Delta G_{\text{ele}}^{\text{bound}}$ without applying any restraints: Comparing the simulation with and without restraints, $\Delta G_{\text{ele}}^{\text{free}} - \Delta G_{\text{ele}}^{\text{bound}}$ increases by 0.6 kJ/mol and the solute part by 0.2 kJ/mol.

It can be seen from Table 6 that all four contributions have sizable errors, compared to the TI simulations. The nonpolar solvent contribution has already been discussed. The polar contribution of the solute is the same for all continuum approaches ($\langle \Delta E_{\text{ele}} \rangle$), providing a 7 kJ/mol too negative estimate. Likewise, the nonpolar contribution of the solute is the same for all continuum methods ($\langle \Delta E_{\text{vdW}} \rangle - \langle T\Delta S \rangle$), −11 kJ/mol, which is 12 kJ/mol less negative than the TI estimate. Thus, the errors in these two terms partly cancel. However, the polar solvation free energy depends on whether we use GB or PCM. Both methods give too positive values, 26 kJ/mol for GB and 12 kJ/mol for PCM, compared to 5 kJ/mol for TI. Consequently, the

conventional MM/GBSA approach estimates a binding free energy that is ~ 20 kJ/mol too unfavorable, as reported previously.⁴³ Replacing $\Delta G_{\text{nonpolsolv}}$ from SASA with the CD estimate, results in a slightly better agreement with the experimental ΔG_{bind} , but only if we assume that the cavity in lysozyme is not filled with waters (MM/GBCD(P0)). Otherwise, this method gives a positive ΔG_{bind} . Replacing GBSA with PCM results in a binding free energy that is 60 kJ/mol too unfavorable. However, this result is strongly improved if we assume that the cavity in lysozyme is not filled with water molecules (MM/PCM(P0) in Table 6), which gives an estimate that is 10 kJ/mol too positive. Likewise, the MM/GBSA result is improved to −9 kJ/mol if the cavity is not filled with water molecules, mainly owing to the GB solvation energy. Altogether, PCM without a water-filled cavity gives the best results, with a maximum error compared to the TI components of 12 kJ/mol (for the nonpolar solute term) and a mean absolute deviation of 6 kJ/mol for the four terms, whereas the SASA(P0) and CD(P0) methods give maximum component errors of 13 kJ/mol (polar solvent) and mean absolute deviations of 8–9 kJ/mol. Of course, it is likely that some of the differences between TI and MM/GBSA comes from effects other than the solvation model, e.g., that TI uses alchemical transformations, whereas MM/GBSA only uses end-point simulations.

Conclusions

In this work, we explain why the SASA and PCM continuum solvation methods give qualitatively different estimates of

the nonpolar solvation energy. Using the binding of benzene to the Leu99Ala T4 lysozyme mutant as a test case and absolute binding free energies estimated with TI and a double-decoupling scheme with restraints as a benchmark, we show that the simple SASA approach gives an appreciably more accurate estimate of the nonpolar solvation energy than the PCM approach. However, the reason for the poor performance of PCM for this case is that it assumes that the benzene cavity in the free protein is filled with water. In general, this is a proper assumption of a continuum solvation method, but in the present case, both calculations and experiments indicate that the cavity is water-free when the ligand is not bound.⁵⁴ We suggest a simple way to correct this problem, and then PCM gives the best results among the continuum methods, providing better estimates of each component than the other methods. Similar approaches with inserted dummy atoms have been used before to avoid ligand cavities to be filled with continuum water.^{55,58,59} However, PCM uses the van der Waals surface for the calculation of the cavitation energy, which is questionable for a protein, because it leads to numerous small crevices and cavities inside the protein, which are too small to house a solvent molecule, and it also leads to opposite trends regarding the change in surface area of the protein upon ligand binding.

We have also included in the investigation a recent two-term expression consisting of a cavitation and a dispersion energy (CD method). This method also assumes a water-filled cavity for the free protein and the results are strongly improved if this problem is cured.

Interestingly, the SASA method makes the same assumption of a water-filled cavity, but the SASA terms are much smaller, so it does not significantly affect the nonpolar solvation energy (however, the polar solvation energy is improved, giving a better net result). In fact, for this test case, the SASA and CD methods give similar results, and that with PCM is only slightly better. The reason for this is most likely that the SASA terms are smaller and therefore more stable. As can be seen from Table 3, the three terms involved in the nonpolar solvation energies are large but of opposite signs. Therefore, it is a formidable task to obtain a sum that is stable and accurate. It is easier to obtain stable results for the sum, which is much smaller in magnitude. The integration over Lennard-Jones terms as used for the dispersion term in both CD and PCM, and for the exchange-repulsion term in PCM are probably as accurate as you can come with a continuum approximation at the MM level. The main problem is most likely the cavity term, which is not a plain average from a simulation (it involves the entropy).

Of course, our results are valid strictly for the present lysozyme test case only, with its small and buried ligand-binding cavity that is void of water molecules in the apo state. However, the problem with the empty cavity in the free protein is general and emphasizes how important it is to know the hydration state of the ligand cavity. Even if the cavity contains water in the free state, it is most likely that these water molecules are not bulk-like. This can be expected to give major problems for continuum models, but also for

TI and FEP methods if the cavity is buried in the protein so that the equilibration of waters between the cavity and bulk is slow.

Acknowledgment. This investigation has been supported by grants from the Swedish Research Council and from the Research School in Pharmaceutical Science. It has also been supported by computer resources of Lunarc at Lund University, C3SE at Chalmers University of Technology and HPC2N at Umeå University. J.K. thanks the Danish Research Council for financial support.

References

- (1) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 4–16.
- (2) Huang, S.-Y.; Zou, X. *J. Chem. Inf. Model.* **2010**, *50*, 262–273.
- (3) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187–4225.
- (4) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.
- (5) Michael, J.; Verdonk, M. L.; Essex, J. W. *J. Med. Chem.* **2006**, *49*, 7427–7439.
- (6) Chen, J.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18* (18), 140–148.
- (7) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci.* **2006**, *103*, 8331–8336.
- (8) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.
- (9) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238–250.
- (10) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. *J. Phys. Chem. B* **2010**, *114*, 8505–8516.
- (11) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754–2759.
- (12) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
- (13) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.
- (14) Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872–1882.
- (15) Tan, C.; Tan, Y.-H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.
- (16) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210–3221.
- (17) Cossi, M.; Tomasi, J.; Cammi, R. *Int. J. Quant. Chem. Quant. Chem. Symp.* **1995**, *29*, 695–702.
- (18) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.
- (19) Söderhjelm, P.; Kongsted, J.; Ryde, U. *J. Chem. Theory Comput.* **2010**, *6*, 1726–1737.
- (20) Morton, A.; Baase, W. A.; Matthews, B. W. *Biochem.* **1995**, *34*, 8564–8575.
- (21) Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.
- (22) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- (23) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.

- (24) Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- (25) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B.* **2003**, *107*, 9535–9551.
- (26) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–314.
- (27) Steinbrecher, T.; Mobley, D. L.; Case, D. A. *J. Chem. Phys.* **2007**, *127*, 214108–2141021.
- (28) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *37*, 9401–9809.
- (29) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham III, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (30) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.
- (31) Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786–3791.
- (32) Tan, C. H.; Yang, L. J.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680–18687.
- (33) Wang, J.; Cai, Q.; Ye, X.; Hsieh, M.-J.; Tan, C.; Luo, R. *Amber Tools User's Manual, Version 1.4* **2010**, 143–150.
- (34) Kongsted, J.; Ryde, U. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 63–71.
- (35) Genheden, S.; Ryde, U. *J. Comput. Chem.* **2010**, *31*, 837–846.
- (36) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *Amber 10*, University of California: San Francisco, 2008.
- (37) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.
- (38) Frisch, A. E.; Frisch, M. J.; Trucks, G. W. *Gaussian 03 User's Reference*; Gaussian, Inc.: Wallingford, CT, USA, 2003, p. 205.
- (39) Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 71–84.
- (40) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (41) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237–5247.
- (42) Morton, A.; Matthews, B. W. *Biochem.* **1995**, *34*, 8576–8588.
- (43) Genheden, S.; Ryde, U. *J. Comput. Chem.* 2010 available online, DOI: 10.1002/jcc.21546.
- (44) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinform.* **2006**, *65*, 712–725.
- (45) Wang, J. M.; Wolf, R. M.; Caldwell, K. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (46) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.
- (47) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (48) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (49) Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512–518.
- (50) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (51) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (52) Hermans, J.; Wang, L. *J. Am. Chem. Soc.* **1997**, *119*, 2707–2714.
- (53) Rodinger, T.; Howell, P. L.; Romes, R. *J. Chem. Phys.* **2008**, *129*, 155102–155113.
- (54) Collins, M. D.; Hummer, G.; Quillin, M. L.; Matthews, B. W.; Gruner, S. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16668–16671.
- (55) Qvist, J.; Davidovic, M.; Hamelberg, D.; Halle, B. *Proc. Natl. Ac. Sci. U.S.A.* **2008**, *105*, 6296–6301.
- (56) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1990**, *255*, 327–335.
- (57) Floris, F. M.; Tomasi, J. *J. Comput. Chem.* **1991**, *12*, 784–791.
- (58) Zou, X.; Sun, Y.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
- (59) Lie, H.-Y.; Kuntz, I. D.; Zou, X. *J. Phys. Chem. B* **2004**, *108*, 5453–5462.

CT100272S

Explicit Water Models Affect the Specific Solvation and Dynamics of Unfolded Peptides While the Conformational Behavior and Flexibility of Folded Peptides Remain Intact

Petra Florová, Petr Sklenovský, Pavel Banáš, and Michal Otyepka*

Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, tr. 17 listopadu 12, 771 46 Olomouc, Czech Republic

Received June 30, 2010

Abstract: Conventional molecular dynamics simulations on 50 ns to 1 μ s time scales were used to study the effects of explicit solvent models on the conformational behavior and solvation of two oligopeptide solutes: α -helical EK-peptide (14 amino acids) and a β -hairpin chignolin (10 amino acids). The widely used AMBER force fields (ff99, ff99SB, and ff03) were combined with four of the most commonly used explicit solvent models (TIP3P, TIP4P, TIP5P, and SPC/E). Significant differences in the specific solvation of chignolin among the studied water models were identified. Chignolin was highly solvated in TIP5P, whereas reduced specific solvation was found in the TIP4P, SPC/E, and TIP3P models for kinetic, thermodynamic, and both kinetic and thermodynamic reasons, respectively. The differences in specific solvation did not influence the dynamics of structured parts of the folded peptide. However, substantial differences between TIP5P and the other models were observed in the dynamics of unfolded chignolin, stability of salt bridges, and specific solvation of the backbone carbonyls of EK-peptide. Thus, we conclude that the choice of water model may affect the dynamics of flexible parts of proteins that are solvent-exposed. On the other hand, all water models should perform similarly for well-structured folded protein regions. The merits of the TIP3P model include its high and overestimated mobility, which accelerates simulation processes and thus effectively increases sampling.

Introduction

Molecular dynamic (MD) simulations of biomacromolecules are based on empirical force fields, which relate potential energy and molecular structure. Several in-depth overviews of current trends in the field of empirical potentials have been published.^{1–5} Among all of the available empirical potentials, the AMBER,^{6,7} GROMOS,⁸ CHARMM,⁹ and OPLS¹⁰ families of force fields have performed well for biomacromolecules. However, in this study we deal only with the AMBER family of force fields, which have been thoroughly tested on a variety of biomolecular systems. The AMBER family of force fields perform well for a range of

systems including proteins,¹¹ RNA,¹² and DNA^{13,14} and their homo- and heterocomplexes.¹⁵

One of the AMBER force fields is ff99 (also referred to as parm99),¹⁶ which is a second-generation force field based on the Cornell et al. set of effective parameters¹⁷ employing the pair-additive potential used for condensed-phase all-atomic simulations. Simmerling et al. have shown that the ff99 force field does not accurately represent glycine behavior in protein simulations and that it is prone to other inaccuracies, including overstabilization of α -helical peptide conformations and underestimation of β -bend propensity.¹⁸ Simmerling et al. also suggested a reparametrization of backbone torsion terms, yielding the modified force field ff99SB,¹⁸ which improved the balance in secondary structure propensities. A third-generation AMBER force field, ff03,

* Corresponding author e-mail: michal.otyepka@upol.cz.

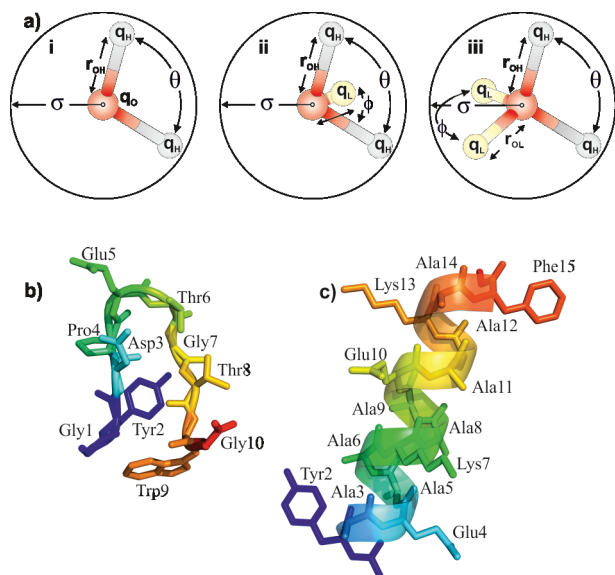


Figure 1. (a) Three-dimensional representations of three-site TIP3P and SPC/E (left), four-site TIP4P (middle), and five-site TIP5P (right) water models. The labels correspond to the parameters listed in Table 1. (b) Chignolin and (c) EK-peptide structures in stick representations.

was introduced by Duan et al., who revised all ψ/ϕ torsion parameters and recalculated atomic partial charges.¹⁹ Despite such efforts, none of these recent force fields are perfect; for example, Hummer et al. showed that ff03 overestimated and ff99SB underestimated α -helical propensities.^{20,21}

The quest for a sufficiently accurate force field has mainly focused on solute behavior, with only marginal consideration of the water environment. However, the water environment plays an essential role in biomolecular processes, and some important effects, such as the hydrophobic effect and Coulomb interaction screening, vanish if the water molecules do not surround a biomolecule.^{22–25} An explicit consideration of water molecules is therefore essential for a reasonably accurate description of, at least, solute–solvent interactions.²⁶ A large number of explicit water models have been developed in attempts to accommodate all of the physicochemical properties of water.^{27–37} Among them, Jorgensen’s TIP3P^{28,29,38} and Berendsen’s SPC/E³⁹ models (Figure 1, Table 1) are the most widely used for biomolecular simulations. It should be noted, however, that although these explicit solvent models have been derived to represent the physicochemical properties of bulk water well, i.e., their solvent–solvent interactions, the balance of solute–solvent interactions in these models remains questionable and requires further study.

All of the popular TIP n P and SPC/E water models agree well with bulk water characteristics at ambient temperatures. The three-site TIP3P model (in which point charges are centered on each of the three atoms) is the most commonly used model in AMBER simulations. TIP3P reproduces well the key features of bulk water at 25 °C and 1 atm (i.e., a density of 0.997 g/cm³ and heat of vaporization of 10.53 kcal/mol), but it underestimates the height of the second (tetrahedral) peak in the O–O radial distribution function and overestimates the diffusion constant (Table 1). Both the four-site TIP4P and five-site TIP5P models give better fits

to the experimental O–O radial distribution function and also behave better than the TIP3P model in many other respects. However, improving the representation of water properties by adding additional extra point charges increases computational costs considerably, with TIP4P being approximately 1.5 times and TIP5P 2.5 times as expensive as TIP3P in terms of the simulation time (Table 1). The original three-site SPC water model has been superseded by the SPC/E model, which includes corrections for self-polarization and improved structural and diffusion properties. The computational demands of the SPC/E model are comparable to those of the TIP3P model.

Despite the development of sufficiently representative force fields and explicit water models, less attention has been paid to the effects of explicit water models on solute behavior, namely, on the structure, dynamics, and kinetics of solute molecules. Moreover, explicit water models differ in ways that are likely to generate differences in solute–solvent interactions, which might then propagate to differences in solute behavior in various explicit solvent models. In response to the call for studies on the effects of explicit water models on solute behavior reported in the literature,^{1,40–43} Nutt and Smith presented CHARMM simulations of *N*-methylacetamide, other small solute molecules, and a small protein (crabrin) in various solvent models, concluding that although TIP3P, TIP4P, and TIP5P differed in solute–solvent interactions, they provided similar overall descriptions of solvation.⁴⁰ However, their results on biomacromolecular solute behavior were, unfortunately, limited by their use of only a 2–3 ns time scale and of an alternative modified TIP3P solvent model (mTIP3P; bearing modified van der Waals parameters for hydrogen, $\sigma_{HH} = 0.449$ Å and $\epsilon_H = 0.046$ kcal/mol).⁴⁴ In a recent study, Vymětal and Vondrášek studied the effects of different explicit water models on the free energy ϕ – ψ profiles of alanine dipeptide using metadynamics. They concluded that the choice of solvent model had no significant effect on the conformational preferences of alanine dipeptide.⁴⁵ Molecular dynamics simulations of small proteins carried out by Wong and Case in ff99SB showed that protein diffusion occurs rapidly due to the high self-diffusion constant of TIP3P, whereas the SPC/E and TIP4P water models with more realistic self-diffusion constants have too large a protein rotation diffusion constant.⁴² Shirts et al. calculated hydration free energies of the amino acid side chain analogues in several water models and concluded that modified TIP3P (TIP3P-MOD; having different $\sigma_O = 3.12171$ Å and $\epsilon_O = 0.190$ kcal/mol parameters compared to the standard TIP3P parameters; see Table 1)⁴⁶ gave the closest match to the experimental data. On the other hand, the ability of explicit solvent models to accurately represent pure water properties did not necessarily determine the ability to correctly predict solute/solvent behavior.^{41,47}

In the present study we examine the structure and dynamics of the small peptide solutes chignolin and EK-peptide (Figure 1) in MD simulations on 50 ns to 1 μ s time scales (20.7 μ s in total). Chignolin is an artificial peptide consisting of 10 residues (GYDPETGTWG) which adopts a β -hairpin conformation in solution under in vitro conditions (Figure 1b). The peptide is stabilized by H-bonds between

Table 1. Parameters and Physicochemical Properties of the Explicit Water Models Studied, As Used in the AMBER Package^a

param (units)	TIP3P	TIP4P	TIP5P	SPC/E	exptl
q_H^b (e)	0.417	0.520	0.241	0.424	
q_O^b (e)	-0.834	0.000	0.000	-0.848	
q_L^b (e)		-1.040	-0.241		
r_{OH}^c (Å)	0.9572	0.9572	0.9572	1.0000	0.9572 ⁷⁶
r_{OL}^c (Å)		0.1500	0.7000		
θ_{HOH}^d (deg)	104.520	104.520	104.520	109.470	104.474 ⁷⁶
ϕ_{LOL}^d (deg)			109.470		
ϵ^e (kcal/mol)	0.1521	0.1550	0.1600	0.1554	
σ_O^e (Å)	3.1506	3.1537	3.1200	3.1656	
cost ^f	100	140	240	100	
dipole moment (D)	2.35 ²⁸	2.18 ³⁰	2.29 ²⁸	2.35 ³⁰	2.95 ⁷⁷
dielectric constant	82.0 ³⁰	53.0 ³⁰	81.5 ²⁸	71.0 ³⁰	78.4 ⁷⁸
density(298 K, 1 atm) (g/cm ³)	0.986 ± 0.010 ^g	0.994 ± 0.010 ^g	0.985 ± 0.010 ^g	0.999 ± 0.010 ^g	0.997
self-diffusion(298 K, 1 atm) (10 ⁻⁵ cm ² /s)	5.5 ^g	3.5 ^g	2.7 ^g	2.5 ^g	2.3 ⁵⁰
density maximum (K)	~182.15 ³¹	~248.15 ²⁸	~277.15 ²⁸	~235.15 ³²	277.134 ⁷⁹
melting temperature (K)	145.55 ³²	232.45 ³²	273.95 ³²	214.95 ³²	273.15

^a See Figure 1. ^b q_H , q_O , and q_L are the partial charges of hydrogen and oxygen and that on the lone pair, respectively. ^c r_{OH} and r_{OL} are the oxygen-hydrogen and oxygen-lone pair distances, respectively. ^d θ_{HOH} and ϕ_{LOL} are the hydrogen-oxygen-hydrogen and lone pair-oxygen-lone pair angles, respectively. ^e ϵ and σ are the well depth and van der Waals radius Lennard-Jones parameters, respectively. ^f Computer cost in percent with respect to TIP3P. ^g Data from this study.

Table 2. List of Simulations Performed for Each Studied System

system		no. of simulations run in tested force fields		
		ff99	ff99SB	ff03
chignolin	TIP3P	6 × 100 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs
	TIP4P	6 × 50 ns	6 × 50 ns	6 × 50 ns
	TIP5P	6 × 50 ns	6 × 50 ns	6 × 50 ns
	SPC/E	6 × 100 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs
EK-peptide	TIP3P	6 × 100 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs	6 × 100 ns; 1 × 1 μs
	TIP4P	6 × 50 ns	6 × 50 ns	6 × 50 ns
	TIP5P	6 × 50 ns	6 × 50 ns	6 × 50 ns
	SPC/E	6 × 100 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs	6 × 50 ns; 1 × 1 μs

atoms Asp3(O) and Gly7(N) (the strongest), Asp3(N) and Thr8(O), and Asp3(O^δ) and Glu5(N). Gly7 plays a further key role in chignolin structure stability because its left-handed α -helical conformation enables propagation of the C-terminal strand. It has been suggested that the side chain interaction between Tyr2 and Trp9 also stabilizes chignolin. The experimental data concerning chignolin stability suggest that the peptide has a ratio of folded to unfolded states equal to ~60:40 at 300 K.⁴⁸ EK-peptide is an artificial 14 residue long α -helix (YAEAAKAAEAAKAF).⁴⁹ An α -helicity of 40% at 273 K was measured for EK-peptide using circular dichroism (CD) spectroscopy. Ghosh and Dill found that the α -helicity of EK-peptides generally decreases with increasing temperature and amounts to ~20% at 300 K for the studied EK-peptide.⁵⁰ In the present study, four explicit solvent models (TIP3P, TIP4P, TIP5P, and SPC/E) combined with three AMBER family force fields (ff99, ff99SB, and ff03) are analyzed to elucidate the role of explicit solvent models on the behavior of both oligopeptides.

Methods

Studied Systems. A designed β -hairpin peptide, chignolin, and an artificial α -helical EK-peptide were chosen as test systems for the MD simulations. The chignolin NMR structure (PDB ID 1UAO) was used as a starting structure in MD simulations, and both chignolin termini were charged

(N-terminus positively and C-terminus negatively) in all MD simulations. The starting structure of EK-peptide was modeled as an α -helix in accordance with CD spectra.⁴⁹ Although, to our best knowledge, no structural data such as X-ray or NMR spectra are available for EK-peptide, the data from CD spectra are sufficient to provide relevant information about its structure. The N-terminus of EK-peptide was capped by an acetyl group and the C-terminus by *N*-methylamide for MD simulations.

Molecular Dynamics Simulations. All MD simulations were carried out using the AMBER suite of programs with the all-atomic force fields ff99,¹⁶ ff99SB,¹⁸ and ff03.¹⁹ The simulation protocol, which has been repeatedly shown to perform well for proteins,^{51–55} was set up as follows. The hydrogen atoms were added by the LEaP module of AMBER. Systems were then neutralized by adding counterions (Na⁺ or Cl⁻, according to the solute charge) and immersed into a rectangular box of explicit water molecules. Each system was solvated by four explicit water models—TIP3P, TIP4P, TIP5P, and SPC/E (Table 1). The initial coordinates of solute, as well as solvent, atoms were identical for a given peptide system in all water models and all force fields. Consequently, we examined the systems' dynamics in all 12 possible combinations of solute force field and solvent type (Table 2). In each simulation, the minimal distance between the solute and the box wall was set to

10 Å. Prior to the production phase of the MD run, each system was minimized by first optimizing the positions of the hydrogen atoms while the heavy atoms remained constrained; then all protein atoms were constrained, and the solvent molecules with counterions were allowed to move during a 1000-step minimization, followed by 10 ps long MD runs under $[NpT]$ conditions ($p = 1$ atm, $T = 298.16$ K). After this, the side chains were relaxed by several minimization runs, with decreasing force constants applied to the backbone atoms. After the relaxation, each system was heated from 10 to 298.16 K for 100 ps. The particle-mesh Ewald (PME) method was used for treating electrostatic interactions, and all simulations were performed under periodic boundary conditions in the $[NpT]$ ensemble at 298.16 K and 1 atm using a 2 fs integration step. The SHAKE algorithm, with a tolerance of 10^{-5} Å, was used to fix the positions of all hydrogen atoms, and a 9.0 Å cutoff was applied to nonbonding interactions. The Berendsen thermostat was used.⁵⁶

All systems studied are listed in Table 2. Generally, the dynamics of each system in respective combinations of solute–solvent description were modeled via six MD simulations on 50–100 ns time scales, in which the coordinates were stored every picosecond. The number of parallel MD simulations in each run was chosen according to Day and Daggett,⁵⁷ who suggested that, for capturing average properties of simulated systems, 5–10 simulations in multiple MD runs are sufficient. Moreover, additional extensive 1 μ s long MD runs with coordinates stored every 10 ps were performed for all force fields in combinations with the TIP3P and SPC/E models. The total simulation time in the present study (the sum of the simulation times of all systems) reached 20.7 μ s.

In addition to the peptide simulations, we carried out water box simulations (i.e., only water molecules without solute) for the TIP3P, TIP4P, TIP5P, and SPC/E models under $[NpT]$ conditions ($p = 1$ atm, $T = 298.16$ K) and using an 8.0 Å cutoff for nonbonding interactions on 10 ns time scales. A periodic rectangular cubic box with dimensions of $20 \times 20 \times 20$ Å filled with 375 explicit water molecules was used in each water box simulation.

Analyses of Trajectories. The α -helix dynamics were monitored using the secondary structure analysis (implemented in ptraj from the AMBER package), the time evolution of the root-mean-square deviations (rmsd's) of the backbone atoms (C, N, C_α) with respect to the initial structure, and the time evolution of the distances between atoms forming salt bridges (Glu3(C_δ)–Lys6(N_ϵ), Lys6(N_ϵ)–Glu9(C_δ), and Glu9(C_δ)–Lys12(N_ϵ)). The numbers of water molecules in the first solvation shell (<3.4 Å) around the carbonyl oxygen of each residue were calculated for ff03 EK-peptide simulations using ptraj. The α -helix was said to be unfolded if there were fewer than four neighboring helical residues for more than 100 ps (in accordance with Daggett⁵⁸). The following structural analyses for chignolin simulations were performed: the time evolution of the rmsd's of the backbone atoms (C, N, C_α) from the initial structure and the distances between C_γ atoms of Tyr2 and Trp9. In this study, all structures with an rmsd of main chain atoms up to 1.7 Å were considered to be native-like, because all

chignolin structures from the NMR ensemble fitted into this interval. In addition, structures from the MD ensemble with rmsd below 1.7 Å displayed a native H-bond network and native distance between Tyr2 and Trp9 (C_γ atoms). We considered the chignolin molecule as unfolded if the rmsd of the backbone atoms was above 1.7 Å for more than 100 ps.

The effect of water models on the unfolding or refolding rates in ff99 chignolin simulations were analyzed by the following statistical model. Unfolding and refolding are stochastic processes having Bernoulli distributions with associated probabilities p_{unfold} and p_{refold} . The p_{unfold} probability can be estimated from the number of unfolding events (within a 1 ps time frame) divided by the number of snapshots where the system was folded. Similarly, the estimate of p_{refold} equals the number of refolding events divided by the number of snapshots where the system is unfolded. These estimates have a binomial distribution, which can be approximated by a normal distribution, and thus, the confidence intervals documenting statistical relevance of these estimated probabilities can be expressed by the Wilson score interval:⁵²

$$p_{\text{min,max}} = \frac{p + \frac{1}{2n}z_{1-\alpha/2}^2 \pm \sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

where p is the estimated probability, p_{min} and p_{max} are the lower and upper limits of the Wilson score interval of the estimated probability, n is the number of realizations (i.e., total number of snapshots) where the system is folded/unfolded, and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of the normal distribution (the α value used was 5%). Finally, the corresponding kinetic constants ($k_{\text{unfold/refold}}$) were derived from the estimated probabilities of unfolding and refolding ($p_{\text{unfold/refold}}$) using the first-order kinetic equation

$$P_{\text{unfold/refold}} = \exp(-k_{\text{unfold/refold}}\Delta t)$$

where Δt denotes the time interval between two consecutive snapshots. Subsequently, these kinetic constants were transformed to the corresponding potential free energy barriers ($\Delta G_{\text{unfold/refold}}^\ddagger$) using the Eyring equation

$$k_{\text{unfold/refold}} = \left(\frac{k_B T}{h}\right) \exp\left(-\frac{\Delta G_{\text{unfold/refold}}^\ddagger}{RT}\right)$$

in which T is the absolute temperature and k_B , R , and h are the Boltzmann, universal gas, and Planck constants, respectively.

The population and evolution of chignolin conformations in various water models in the ff99 simulations were monitored by the Bayesian clustering algorithm⁵⁹ implemented in ptraj, with all variables set to the default (i.e., without the critical distance metric). The optimal cluster number was found iteratively with visual inspection of indices measuring the clustering performance.

The Ramachandran plots of peptide residues were depicted as scatter plots and density plots using an in-house script. The names of canonical regions and their positions in the Ramachandran plot were assigned according to Schlick.⁶⁰

The following analyses of specific solvation in various solvents were performed only for ff99SB chignolin simulations, given the sufficient stability of chignolin in ff99SB; density maps of water models were calculated using ptraj (grid analysis). Prior to the analysis, chignolin was image-centered and rms-fitted, and then finally the density of water molecules was calculated using cubic grids spread over the entire box volume. The output file was visualized using VMD. Further, the numbers of water molecules in the first and second solvation shells around the chignolin molecule were calculated using ptraj. The distance of the chignolin first hydration shell was set to 3.4 Å and the second to 5.0 Å. The analysis of water residence times on the chignolin surface was based on a survival probability correlation function as implemented in ReTiNal (Resident Time Analyzer v1.0, Petr Kulhanek, NCBR Brno, <http://troll.chemi.muni.cz/whitezone/development/root/>) software (see ref 61 for details). Water molecules interacting with oxygen or nitrogen atoms of chignolin with residence time over 0.5 ns were further analyzed.

Self-diffusion coefficients (Table 1) were calculated from the last 5 ns of 10 ns long water box simulations using the Einstein equation

$$D = \frac{1}{6tN} \lim_{t \rightarrow \infty} \left\langle \sum_{i=1}^N [r_i(t) - r_i(0)]^2 \right\rangle$$

where D is the self-diffusion coefficient, t is time, N is the total number of atoms, and $r_i(t)$ is the displacement vector of the i th atom at time t . The radial distribution functions of each water model (Figure S1, Supporting Information) were calculated over the entire time scale of the water box simulations by ptraj.

Results

Chignolin. ff99 Force Field. The simulations showed that chignolin did not maintain its native-like structure in the ff99 force field and melted readily in almost all simulations with different solvent models (Table 3). The differences among estimated probabilities and corresponding free energy barriers of unfolding/refolding processes in different solvation models were not statistically significant ($\alpha = 0.05$). Nonetheless, the free energy barriers of unfolding and refolding differed significantly in different force fields, as discussed below (Table 4). The agreement between free energy barriers of unfolding estimated from a series of 100 ns simulations, and from the 1 μ s long simulation in the TIP3P and SPC/E water models, shows that the simulations sufficiently converged on a 100 ns time scale, in terms of estimation of unfolding probability. On the other hand, this does not apply for the probability of refolding, because, on longer simulation time scales, higher free energy barriers of refolding were in most cases estimated. This could have been a consequence of further structural relaxation of the unfolded state toward some energetically deeper minima.

After chignolin lost its native-like structure, it adopted one of five misfolded conformations (designated clusters 1–5; Figure S2, Supporting Information), which were subsequently identified by a cluster analysis (Table 5). We did not observe

Table 3. Populations (%) of the Native-like Chignolin Structures and α -Helicity (%) of EK-Peptide in the Respective MD Simulations

force field	water model	native-like ^a		α -helicity ^b	
		multiple ^c	long ^d	multiple ^c	long ^d
ff99	TIP3P	18 ± 26	3	16 ± 20	15 ± 20
	TIP4P	24 ± 26		14 ± 19	
	TIP5P	53 ± 36		20 ± 23	
	SPC/E	36 ± 35	11	21 ± 23	17 ± 21
ff99SB	TIP3P	95 ± 7	74	21 ± 26	4 ± 12
	TIP4P	94 ± 7		17 ± 24	
	TIP5P	94 ± 6		35 ± 30	
	SPC/E	96 ± 6	83	20 ± 29	4 ± 12
ff03	TIP3P	96 ± 2	46	70 ± 22	68 ± 23
	TIP4P	80 ± 22		74 ± 17	
	TIP5P	96 ± 2		73 ± 22	
	SPC/E	78 ± 19	79	79 ± 15	70 ± 22

^a The mean populations of the native chignolin structure with their standard deviations were calculated from multiple simulations. ^b The α -helicities are mean values with standard deviations, calculated from all multiple MD simulations or from one 1 μ s run, respectively. ^c Six independent simulations. ^d A 1 μ s MD simulation.

any significant difference in populations of unfolded clusters between simulations with the TIP3P and TIP4P models. However, the TIP5P model produced simulations significantly different from those of both TIP3P and TIP4P. The cluster 1 conformation, which was the dominant unfolding state in the TIP3P and TIP4P simulations, was hardly populated in the TIP5P simulations. Instead, the cluster 2 conformation was significantly preferred in the TIP5P simulations, but negligible in populations from the TIP3P and TIP4P simulations (Table 5). Note that the cluster 2 conformation, despite being unfolded, retained a left-handed α -helical conformation of Gly7, which seems to be the critical residue for proper chignolin folding.⁴⁸ Both clusters 1 and 2 were occupied in the series of 100 ns simulations with SPC/E, and their populations did not significantly differ from those of any of the TIP n P simulations. When we extended sampling of simulations with TIP3P and SPC/E, the most widely used models, to a 1 μ s time scale, we found the cluster populations in TIP3P 100 ns simulations were more or less converged, while further relaxation was observed with the SPC/E model. Specifically, cluster 1 became the most populated unfolded state while cluster 2 was no longer observed in 1 μ s long SPC/E simulations. Thus, chignolin in the SPC/E water model on microsecond time scales occupied the same unfolded clusters as in the TIP3P and TIP4P models, which were, however, significantly different from those clusters populated in the TIP5P model. It is worth noting that the transitions between unfolded clusters happened more rapidly in TIP3P than in the SPC/E model, which seems to be a consequence of the higher self-diffusion coefficient of TIP3P (Table 1) and could be the reason for better convergence of cluster populations in TIP3P 100 ns simulations compared to SPC/E simulations (see Figure S3, Supporting Information).

Ramachandran density and scatter plots revealed that some residues moved apart from the native regions (defined from the NMR structure;⁴⁸ Figure S2, Supporting Information) and also populated the non-native ones (Figure S4, Supporting

Table 4. Estimated Free Energy Barriers (kcal/mol) of Unfolding ($\Delta G_{\text{unfold}}^{\ddagger}$) and Refolding ($\Delta G_{\text{refold}}^{\ddagger}$) and Corresponding Free Energy Differences (kcal/mol) between Unfolded and Folded States ($\Delta G = G_{\text{unfold}} - G_{\text{fold}}$) of Chignolin in Various Solvent Models and Force Fields (298 K, 1 atm)^a

force field	water model	$\Delta G_{\text{unfold}}^{\ddagger}$	$\Delta G_{\text{refold}}^{\ddagger}$	ΔG
ff99	TIP3P	6.8 ± 0.4 (6.3 ± 0.5)	8.5 ± 0.8 (8.3 ± 0.5)	-1.7 ± 0.9 (-2.1 ± 0.7)
	TIP4P	6.5 ± 0.4	7.8 ± 0.6	-1.3 ± 0.7
	TIP5P	7.1 ± 0.4	7.3 ± 0.6	-0.2 ± 0.7
ff99SB	SPC/E	6.7 ± 0.3 (6.9 ± 0.4)	7.2 ± 0.3 (8.2 ± 0.5)	-0.5 ± 0.4 (-1.3 ± 0.6)
	TIP3P	7.5 ± 0.5 (7.3 ± 0.3)	5.7 ± 0.5 (6.8 ± 0.3)	1.9 ± 0.7 (0.6 ± 0.4)
	TIP4P	7.4 ± 0.4	5.4 ± 0.4	1.9 ± 0.6
	TIP5P	7.5 ± 0.5	5.5 ± 0.5	2.0 ± 0.7
ff03	SPC/E	8.2 ± 0.8 (7.8 ± 0.3)	6.2 ± 0.8 (6.9 ± 0.4)	2.0 ± 1.1 (0.9 ± 0.5)
	TIP3P	7.8 ± 0.6 (7.2 ± 0.3)	4.5 ± 0.6 (7.3 ± 0.3)	3.3 ± 0.8 (-0.1 ± 0.4)
	TIP4P	7.1 ± 0.4	6.3 ± 0.4	0.9 ± 0.5
	TIP5P	8.2 ± 0.8	5.6 ± 1.0	2.6 ± 1.3
	SPC/E	7.0 ± 0.3 (7.0 ± 0.2)	6.5 ± 0.4 (6.1 ± 0.2)	0.6 ± 0.5 (0.8 ± 0.3)

^a The estimates in parentheses relate to 1 μ s long simulations.

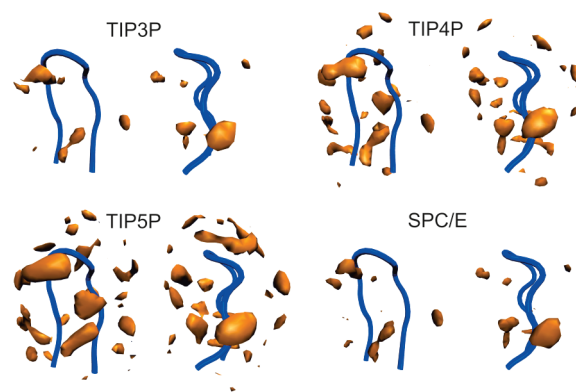
Table 5. Population (%) of the Native-like Chignolin Structure and Relative Populations (%) of Unfolded Chignolin Clusters in Multiple ff99 Simulations with the TIP3P, TIP4P, TIP5P, and SPC/E Models^a

cluster number	cluster				cluster	
	TIP3P ^b	TIP4P ^b	TIP5P ^b	SPC/E ^b	TIP3P ^c	SPC/E ^c
native-like	18 ± 26	24 ± 26	53 ± 36	36 ± 35	3	11
1	67 ± 19	44 ± 31	5 ± 4	19 ± 32	72	71
2	8 ± 5	9 ± 15	50 ± 29	30 ± 35	0	10
3	1 ± 1	4 ± 5	8 ± 10	22 ± 37	1	1
4	19 ± 13	32 ± 39	24 ± 9	12 ± 13	12	6
5	9 ± 12	14 ± 21	17 ± 16	22 ± 36	13	12

^a Note that the population of unfolded clusters is normalized, so it represents the portion of unfolded states. ^b Six independent simulations; the percentages of native-like structures and misfolded chignolin clusters of particular simulation runs are listed in Table S1 (Supporting Information). ^c A 1 μ s MD simulation.

Information). All simulations showed that the residues Thr6 and Glu5 could be found not only in the native right-handed α -helical region, but also in a region of noncanonical artificial structures ($\varphi = \sim -145^\circ$ and $\psi = \sim 0^\circ$), which is a known artifact of ff99.¹⁸ The most flexible Gly7 readily lost its native left-handed α -helical conformation in TIP3P, TIP4P, and (eventually) SPC/E simulations and shifted to the artificial region around $\varphi = \sim -165^\circ$ and $\psi = \sim 35^\circ$ or to the glycine β_s region.⁶² The population of the non-native and artificial region ($\varphi = \sim -145^\circ$ and $\psi = \sim 0^\circ$) was significantly reduced in TIP5P simulations, due to either slightly higher stability of chignolin in the TIP5P model or its favoring of cluster 2, causing less dense population of this artificial region.

ff99SB Force Field. In contrast to the ff99 simulations, chignolin was highly stable in all ff99SB simulations (Table 3). In most simulations, no unfolding event occurred, and even if the structure melted, it adopted near-native structures that swiftly refolded to the native chignolin. Both increased unfolding and decreased refolding free energy barriers contributed to significantly higher stability of the chignolin native structure in ff99SB simulations. On the other hand, neither stability nor estimated free energy barriers of unfolding and refolding differed in the set of four explicit solvent models (Table 4). The artificial population of the non-native region ($\varphi = \sim -145^\circ$ and $\psi = \sim 0^\circ$) of the Ramachandran

**Figure 2.** Water density maps (orange) around chignolin (the blue tube represents the chignolin backbone) for ff99SB simulations.

plots vanished in all ff99SB simulations, and backbone torsions of chignolin residues fluctuated around the NMR native values (Figure S5, Supporting Information).

The finding that chignolin was highly stable in all ff99SB simulations allowed us to monitor the behavior of water molecules in the vicinity of chignolin and to consider the effects of various solvent models on the specific solvation of chignolin. Both thermodynamic (the number and position of binding sites of long-residency water molecules) and kinetic (the mobility of these long-residency water molecules) perspectives on the preferential hydration of the chignolin surface were taken into account.⁶³ Similar local densities of water molecules were observed for simulations with the TIP4P and TIP5P models, while identifying large numbers of specific hydration sites. The three-site TIP3P and SPC/E water models also yielded similar density maps, but with significantly fewer specific hydration sites and with lower densities, in comparison with the TIP4P and TIP5P models (Figure 2). In addition, the analysis of long-residency water molecules (here, those with residence times longer than 0.5 ns) identified significant differences among solvation models, such that the number of long-residency water molecules decreased in the order TIP5P \gg SPC/E \approx TIP4P $>$ TIP3P (Table 6).

These findings show that the specific hydration of chignolin is determined by the mobility of a given water model in terms of its self-diffusion coefficient (Table 1) and its

Table 6. Average Number of Water Molecules in the First Two Solvation Shells of Chignolin and Average Number of Long-Residency Water Molecules (Residence Times >0.5 ns) in ff99SB Chignolin Simulations with Corresponding Standard Deviations

	TIP3P	TIP4P	TIP5P	SPC/E	TIP3P ^{1 μs}	SPC/E ^{1 μs}
first shell, <3.4 Å	93 ± 8	93 ± 6	100 ± 8	94 ± 7	90 ± 6	95 ± 8
second shell, <5.0 Å	185 ± 14	182 ± 12	196 ± 15	186 ± 13	179 ± 11	189 ± 15
long-residency water molecules	0.0 ± 0.0	2.0 ± 1.3	12.8 ± 4.5	3.7 ± 2.7		

propensity to bind to specific hydration sites. Thus, the highest specific solvation in TIP5P simulations was caused by a large number of specific hydration sites and the low (and most realistic) mobility of TIP5P water molecules. On the other hand, no long-residency water molecules were identified for the TIP3P model, due to the small number of preferred hydration sites and rapid exchange of TIP3P water molecules at these sites. Interestingly, although the SPC/E model has a self-diffusion coefficient similar to that of TIP5P, the significantly smaller quantity of long-residency SPC/E water molecules was caused by a low number of specific hydration sites, most likely stemming from solute–solvent interactions and thermodynamic factors. Similarly, although the water density map of TIP4P was similar to that of TIP5P, with the same number of specific hydration sites, the higher self-diffusion coefficient of TIP4P resulted in significantly more rapid dynamics at these hydration sites, and thus, the specific hydration was reduced for kinetic reasons.

ff03 Force Field. As in ff99SB simulations, the chignolin native structure was highly populated in ff03 simulations. The rare unfolding events were frequently quickly followed by chignolin refolding in almost all cases, except in some SPC/E and TIP4P simulations and the 1 μs TIP3P simulation, where we observed minor but apparent populations of unfolded chignolin with Gly7 shifted from the left-handed α-helical region to the β_{PR} region with φ = ~80° and ψ = ~-150° (Table 3; Figure S6, Supporting Information).⁶² These non-native chignolin conformations closely match the β-hairpin topology, but lack the native hydrogen bond network and contacts between Tyr2 and Trp9 residues. On the other hand, the shift of Gly7 to the β_{PR} region was always fully reversible, and refolding was sooner or later observed. It seems that the shift of Gly7 does not depend on the solvation model and can be considered as an ff03 force field effect. The difference in the balance between left-handed α-helical and β_{PR} conformations of Gly7 in the ff03 and ff99SB force fields can be explained by differences in the free energy landscapes of the φ and ψ dihedrals in ff03 and ff99SB.⁴⁵

EK-Peptide. *ff99 Force Field.* The EK-peptide readily lost its α-helical structure in all ff99 simulations (Table 3). Melting that propagated from the termini occurred within a time scale of hundreds of picoseconds to several nanoseconds. Melting times were progressively longer in the TIP5P and SPC/E water models, whose self-diffusion coefficients were smaller (and more realistic) than those of TIP3P, which had the highest self-diffusion coefficient of the explicit models in this study (Figures S13 and S14, Supporting Information). In the unfolded state, the peptide was highly dynamic, often switching among the ₃₁₀-helical structure, turn, random coil, and α-helical conformations. Some

Table 7. Salt Bridge Probabilities (Distance up to 4.5 Å) Calculated from EK-Peptide Simulations with ff03^a

water model	Glu3...Lys6	Lys6...Glu9	Glu9...Lys12
TIP3P	0.36 ± 0.04 (0.35)	0.06 ± 0.02 (0.06)	0.31 ± 0.02 (0.33)
TIP4P	0.33 ± 0.07	0.03 ± 0.01	0.36 ± 0.04
TIP5P	0.49 ± 0.20	0.08 ± 0.10	0.51 ± 0.15
SPC/E	0.33 ± 0.07 (0.29)	0.06 ± 0.06 (0.06)	0.36 ± 0.07 (0.29)

^a The mean probabilities and corresponding standard deviations were calculated from multiple simulations having helicity above 70%. The estimates in parentheses relate to 1 μs simulations.

residues also populated the artificial region around φ = ~-145° and ψ = ~0°, which is a known artifact of the ff99 force field.¹⁸ Notably, internal residues of the EK-peptide had a propensity to refold for up to ~10 ns in the TIP5P and SPC/E models.

ff99SB Force Field. As in the previous case, the native α-helical fold was unstable throughout the time scale of the conducted simulations (Figures S15 and S16, Supporting Information). Melting of the helical structure occurred early, within a time scale of several nanoseconds, in simulations with the most mobile models, TIP3P and TIP4P. In contrast, in simulations using the water models with lower (and more realistic) mobility, TIP5P and SPC/E, the system melted later, generally on a time scale of tens of nanoseconds. The residues of the unfolded EK-peptide populated the left- and right-handed helical regions, the region of noncanonical structures around φ = ~-145° and ψ = ~0°, and the antiparallel β-sheet and triple-stranded collagen helix regions of the Ramachandran plot (Figures S8 and S11, Supporting Information). Nonetheless, the 1 μs TIP3P and SPC/E MD runs characterized the behavior of the system after melting, revealing that the antiparallel β-sheet, triple-stranded collagen helix, and left-handed regions were significantly populated in this state.

ff03 Force Field. In all simulations, the EK-peptide displayed a significant level of intrinsic α-helicity lying in the interval from ~70% to ~80% (Table 3; Figures S17 and S18, Supporting Information). The terminal residues had less helical propensity than internal residues and in the nonhelical state populated the triple-stranded collagen helix, antiparallel β-sheet, left-handed α-helix, and φ = ~-160° and ψ = ~-25° regions (Figures S9 and S12, Supporting Information).

The high stability of α-helical content in ff03 simulations of EK-peptide led us to monitor the stability of salt bridges between Lys and Glu side chains (Table 7) and the specific hydration of backbone carbonyl oxygen atoms (Figure 3). The formation of salt bridges and shielding of backbone H-bonds forming group (-C=O, -N-H) from water molecules were suggested to contribute to the stability of

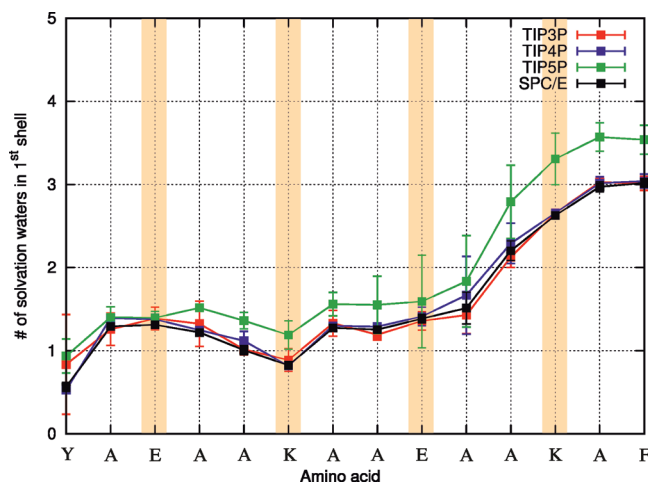


Figure 3. Average numbers of water molecules within the first hydration shell ($<3.4 \text{ \AA}$) of backbone carbonyl oxygen atoms calculated from ff03 simulations of EK-peptide.

alanine-based α -helical peptides (see refs 64 and 65 and references therein). The TIP3P, TIP4P, and SPC/E models behave similarly; i.e., the stability of salt bridges and hydration of backbone carbonyl oxygen atoms do not significantly differ among these three solvent models. On the other hand, MD simulations with TIP5P show higher stabilities of salt bridges (Table 7) and more extensive hydration of backbone carbonyl oxygen atoms (Figure 3).

Discussion

In the present study we aimed to elucidate the effects of four popular explicit solvent models (TIP3P, TIP4P, TIP5P, and SPC/E) on the behavior of two oligopeptides: a β -hairpin chignolin and an α -helical EK-peptide. We combined these solvent models with three recent protein force fields (ff99, ff99SB, and ff03) from the AMBER family. Our results suggest that the choice of solvent model does not significantly affect the stability of the studied peptides, which was completely governed by the force field employed. Furthermore, even the conformational behavior of the peptides in their native folded state was not influenced by the choice of solvent model.

We found that the ff99 force field destabilized the native fold of both peptides to an extent that significantly underestimated peptide stabilities in comparison with experimentally observed values.^{48,50} In addition, in ff99 simulations we observed a significant population of noncanonical conformations in the form of φ - ψ torsions ($\varphi = \sim -145^\circ$ and $\psi = \sim 0^\circ$), which are a known ff99 force field artifact.¹⁸ Elimination of this artifact motivated the development of a reparametrized force field named ff99SB.¹⁸ The ff99SB force field overstabilized the chignolin structure, while it did not improve the stability of the EK-peptide, destabilizing helical EK-peptide in favor of β -structured states. This indicates that although the ff99SB force field slightly improves the behavior of β -structured chignolin, it yields unbalanced force fields biased toward β -structures. Finally, the ff03 force field seemed the best choice among the tested force fields, as the chignolin stability estimated from all ff03 simulations was

$72 \pm 21\%$, in good agreement with the experimentally observed chignolin stability of 60%. Moreover, the ff03 simulations did not cause any force field artifacts. They displayed folding/unfolding processes in chignolin connected with shifts of the flexible Gly7 between the left-handed α -helix and β_{PR} conformations, which represents the tolerated conformational variability of this Gly residue. Nonetheless, the ff03 force field still overestimated the stability of α -helical EK-peptide, as the α -helicity observed in the ff03 simulations amounted to ~ 70 – 80% , while the experimental helical content equals $\sim 20\%$ at 300 K. Thus, the ff03 force field seems the best, but still not perfect, choice from the studied set.

While we did not observe any effect of explicit solvent models on the stability of the studied peptides, we found substantial differences in specific solvation of the backbone carbonyls of the EK-peptide α -helix and of the chignolin native structure in the set of studied water models. The TIP5P model showed more extensive hydration of backbone carbonyl oxygen atoms in comparison with TIP3P, TIP4P, and SPC/E, which gave almost the same results (Figure 3). The higher hydration of H-bonds between the backbone amide and carbonyl groups of the α -helix was suggested to destabilize the α -helicity of alanine-based peptides.^{64,65} Thus, the lower α -helicity of EK-peptide in the TIP5P simulations in comparison with other solvation models was anticipated as an effect of the more extensive backbone hydration. However, the α -helicity of EK-peptide did not significantly differ among various solvent models (Table 3). This finding can be most likely explained by the higher probabilities of salt bridges in TIP5P water (Table 7) and compensation between the specific solvation of backbone carbonyl oxygens and salt bridge stabilities.

Chignolin was highly solvated in the TIP5P simulations, while significantly lower solvation was observed in other water model simulations. We analyzed the reasons for decreased specific solvation in the remaining three models and found that specific solvation by TIP4P was reduced by rapid exchange of water molecules in the solvation binding sites, i.e., for kinetic reasons. This finding can be explained by the larger self-diffusion coefficient of TIP4P in comparison with the TIP5P model. The specific solvation by SPC/E was reduced due to the small number of specific solvent binding sites compared to those in the TIP5P and TIP4P models, i.e., for thermodynamic reasons. Finally, the reduced specific solvation in TIP3P simulations was a consequence of both the aforementioned kinetic and thermodynamic factors (Figure 4). Consequently, assuming that the studied solvation models differ significantly in the specific solvation of solute and do not concurrently affect the stability of the studied peptide solutes, the question arises of whether these differences in specific solvation can influence the solute and, if so, how.

As the differences in specific solvation of the studied models did not affect the behavior of the studied peptides in their folded states, we focused rather on their behavior in unfolded states in ff99 simulations. We found that the conformational behavior of chignolin in unfolded states is similar in all three water models with lower specific solvation, i.e., in TIP3P, TIP4P, and SPC/E, while a different

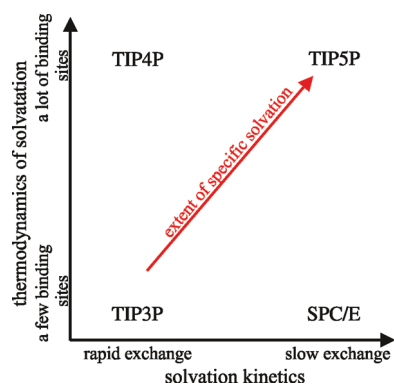


Figure 4. Schematic relationship among the solvation kinetics, thermodynamics of specific solvation (expressed by the number of specific binding sites), and extent of specific solvation of chignolin peptide in four studied water models.

conformational behavior of unfolded chignolin was observed in simulations with TIP5P, demonstrating a high degree of specific solvation. Specifically, we observed different preferences in the population of clusters of unfolded structures. The TIP3P, TIP4P, and SPC/E simulations resulted in preferred occupancy of the unfolded cluster with an artificial conformation of Gly7 ($\varphi = \sim -165^\circ$ and $\psi = \sim 35^\circ$), while in TIP5P simulations an alternative cluster with a left-handed α -helical conformation of Gly7 was preferred. This result implies that the solvation models will most likely not significantly affect the structure, stability, and flexibility of structured segments of folded proteins. However, the differences in specific solvation might contribute to conformational variability and flexibility of solvent-exposed flexible parts, such as flexible glycine-rich loops. Notably, these most flexible parts of proteins are frequently connected with specific biological functions and are at the focus of attention of many MD simulations.^{51,52,66–72} Moreover, specific solvation is also the subject of many other studies concerning, e.g., drug design.^{67,73–75} Therefore, in these cases, the correct description of specific solvation might be critical for correct results.

Unfortunately, we are not able to ultimately conclude which model provides the most realistic description of specific solvation. On the other hand, our data suggest that the TIP5P model provides a significantly different specific solvation that could propagate to solute behavior via stability of salt bridges and modification of conformational preferences in unfolded states. As the TIP5P model has the most realistic solvent parameters (among studied solvent models), we suggest that TIP5P is the most promising candidate for the realistic description of solute–solvent interactions and specific solvation. Nonetheless, this implication is not straightforward and might be questionable, so direct comparison with reference experimental data of the specific solvation or of the solute–solvent interaction is needed to determine which model performs the best.

In addition, we observed a relationship between the mobility of a water model, represented by its self-diffusion coefficient, and the speed of processes in our simulations, namely, the kinetics of conversion between the unfolded clusters and the structural relaxation of unfolded states in

ff99 simulations. Thus, we conclude that the mobility of solvent models can accelerate the kinetics of processes in solutes.

Taken together, in sharp contrast to the choice of force field, choosing the most reliable water model is not straightforward. In cases where it can be assumed that specific solvation will not play an important role (i.e., in well-structured proteins lacking highly flexible loops), all the model performances may be similar in quality. Furthermore, in such cases, simulations can benefit from the high (overestimated) mobility of the TIP3P model, because it does not significantly affect the thermodynamic properties of the solutes, such as the structure of thermodynamically stable conformers, but it significantly reduces the barriers between these states and thus accelerates the kinetics and effectively increases sampling in the simulations. On the other hand, if the solute contains flexible and solvent-exposed loops, it would be better to use the more realistic description of solvation by the TIP5P model, even though this would entail an increased simulation time due to the relatively slower processes and lower sampling.

Acknowledgment. Support through the Ministry of Youth, Sports and Education of the Czech Republic (Grants CZ.1.05/2.1.00/03.0058, LC512, and MSM6198959216), the Grant Agency of the Czech Republic (Grant 203/09/H046), and Student Project PrF_2010_025 of Palacký University is gratefully acknowledged. We thank Sees-Editing, Ltd. (U.K.) for linguistic revision.

Supporting Information Available: Figures S1–S21 and Tables S1–S4 giving the OO radial distribution function for explicit water models, three-dimensional structure and scatter plots of chignolin misfolded structures (ff99) and scatter plot of chignolin NMR structures, time courses of rmsd colored according to chignolin misfolded structures (ff99), population of the chignolin clusters and populations of the native-like chignolin structures in TIP3P, TIP4P, TIP5P, and SPC/E models for chignolin simulations, Ramachandran density and scatter plots of chignolin simulations, secondary structure of EK-peptide simulations, Ramachandran density and scatter plots of EK-peptide long simulations, secondary structure of EK-peptide (ff99SB), secondary structure of EK-peptide (ff03), α -helicity of EK-peptide, salt-bridge probabilities (ff03), normalized histograms of Lys–Glu distances, cumulative histograms of the probability of salt-bridge formation, and number of waters within the first hydration shell around the carbonyl oxygen of each residue in the EK-peptide (ff03). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584.
- (2) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27.
- (3) van der Kamp, M. W.; Shaw, K. E.; Woods, C. J.; Mulholland, A. J. *J. R. Soc. Interface* **2008**, *5* (3), S173.
- (4) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke,

- D. P.; Glattli, A.; Hunenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064.
- (5) Sponer, J.; Spackova, N. *Methods* **2007**, *43*, 278.
- (6) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1.
- (7) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.
- (8) Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (9) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (10) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
- (11) Guvench, O.; MacKerell, A. D., Jr. *Methods Mol. Biol.* **2008**, *443*, 63.
- (12) Ditzler, M. A.; Otyepka, M.; Sponer, J.; Walter, N. G. *Acc. Chem. Res.* **2010**, *43*, 40.
- (13) Perez, A.; Lankas, F.; Luque, F. J.; Orozco, M. *Nucleic Acids Res.* **2008**, *36*, 2379.
- (14) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Loughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. *Nucleic Acids Res.* **2010**, *38*, 299.
- (15) MacKerell, A. D.; Nilsson, L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 194.
- (16) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (17) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1357.
- (18) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.
- (19) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.
- (20) Best, R. B.; Buchete, N. V.; Hummer, G. *Biophys. J.* **2008**, *95*, 4494.
- (21) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004.
- (22) Chandler, D. *Nature* **2005**, *437*, 640.
- (23) Roca, M.; Messer, B.; Warshel, A. *FEBS Lett.* **2007**, *581*, 2065.
- (24) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647.
- (25) Leach, A. R. *Molecular Modelling: Principles and Applications*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 2001; p 768.
- (26) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187.
- (27) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219.
- (28) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.
- (29) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *114*, 363.
- (30) Kusalik, P. G.; Svishchev, I. M. *Science* **1994**, *265*, 1219.
- (31) Vega, C.; Abascal, J. L. F. *J. Chem. Phys.* **2005**, *123*, 144504.
- (32) Vega, C.; Sanz, E.; Abascal, J. L. F. *J. Chem. Phys.* **2005**, *122*, 114507.
- (33) Baez, L. A.; Clancy, P. *J. Chem. Phys.* **1994**, *101*, 9837.
- (34) Vega, C.; McBride, C.; Sanz, E.; Abascal, J. L. F. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1450.
- (35) Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954.
- (36) Zielkiewicz, J. *J. Chem. Phys.* **2005**, *123*, 104501.
- (37) Hess, B.; van der Vegt, N. F. *J. Phys. Chem. B* **2006**, *110*, 17616.
- (38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (39) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (40) Nutt, D. R.; Smith, J. C. *J. Chem. Theory Comput.* **2007**, *3*, 1550.
- (41) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 1.
- (42) Wong, V.; Case, D. A. *J. Phys. Chem. B* **2008**, *112*, 6013.
- (43) Glass, D. C.; Krishnan, M.; Nutt, D. R.; Smith, J. C. *J. Chem. Theory Comput.* **2010**, *6*, 1390.
- (44) Price, D. J.; Brooks, C. L. *J. Chem. Phys.* **2004**, *121*, 10096.
- (45) Vymetal, J.; Vondrasek, J. *J. Phys. Chem. B* **2010**, *114*, 5632.
- (46) Sun, Y. X.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1164.
- (47) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.
- (48) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. *Structure* **2004**, *12*, 1507.
- (49) Scholtz, J. M.; Barrick, D.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 185.
- (50) Ghosh, K.; Dill, K. A. *J. Am. Chem. Soc.* **2009**, *131*, 2306.
- (51) Bartova, I.; Koca, J.; Otyepka, M. *J. Mol. Model.* **2008**, *14*, 761.
- (52) Wilson, E. B. *J. Am. Stat. Assoc.* **1927**, *22*, 209.
- (53) Sklenovsky, P.; Banas, P.; Otyepka, M. *J. Mol. Model.* **2008**, *14*, 747.
- (54) Skopalik, J.; Anzenbacher, P.; Otyepka, M. *J. Phys. Chem. B* **2008**, *112*, 8165.
- (55) Sklenovsky, P.; Otyepka, M. *J. Biomol. Struct. Dyn.* **2010**, *27*, 521.
- (56) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (57) Day, R.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13445.
- (58) Scott, K. A.; Alonso, D. O. V.; Pan, Y. P.; Daggett, V. *Biochemistry* **2006**, *45*, 4153.
- (59) Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. *J. Chem. Theory Comput.* **2007**, *3*, 2312.
- (60) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*; Springer Science + Business Media LLC: New York, 2002; Vol. 21.
- (61) Kriz, Z.; Otyepka, M.; Bartova, I.; Koca, J. *Proteins* **2004**, *55*, 258.
- (62) Ho, B. K.; Brasseur, R. *BMC Struct. Biol.* **2005**, *5*, 14.

- (63) Priya, M. H.; Shah, J. K.; Asthagiri, D.; Paulaitis, M. E. *Biophys. J.* **2008**, *95*, 2219.
- (64) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782.
- (65) Ghosh, T.; Garde, S.; Garcia, A. E. *Biophys. J.* **2003**, *85*, 3187.
- (66) Dvorakova-Hola, K.; Matuskova, A.; Kubala, M.; Otyepka, M.; Kucera, T.; Vecer, J.; Herman, P.; Parkhomenko, N.; Kutejova, E.; Janata, J. *J. Mol. Biol.* **2010**, *396*, 1197.
- (67) Bartova, I.; Otyepka, M.; Kriz, Z.; Koca, J. *Protein Sci.* **2005**, *14*, 445.
- (68) Bartova, I.; Otyepka, M.; Kriz, Z.; Koca, J. *Protein Sci.* **2004**, *13*, 1449.
- (69) Wong, L.; Jennings, P. A.; Adams, J. A. *Acc. Chem. Res.* **2004**, *37*, 304.
- (70) Narayanan, A.; Jacobson, M. P. *Curr. Opin. Struct. Biol.* **2009**, *19*, 156.
- (71) Prudent, R.; Sautel, C. F.; Cochet, C. *BBA—Proteins Proteomics* **2010**, *1804*, 493.
- (72) Otyepka, M.; Bartova, I.; Kriz, Z.; Koca, J. *J. Biol. Chem.* **2006**, *281*, 7271.
- (73) de Beer, S. B.; Vermeulen, N. P.; Oostenbrink, C. *Curr. Top. Med. Chem.* **2010**, *10*, 55.
- (74) Yan, A.; Grant, G. H.; Richards, W. G. *J. R. Soc. Interface* **2008**, *5* (3), S199.
- (75) Zhang, B.; Tan, V. B.; Lim, K. M.; Tay, T. E. *J. Chem. Inf. Model.* **2007**, *47*, 1877.
- (76) Hasted, J. B. *Liquid Water: Dielectric Properties. Water: A Comprehensive Treatise*; Plenum Press: New York, 1972; Vol. 1.
- (77) Gubskaya, A. V.; Kusalik, P. G. *J. Chem. Phys.* **2002**, *117*, 5290.
- (78) Fernandez, D. P.; Mulev, Y.; Goodwin, A. R. H.; Sengers, J. M. H. L. *J. Phys. Chem. Ref. Data* **1995**, *24*, 33.
- (79) Jancso, G.; van Hook, W. A. *Chem. Rev.* **1974**, *74*, 689.

CT1003687

JCTC

Journal of Chemical Theory and Computation

QM/MM Studies of the Matrix Metalloproteinase 2 (MMP2) Inhibition Mechanism of (*S*)-SB-3CT and its Oxirane Analogue

Jia Zhou,[†] Peng Tao,[†] Jed F. Fisher,[‡] Qicun Shi,[‡] Shahriar Mobashery,[‡] and H. Bernhard Schlegel^{*,†}

Department of Chemistry, Wayne State University, 5101 Cass Ave, Detroit, Michigan 48202, United States, and Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

Received July 8, 2010

Abstract: SB-3CT, (4-phenoxyphenylsulfonyl)methylthiirane, is a potent, mechanism-based inhibitor of the gelatinase subclass of the matrix metalloproteinase (MMP) family of zinc proteases. The gelatinase MMPs are unusual in that there are several examples where both enantiomers of a racemic inhibitor have comparable inhibitory abilities. SB-3CT is one such example. Here, the inhibition mechanism of the MMP2 gelatinase by the (*S*)-SB-3CT enantiomer and its oxirane analogue is examined computationally and compared to the mechanism of (*R*)-SB-3CT. Inhibition of MMP2 by (*R*)-SB-3CT was shown previously to involve enzyme-catalyzed C–H deprotonation adjacent to the sulfone, with concomitant opening by β -elimination of the sulfur of the three-membered thiirane ring. Similarly to the *R* enantiomer, (*S*)-SB-3CT was docked into the active site of MMP2, followed by molecular dynamics simulation to prepare the complex for combined quantum mechanics and molecular mechanics (QM/MM) calculations. QM/MM calculations with B3LYP/6-311+G(d,p) for the QM part (46 atoms) and the AMBER force field for the MM part were used to compare the reaction of (*S*)-SB-3CT and its oxirane analogue in the active site of MMP2 (9208 atoms). These calculations show that the barrier for the proton abstraction coupled ring-opening reaction of (*S*)-SB-3CT in the MMP2 active site is 4.4 kcal/mol lower than that of its oxirane analogue, and the ring-opening reaction energy of (*S*)-SB-3CT is only 1.6 kcal/mol less exothermic than that of its oxirane analogue. Calculations also show that the protonation of the ring-opened products by water is thermodynamically much more favorable for the alkoxide obtained from the oxirane than for the thiolate obtained from the thiirane. In contrast to (*R*)-SB-3CT and the *R*-oxirane analogue, the double bonds of the ring-opened products of (*S*)-SB-3CT and its *S*-oxirane analogue have the *cis* configuration. Vibrational frequency and intrinsic reaction path calculations on a reduced size QM/MM model (2747 atoms) provide additional insight into the mechanism. These calculations yield 5.9 and 6.7 for the deuterium kinetic isotope effect for C–H bond cleavage in the transition state for the *R* and *S* enantiomers of SB-3CT, in good agreement with the experimental results.

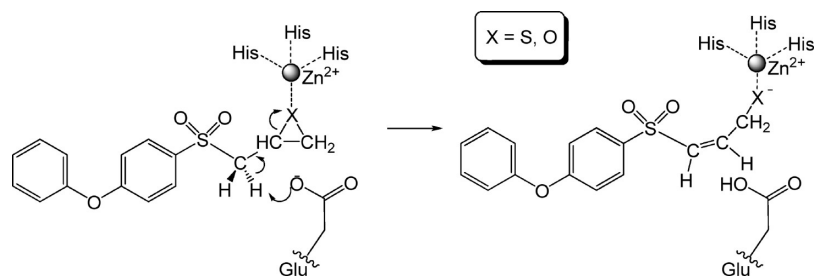
The matrix metalloproteinases (MMPs) are key proteolytic regulators of the integrity of the extracellular matrix. MMPs

are implicated in embryonic development,^{1–3} tissue remodeling and repair,^{4–6} neurophathic pain processes,⁷ cancer,^{8–11} and other diseases.^{12–15} MMP2 (gelatinases A), one of these zinc-dependent proteolytic enzymes, digests type IV collagens.¹⁶ The structure and function of this protein have been studied extensively for the purpose of selective inhibitor

* Corresponding author phone: (313) 577-2562; fax: (313) 577-8822; e-mail: hbs@chem.wayne.edu.

[†] Wayne State University.

[‡] University of Notre Dame.

Scheme 1. MMP2 Inhibition Mechanisms by SB-3CT (**3**, X = S) and its oxirane analogue (**4**, X = O)

design.^{17–29} One of these inhibitors, (4-phenoxyphenylsulfonyl)methylthiirane (SB-3CT), selectively inhibits MMP2 with high potency.^{30,31} The inhibition mechanism of MMP2 by (*R*)-SB-3CT is coupled deprotonation of the methylene group juxtaposed between the sulfone and the thiirane and the opening of the thiirane ring.^{31,32} This reaction creates a thiolate anion, which strongly coordinates with the zinc at the active site.

It is remarkable that the *R* and *S* enantiomers of SB-3CT display similar potency as inhibitors of MMP2, even though they are expected to have rather different binding modes in the active site.³³ The kinetic parameters for the *R* enantiomer are $K_{\text{on}} = 2.2 \pm 0.5 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$, $K_{\text{off}} = 5.3 \pm 0.5 \times 10^{-4} \text{ s}^{-1}$, and $K_i = 24 \pm 6 \text{ nM}$, and the corresponding values for the *S* enantiomer are $K_{\text{on}} = 1.7 \pm 0.4 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$, $K_{\text{off}} = 4.0 \pm 0.3 \times 10^{-4} \text{ s}^{-1}$, and $K_i = 23 \pm 6 \text{ nM}$. Their similar potency suggests that both enantiomers have a similar inhibitory mechanism, despite this anticipated difference in their binding modes. In this study we used methods similar to our previous study of MMP2•(*R*)-SB-3CT³² to investigate these binding modes and to compare the inhibition mechanism of (*S*)-SB-3CT (**3**) and its oxirane analogue (**4**) (Scheme 1). The MMP2•(*S*)-SB-3CT complex was constructed by docking and molecular dynamics studies. The details of the deprotonation/ring-opening mechanism for inhibition were examined by combined quantum mechanics and molecular mechanics (QM/MM) methods and compared to those of (*R*)-SB-3CT. Vibrational frequencies, intrinsic reaction paths, and kinetic isotope effects were calculated for both the *R* and *S* enantiomers of SB-3CT.

Computational Methods

As in the previous study of the *R* isomer of SB-3CT, the initial structures of the MMP2 complex with (*S*)-SB-3CT and its oxirane analogue were built by docking and molecular dynamics (MD) methods.³² A two-layer ONIOM method^{34–41} was used for the QM/MM study of the inhibition mechanism of **3** and **4**. The QM region (46 atoms) consists of the zinc ion; the three imidazole rings from His403, His407, and His413; the CH_2CO_2^- part of the Glu404 side chain; the thiirane and the SO_2CH_2 group of the inhibitor; and one water molecule. The B3LYP/6-31G(d) level of density functional theory (DFT) described the QM part of the system, and the Amber force field⁴² described the MM part of the system. QM/MM geometry optimization was carried out with a mechanical embedding scheme. The QM part and all residues and solvent molecules in the MM part within 6 Å of the QM part were fully optimized (936 atoms), while the

remaining atoms were held fixed. Similar cutoffs have been used previously in QM/MM studies of enzymatic systems to avoid spurious changes in the energy due to remote fluctuation in the geometry.^{43,44} Because the MMP2 active site is rather open to the solvent, a smaller cutoff of 6 Å was used. The partial charges for the reactive system were refined by alternating between QM/MM geometry optimization and RESP^{45,46} charge fitting.³² With mechanical embedding, the electrostatic interactions between the QM and MM regions are calculated using these partial charges. Single point calculations with electronic embedding⁴¹ were used for the final QM/MM energies calculated at the ONIOM(B3LYP/6-311+G(d,p):AMBER) level of theory. When calculated with the same basis set, the mean absolute difference in the relative energies with electronic embedding vs mechanical embedding is 3 kcal/mol (see Supporting Information Table S1 for details). All ONIOM calculations were performed with the development version of Gaussian.⁴⁷ The ONIOM toolkit⁴⁸ facilitated the QM/MM calculations.

Results and Discussion

The MMP2•(*R*)-SB-3CT and MMP2•(*S*)-SB-3CT complexes are compared in Figure 1. These complexes are obtained by docking, followed by MD simulation and QM/MM geometry optimization. Similar to (*R*)-SB-3CT in the MMP2 active site,³² the phenoxyphenyl side chain of (*S*)-SB-3CT fits into the S1' pocket, and the hydrogen bond (1.83 Å) between the backbone NH of Leu191 to the *pro-S* oxygen of the sulfone is preserved. The second oxygen of the sulfone is exposed to solvent. In contrast to (*R*)-SB-3CT, the plane of the thiirane ring of (*S*)-SB-3CT points away from the zinc. The Zn–S distance in MMP2•(*S*)-SB-3CT (4.55 Å) is significantly longer than in MMP2•(*R*)-SB-3CT (2.91 Å). As a consequence, a water molecule coordinates with the zinc in the (*S*)-SB-3CT complex (Zn–O distance is 2.11 Å) but not in the (*R*)-SB-3CT complex (3.58 Å). The MMP2 complex of the oxirane analogue of (*S*)-SB-3CT is generated by replacement of the sulfur of the thiirane with oxygen. The QM part and adjacent regions were examined visually (see Supporting Information Figure S1 for a superposition of the reactants and transition states for **3** and **4**) to ensure that the reactant complex and the transition state have similar hydrogen-bonding patterns with the solvent water molecules, to avoid spurious differences in barrier heights caused by minor changes in the solvent.

A. Deprotonation and Ring-Opening of the Inhibitor. The reactant, transition state (TS), and product structures for the inhibition of MMP2 by (*S*)-SB-3CT are shown in

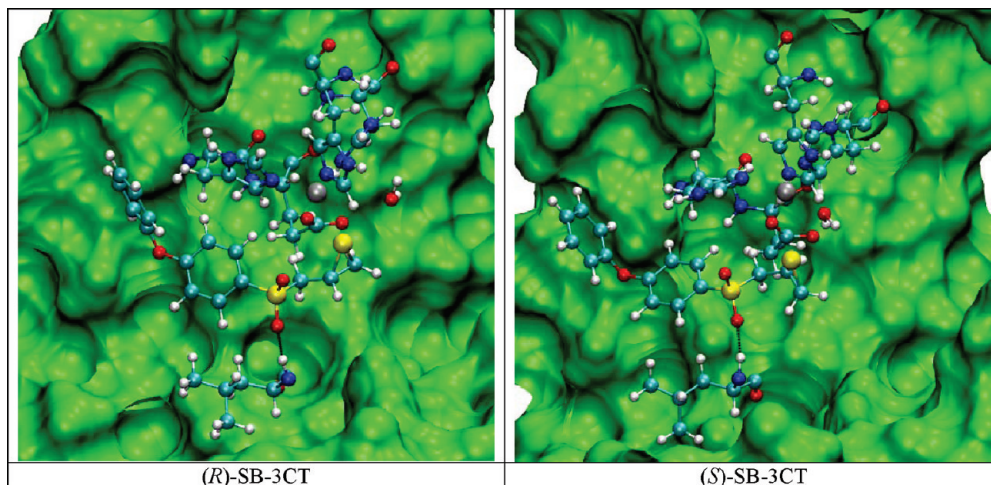


Figure 1. Structures of (*R*)-SB-3CT and (*S*)-SB-3CT in the MMP2 active site optimized at the ONIOM(B3LYP/6-31G(d):AMBER) level of theory. Residues of MMP2·(*R*)-SB-3CT and MMP2·(*S*)-SB-3CT are shown in ball-and-stick representation with atom colored according to atom types (H, C, N, O, S, and Zn, shown in white, cyan, blue, red, yellow, and gray, respectively). The same color scheme is used in all the figures.

the top two rows of Figure 2 and selected geometrical parameters are given in Figure S2 (Supporting Information). The respective structures for the oxirane analogue are shown at the bottom of these figures. In both reactant structures, the zinc at the MMP2 active site is coordinated with Glu404, three histidines, and one water molecule. Depending on how the Glu404 coordinates with the zinc, two local minima are identified for the reactant complex of MMP2·(*S*)-SB-3CT and of its oxirane analogue (**3-R** and **3-R'**; **4-R** and **4-R'** in Figure 2). In **3-R**, the oxygen of the Glu404 that will accept the transferring proton of (*S*)-SB-3CT is coordinated (2.05 Å) to the zinc, while the other Glu404 oxygen is not (3.29 Å). For **3-R'**, the $-\text{CO}_2^-$ group of Glu404 is shifted so that the acceptor oxygen is farther from the zinc (3.60 Å) and the other oxygen is coordinated to the zinc (1.96 Å). **3-R'** is only 2.6 kcal/mol higher in energy than **3-R**. Likewise, the oxirane analogue **4-R'** is 3.9 kcal/mol higher than **4-R**. In both **3-R** and **3-R'**, the sulfur of the thiirane is significantly further from the zinc (by 4.55 and 4.11 Å, respectively) than the distance between the oxygen of oxirane and the zinc in **4-R** and **4-R'** (2.32 and 2.26 Å, respectively). The energy of MMP2·(*S*)-SB-3CT increases by 4.9 kcal/mol if the sulfur is constrained to be closer to the zinc, similar to the distance in MMP2·(*R*)-SB-3CT (2.86 Å).

In the TS identified in the QM/MM calculations, **3-TS** (Figure 2 and Supporting Information Figure S2), the transferring proton is 1.50 Å from the donor carbon and 1.18 Å from the acceptor oxygen, indicating that the TS is a little earlier compared to that of MMP2·(*R*)-SB-3CT (C–H and H–O distances of 1.57 Å and 1.14 Å, respectively). The breaking C–S bond of the thiirane is elongated to 2.04 Å in **3-TS**. The Glu404 side chain moves away from the zinc in order to abstract the proton. The TS of the oxirane analogue (**4-TS**) is similar to the thiirane system (C–H and H–O distances are 1.58 and 1.13 Å), but a little later than the (*R*)-oxirane TS (C–H and H–O distances are 1.43 Å and 1.24 Å). In the transition states, the thiirane sulfur and oxirane oxygen are strongly coordinated to the zinc (Zn–S and

Zn–O distances are 2.38 and 1.94 Å, respectively), but the glutamate is not.

B. Vibrational Frequencies and Reaction Path Following. While the full frequency calculations at the current level of theory are not feasible for the whole system (9208 atoms), the two lowest vibrational modes can be calculated. Only one imaginary frequency is found for **3-TS** and **4-TS** (1018i and 540i cm^{-1} , respectively), verifying that these structures are transition states. Two reduced size (partial) models were built from the full thiirane and oxirane TS complexes by extracting all the QM atoms (46 atoms), the MM atoms that are allowed to move during geometry optimization (936 atoms), and enough frozen MM atoms to surround the first two parts, for a total of 2747 atoms. When the transition states of the reduced systems were reoptimized, the key parameters changed very little (rmsd = 0.04 Å, C–H and O–H distances involving the transferring proton changed by less than 0.01 Å). The fact that these changes are quite small suggests that the reduced systems are good models of the full protein–inhibitor complexes. Using the optimized transition states of these reduced size models, the full frequency analysis and calculation of the intrinsic reaction coordinate (IRC) are both practical with the latest version of the code.⁴⁹ The frequency analysis for the partial models of the TSs gives imaginary frequencies close to the full system for both the thiirane (1031i cm^{-1}) and oxirane (601i cm^{-1}). As a part of the present work, similar calculations were carried out for the TS structure for the inhibition of MMP2 by (*R*)-SB-3CT.³² The imaginary frequencies are 767i and 642i for the full and reduced size thiirane systems, respectively. For its oxirane analogue, the imaginary frequencies are 1217i and 1234i for the full and reduced size systems, respectively.

The IRC was calculated using the reduced size model to generate the reaction path from the reactants to the products through the TSs. The profiles of the IRC paths, and the key structures, are shown in Figure 3 (see Supporting Information Figure S3 for *R* enantiomer). The reaction path calculations

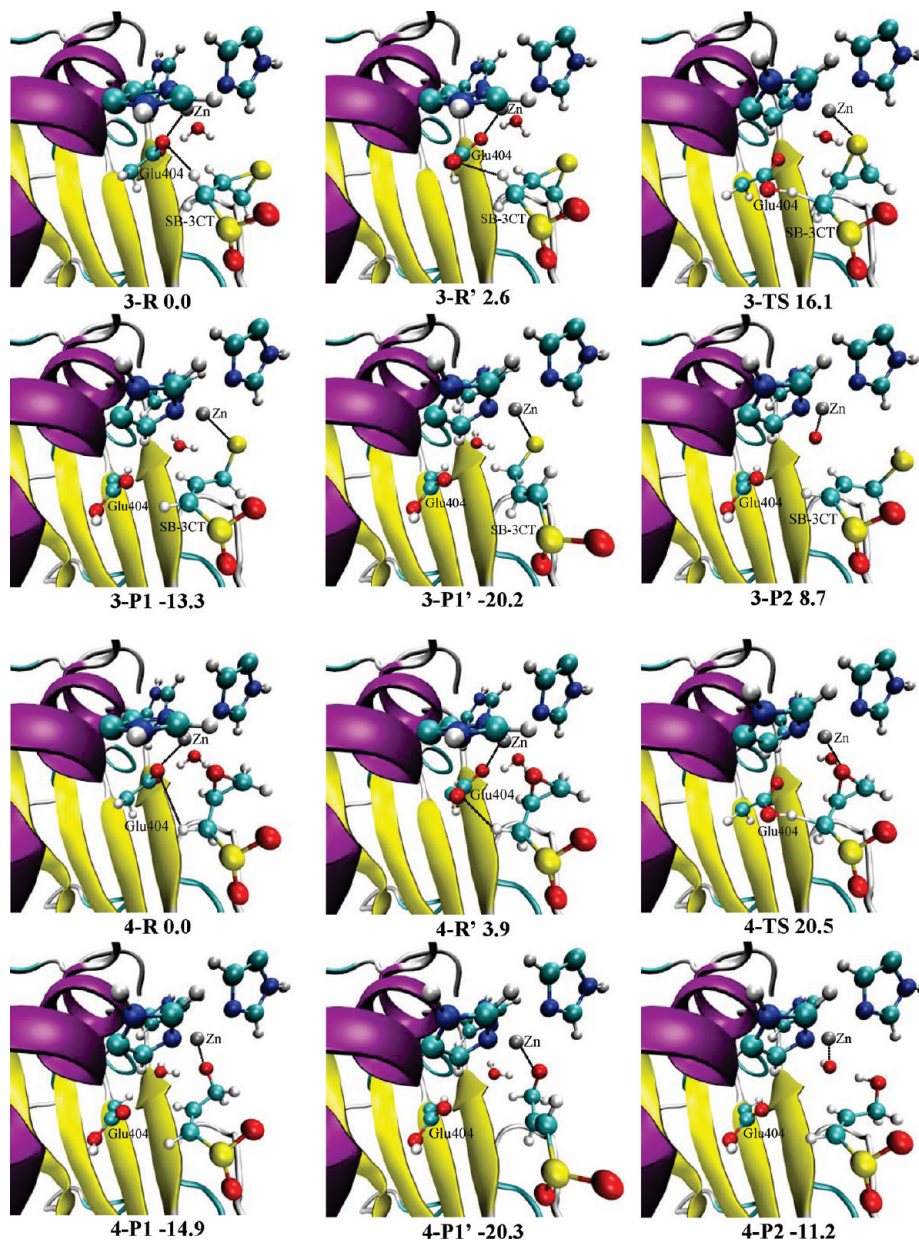


Figure 2. Reactants, transition states, and products for (*S*)-SB-3CT (**3**) and its oxirane analogue (**4**) in the MMP2 active site optimized at the ONIOM(B3LYP/6-31G(d):AMBER) level of theory. Energies (in kcal/mol) were calculated at the ONIOM(B3LYP/6-311+G(d,p):AMBER) level using electronic embedding with the reactant complexes used as the reference states. **3-P1** and **4-P1** are the cis isomers of the unprotonated ring-opening products. **3-P1'** and **4-P1'** are the trans isomers. In **3-P2** and **4-P2**, the cis ring-opening products are protonated by the water molecule, and the resulting hydroxide anion coordinates with the zinc.

clearly demonstrate that there are no additional barriers between the TSs and the reactants, and TSs and the products, for both MMP2•(*S*)-SB-3CT and its oxirane analogue. The reaction path shows that the *R* enantiomer of SB-3CT and its oxirane analogue open to the trans products. The corresponding reaction paths for the *S* enantiomer of the thiirane and oxirane lead to the cis products, **3-P1** and **4-P1**. The cis products are 6.9 and 5.4 kcal/mol higher than the trans products, **3-P1'** and **4-P1'**, partly because of less favorable interactions with the active site. The different stereochemical outcome for the two enantiomers demonstrates that stereoelectronic control exists in the transition state. Animations of the normal mode corresponding to the imaginary frequencies and the IRC paths for MMP2•(*S*-

SB-3CT and its oxirane analogue are provided in the Supporting Information.

C. Thermodynamics. The energy profiles for (*S*)-SB-3CT (**3**) and its oxirane analogue (**4**) in the MMP2 active site are shown in Figure 4 and summarized in Table 1. The ring-opening barriers for the *S*-thiirane and *S*-oxirane in the active site of MMP2 are 16.1 and 20.5 kcal/mol, respectively. The thiolate ring-opened product from the thiirane (**3-P1**) and the alkoxide ring-opened product from the oxirane (**4-P1**) both show tight coordination with the zinc in these product complexes. The protonated Glu404 moves significantly away from the zinc. The water molecule in both **3-P1** and **4-P1** moves away from the zinc to distances of 3.16 and 3.37 Å,

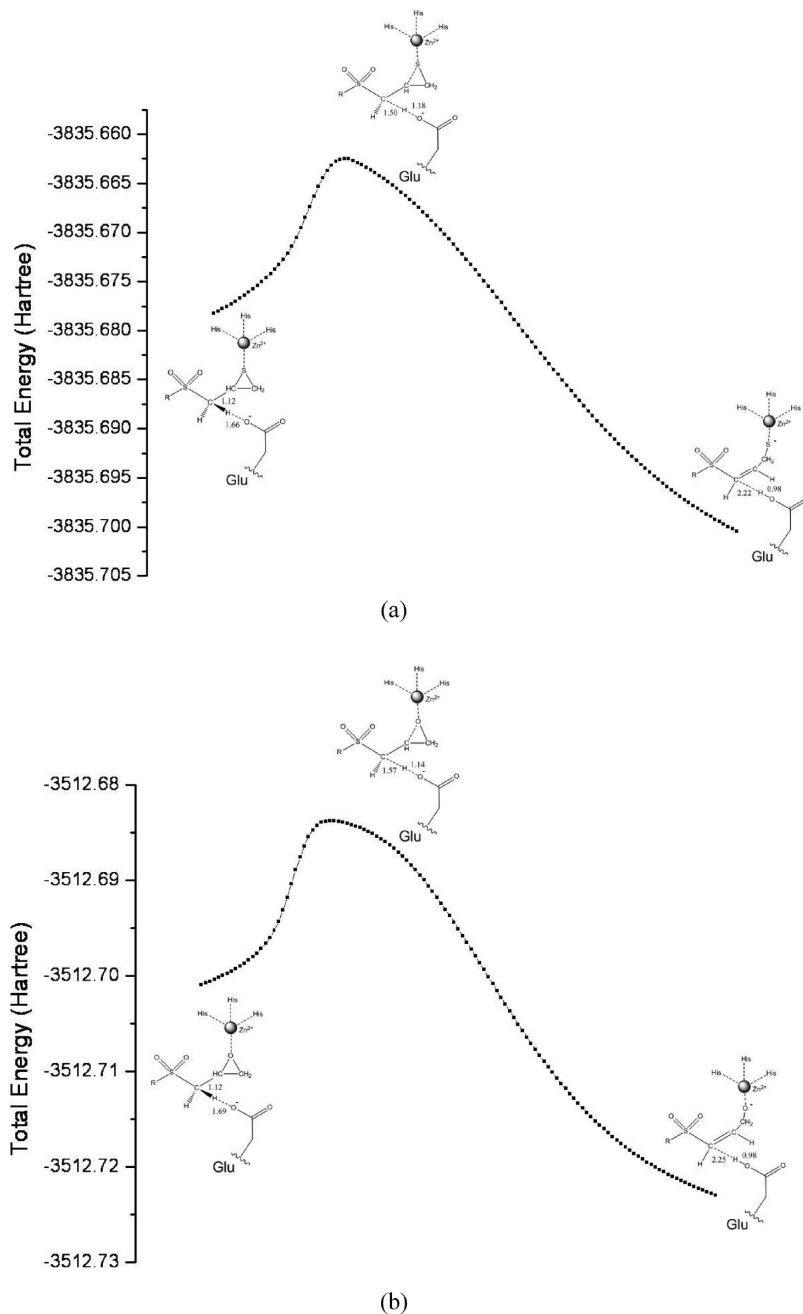


Figure 3. IRC profiles for (*S*)-SB-3CT (a) and its oxirane analogue (b) in the MMP2 active site using the partial model at the ONIOM(B3LYP/6-31G(d):AMBER) level of theory. Key bond lengths are in angstroms.

respectively. The reaction energies for thiirane and oxirane are -13.3 and -14.9 kcal/mol, respectively.

D. Product Protonation by Water. Since solvent has substantial access to the active site of MMP2, the ring-opened products can be protonated by solvent. Proton transfer from a nearby water to **3-P1** and **4-P1** gives structures **3-P2** and **4-P2**, respectively (Figure 2 and Supporting Information Figure S2). Proton transfer from the water molecule generates a hydroxide ion, which coordinates with the zinc, replacing the thiolate and alkoxide as zinc ligands. For the thiirane system, the protonated product complex **3-P2** is 8.7 kcal/mol endothermic compared to the reactant complex **3-R**. For the oxirane system, by contrast, the protonated product complex **4-P2** is 11.2 kcal/mol exothermic compared to the reactant complex **4-R**. A similar result was observed in the

study of the *R* enantiomer.³² This difference between the thiirane and oxirane may explain why both the *R* and *S* enantiomers of SB-3CT are slow-binding inhibitors of MMP2, while their oxirane analogues are linear competitive inhibitors.³⁰

E. Kinetic Isotope Effect (KIE) Calculations Using the QM/MM Model. In our earlier studies, the KIE calculation for the *R*-thiirane could only be carried out for the solution model.³² However, recent improvements in computer code and resources have made it feasible to calculate the full set of frequencies for QM/MM systems. Using the frequency calculations for the 2747 atom QM/MM models, deuterium KIEs were calculated for both the MMP2•(*R*)-SB-3CT and MMP2•(*S*)-SB-3CT complexes. The Wigner tunneling correction⁵⁰ contributes a factor of

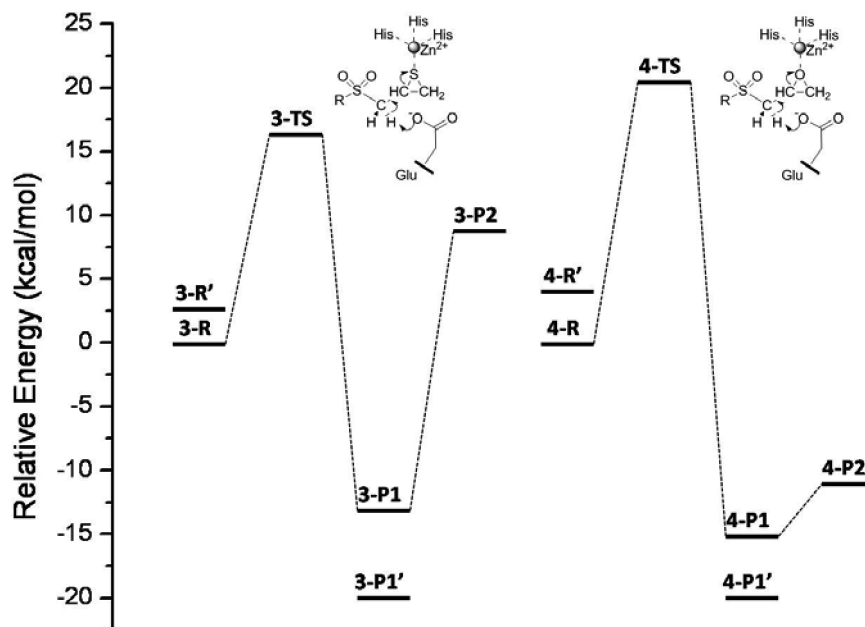


Figure 4. Energy profiles for (*S*)-SB-3CT (**3**) and its oxirane analogue (**4**) in the MMP2 active site. Relative energies (in kcal/mol) were calculated at the ONIOM(B3LYP/6-311+G(d,p):AMBER) level using electronic embedding with the reactant complexes used as reference states.

Table 1. QM/MM Calculations of the Energetics for the Ring-Opening Reactions of Inhibitions in the Active Site of MMP2^a

inhibitor	barrier height	reaction enthalpy		
		P1 (unprotonated cis product)	P1' (unprotonated trans product)	P2 (protonated cis product)
(<i>S</i>)-SB-3CT (3) ^b	16.1	-13.3	-20.2	8.7
(<i>S</i>)-oxirane analogue (4) ^b	20.5	-14.9	-20.3	-11.2

^a ONIOM(B3LYP/6-311+G(d,p):AMBER)//ONIOM(B3LYP/6-31G(d):AMBER) with electronic embedding; energies in kcal/mol. ^b See Figure 2.

1.15 for *R*-thiirane and 1.27 for *S*-thiirane. The calculated KIE (k_H/k_D) is 5.9 for the inhibition of MMP2 by (*R*)-SB-3CT and 6.7 for (*S*)-SB-3CT. These numbers agree well with the experimental result for the racemic mixture ($k_H/k_D = 5.0$).³¹

Conclusions

The inhibition of MMP2 by (*S*)-SB-3CT (**3**) and its oxirane analogue (**4**) involves the concurrent deprotonation of the inhibitor by the active-site glutamate, opening of the respective three-membered heterocycle, and coordination of the heteroatom anion product to the active-site zinc. For the present QM/MM calculations, the barrier for this ring-opening reaction of (*S*)-SB-3CT in the MMP2 active site is 16.1 kcal/mol, 4.4 kcal/mol lower than the barrier for the oxirane analogue. Abstraction of a proton from the inhibitor by glutamate is the key event in the inhibition reaction, as indicated by the kinetic isotope effects, and is directly coupled with the ring-opening. In the transition state, the heteroatom of the three-membered ring moves closer to the zinc, facilitating completion of the deprotonation and ring-opening events for progress toward the product complexes. The reaction enthalpies are quite similar (-13.3 kcal/mol for **3** and -14.9 kcal/mol for **4**). Reaction path following calculations show that ring-opening of (*S*)-SB-3CT by MMP2 yields the cis product, while ring-opening of (*R*)-SB-3CT in

the MMP2 produces the trans product. The calculations show that the protonation of the alkoxide product from the ring-opening of **4** by a water molecule in the active site is exothermic, whereas protonation of the thiolate from ring-opening of **3** is endothermic. The calculated KIEs of the reaction agree well with the experimental results, increasing the confidence in the present study. The previous and present studies provide the solid theoretical support for the inhibition mechanism of MMP2.

Acknowledgment. This work is supported at Wayne State University by the National Science Foundation (CHE0910858) and at the University of Notre Dame by the National Institute of Health (CA122417). Computer time on Wayne State University computer grid is gratefully acknowledged.

Supporting Information Available: Superposition of the reactant and transition state for **3** and **4**, QM/MM geometries for the ring-opening reaction of **3** and **4** in the active site of MMP2, IRC profiles for (*R*)-SB-3CT and its oxirane analogue in the MMP2 active site, ONIOM energetics with mechanical embedding, animations of the imaginary frequency modes, and intrinsic reaction paths using the partial QM/MM model for **3-TS** and **4-TS**. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Lei, H.; Furth, E. E.; Kalluri, R.; Wakenell, P.; Kallen, C. B.; Jeffrey, J. J.; Leboy, P. S.; Strauss, J. F., III. *Biol. Reprod.* **1999**, *60*, 183–189.
- (2) Liu, C.-H.; Wu, P.-S. *Biotechnol. Lett.* **2006**, *28*, 1725–1730.
- (3) Shi, Y.-B.; Fu, L.; Hasebe, T.; Ishizuya-Oka, A. *Pharmacol. Ther.* **2007**, *116*, 391–400.
- (4) Woessner, J. F., Jr. *FASEB J.* **1991**, *5*, 2145–2154.
- (5) Smith, M. F.; Ricke, W. A.; Bakke, L. J.; Dow, M. P.; Smith, G. W. *Mol. Cell. Endocrinol.* **2002**, *191*, 45–56.
- (6) Homandberg, G. A.; Ummadi, V.; Kang, H. *Inflammation Res.* **2004**, *53*, 534–543.
- (7) Kawasaki, Y.; Xu, Z.-Z.; Wang, X.; Park, J. Y.; Zhuang, Z.-Y.; Tan, P.-H.; Gao, Y.-J.; Roy, K.; Corfas, G.; Lo, E. H.; Ji, R.-R. *Nat. Med.* **2008**, *14*, 331–336.
- (8) Coussens, L. M.; Werb, Z. *Chem. Biol.* **1996**, *3*, 895–904.
- (9) Vihinen, P.; Kähäri, V. M. *Int. J. Cancer* **2002**, *99*, 157–166.
- (10) Egeblad, M.; Werb, Z. *Nat. Rev. Cancer* **2002**, *2*, 161–174.
- (11) Noël, A.; Jost, M.; Maquoi, E. *Semin. Cell Dev. Biol.* **2008**, *19*, 52–60.
- (12) Dollery, C. M.; McEwan, J. R.; Henney, A. M. *Circ. Res.* **1995**, *77*, 863–868.
- (13) Luft, F. C. *J. Mol. Med.* **2004**, *82*, 781–783.
- (14) Janssens, S.; Lijnen, H. R. *Cardiovasc. Res.* **2006**, *69*, 585–594.
- (15) Chow, A. K.; Cena, J.; Schulz, R. *Br. J. Pharmacol.* **2007**, *152*, 189–205.
- (16) Liotta, L. A.; Tryggvason, K.; Garbisa, S.; Robey, P. G.; Abe, S. *Biochemistry* **1981**, *20*, 100–104.
- (17) Morgunova, E.; Tuuttila, A.; Bergmann, U.; Isupov, M.; Lindqvist, Y.; Schneider, G.; Tryggvason, K. *Science* **1999**, *284*, 1667–1670.
- (18) Briknarova, K.; Gehrman, M.; Banyai, L.; Tordai, H.; Patthy, L.; Llinas, M. *J. Biol. Chem.* **2001**, *276*, 27613–27621.
- (19) Feng, Y.; Likos, J. J.; Zhu, L.; Woodward, H.; Munie, G.; McDonald, J. J.; Stevens, A. M.; Howard, C. P.; De Crescenzo, G. A.; Welsch, D.; Shieh, H.-S.; Stallings, W. C. *Biochim. Biophys. Acta, Proteins Proteomics* **2002**, *1598*, 10–23.
- (20) Lukacova, V.; Zhang, Y.; Mackov, M.; Baricic, P.; Raha, S.; Calvo, J.; Balaz, S. *J. Biol. Chem.* **2004**, *279*, 14194–14200.
- (21) Diaz, N.; Suarez, D.; Sordo, T. *J. Phys. Chem. B* **2006**, *110*, 24222–24230.
- (22) Tochowicz, A.; Maskos, K.; Huber, R.; Oltenfreiter, R.; Dive, V.; Yiotakis, A.; Zanda, M.; Bode, W.; Goettig, P. *J. Mol. Biol.* **2007**, *371*, 989–1006.
- (23) Diaz, N.; Suarez, D. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 50–61.
- (24) Ikejiri, M.; Bernardo, M. M.; Bonfil, R. D.; Toth, M.; Chang, M.; Fridman, R.; Mobashery, S. *J. Biol. Chem.* **2005**, *280*, 33992–34002.
- (25) Fisher, J. F.; Mobashery, S. *Cancer Metastasis Rev.* **2006**, *25*, 115–136.
- (26) Khandelwal, A.; Balaz, S. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 131–137.
- (27) Khandelwal, A.; Balaz, S. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 326–339.
- (28) Gupta, S. P. *Chem. Rev.* **2007**, *107*, 3042–3087.
- (29) Zhang, Y.; Lukacova, V.; Bartus, V.; Nie, X.; Sun, G.; Manivannan, E.; Ghorpade, S.; Jin, X.; Manyem, S.; Sibi, M.; Cook, G.; Balaz, S. *Chem. Biol. Drug. Des.* **2008**, *72*, 237–248.
- (30) Brown, S.; Bernardo, M. M.; Li, Z.-H.; Kotra, L. P.; Tanaka, Y.; Fridman, R.; Mobashery, S. *J. Am. Chem. Soc.* **2000**, *122*, 6799–6800.
- (31) Forbes, C.; Shi, Q. C.; Fisher, J. F.; Lee, M.; Heseck, D.; Llarrull, L. I.; Toth, M.; Gossing, M.; Fridman, R.; Mobashery, S. *Chem. Biol. Drug. Des.* **2009**, *74*, 527–534.
- (32) Tao, P.; Fisher, J. F.; Shi, Q. C.; Vroven, T.; Mobashery, S.; Schlegel, H. B. *Biochemistry* **2009**, *48*, 9839–9847.
- (33) Lee, M.; Bernardo, M. M.; Meroueh, S. O.; Brown, S.; Fridman, R.; Mobashery, S. *Org. Lett.* **2005**, *7*, 4463–4465.
- (34) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.
- (35) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- (36) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.
- (37) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct.—THEOCHEM* **1999**, *461–462*, 1–21.
- (38) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419–1432.
- (39) Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760–769.
- (40) Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701–714.
- (41) Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.
- (42) 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.
- (43) Lundberg, M.; Kawatsu, T.; Vreven, T.; Frisch, M. J.; Morokuma, K. *J. Chem. Theory Comput.* **2009**, *5*, 222–234.
- (44) Meroueh, S. O.; Fisher, J. F.; Schlegel, H. B.; Mobashery, S. *J. Am. Chem. Soc.* **2005**, *127*, 15397–15407.
- (45) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (46) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.
- (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.; Scalmani, G.; Mennucci, B.; Barone, V.; Petersson, G. A.; Caricato, M.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Kudin, K. N.; Heyd, J. J.; Brothers, E.; Staroverov, V.; Zheng, G.; Kobayashi, R.; Normand, J.; Sonnenberg, J. L.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo,

J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; 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.; Chal-

lacombe, M.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian, Revision F.02 ed.*; Gaussian, Inc.: Wallingford, CT, 2007.

- (48) Tao, P.; Schlegel, H. B. *J. Comput. Chem.* **2010**, *31*, 2363–2369.
- (49) Tao, P.; Fisher, J. F.; Shi, Q.; Mobashery, S.; Schlegel, H. B. *J. Phys. Chem. B* **2010**, *114*, 1030–1037.
- (50) Wigner, E. *Z. Phys. Chem.* **1932**, *B19*, 203–216.

CT100382K

Simulation of Coarse-Grained Protein–Protein Interactions with Graphics Processing Units

Ian Tunbridge,[†] Robert B. Best,[‡] James Gain,[†] and Michelle M. Kuttel^{*,†}

*Department of Computer Science, University of Cape Town, Cape Town, South Africa
and Department of Chemistry, Cambridge University, Cambridge, United Kingdom*

Received July 12, 2010

Abstract: We report a hybrid parallel central and graphics processing units (CPU–GPU) implementation of a coarse-grained model for replica exchange Monte Carlo (REMC) simulations of protein assemblies. We describe the design, optimization, validation, and benchmarking of our algorithms, particularly the parallelization strategy, which is specific to the requirements of GPU hardware. Performance evaluation of our hybrid implementation shows scaled speedup as compared to a single-core CPU; reference simulations of small 100 residue proteins have a modest speedup of 4, while large simulations with thousands of residues are up to 1400 times faster. Importantly, the combination of coarse-grained models with highly parallel GPU hardware vastly increases the length- and time-scales accessible for protein simulation, making it possible to simulate much larger systems of interacting proteins than have previously been attempted. As a first step toward the simulation of the assembly of an entire viral capsid, we have demonstrated that the chosen coarse-grained model, together with REMC sampling, is capable of identifying the correctly bound structure, for a pair of fragments from the human hepatitis B virus capsid. Our parallel solution can easily be generalized to other interaction functions and other types of macromolecules and has implications for the parallelization of similar N-body problems that require random access lookups.

1. Introduction

The formation of multiprotein complexes, either transient or permanent, is integral to many biological processes. Some examples are antibody–antigen and protease–inhibitor complexes, protein complexes involved in cellular signal transduction processes, structural proteins that maintain the shape of a biological cell, and the very large multiprotein complexes represented by the proteasome, the nuclear pore complex and viral capsids. Identification of the site and strength of interaction (binding) between proteins involved in common cellular functions is integral to a comprehension of how they work cooperatively. This can improve our grasp of disease mechanisms and provide the basis for new therapeutic approaches. Consequently, the prediction of

protein binding sites has been identified as one of the 10 most sought-after solutions in protein bioinformatics.¹ This problem is closely related to the well-known NP-hard “protein folding problem” of predicting the three-dimensional structure of a protein from its primary sequence.

In the absence of sufficient experimental data on the atomic structure of protein complexes, molecular dynamics (MD) or Monte Carlo (MC) simulations of protein complex components can assist in determining both their mode of interaction and the location of the interaction site(s).² Molecular simulations generate an ensemble of configurations, from which both structural and thermodynamic data can be extracted. The configurations representing bound protein complexes enable identification of both the docking sites and the relative orientation of the proteins, while the binding affinity of a complex can be estimated from the proportion of bound samples occurring in the ensemble. However, all-atom simulations of multiprotein complexes are highly computationally expensive and are therefore limited

* Corresponding author. E-mail: mkuttel@cs.uct.ac.za. Telephone: +27 (0)21 6505107.

[†] University of Cape Town.

[‡] Cambridge University.

in scale by the available computing resources. Simulations are typically restricted to simple biological systems (e.g., small binary protein complexes without solvent) and to nanosecond time scales. Accurate coarse-grained models have helped to extend molecular simulations to more biologically relevant length- and time-scales.^{3–11} These reduced molecular models aggregate single atoms into large spherical beads to significantly decrease the computational cost of a simulation. There is considerable potential for further accelerating molecular simulations by combining coarse-grained models with the computational power of massively parallel graphics processing units (GPUs).

Modern GPUs have floating-point computational capabilities far in excess of current central processing units (CPUs). Essentially, GPUs fall into the category of single program multiple data (SPMD) compute devices; organizing data into homogeneous streams of elements and executing a function, or *kernel*, on all elements of a stream simultaneously. Current high-end GPUs also have high memory bandwidth compared to CPUs. For example, the nVIDIA GTX280 has 240 fragment or stream processors and a theoretical memory bandwidth of 141 GB/s. As a consequence, these compact devices are capable of rapid high-throughput numeric operations and can be employed effectively by nongraphical memory-bound algorithms of high arithmetic intensity, such as the N-body problem inherent in molecular simulations. For coarse-grained potentials, evaluation of the total interaction potential between all beads, N , in a protein molecule is an $O(N^2)$ operation and the chief performance bottleneck, a common feature of N-body simulations in general. The independence of each pairwise interaction means that the calculation of all such potentials suits the SPMD GPU architecture, promising good speedups over CPU-only implementations.

The difficult task of porting algorithms to the GPU architecture, while maintaining effective use of the CPU, has been made easier with the development of general application programming interfaces. In 2007, nVIDIA released the compute unified device architecture (CUDA) API, which allows the general programmer direct access to the nVIDIA GPU hardware. CUDA allows for operations not supported by graphics APIs, such as local data communication between kernels and scatter and gather operations. However, CUDA GPU programming is not trivial. Programmers must be mindful of the GPU memory hierarchy, which requires explicit management to minimize access latency and effective packing of data to enable a coalesced memory access pattern.¹² In addition, maximizing GPU performance often requires latency hiding through exploitation of the multithreading capabilities of the CPU cores,¹³ adding the difficulties of conventional multithreaded asynchronous programming to the GPU-specific programming techniques.

However, despite these difficulties, there are increasing reports of successful CUDA implementations of N-body algorithms achieving good speedups over CPU implementations.^{14–16} Specifically, GPU-based calculations of the expensive long-range electrostatics and other nonbonded forces necessary for molecular mechanics simulations are typically 10–100 times faster than heavily optimized CPU-

based implementations.^{17–19} Friedrichs et al. showed speedups over a single CPU implementation of up to 700 times for large all-atom protein MD running entirely on the GPU.¹⁶ However, such great speedups are not always achievable. A recent implementation of an acceleration engine for the solvent–solvent interaction evaluation of MD simulations shows speed-ups of up to a factor of 54 for the solvent–solvent interaction component but only 6–9 for the simulations as a whole.²⁰

Here we report our hybrid CPU–GPU parallel implementation of a coarse-grained model and an energy function for simulation of multiprotein complexes recently developed by Kim and Hummer¹¹ together with a replica exchange Monte Carlo (REMC) simulation protocol to enhance sampling. We leave the original model and simulation methods unchanged; our primary focus is on development of a general, highly scalable parallel implementation, with the ultimate goal of increasing the size of tractable simulations to the point where far more biologically relevant simulations can be attempted.

Previous implementations of related N-body dynamics on a GPU, such as the GRAPE implementation,¹⁵ translate the potential evaluation into a convenient map-reduce problem.^{14–16} However, in our case this approach is not feasible, as the Kim–Hummer coarse-graining model requires very frequent random-access lookups in evaluation of the interaction potential, and the memory resource limitations of the GPU prevent the use of standard optimal GPU memory-access models for this problem.^{21,22} Therefore, in order to optimize the parallel performance of our implementation, we assessed the performance impact of storing the structural data and the potential lookup table in the various different types of memory available on a GPU to establish the optimal configuration. We validate our final CPU–GPU implementation against the original, demonstrating that our simulations reproduce the reference results of Kim and Hummer,¹¹ while showing a factor of 4 speedup for small systems. Further benchmarking analysis demonstrates that our implementation of this model achieves excellent speedups of over 1400 for large simulations. We demonstrate that the potential is capable of identifying the correct bound structure for a pair of protein fragments from the human hepatitis B (HBV) virus capsid, although there are also minor populations of incorrectly bound structures. This paves the way for larger scale simulation studies of capsid assembly mechanisms.

The parallelization approach developed in this work is generally applicable to N-body problems that require similar random access lookups. This often occurs where the aspects of the interaction between bodies are dependent on their type or state. One instance is the commonly used energy functions in all-atom MD simulations, in which the interactions depend on the type of each atom.²³

2. Methods

In this section, we present details of the coarse-grained protein–protein simulation method and some high-level considerations for its parallelization, validation, profiling, and benchmarking.

2.1. The Kim–Hummer Coarse-Grained Model. In common with many current coarse-grained models, the

Kim–Hummer model (hereafter referred to as the model) represents a protein molecule as a chain of beads corresponding to specific amino acid residue types. Coarse-grain representations are generated from a protein's atomic structure by centering a bead on the C_α atom in each amino acid residue, with the radius of the bead being the van der Waals radius of the specific residue.

In the simplest version of the model, proteins are treated as rigid bodies, with the interaction potential, U_{tot} , comprising only the pairwise sum of short-range amino-acid-dependent Lennard-Jones potentials, $u_{ij}(r)$, and long-range electrostatic Debye–Hückel interactions, $u_{ij}^{\text{el}}(r)$:

$$U_{\text{tot}} = \sum_{ij} f_i f_j (u_{ij}(r) + u_{ij}^{\text{el}}(r)) \quad (1)$$

where r is the distance between residues i and j , and $0 \leq f_i \leq 1$ is the weighting factor for residue i . The weighting factor scales the contribution of a particular residue, allowing residues on the surface of the molecule to contribute more than residues buried within the protein.²⁴ In the simplest case, all interactions are weighted equally ($f_i = 1$ for all i). Since we use rigid bodies to represent proteins, these weights remain constant for each residue throughout a simulation.

Based on known contact potentials, ϵ_{ij} , between residues, short-range interactions are either attractive ($\epsilon_{ij} < 0$) or repulsive ($\epsilon_{ij} > 0$) and are defined as

$$u_{ij}(r) = \begin{cases} 4|\epsilon_{ij}|[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^6], & (\epsilon_{ij} < 0) \\ 4\epsilon_{ij}[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^6] + 2\epsilon_{ij}, & (\epsilon_{ij} > 0, r < r_{ij}^0) \\ -4\epsilon_{ij}[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^6], & (\epsilon_{ij} > 0, r \geq r_{ij}^0) \end{cases} \quad (2)$$

where σ_{ij} is the distance between residues i and j , determined from the average of the respective van der Waals diameters of each residue,¹¹ and $r_{ij}^0 = 2^{1/6}\sigma_{ij}$ is the lowest energy separation.

The long-range electrostatic potential between residues is defined as

$$u_{ij}^{\text{el}} = \frac{q_i q_j \exp\left(-\frac{r}{\xi}\right)}{4\pi D r} \quad (3)$$

where q_i is the charge of residue i , D is the dielectric constant of the solvent, and ξ is the Debye screening length. Kim and Hummer use $D = 80$ as the dielectric constant of water and $\xi \approx 10 \text{ \AA}$, corresponding to physiological salt concentrations.

Note that our implementation currently does not allow for the flexible linker peptides connecting rigid-protein domains in the original model. However, these can be included relatively cheaply, with the contribution of the appropriate stretching, bending, and torsion-angle potentials for the flexible linkers calculated in $O(N)$ time complexity on the CPU.

2.2. Methods for REMC Simulations. In our implementation, we use the REMC protocol reported for the original model.¹¹ MC methods employ a sequence of random geometric mutations (moves) of the system (usually rotations or translations) to explore and sample configurational space in a desired ensemble. In our case, the Metropolis MC

algorithm^{25,26} is most suitable. In Metropolis sampling, the energy of the system is evaluated after each move, and mutations are accepted or rejected in such a way that configurations are sampled with probabilities given by the Boltzmann distribution (i.e., the probability of configuration X , $P(X) \propto \exp(-E(X)/k_B T)$, with k_B being the Boltzmann constant).

In each MC mutation, a randomly selected protein is either translated (by 0.5 \AA) or rotated (by 0.2 radians about a protein's centroid) along a random axis within a periodic bounding box. Varying the volume of the box alters the concentration at which the simulation is performed. For these simulations, a value of less than $2k_B T = -1.186 \text{ kcal/mol}$ for the interaction energy is used to determine that the proteins are in a bound state,² although the results are not very sensitive to this value.¹¹ To determine the binding affinity of a complex, the fraction bound y (the proportion of bound samples out of the total number of samples) is determined as a function of the protein concentration $[A] = 1/V$, and the dissociation constant K_d is evaluated from the relation:¹¹

$$y = \frac{[A]}{[A] + K_d} \quad (4)$$

The sampling properties of MC simulation are further improved by using replica exchange, in which concurrent simulations are run at different temperatures.²⁷ Such a REMC search maintains χ independent replicas of the system, with each replica run at a different temperature value (T_1, T_2, \dots, T_χ). During replica exchange, configurations are swapped between replicas at neighboring temperature values with a probability proportional to their energy and temperature differences. Specifically for a pair of replicas at temperatures T_1 and T_2 , with current coordinates X_1 and X_2 , respectively, configurations are swapped using the criterion $\Delta = \exp[(\beta_1 - \beta_2)(U(X_1) - U(X_2))]$, with $\beta_1 = 1/k_B T_1$, $\beta_2 = 1/k_B T_2$, and $U(X)$ the potential energy. Analogous to standard Metropolis MC, moves are always accepted if $\Delta > 1$, otherwise they are accepted if $\rho < \Delta$, with ρ random numbers on the interval $[0,1)$. This process runs until sufficient samples have been generated for the calculated fraction bound metric, y , to converge, typically requiring between 10^8 and 10^9 MC mutations per replica.

2.3. General Parallelization Approach. In our approach to parallelization of the Kim–Hummer model, we aim to exploit the various levels of concurrency inherent in the REMC algorithm, from fine-grained parallelism in the costly potential evaluation to relatively coarse-grained parallelism in the multiple replicas.

As we only had access to an implementation of the model within the CHARMM package,²⁸ we began with development of a serial implementation with full simulation functionality to provide a reference point from which to perform validation, profiling, and benchmarking of the subsequent parallel GPU implementation. We note that our CPU implementation shows slightly faster run times than the CHARMM version.

At a high level, a parallel version of the $O(R)$ REMC algorithm (where R is the number of replicas) is relatively

straightforward to implement, and the many concurrent independent replicas make this algorithm highly suitable for parallel execution across many CPU cores. From a multithreading perspective, replicas encapsulate parallel work units; each replica performs a self-contained MC simulation. The Markov chain nature of these simulations means that an individual simulation cannot be split across threads, as each step of the MC simulation is dependent on the previous step. We therefore employ a multiple producer–consumer model, with R replica MC simulation threads and one replica exchange thread. Control is initially passed to the MC threads, with the replica exchange thread waiting until the MC threads complete their allotted iterations. At this point, control is passed to the replica exchange thread, and replica exchange is performed. This process is repeated for the specified number of MC steps. This high-level approach scales well to multiple CPU cores, with or without a GPU.

At a lower level, the $O(N^2)$ potential evaluation is by far the most costly operation and is therefore exported to GPU. Although it is considered best to associate thread contexts one-to-one with GPU runtime contexts,²² our model also allows for sharing of one or more GPU's between multiple threads. Such concurrent potential evaluation by multiple replica threads can be handled in two ways: either by sharing the GPU between threads through multiple contexts or by using CUDA streams to perform asynchronous computation on the GPU. Asynchronous calls maximize resource utilization of both the GPU and CPU but require explicit management of the streaming process, whereas the use of pthreads to run concurrent MC simulations implicitly manages processing overlap between the GPU and CPU. For asynchronous GPU calls, each replica is assigned to a stream, and the application is configured to overlap MC mutations and acceptance/rejection sampling with interaction potential calculations. This allows for maximum resource utilization in the application, ensuring that the GPU is busy at all times, thereby maximizing the simulation throughput. For purposes of comparison, we implemented both approaches.

2.4. Simulations. *2.4.1. Benchmarking.* For benchmarking purposes, 20 replicas were simulated for 1000 MC steps each, for system sizes ranging from 100 to 7668 residues. (Note that a typical simulation of 20 replicas will realistically run for 10^7 steps per replica, 10 000 times longer than this benchmark.) The two smallest benchmark cases are the UIM/Ub (100 residues) and Cc/CcP (402 residues) systems also used for validation. For the larger systems, we simulate the interactions of viral capsid proteins from the human hepatitis B virus (HBV, PDB ID 2G33), where each protein dimer comprises 284 residues. We simulate systems consisting of 2 (568 residues) up to 27 (7668 residues) dimers. This system was chosen with the eventual application of virus assembly in mind, although for benchmarking, we are only concerned with the performance aspects.

All benchmarking simulations were run using an nVIDIA GTX280 (Asus ENGTX280) running at a stock clock speed of 600 MHz paired with a Intel Core 2 Duo 3 GHz E8400 CPU and 4GB of DDR2–800 MHz RAM.

2.4.2. Validation. Our implementation was validated in two stages. First, the single point interaction energies

produced by both our single CPU and the GPU implementations were validated against those from the original model implementation for 10 reference structures (Youngchan Kim, personal communication). We note that the serial CHARMM implementation gives identical results to the original model. Second, to verify the MC simulation implementation, we reproduce the binding affinity for simulations of two protein complexes originally used in the development of the model,¹¹ namely: (i) the binding of ubiquitin to the UIM1 domain of the Vps27 protein (abbreviated to UIM/Ub, PDB ID 1Q0W)²⁹ and (ii) the binding of yeast cytochrome c to cytochrome c peroxidase (Cc/CcP, PDB ID 2PCC),³⁰ both at 300K. We reproduce the simulation protocol of the original study, using replica exchange with 20 replicas at various temperatures between 250 and 600 K, exchanged every 1000 MC steps. Calculation of the equilibrium binding affinity of each complex allows for direct comparison with the original results. Simulations were run for 10^7 steps per replica, with a total of 2×10^8 simulation steps for each of 6 concentrations between 100 and 1000 μM . The GROMACS utility `g_cluster`³¹ was used to cluster bound configurations via the linkage algorithm using root-mean-squared deviations (rmsd) as a metric, with a cutoff of 0.1 nm.

2.4.3. Initial Application. As a first step toward the simulation of the assembly of an entire viral capsid, we simulate the interaction of two identical fragments from the human HBV capsid. Each unit comprises 4 chains, or a total of 582 residues, with the structure taken from PDB ID 2G34 (substructure of one fragment). Sampling is done by REMC with 10 replicas at temperatures between 300 and 416 K (temperatures follow a geometric progression), exchanged every 1000 MC steps. Average acceptance ratios are between 16 and 97% for the MC mutations. Simulations were run for 10 million steps per replica for a total of 100 million simulation steps. This produced $\sim 40\,000$ bound instances at 300 K which were clustered according to rmsd, as described above.

3. Software Implementation

Here, we present specifics of the algorithm implementation: the object-based decomposition, the data bundling and transfer strategies, and the kernel implementation.

All code was written in C and C++ using the GNU Scientific Library, the C++ posix thread library, the C++ Standard Template Library and Linux 64-bit operating system. Versions 2.0–2.3 of the CUDA toolkit were used for development.

Our application accepts Protein Data Bank (PDB) format files of atomic protein structures. These files are parsed, and coarse-grained representations of each protein generated.

3.1. Multithreaded REMC Simulations. Simulations are initialized with the number of MC steps, the intervals at which replica exchange is performed, and the sampling interval. MC mutations were implemented with the aid of the GSL Mersenne twister pseudorandom number generator.³² We ensure that the mutation operations preserve the integrity of molecular structures by performing only rigid-body rotations or translations. Uniform random selections of both molecule and mutation type are performed, followed

either by translation by 0.5 Å or by rotation about a molecule centroid by 0.2 radians, along a random axis. Translations are implemented trivially as a vector translation in 3D with single precision, but rotations, given the accumulation of error in simulations of such length, require full double precision. Since the relative error in the displacement of residues from their intended position is approximately 10^{-4} after 1 million rotations, the error accumulated from single-precision rotation would quickly render the data from simulations unusable. Rotations are thus performed by generating a quaternion representation along the axis of rotation and applying it to the residues in the selected protein. Residue positions are cast to double precision for the rotation operation and cast back to single precision afterward.

Our implementation of replica exchange is extremely lightweight, with a negligible contribution to simulation time. Each replica contains counters for various metrics, such as the fraction bound and the MC acceptance/rejection ratio, which are in turn written at the sampling interval to simulation output files. Because we perform rigid-body docking, structural information is stored as the rotation and displacement of each molecule relative to its initial state.

3.2. Data Representation. We employed a hierarchy of classes to encapsulate the simulation. At the top of the hierarchy, a replica object encapsulates the data specific to a replica instance, such as pointers to the specific arrays in GPU memory, where the replica residue data is stored for the duration of a simulation. The MC simulations are encapsulated within each replica. To reduce transfer times, after any mutation, only the altered molecule is updated on the GPU before invocation of the GPU kernel.

Each residue bead is represented as a data tuple comprising the position (relative and absolute) of each residue, the residue type, the van der Waals radius, and the electrostatic charge. Proteins consist of a contiguous array of residues, which ensures that as many residues as possible occur in the same cache line, consequently improving the rate at which residues are loaded into cache memory on the CPU for mutation.

The short-range contact potential, ε_{ij} , values for the 210 unique residue interaction pairs are stored in a lookup table. Both the residue array and this look-up table are transferred once to the GPU, where they persist for the duration of the simulation. The residue data is stored in the GPU global memory in arrays of type float4 and float3 to ensure coalesced reads when kernels access residue data.²¹ The look-up table can be stored in several ways: as a noncached global array, a cached array using constant memory, or a texture mapped to global memory. We determined the best among these storage locations via careful profiling of the algorithm, as discussed above.

3.3. GPU Kernel Implementation. The kernel is the fragment of code executed by every GPU thread. Our kernel calculates the pairwise interaction potential (algorithm 1) which, using the implicit addressing model provided by the GPU, is applied to residues according to their position in global memory, as determined by thread and block indexes and by block dimension. Residue pairs are allocated across a grid of thread blocks, partitioning the work done by the

GPU into tiles containing subsets of the pairwise interactions. We use a GPU-optimized parallel reduction³³ to compute a partial sum of the interaction potential for the tile assigned to each thread block. As the matrix is symmetric, only thread blocks above the major diagonal contribute to the interaction potential, and the partial sum is halved for thread blocks on the diagonal. Partial sums are written back to global memory on the CPU host, for addition to the total interaction potential, U_{tot} (eq 1). We found no benefit to performing this accumulation on the GPU for larger simulations, and this shift of work to the CPU speeds up the single GPU/multiple pthread configuration.

Algorithm 1 Residue Pairwise Interaction

```

r ← eps + || positioni − positionj ||
eij ← λ(LJ(typei, typej) − e0)
σij = ½(radiusi + radiusj)
uij ← −4 · eij · (σij/r)6 ((σij/r)6 − 1)
if eij > 0 and r < √2 · σij then
    uij ← −uij + 2eij
end if
uijel ← chargei · chargej · e−r/ξ / r
Uij = uij + uijel

```

Our default implementation stores the data for residues assigned to a thread block in the shared memory for the associated symmetric multiprocessor (SM) in a coalesced manner, as in astronomical N-body CUDA implementations.^{14,15} However, as we require more shared memory than previous implementations, we experience lower occupancy on each processor because all threads assigned to a processor jointly access its 16 KB of shared memory. An alternative is the use of texture or global memory for these residues.

Further, the short-range interaction potential requires random access retrieval from a lookup table of *LJ*(*type*_{*i*}, *type*_{*j*}), the van der Waals contact potential for a pair of residues. Unfortunately, this table consists of 210 distinct values, which may be required in any order, which does not fit the ideal data model for a GPU.²² Texture memory tolerates random memory accesses better than other types of memory on the GPU,^{21,22} providing us with an alternate mechanism for implementing the lookup table. However, the latency hiding effects of multiple threads and caching afforded by constant memory do not necessarily make this the best choice. The impact of the type of GPU memory used for both the residue data and the ε_{ij} (eq 2) contact potential lookup table on application performance is discussed below.

4. Performance Tuning

We followed a staged approach to tuning our application for optimal performance. First, kernel parameters were adjusted to ensure that the kernel execution time is as fast as possible. As the CUDA kernel is an indivisible unit of work, its optimization is independent of multithreading and

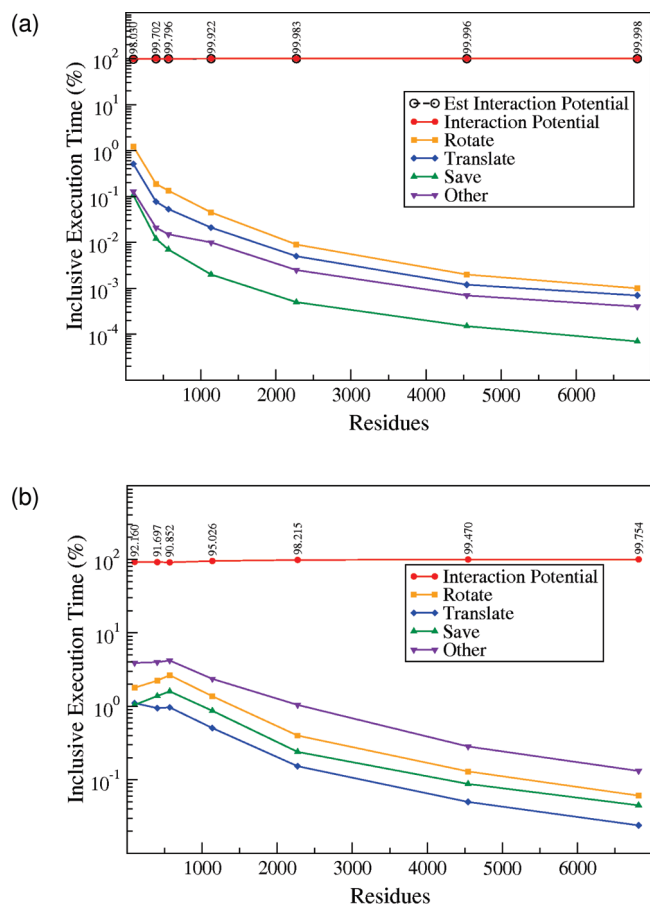


Figure 1. (a) Serial CPU and (b) parallel GPU profiles of the percentage execution time consumed by various operations (logarithmic scale). The interaction potential calculation accounts for over 97% of the simulation time, approaching almost 100% for large systems. Analytical estimates for the expected proportion of time devoted to interaction potential (Est.) are close to the observed values.

asynchronous calls. (Note, however, that the most recently released CUDA-based compute devices can launch different kernels concurrently.) Once an optimal kernel configuration was found, multithreading was used to divide work between CPU cores and overlap CPU and GPU computation. Streaming was undertaken as the last stage of application tuning.

As expected, profiles of a single-threaded CPU implementation for each of the benchmark cases (Figure 1a) demonstrate that the $O(N^2)$ evaluation of the interaction potential consumes the vast majority of simulation time; rotation, translation, and other operations in the MC simulation are relatively inexpensive in comparison. For this reason there is no real benefit in improving the runtime of these operations.

For the kernel optimization, the balance of random access lookup, occupancy, shared memory allocation and the block size provide a myriad of options. We found best performance in most cases for a kernel comprising a thread block size of 64 threads, texture memory for constant potential lookups, and a shared memory model mimicking the GRAPE hardware model.¹⁵ We therefore used this configuration to profile our GPU implementation.

Exporting the interaction potential evaluation to the GPU leads to a massive reduction in execution time, in the range

of one to three orders of magnitude improvement. Predictably, larger problem sizes benefit most, due to the greater degree of parallelism in their computation. However, interaction potential calculations still require 90%+ of a simulation's runtime (see Figure 1). Note that the GPU implementation requires that data be copied to GPU memory before and after kernel invocation, which accounts for the increased "other" operations profile on the GPU as compared to the CPU.

Instruction-level optimization yielded little improvement in our kernels due the dependence of the kernel on lookup table access. Reordering of the pairwise interaction formulas (eqs 2 and 3, as listed in algorithm 1) results in a few key optimizations. For one, $(\sigma_{ij}/r)^6$ can be calculated once and stored in a local variable, allowing the powf function to be called once, as opposed to six times in the original equation (eq 2). Algorithm 1 also minimizes the divergence in the calculation by ensuring the branching required to calculate u_{ij} only contains two multiplies and one add operation. Further, we use the intrinsic GPU functions expf and rsqrtf to increase instruction throughput. The drawback of this type of reordering is that more registers are required to store temporary values, but since it is shared memory rather than register memory that limits our occupancy, this does not negatively impact performance.

4.1. Kernel Tuning. The optimal kernel configuration for a GPU application is dependent on many factors, ranging from high (thread block size and the GPU grid size) to very low (use of specific memory types and instruction-level code optimization) levels.

We load residues within a kernel in the conventional N-body manner,¹⁴ using an array of shared memory within each thread block to store the residues so that they can be accessed quickly. However, we require seven 32-bit values per residue, three more than required for gravitational N-body simulations, which only store position and mass in shared memory.^{14,15} In addition, a patch of shared memory is required for the parallel sum at the end of each thread block to produce an overall interaction potential value for all of the electrostatic interactions of the block. In the gravitational case, each thread computes and stores the updated velocity and position of a single body resulting in a much simpler kernel.

This increased shared memory requirement proves to be the limiting factor on the kernel performance, as the occupancy of the SMs on the GPU is limited by the amount of shared memory available to each thread block. This is illustrated in Figure 2a. For this configuration, occupancy on the GPU is constrained to 37.5% in the cases of 64 or 128 threads per block and 25% occupancy in the cases of 32 or 256 threads per block. A thread block of 512 threads will not execute due to insufficient memory. For smaller problem sizes, where very few thread blocks are required, the amount of padding can have a negative impact on the simulation. For example, 32 threads per block results in the least padding and thread blocks to execute the kernel, making it fastest for small problems. Conversely, 256 threads per block requires the most padding and results in times much longer than the other configurations. Predictably, performance

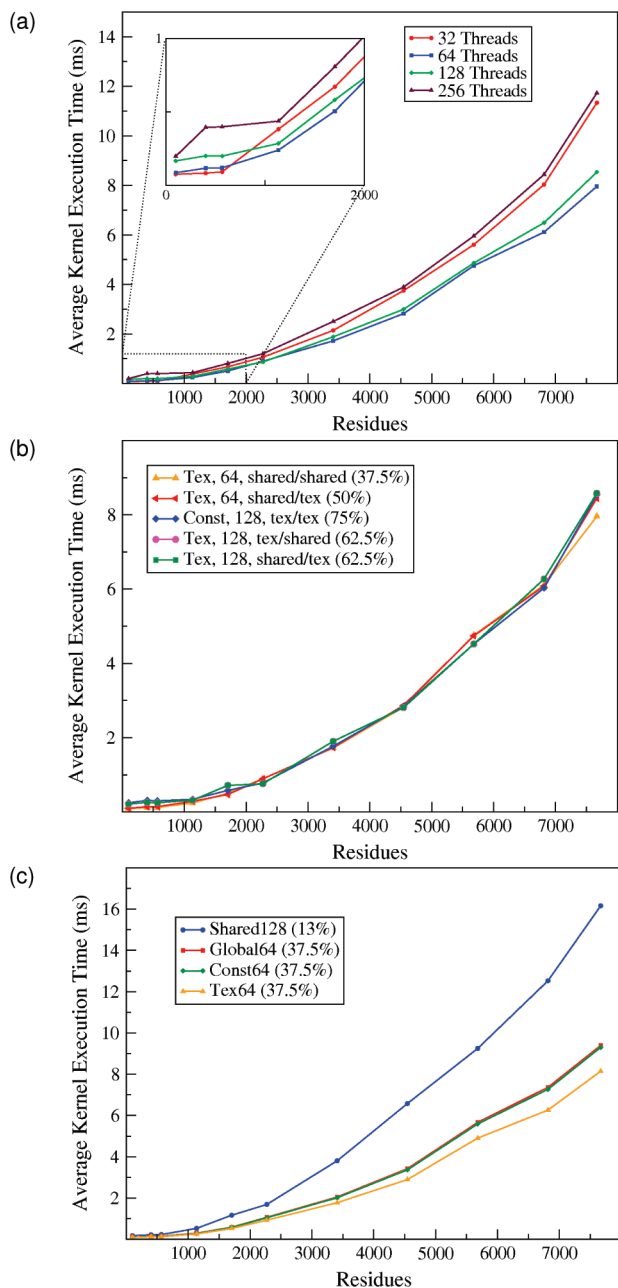


Figure 2. (a) Kernel execution time with varying number of threads per block, using texture memory for the contact potential lookups. Configurations of 32, 64, 128, and 256 threads per block execute in times indicative of their percentage occupancy. (b) Relative performance of alternate memory locations for storing the residue data. Each entry in the key lists the location for contact potential lookups storage, the number of threads per block, the storage location for the residue data (molecule identifier and position vs rest of data), with kernel occupancy values in brackets, e.g., Tex, 64, shared/shared refers to texture memory for contact potential lookups, 64 threads per block, and shared memory for all residue data (shared/shared). (c) Relative performance of various GPU memory locations (global, constant or shared) for storing the potential lookup table.

for larger residues is linked to occupancy in this configuration, as illustrated in Figure 2a. With occupancies of 37.5%, 64 and 128 threads per block perform best, outperforming 32 and 256 threads per block, which can only achieve 25%

occupancy. Thread blocks of 64 threads run the fastest, as they strike the best balance between occupancy, warps per block and blocks per SM. This is because, ultimately, the random access of the contact potentials is a bottleneck on the GPU, and the fastest configuration is the one that best hides the cost of serializing this operation. For this reason, it is important to maximize kernel occupancy.

We experimented with the use of the different GPU memory types—texture, global, constant, or shared—for storing the contact potential lookup tables on the GPU. Figure 2b shows the best benchmark performance for each type of memory. As expected, the poorest performance is experienced using shared memory, which is already limited in availability. Using shared memory for the lookup table as well as the residue data results in an even lower occupancy—6% for 32 and 64 threads and 12.5% for 128 threads, with a proportional degradation of performance. However, if there were sufficient shared memory on each SM to fit many blocks, each containing a constant potential lookup table, then we would expect performance to improve due to increased occupancy and faster memory access, even though this does not accord with the prescribed memory access model.²² The performance of constant and global memory is interesting. Since they perform almost identically, it is clear that the diversity of lookup values prevents the single value caching afforded by constant memory from having any impact. This is also the reason why texture performs best; because of the spacial caching feature of texture memory, any read from texture memory results in the entire table being cached on the GPU for 3 SMs. This results in faster lookup times for all thread blocks running on the SMs which belong to the texture unit. Both global and constant memory performance is generally 10–20% slower than its texture memory lookup equivalent because of this caching effect.

Unexpectedly, some alternative configurations perform almost as well as the best configuration (using texture memory for contact potential lookups and shared memory for prefetching residues for fast access in thread blocks, with 64 threads per block). As shown in Figure 2c, performance may be improved by increasing the occupancy, using texture instead of shared memory for the residue data, in part or entirely. For example, storing only the molecule identifier and position in shared memory and the rest of the residue data in texture memory results in a negligible drop in kernel performance. In this case, the higher access latency of texture memory is offset by the benefit of higher occupancy (50–75%). A possible advantage of this configuration is that large simulations (which we are aiming at) of up to 131 072 residues can be performed using a thread block size of 512 (previously prohibited because of shared memory constraints allowing a maximum simulation size of only 65 536 residues). In the case of constant memory, this strategy improves the performance of all configurations, but not to the extent that they are faster than the original configuration. As another example, using constant memory for lookups performs surprisingly well if all residue data is moved to texture memory (Figure 2b, red line). With a block size of 128 threads per block, this increases the SMs occupancy to 75%,

Table 1. Conformation Energies^a

Conf.	CHARMM	CPU	GPU	η
1	-0.294085	-0.293705	-0.293705	2.03×10^{-7}
2	-1.056417	-1.056291	-1.056291	2.26×10^{-7}
3	-10.278304	-10.277435	-10.277431	4.64×10^{-7}
4	-7.584171	-7.580382	-7.580391	1.20×10^{-6}
5	-0.000079	-0.000079	-0.000079	3.67×10^{-7}
6	-5.564564	-5.562238	-5.562239	2.57×10^{-7}
7	-5.452568	-5.480216	-5.480217	2.61×10^{-7}
8	-10.670303	-10.711964	-10.711967	2.67×10^{-7}
9	-9.904111	-9.900359	-9.900359	0.00
10	-8.518124	-8.527744	-8.527749	5.59×10^{-7}

^aComparison of U_{tot} values (kcal/mol) for 10 reference conformations in the CHARMM implementation (equivalent to the original implementation)¹¹ and our serial CPU and GPU implementations. Relative errors (η) are on the order of 10^{-7} , with an average GPU error of 3.8×10^{-7} .

almost matching our fastest kernel and affording the best latency hiding of any configuration.

Generally we observe that block sizes of 32 threads per block perform best for simulations of fewer than 1000 residues, and for larger simulations a block size of 64 threads is optimal. As a result, we incorporated a simple autotuning feature to achieve better kernel occupancy, simply selecting a dynamic block size on-the-fly, based on prior kernel performance. We limit autotuning to block sizes of 32, 64, 128, 256, and 512 threads to satisfy the reduction algorithm.³³ We find that the relative differences between our kernels is minor, provided that a sensible kernel configuration is selected (Figure 2b and c).

5. Results and Discussion

5.1. Accuracy of the GPU Implementation. Of particular importance for a successful CUDA implementation of an algorithm is the accuracy of the mathematical functions employed by the GPU kernel. A limitation of the GTX280 architecture is that a double precision arithmetic is performed at one-eighth of the speed of a single precision (30 double vs 240 single precision units). Although the latest nVIDIA architectures have better double precision performance, single precision is still significantly faster. On the Fermi GF100, double precision is performed at half the rate of single precision with 256 FMAs per clock vs 512 single precision FMAs.³⁴ Therefore the use of single precision arithmetic is still required to achieve the best performance on GPU architectures.

However, care must be taken to ensure that the use of single precision arithmetic does not result in an unacceptable reduction in computational accuracy. We find a mean relative error of 0.00146 for interaction potential calculations in our GPU implementation and CHARMM, almost identical to the mean relative error between our CPU implementation and CHARMM (0.00143). The largest differences occur in the calculation of the van der Waals component of the interaction potential. The mean relative error between our double precision CPU and single precision GPU implementations is 3.8×10^{-7} (Table 1). In cases where the GPU values differ most from the CPU, use of the powf function in calculating $(\sigma_{ij}/r)^6$ proved to be the cause, due to its maximum ULP error of 8.²¹ A benefit of our GPU kernel is

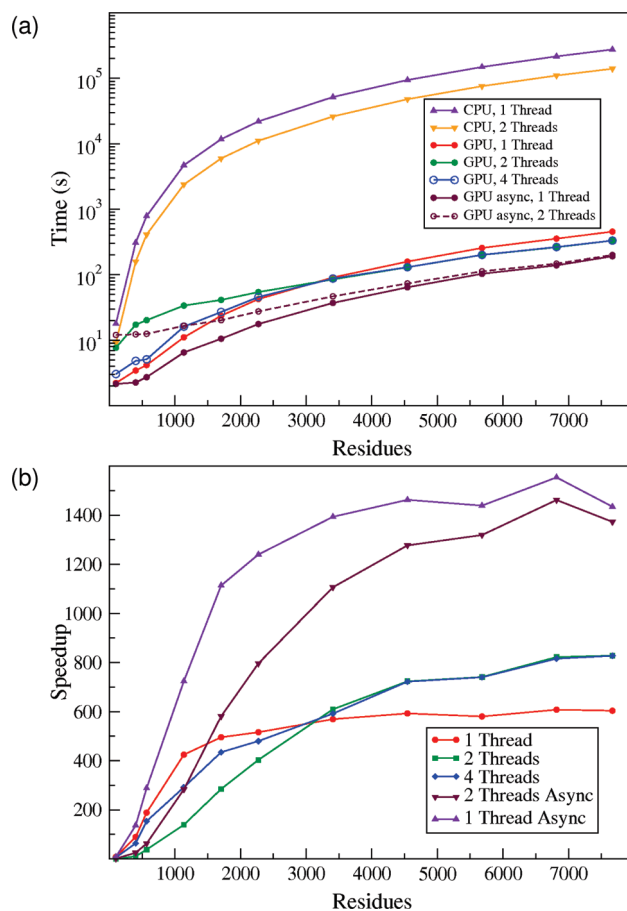


Figure 3. (a) Simulation times for various configurations of GPU and CPU. The GPU simulations generally outperform the CPU simulations by two orders of magnitude from as little as 500 residues. Asynchronous GPU usage proves to be fastest in all cases, provided the GPU is not shared between contexts. The overhead of context switching is evident in the constant difference of 10 s between the multithreaded and serial asynchronous benchmarks. (b) Speedup of the GPU implementation over our serial implementation. The use of CUDA streams results in the best speedup of our implementation, with the serial asynchronous GPU solution performing up to 1400 times faster than the serial CPU solution for larger benchmarks. A clear crossover point occurs for the synchronous multithreaded benchmarks at approximately 3000 residues; for larger simulations, the cost of context switches is amortized by the thread level parallelism of the simulation.

that pairwise summation is implemented implicitly in the reduction used to calculate the total interaction potential from each pairwise potential. This results in the error rate due to round-off errors growing at a low rate proportional to $O(\epsilon\sqrt{\log n})$, as compared to $O(n)$ for a simple summation.

5.2. Performance of the GPU Implementation. We benchmarked simulation times for various options ranging from a thread on the CPU to multiple threads in combination with asynchronous calls to the GPU kernel. Multithreading across CPU cores results in performance improvement proportional to the number of cores (Figure 3 a) on a dual core machine, two threads perform twice as fast as a single thread, illustrating the high scalability of parallel MC simulations. The benefits of the GPU implementation are clear. Even a straightforward, single-threaded, single GPU

configuration shows a remarkable peak speedup of 600 times that of the serial CPU solution (Figure 3b, red line). For large problems (>3000 residues), sharing the GPU among multiple CPU threads provides an even better performance of up to 800 times that of the CPU (Figure 3b, green line). This is because multiple threads ensure that the GPU is better utilized, even though the card has to perform costly CUDA context switching between threads. It is clear that the benefits of using a GPU are only fully realized when simulating large enough problems. For example, our smallest benchmark shows a speedup factor of 8 for a single thread and actually performs 30% slower for two threads.

However, the best performance was obtained not with multithreading but with asynchronous calls to the GPU with CUDA streams. For streaming with one thread of execution, we found that any configuration of 2, 4, 5, or 10 streams performed equally well. Specifically, a configuration comprising a single thread with asynchronous GPU calls (Figure 3b, purple line) performs best across all benchmarks, over 1400 times for systems larger than 4000 residues. The relatively poor performance of multithreaded configurations is partly attributable to the cost of context switches between threads, a fact highlighted by the constant 10 s performance difference between the streaming and multithreaded benchmarks for all system sizes (Figure 3a). Indeed, for smaller simulations, where the relative penalty of performing a context switch between threads is greater, a single-thread configuration (Figure 3b, red and purple lines) always outperforms the equivalent multithreaded code. For multithreaded asynchronous configurations (Figure 3b, maroon line), the cost of context switching is largely hidden by the overlap in GPU and CPU computation. Remaining differences in performance are governed by the kernel efficiency of a particular implementation. For example, the “kinks” in the speedup curves are due to the scheduling of kernels on the GPU, occurring for various block dimensions at differing simulation sizes. Although seemingly large in Figure 3b, the speedup graph merely exaggerates the effect of the variations in Figure 2b.

We conclude that the GPU is fully utilized if streams are used for relatively small problem sizes. However, if streams are unavailable, a configuration with as many CPU threads as cores is likely to provide optimal performance. Macromolecular simulations performed using this system run successfully on a cluster of 1.0 T C870 (G80) cards, where asynchronous CUDA calls are not supported and each card only possesses 128 CUDA cores, with half the amount of shared memory per SM of the GTX280.

5.2.1. Comparison with other CUDA Implementations.

The performance of our GPU kernel is consistent with that achieved by Friedrichs et al.¹⁶ for similar molecular potential evaluations on the same GTX280 architecture (Figure 4). The potential evaluation components of our REMC MC simulations and their CUDA implementation of MD map similarly to the GPU. However, there are some key differences in the MC algorithm that makes it more computationally expensive and accounts for the better performance of the MD code. Our kernel is required to perform a $\log_2 n$ reduction of the n^2 pairwise potential to a single energy value

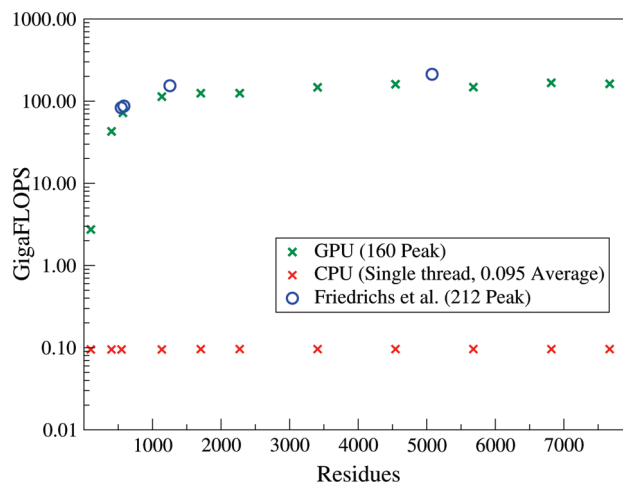


Figure 4. Calculation of the interaction potential using a heterogeneous GPU–CPU calculation achieves peak performance of 160 GFLOPS. A comparable MD simulation by Friedrichs et al. achieves peak performance of 212 GFLOPS. Branching and reduction operations required to generate our potential result in lower overall throughput than the MD simulations, where integration of each atom’s velocity and position ports more amenable to the GPU architecture.

(which is used to either accept or reject the MC move). Reduction requires synchronization before each of its $\log_2 n$ iterations and performs one arithmetic operation to two loads and one store, which impacts negatively on the kernel performance relative to the MD code. Reduction operations are memory bound with low algorithmic intensity and therefore have relatively poor gigaFLOP performance on the GPU architecture.³³ Conversely, the MD kernel includes an $O(n)$ integration step after the pairwise accumulation, the higher the algorithmic intensity, separability of these force calculations and lack of synchronization is better suited to the GPU architecture and results in slightly better GFLOP performance for this kernel.

Indeed, it has been noted that, while MD simulation techniques can be fully implemented on the GPU with relative ease, this does not hold for MC methods.³⁵ Many particle MC simulations are difficult to parallelize, both on conventional parallel architectures and on SIMD hardware, because the random acceptance moves cause unpredictable branching and the requirement for global synchronization.³⁶ We do not implement the MC mutation in the kernel but leave it to the CPU, for reasons of higher double precision accuracy in the geometric transformations. This also enables concurrent execution on both GPU and CPU for multiple replicas, achieving higher overall performance for multiple replicas as opposed to single replica performance.

In addition, the LJ potential employed coarse-grained model (eq 2) is more complex than that used by Friedrichs et al. It is either attractive or repulsive and thus dependent not only on the distance between residues but also on the sign of the contact potential. This additional branching conditional makes the LJ computation more costly to compute on a GPU. We also do not employ a cut-off radius for nonbonded interactions, as is the case for the MD code (on the GPU, a branch in which no computation occurs is effectively free), our kernel always has to calculate the LJ

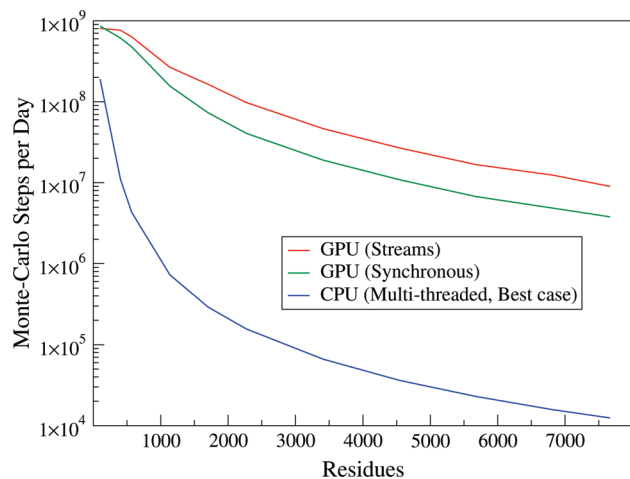


Figure 5. On a dual core machine with one GT280 card, our hybrid CPU–GPU implementation performs almost 1 billion MC iterations per day, outperforming the CPU by 2 orders of magnitude for as little as 500 residues.

force. These differences explain why, though our performance follows a similar profile to the MD code, we do not achieve their same maximum performance level (160 vs 212 GFLOPS, Figure 4). However, measures of FLOPS do not provide a reasonable metric by which to evaluate the real-world utility of an implementation. The overall performance of the simulation is critical, not merely the performance of the most parallel component. The comparatively lower performance of our potential evaluation kernel on the GPU is offset by streaming and multiple replicas in the REMC algorithm. We keep the branching and synchronization off the GPU as much as possible, sacrificing single stream performance for overall throughput. That this is an effective strategy for MC algorithms is clearly shown by the number of MC iterations per day achieved by our system (Figure 5). On a dual core machine with one GTX280 card, our GPU-accelerated code can perform almost 1 billion iterations per day for a simulations of 2 molecules totalling 100 residues. (In general, an unbound simulation requires on the order of 1 million to over 10 million iterations per replica before it reaches equilibrium, depending on the complexity and number of structures and on at least as many iterations more to perform sufficient sampling.) The largest simulations of 7668 residues are capable of up to 9 million iterations per day, whereas the CPU, fully utilizing both cores, manages only 12 000.

5.3. Validation and Initial Applications. To validate the correctness of the entire system (the energy function implementation, the MC algorithms for moving the proteins, and the replica exchange protocol), we have tested the code on two systems studied before: the binding of yeast cytochrome c to the cytochrome c peroxidase complex (Cc/CcP, Figure 6) and ubiquitin to the UIM1 domain of the Vps27 protein (UIM/Ub, Figure 7).²⁹ Binding curves from the present implementation closely match those from the previous CHARMM code; a clear validation of the correctness of our GPU implementation.

For the UIM/Ub system, we have simulated a “full-length” version of ubiquitin (residues 1–76) in addition to the truncated ubiquitin (residues 1–72 with flexible C-terminus

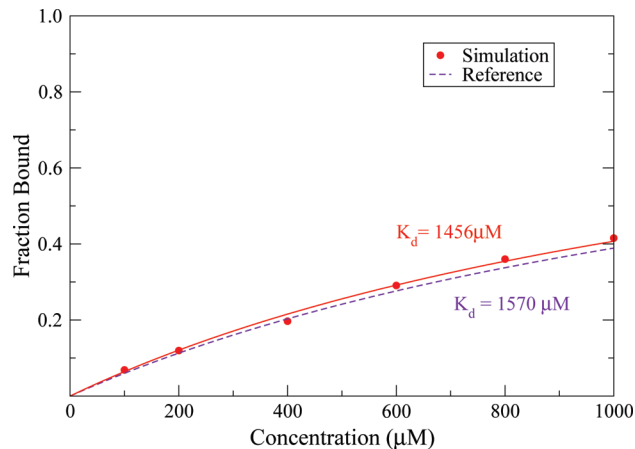


Figure 6. Binding affinity for Cc/CcP. Titration curves obtained with simulations using the GPU implementation are compared with the binding curve from the original reference implementation.¹¹

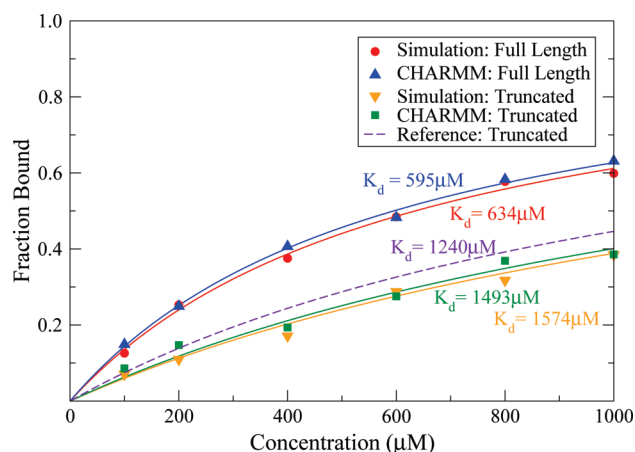


Figure 7. Binding affinity for UIM1/Ub. Binding data are given for a “full length” version of ubiquitin (residues 1–76) as well as the truncated ubiquitin (residues 1–72) in which the flexible C-terminus was removed, as in the original publication.¹¹ “Simulation” refers to the present implementation, “CHARMM” refers to the CHARMM implementation, and “Reference” refers to the original publication. The small discrepancy from the reference implementation is probably due to the use of spherical boundaries in that case, in contrast to the other implementations which use periodic boundaries.

removed) reported originally by Kim and Hummer.¹¹ The results for this system are interesting. The original study used a truncated form of ubiquitin because the C-terminus has a flexible tail. This resulted in a dominant population of bound complexes with native-like structures, with a population exceeding 40%. In addition, however, approximately 20% of the bound population comprised structures occupying the native binding pocket in the ubiquitin but with an inverted orientation of the UIM1 helix. In our study, we find a similar result. A cluster analysis of the bound conformations of the truncated ubiquitin simulations with a 1 Å rmsd cut-off gives a 1.0:4.0 ratio of incorrect:correct helix orientations. For the full length version of ubiquitin, however, there is a slight increase in overall binding affinity (Figure 7), suggesting additional favorable interactions between the UIM1 and the flexible C-terminus of ubiquitin. Furthermore, we find that

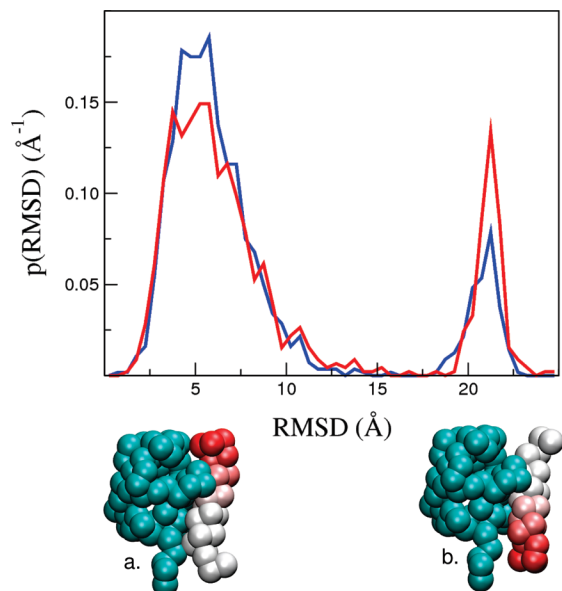


Figure 8. The distribution of rmsd to the correct binding site for the UIM1/Ub system in the full length version of ubiquitin (blue line) as well as the truncated ubiquitin (red line) in which the flexible C-terminus was removed, as in the original publication.¹¹ The small but significant population shift from the incorrectly (b) toward the correctly (a) bound conformation is clear for the full-length version of ubiquitin.

inclusion of the C-terminus slightly improves the specificity of binding. The ratio of incorrect:correct bound orientations improves to 1.0:6.2, indicating that the C-terminal interactions also preferentially stabilize the correctly bound conformation. Plots of the distribution of rmsd to the correct binding site for the both full length and the truncated ubiquitin make the shift in population clear (Figure 8). Since our study still somewhat unrealistically treats the C-terminus as rigid even when it is included, it will be interesting in the future to examine whether allowing flexibility in the tail further favors the correctly bound state.

One of our goals for the application of this code was to study the assembly mechanism of virus capsids from their constituent proteins. To test the feasibility of such a study, we have carried out trial simulations of a pair of protein fragments (each corresponding to the four protein chains from the PDB entry 2G33). The results for our trial simulations of two fragments from the HBV capsid are very encouraging. The bound conformations were initially clustered using the same method as for the UIM/Ubq complex; some of the clusters were then combined once the symmetry of the complex (i.e., AB:CD is equivalent to CD:AB) was accounted for, resulting in five clusters overall. In Figure 9 we have plotted the distance root-mean-square (DRMS) of structures from each cluster against their energy. We find that by far the largest cluster is that representing the correctly bound conformation, with a DRMS of around 4 Å from the experimental structure, comparable to that obtained for other complexes with this potential.¹¹ The correctly bound structures also have the lowest potential energy. A number of subsidiary, higher energy, bound clusters was also identified, indicative of mild frustration on the binding energy landscape.

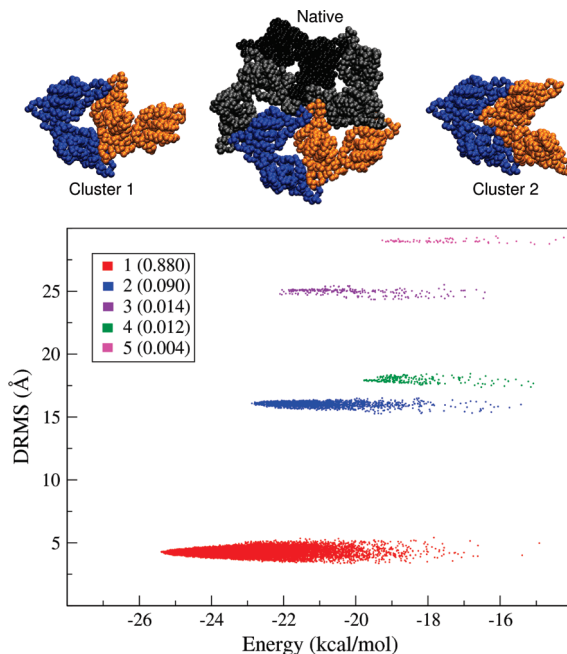


Figure 9. Bound structures of two viral capsid fragments. The DRMS from the bound structure is plotted against the potential energy for two fragments from the HBV virus capsid. Colors identify the different clusters of bound conformations. Population relative to the total bound conformations appears in brackets in the key. Conformations representative of clusters 1 and 2 are shown above the graph, together with a five-unit fragment of the HBV native structure.

6. Conclusions

We report a successful parallel CPU–GPU CUDA implementation of the Kim–Hummer coarse-grained model for replica exchange Monte Carlo (REMC) protein simulations. Our software is designed for new hybrid high-performance computing architectures combining multicore with GPU accelerators. This type of relatively low-cost parallel architecture has great relevance for researchers in developing countries, where High Performance Computing centers are generally not available.

Our hybrid parallel implementation employs multithreading for the Monte Carlo (MC) replicas and asynchronous calls to the potential evaluation kernel on the GPU. We did not explore CPU optimization further than multithreading, as the real performance gains are to be found in optimizing the GPU code. The simulation runtime is wholly dependent on the interaction potential calculations, which accounts for upward of 98% of the runtime. For the GPU kernel implementation, we found that the random access contact potential lookups are the chief performance bottleneck, as they prohibit defined optimal GPU memory usage patterns.

In general, best performance of our GPU kernel was achieved through adherence to the three chief tabled “CUDA best practices”: maximizing parallel execution and optimizing memory and instruction usage,²² although our solution illustrates that, for algorithms that do not fit the perfect GPU programming model, counterintuitive configurations may perform surprisingly well. Careful assessment of the performance of various memory locations for storing contact potential lookup table was key to achieving optimal perfor-

mance of our code on the nVIDIA GTX280 GPU architecture, which we finally achieved by bypassing the use of shared memory entirely in favor of texture memory.

Although the optimal balance between occupancy, memory usage, and type of memory used we identified is specific to the GTX280 hardware, in general nVIDIA's adherence to their architecture ensures that the optimizations performed on our code are applicable to past and future hardware.²¹ With respect to the general applicability of our code, we note that we found that different memory resource configurations do not cause extreme changes to the overall simulation runtime and to our implementation autotunes to achieve a block size for optimal kernel occupancy. Tuning of memory usage and problem decomposition remains relevant for the older GPU cards, which are likely to be around for a while. The new nVIDIA Fermi architecture has altered the resources on the GPU. There are four times as many cores per SM, and L2 cache has been introduced, along with coherent caching, making the GPU caching model more like the CPU caching model.³⁴ As the CUDA model has not changed for the Fermi architecture, performance of our code is likely to improve because of the increase in the number of cores. Our simulations require a high degree of accuracy for rotational transformations, which involves square roots and sine and cosine functions. These functions still do not meet the Institute of Electrical and Electronics Engineers (IEEE) 754 compliance on GF100 nVIDIA GPUs. Therefore, it is still necessary to perform the MC mutations on the CPU with the new generation of GPU hardware. The adoption of OpenCL for general GPU programming means that the same code, when ported to OpenCL, can be compiled to run on both nVIDIA and ATI devices. The generalization of OpenCL removes vendor concepts, such as texture memory, as the effects of texture memory are afforded by the true caching abilities of new hardware. Porting of our code to OpenCL is likely to be straightforward. The CUDA Driver API and OpenCL are very similar, with a high correspondence between functions and most differences relating to syntax.³⁷ Critically, the model and manner in which a GPU must be used and the topics discussed here in macro-tuning our kernels are relevant for both Cuda and OpenCL. We found very good performance of our accelerated parallel implementation, achieving over 1400 times speedup over a serial solution for simulations of systems larger than 4000 residues. On a dual-core machine with one GTX280 card, our GPU-accelerated code is capable of up to 9 million MC iterations per day with our largest benchmark simulation of 7668 residues, whereas the CPU, fully utilizing both cores, manages only 12 000. This allows for more thorough testing of the Kim–Hummer coarse-grained model. We find that inclusion of the ubiquitin C-terminus tail increases both the binding affinity and the specificity of binding for the UIM/Ub system, even with a rigid model. We also report successful preliminary simulations using the Kim–Hummer potential of the binding of two HBV capsid components, where by far the largest cluster is that representing the correctly bound conformation.

Simulations of massive protein structures are now within reach. We intend to use our system for a novel complete

simulation of the assembly of viral proteins into a viral capsid, a task previously prohibited by the excessive computation time required. Investigation of this process will give new insights into molecular self-assembly and may yield insights useful in the development of therapeutic drugs.

Finally, we note that, although we have only considered a specific model for protein–protein interactions, our implementation could easily be generalized to other types of interaction functions, as well as to other types of coarse-grained macromolecules (e.g., DNA). Finally, the effective parallelization approach developed in this work is generally applicable to N-body problems that require similar random access to lookup tables, where aspects of the interaction between bodies are dependent on their type or state.

Acknowledgment. The authors thank the Royal Society and the South African National Research Foundation (NRF) for a SA-UK Science Networks award. R.B. is supported by a Royal Society University Research Fellowship.

References

- (1) Tramontano, A. Problem 7. In *The Ten Most Wanted Solutions in Protein Bioinformatics; Chapman & Hall/CRC mathematical biology and medicine series*; Etheridge, A., Gross, L., Lenhart, S., Maini, P., Safer, H. Voit, E., Eds.; CRC Press: Boca Raton, FL, 2005; pp 117–139.
- (2) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409–443.
- (3) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **1976**, *104*, 59–107.
- (4) Friedrichs, M. S.; Wolynes, P. G. Toward protein tertiary structure recognition by means of associative memory Hamiltonians. *Science* **1989**, *246*, 371–373.
- (5) Shakhovich, E. I.; Gutin, A. M. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **1990**, *346*, 773–775.
- (6) Skolnick, J.; Kolinski, A. Simulations of the folding of a globular protein. *Science* **1990**, *250*, 1121–1125.
- (7) Karanicolas, J.; Brooks, C. L., III. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **2002**, *11*, 2351–2361.
- (8) Buchete, N.-V.; Straub, J. E.; Thirumalai, D. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J. Chem. Phys.* **2003**, *118*, 7658–7671.
- (9) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (10) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (11) Kim, Y. C.; Hummer, G. Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J. Mol. Biol.* **2008**, *375*, 1416–1433.
- (12) Kirk, D. B.; Hwu, W.-M. W. Performance Considerations. In *Programming Massively Parallel Processors: A Hands-on Approach*, 1st ed.; Morgan Kaufmann: Burlington, MA, 2010.

- (13) Tomov, S.; Nath, R.; Ltaief, H.; Dongarra, J. Dense Linear Algebra Solvers for Multicore with GPU Accelerators. In *Proceedings of IPDPS '10*; Sigarch, A., Ed.; IEEE Computer Society Press: Washington, D.C., 2010.
- (14) Nyland, L.; Harris, M.; Prins, J. Fast N-Body Simulation with CUDA. In *GPU Gems 3*; Nguyen, H., Ed.; Addison Wesley Professional: Reading, MA, 2007; pp677–695.
- (15) Belleman, R. G.; Bedorf, J.; Zwart, S. P. High Performance Direct Gravitational N-body Simulations on Graphics Processing Units II: An implementation in CUDA. *New Astron.* **2007**, *13*, 103–112.
- (16) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* **2009**, *30*, 864–872.
- (17) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. Accelerating Molecular Modeling Applications with Graphics Processors. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (18) Rodrigues, C. I.; Hardy, D. J.; Stone, J. E.; Schulten, K.; Mei, W. GPU acceleration of cutoff pair potentials for molecular modeling applications. In *Proceedings of the 5th conference on Computing frontiers (CF'08)*; ACM: New York, 2008; pp 273–282.
- (19) Hardy, D. J.; Stone, J. E.; Schulten, K. Multilevel summation of electrostatic potentials using graphics processing units. *Parallel Comput.* **2009**, *35*, 164–177.
- (20) Schmid, N.; Bötschi, M.; van Gunsteren, W. F. A GPU Solvent-Solvent Interaction Calculation Accelerator for Biomolecular Simulations Using the GROMOS Software. *J. Comput. Chem.* **2010**, *31*, 1636–1643.
- (21) CUDA Programming Guide, version 2.3; NVIDIA Corporation: Santa Clara, CA, 2009; <http://developer.download.nvidia.com/compute/cuda/2/toolkit/docs/NVIDIAUDARogramminguide.3.pdf>. Accessed: August 19, 2010.
- (22) NVIDIA, CUDA Best Practices Guide 2.3; NVIDIA Corporation: Santa Clara, CA, 2009; http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_BestPracticesGuide_2.3.pdf. Accessed: August 19, 2010.
- (23) Mackerell, A. D.; Feig, M.; Brooks, C. L. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- (24) Fraczekiewicz, R.; Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **1998**, *19*, 319–333.
- (25) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (26) Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.
- (27) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics methods for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (28) Brooks, B. R.; Brucoleri, 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.
- (29) Swanson, K. A.; Kang, R. S.; Stamenova, S. D.; Hicke, L.; Radhakrishnan, I. Solution structure of Vps27 UIM-ubiquitin complex important for endosomal sorting and receptor down-regulation. *EMBO J.* **2003**, *22*, 4597–4606.
- (30) Pelletier, H.; Kraut, J. Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. *Science* **1992**, *258*, 1748–1755.
- (31) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (32) Matsumoto, M.; Nishimura, T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8*, 3–30.
- (33) Harris, M. Optimizing Parallel Reduction in CUDA; NVIDIA Corporation: Santa Clara, CA, 2007; http://developer.download.nvidia.com/compute/cuda/1_1/Website/projects/reduction/doc/reduction.pdf. Accessed August 19, 2010.
- (34) nVIDIA's Next Generation CUDA Compute Architecture: Fermi. 2010; NVIDIA Corporation: Santa Clara, CA, 2010; http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf. Accessed August 19, 2010.
- (35) van Meel, J. A.; Arnold, A.; Frenkel, D.; Portegies; Belleman, R. G. Harvesting graphics power for MD simulations. *Mol. Simul.* **2008**, *34*, 259–266.
- (36) Tomov, S.; McGuigan, M.; Bennett, R.; Smith, G.; Spiletic, J. Benchmarking and implementation of probability-based simulations on programmable graphics cards. *Comput. Graph.* **2005**, *29*, 71–80.
- (37) nVIDIA OpenCL Jumpstart Guide; NVIDIA Corporation: Santa Clara, CA, 2009; http://developer.download.nvidia.com/OpenCL/NVIDIA_OpenCL_JumpStart_Guide.pdf. Accessed August 18, 2010.

Internal Dynamics of an Analytically Coarse-Grained Protein

Michael J. M. Mazack, Alessandro Cembran, and Jiali Gao*

*Department of Chemistry, Digital Technology Center, and Supercomputing Institute,
University of Minnesota, Minneapolis, Minnesota 55455, United States*

Received July 31, 2010

Abstract: An analytically coarse-grained model (ACG) is introduced to represent individual macromolecules for the simulation of dynamic processes in cells. In the ACG model, a macromolecular structure is treated as a fully coarse-grained entity with a uniform mass density without the explicit atomic details. The excluded volume and surface of the ACG macromolecular species are explicitly treated by a spherical harmonic representation in the present study (although ellipsoidal, solid, and radial augmented functions can be used), which can provide any desired accuracy and detail depending on the problem of interest. The present paper focuses on the description of the internal fluctuations of a single ACG macromolecule, modeled by the superposition of low frequency quasiharmonic modes from explicit molecular dynamics simulation. A procedure for estimating the amplitudes, time scales of the quasiharmonic motions, and the corresponding phases is presented and used to synthesize the complex motion. The analytical description and numerical algorithm can provide an adequate representation of the internal protein fluctuations revealed from the corresponding atomistic simulations, although the internal motions of ACG macromolecules do not explore motions not exhibited in the dynamic simulations.

1. Introduction

Molecular dynamics simulations of biological macromolecules in explicit aqueous solution offer the most detailed information at the atomic level, which is essential for the understanding of dynamics, binding, and activity of these systems.¹ Tremendous progress has been made both in the advance of computer architecture and in the development of computational algorithms, enabling atomistic dynamics simulations to treat systems containing millions of atoms² and the dynamics lasting up to milliseconds.^{3,4} However, the spatial and temporal scales needed to address questions relevant to cellular processes such as protein–protein and protein–nucleic acid interactions and macromolecular assembly dwarfs the most sophisticated atomistic approaches available today and perhaps in the distant future.⁵ In such a mesoscopic system, it is necessary to use a coarse-grained approach to describe the individual macromolecular components.^{6,7} In this series of papers, we present an analytically coarse-grained (ACG) model to represent macromolecular

entities such as proteins and nucleic acids as single building blocks that can be used to study macromolecular interactions and assembly in biological cells.

In the past decade, significant efforts have been devoted to the development of coarse-grained models to circumvent the need for describing molecular systems with increasing demands in size and time.⁸ Nevertheless, the concepts of atoms and molecules are deeply rooted in our perception of intermolecular interactions; not surprisingly, coarse-grained models typically involve interaction sites in terms of reduced representation of the detailed atomic features of the target system. The early united-atom force field is an example of this type of coarse graining,^{9,10} and recent advances have enabled a much larger number of atoms to be grouped into a united site along with continuum elastic network models.⁸ However, to model macromolecular interactions and assembly such as the mechanism of a virus capsid formation, the detailed sequence and the “united-atom” constituents are no longer critical, and it becomes unnecessary to enumerate the specific pairwise interactions between coarse-grained groups among proteins. On the other hand, the use of regular

* Corresponding author e-mail: gao@jialigao.org.

geometrical shapes seems too crude. Indeed, the key structural components are the specific shape and the excluded volume of each capsid protein along with their intrinsic dynamic fluctuations and the accompanying surface electrostatic potential and surface tension. The detailed atomistic interactions, of course, are essential for recognition and binding when two macromolecular particles are in close contact, but these are not the types of details needed for transport processes in the cell. It appears to be desirable to develop a theoretical method that is not restricted to the detailed features of atoms and molecules or coarse-grained interacting groups in a large system; yet, the individual macromolecular species still retain the information of, and are constructed based on the detailed atomistic coordinates determined experimentally, which also provide all physical and biological properties needed to model the dynamic system. Furthermore, the intrinsic fluctuations of the coarse-grained macromolecules relevant to the time scale of the dynamic model for the mesoscopic system need to be taken into account.⁸ To limit the scope of discussion, the present article is only concerned with the representation of the internal dynamic fluctuations of an ACG macromolecule itself, which are derived from explicit molecular dynamics simulations.

To this end, we make use of the mathematical tools of spherical harmonic analysis to represent the macromolecular particles of interest; spherical harmonic analysis has been extensively applied in a wide range of areas such as geopotential,¹¹ topography,¹² and physics as well as motion picture and gaming animation. Since analytical harmonic basis functions are used, the method that we design for modeling cellular processes is called the analytical coarse-graining (ACG) of macromolecules. The harmonic representation can also be used directly to describe the physical interactions and to model the dynamics of the system, which will be detailed in the next papers. Our strategy provides a single, unifying theory and computational algorithm to study macromolecular systems consisting of thousands of macromolecular particles and entities. In such an approach, each macromolecular unit, such as a protein, is “coarse-grained” as a single moiety of uniform mass density whose excluded volume is encompassed by its solvent-accessible surface that is represented by a set of analytical harmonic basis functions. Furthermore, its physical and biological properties can be treated by exactly the same mathematical procedure as the representation of the macromolecule. Here, we describe the treatment of the intrinsic dynamic fluctuations of a single ACG protein using spherical harmonic functions.

Although the mathematical tools of spherical harmonic computation have been established since the 1780s and modern numerical techniques have greatly enhanced the computational speed,^{13,14} Max and Getzoff,¹⁵ and Olson and co-workers, were among the first to apply spherical harmonic functions to the visualization of molecular surfaces as a graphics rendering tool.^{16–20} Recent years have seen the increased usage of this technique to model protein–ligand interactions and protein docking.^{21–25} Buchete et al. used spherical harmonics to analyze coarse-grained potentials for folding calculations.²⁶ In a subsequent publication, Duncan

and Olson described the possibility of animating the dynamic motion of a protein to render real-time graphics visualization;²⁷ however, it does not appear that specific investigations have been reported. The approach described by Olson and co-workers was aimed to follow the time-dependent Cartesian coordinates of a protein surface as modeled by normal mode dynamics; the method provides an efficient procedure to generate graphics rendering, but it does not guarantee single-valued properties on the surface, nor is it suitable for modeling protein dynamic motions.^{19,27} In contrast, the method described in this article focuses on radial fluctuations of an ACG protein surface that concerns no atomic details, but it is designed to model the most significant dynamic motions revealed from an explicit molecular dynamics trajectory, in which the radial fluctuations are decomposed based on quasiharmonic dynamic analysis. The latter has been extensively explored to characterize large amplitude motions and quasiharmonic vibrational modes of proteins and nucleic acids;^{28–32} it provides an adequate analytical procedure to describe the global large amplitude motions of a fully coarse-grained macromolecule with spherical harmonic representation. In addition, for a set of well-chosen numerical quadrature points in the spherical harmonic analysis, our approach provides an efficient procedure for evaluation of molecular properties.

In the following, we first present the analytically coarse-grained (ACG) model and computational details, focusing on the use of spherical harmonic basis functions. Then, we describe a procedure for constructing a mathematical approach to describe the intrinsic quasiharmonic dynamic fluctuations of a protein based on the information from molecular dynamics simulations in explicit solvent. This is followed by an illustrative example to show the feasibility of modeling protein fluctuations without explicit atomic details using the ACG model. Finally, we summarize the key findings of the present study.

2. Method

Throughout this article, we mainly use the homodimeric enzyme, orotidine 5′-monophosphate decarboxylase (OMP-DC), as an illustrative example, which consists of two β -barrels of eight strands of β -sheet and eight α -helices.³³ In this article, for convenience of discussion, we focus on the use of a spherical harmonic basis, with which the ACG method is applicable to any macromolecular systems that have star-like topology.¹⁵ We note here that this is not a restriction because any geometrical shapes including non-star-like macromolecules can be represented in the ACG model by augmenting radial functions such as the Zernike function or Slater-type radial functions,^{23–25,34,35} but we shall not discuss these approaches here. In the rest of the paper, we interchangeably use the terms of “protein” and “macromolecule”, which include proteins, nucleic acids, lipids, and other components of a macromolecular assembly, without specific distinction. In this section, we first outline a qualitative description of the coarse-grained macromolecular model. Then, we provide a brief summary of spherical harmonic representation of the surface of a macromolecular structure. This is followed by the description of incorporating

internal dynamic fluctuations of the protein based on the information obtained from principal component analysis (PCA)^{29,30} of the atomistic molecular dynamic trajectory of OMPDC in water.³³ The purpose in this article is not aimed at studying the dynamics of proteins using quasiharmonic dynamics; the latter has been thoroughly investigated and its applications and limitations have been characterized.^{29–32} The goal here is to illustrate the capability and procedure of incorporating low-frequency, large-amplitude dynamic fluctuations, as revealed by an explicit dynamics simulation, into a fully coarse-grained protein model.

2.1. Description of a Macromolecular Particle. We consider a macromolecular structure, which can be a protein or a domain of a protein complex, a segment of nucleic acids, or a protein–nucleic acid complex, as a single entity of uniform mass density. The excluded volume of a given macromolecular structure is defined as the cavity enclosed by the solvent-accessible surface (or the van der Waals surface depending on needs), originating from the detailed three-dimensional atomic structure determined experimentally by X-ray crystallography or NMR, or generated computationally by homology modeling and protein-folding prediction in the absence of experimental data. Note that the solvent-accessible surface encloses a molecular volume which may be significantly greater than that defined by its van der Waals surface, the latter of which is more appropriate for evaluating the macromolecular density. All biochemical functions and physical properties of the macromolecule are fully encoded in the three-dimensional structure, necessary for microscopic and mesoscopic modeling of intermolecular interactions, including electrostatics and hydrophobic surface tension. The characteristic features of a macromolecular structure are considered to have distinguishing features both in size and property from small molecules, peptide fragments, ligands and cofactors, ions and solvent molecules, although the specific criteria depends on a particular application.

From the onset, we do not consider the detailed atomic coordinates or interaction sites; the entire macromolecular unit is a single coarse-grained entity. This is justified as a result of statistical averaging over the spatial and temporal scales to be used to model the dynamic system, which is the cell. However, the size, as defined by the excluded volume occupied by the individual atoms, ligands, and perhaps a small number of buried or surface solvent molecules, and the shape, as represented by the solvent-accessible surface, of a given macromolecule are necessary and critical to a physics-based approach; they are well-defined in our ACG model (vide infra) by a single mathematical approach for all types of macromolecular particles of different sizes and shapes, which can be used to systematically provide any desired accuracy and detail of the coarse-grained macromolecule. The definition of a uniform mass density within its excluded volume is akin to the use of a continuum solvent model and a single interior low dielectric constant for a protein in Poisson–Boltzmann calculations and is consistent with our goal of modeling the dynamics of the entire system, which involves the integration of equations of motion at a time step in the order of tens of picoseconds to nanoseconds per iteration. Thus, the representation of the macromolecular

species is an average of the system over the time series of the coarse-grained model,^{8,36,37} involving the internal atomic fluctuation and spatial orientation when the center of mass of the macromolecule is chosen as a reference point.

Throughout this article, the method of Lee and Richards is used to define the macromolecular surface and the excluded volume,³⁸ although other approaches are available.³⁹

2.2. Spherical Harmonic Representation of a Macromolecular Particle. The method of using spherical harmonics functions to represent the surface of globular proteins was described by Max and Getzoff¹⁵ and later by Duncan and Olson and by others.^{16,17,19,20} Here, we briefly summarize the key elements of this approach and highlight the numerical details employed in our implementation.

Arising from the solution of Laplace's equation in spherical coordinates, the spherical harmonic expansion is the spherical coordinate analog of the widely used Fourier series expansion. Spherical harmonic representations of macromolecules provide not only a smooth and aesthetically pleasing surface, but also the ability for evaluating surface properties such as normal vectors and principal curvature. For any star-like surface, which is single valued in the radial direction of (θ, ϕ) with respect to an origin, there exists a spherical harmonic expansion given as follows:

$$S(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l a_{lm} Y_l^m(\theta, \phi) \quad (1)$$

where θ denotes the latitudinal or zenith angle ($0 \leq \theta \leq \pi$), ϕ specifies the longitudinal or azimuth angle ($0 \leq \phi < 2\pi$), $S(\theta, \phi)$ is the radial distance of the surface at angular coordinate (θ, ϕ) , a_{lm} are the expansion coefficients and $Y_l^m(\theta, \phi)$ are the real spherical harmonic basis functions, which are orthonormal under the L^2 inner product. We have used the center of mass as the origin in all calculations, and the local axis is generally chosen to coincide with the principal moments of inertia. In general, $S(\theta, \phi)$ can be any scalar physical or chemical property mapped on to the surface of a unit sphere.

The real spherical harmonic basis functions are defined by

$$Y_l^m(\theta, \phi) = \begin{cases} \frac{1}{\sqrt{2\pi}} \bar{P}_l^0(\cos \theta) & m = 0 \\ \frac{1}{\sqrt{\pi}} \bar{P}_l^m(\cos \theta) \cos m\phi & m > 0 \\ \frac{1}{\sqrt{\pi}} \bar{P}_l^{-m}(\cos \theta) \sin m\phi & m < 0 \end{cases} \quad (2)$$

where $\bar{P}_l^m(\cos \theta)$ denotes the normalized associated Legendre polynomial of the first kind, of order m and degree l . Methods exist for evaluating the Legendre polynomials of eq 2, one of which is discussed in ref 40, and ref 22 lists a technique yielding more numerically stable results.

Making use of the orthonormal property of the real spherical harmonic basis functions under the L^2 inner product, the expansion coefficients in eq 1 are given by

$$a_{lm} = \int_0^{2\pi} \int_0^\pi S(\theta, \phi) Y_l^m(\theta, \phi) \sin \theta \, d\theta \, d\phi \quad (3)$$

In practice, the infinite sum of eq 1 is truncated to a relatively small value L , which produces an approximate surface \mathbf{S} . Thus, it is only necessary to determine $(L + 1)^2$ coefficients and function values for the evaluation of points on the surface:

$$S(\theta, \phi) \approx \sum_{l=0}^L \sum_{m=-l}^l a_{lm} Y_l^m(\theta, \phi) \quad (4)$$

The Cartesian coordinates of vertices used to form a triangulated mesh for graphics display are obtained from the corresponding values in the polar spherical coordinates. This differs from the approach of Duncan and Olson,^{16,17} who used spherical harmonics expansions to directly approximate the Cartesian coordinates of surface points. Although the direct representation of the surface Cartesian coordinates is convenient for graphics display, it is not suitable for computation of molecular properties of the system, including intermolecular interactions.

2.3. Dynamic Motion. *2.3.1. Quasiharmonic Dynamics.* We use the lowest frequency modes from principal component analysis (PCA) of a molecular dynamics trajectory of a solvated protein to represent its dynamic fluctuations; the dimeric enzyme OMPDC is employed as an illustrative example. The PCA results show the directionality and frequency of protein dynamic motions, in which the lowest frequency modes are typically correlated with protein conformational changes and have been used to interpret conformational variations observed experimentally.⁴¹ Although other approaches such as the continuum elastic network model can be used,⁴² for an analytically represented coarse-grained protein without the explicit details of atomic structure, the PCA modes provide the most direct connection to the dynamic motions sampled during an explicit molecular dynamics simulation. The animation of atomic motions following a given normal mode and quasiharmonic dynamics simulation of the complex motions of a macromolecule have been widely used in structure and dynamics analyses at the atomic details. Voth and co-workers described a method to map coarse-grained sites on the basis of PCA modes,⁴³ and the model has been extended to using the low-frequency normal modes of an elastic network model for the protein.⁴⁴ Our method follows a different route of representation than that of Zhang et al.,^{43,44} in ACG, the model is used to represent and animate the low-frequency PCA modes, rather than being derived from PCA. We apply the approach of quasiharmonic dynamics to model the internal fluctuations of coarse-grained macromolecules.

The overall protein fluctuation is obtained by the superposition of individual quasiharmonic modes:

$$\mathbf{R}_j(t) = \mathbf{R}_j(0) + \sum_k^K \mathbf{Q}_{jk} \sigma_k \cos(\omega_k t + \lambda_k) \quad (5)$$

where $\mathbf{R}_j(0)$ and $\mathbf{R}_j(t)$ are the coordinates of atom j at time 0 and t , respectively, and K is the number of quasiharmonic modes used to animate the total dynamic fluctuation of the system. In eq 5, the parameters associated with mode k , ω_k ,

\mathbf{Q}_k , λ_k , and σ_k are the frequency, mode direction eigenvector, phase, and amplitude, respectively. The phase λ_k is associated with the initial atomic positions, and the thermal average of the second moment of the amplitude distribution is given by $\sigma_k^2 = k_B T / \omega_k^2$, where k_B is Boltzmann's constant and T is temperature. The value of K in eq 5 is restricted by the integration time increment, τ , used to propagate the dynamic equations of the coarse-grained system such that $\tau > 2\pi / \omega_K$. Typically, the inclusion of the lowest 10 to 20 modes is more than sufficient to represent the most significant large-amplitude motions.

Here, we use the lowest frequency quasiharmonic motions to represent the internal dynamic fluctuations of analytically coarse-grained macromolecule particles. The limitation of this approach is that it will not produce information for even larger amplitude motions that have not been uncovered in the explicit molecular dynamics simulation. Thus, if the protein undergoes folding and unfolding exchange, it is not appropriate to use the present model; however, it is possible to incorporate into the present treatment conformational transitions for which structures in different conformational substates have been determined experimentally (e.g., by X-ray crystallography or NMR). Nevertheless, our approach is not a simple reproduction of the fluctuation of the molecular dynamics trajectory itself because collision between different coarse-grained macromolecular species can cause random changes in the amplitude and phase of each quasiharmonic mode, resulting in different combinations of modes and trajectories.

Previously, Duncan and Olson proposed a method for shape analysis of protein dynamic surfaces.²⁷ In that approach, the Cartesian coordinate displacements of surface points corresponding to triangulation vertices were obtained from the static surface and the expansion coefficients for the normal mode eigenvectors projected to these points. Although the method is extremely useful for fast visualization of surface motion, it is not designed to model real-time dynamics. Furthermore, the displacements of triangulation vertices in such a shape analysis algorithm are not suited for property evaluation because the quadrature points and weights will have to be recomputed, which is impractical for real time dynamics simulations. In our approach, the surface deformation is restricted to the direction along the radial vector, consequently preserving the angular coordinates (θ_i, ϕ_i) and the precomputed numerical weights.

2.3.2. Definition of Surface Displacement Vector. We begin with a molecular dynamics trajectory $\{\mathbf{R}(t_n); n = 0, 1, \dots, N\}$ that was saved at time slice t_n , where $\mathbf{R}(t_n)$ is a vector of all atomic coordinates. Principal component analysis of this trajectory yields a set of quasiharmonic vibrations with frequencies $\{\omega_k\}$ and eigenvectors $\{\mathbf{Q}_k\}$. The K lowest frequency modes will be used to model the total dynamic motions that have been sampled by the original molecular dynamics simulation.

For each mode k , we use two distorted configurations generated by following the eigenvector direction stretched to $-2\sigma_k$ and $+2\sigma_k$ from its mean, denoted by $\mathbf{R}_k^{-2\sigma}$ and $\mathbf{R}_k^{+2\sigma}$, to represent approximately the "lower" and "upper" bound of an amplitude, respectively. Let $\{\mathbf{S}_k^{-2\sigma}\}$ and $\{\mathbf{S}_k^{+2\sigma}\}$ be the

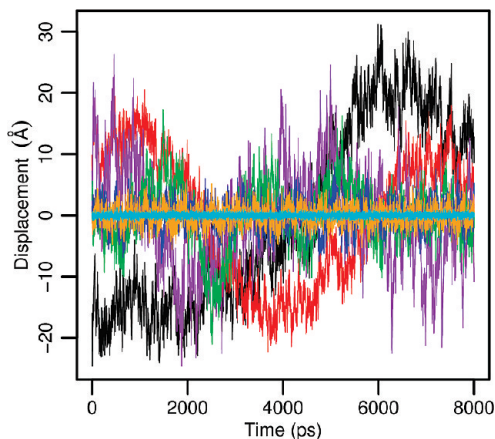


Figure 1. Histogram of the computed projection of instantaneous molecular structure of OMPDC in water onto normalized eigenvector directions of the lowest quasiharmonic mode (black), the second (red), the fourth (purple), the tenth (green), the fiftieth (blue), the one hundredth (orange), and the one thousandth (cyan) modes.

solvent-accessible surfaces (SAS) for the two extreme configurations associated with mode k . Given a set of surface points, $\{u_{ij} = (\theta_i, \phi_j); i = 1, \dots, M_\theta; j = 1, \dots, M_\phi\}$, where M_θ and M_ϕ are the number of quadrature points, the radial displacement, due to quasiharmonic vibration of mode k , at the surface point $u_{ij} = (\theta_i, \phi_j)$ is defined as the $2\sigma_k$ variance:

$$q_{ij}^k = \frac{1}{2} [S_k^{2\sigma}(\theta_i, \phi_j) - S_k^{-2\sigma}(\theta_i, \phi_j)] \quad (6)$$

The vector \mathbf{q}^k , which can be considered as a property of a unit sphere, represents the approximate amplitude (see below) and direction of surface deformation associated with PCA quasiharmonic mode k :

$$\mathbf{q}^k = \begin{pmatrix} q_{11}^k \\ q_{21}^k \\ \dots \\ q_{P_\theta P_\phi}^k \end{pmatrix} \quad (7)$$

Consequently, the radial displacement vector can also be expressed by a spherical harmonic expansion whose coefficients are determined using the same procedure as for the molecular surface itself (eq 3):

$$q_{ij}^k = \sum_{l=0}^L \sum_{m=-l}^l c_{lm}^k Y_l^m(\theta_i, \phi_j) \quad (8)$$

2.3.3. Frequency and Phase. Although the rank of the low frequency modes from principal component analysis is very reasonable, the quantitative values of the lowest quasiharmonic vibrational frequencies and the associated time scales are not expected to be accurate to represent the real time dynamic motion.^{29–32} Thus, one needs to seek a different way to obtain the desired oscillatory frequencies. We examined the time dependence of the projections of the instantaneous coordinate vector onto the normalized quasiharmonic eigenvectors, seven of which are shown in Figure 1, corresponding to PCA mode numbers 1, 2, 4, 10, 50, 100,

and 1000 over a total of 8 ns MD trajectory. Although not unexpected, we are pleased to see the oscillatory behavior of each mode, and the amplitudes and frequencies of these oscillations roughly coincide with the order of the PCA modes. For modes above number 50, the fluctuations illustrated in Figure 1 can be considered as noise (friction) with respect to the motions of the lowest frequency modes. Importantly, it appears that the PCA mode-projection results can be used to estimate the quantitative frequencies as well as the phase with respect to the structure at time $t_0 = 0$ needed to animate the complex motion of the superimposed fluctuations.

To this end, we used a sinusoidal fitting procedure to optimize the amplitude A_k (not used for mode animation, see below), frequency ω_k , and phase λ_k in eq 9 for each mode to best reproduce the time-dependent quasiharmonic mode projection data.

$$M_k(t) = A_k \cos(\omega_k t + \lambda_k) \quad (9)$$

In Figure 2, we depict the fitted curves against the raw data for modes 1, 2, 4, and 10, whereas the results for the first 20 modes are given as Supporting Information. Table 1 lists the optimized amplitudes, frequencies, and phases for the first 20 modes, the first 10 of which are used to represent the overall protein dynamics fluctuations as an illustrative example in this article. An alternative approach is to use the spectral transform of the autocorrelation function of the quasiharmonic mode fluctuations.

2.3.4. Time Evolution of the Dynamic Fluctuation. The SAS surface $\mathbf{S}(t_0)$ corresponding to the structure $\mathbf{R}(t_0)$ at time $t_0 = 0$ in the dynamic trajectory is chosen as the starting configuration and is expressed in terms of spherical harmonics basis as follows:

$$S_{ij}(0) \equiv S(t = 0, \theta_i, \phi_j) = \sum_{l=0}^L \sum_{m=-l}^l \mathbf{a}_{lm}^o Y_l^m(\theta_i, \phi_j) \quad (10)$$

where $S_{ij}(0)$ is the radial distance at an angular coordinate $u_{ij} = (\theta_i, \phi_j)$, and the coefficients $\{\mathbf{a}_{lm}^o\}$ are determined according to eq 3. We assume that the dynamic modulation of the protein surface associated with mode k also has the same frequency. Thus, the atomic coordinates in eq 5 are replaced by protein surface points, and we write the total surface radial displacement at point $u_{ij} = (\theta_i, \phi_j)$ as follows:

$$\begin{aligned} S_{ij}(t) &= S_{ij}(0) + \sum_{k=1}^K W_k q_{ij}^k \cos(\omega_k t + \lambda_k) \\ &= \sum_{l=0}^L \sum_{m=-l}^l [\mathbf{a}_{lm}^o + \sum_{k=1}^K W_k c_{lm}^k \cos(\omega_k t + \lambda_k)] Y_l^m(\theta_i, \phi_j) \end{aligned} \quad (11)$$

where W_k is a mode weighting factor to be determined by least-squares fit to the instantaneous surfaces $\tilde{\mathbf{S}}(t_n)$ of the structures sampled by the explicit molecular dynamics simulations in the entire trajectory, $\{\mathbf{R}(t_n) \rightarrow \tilde{\mathbf{S}}(t_n); n = 0, 1, \dots, N\}$. Equation 11 preserves the angular coordinates, consequently all precomputed values of the spherical harmonic functions and quadrature weights needed for

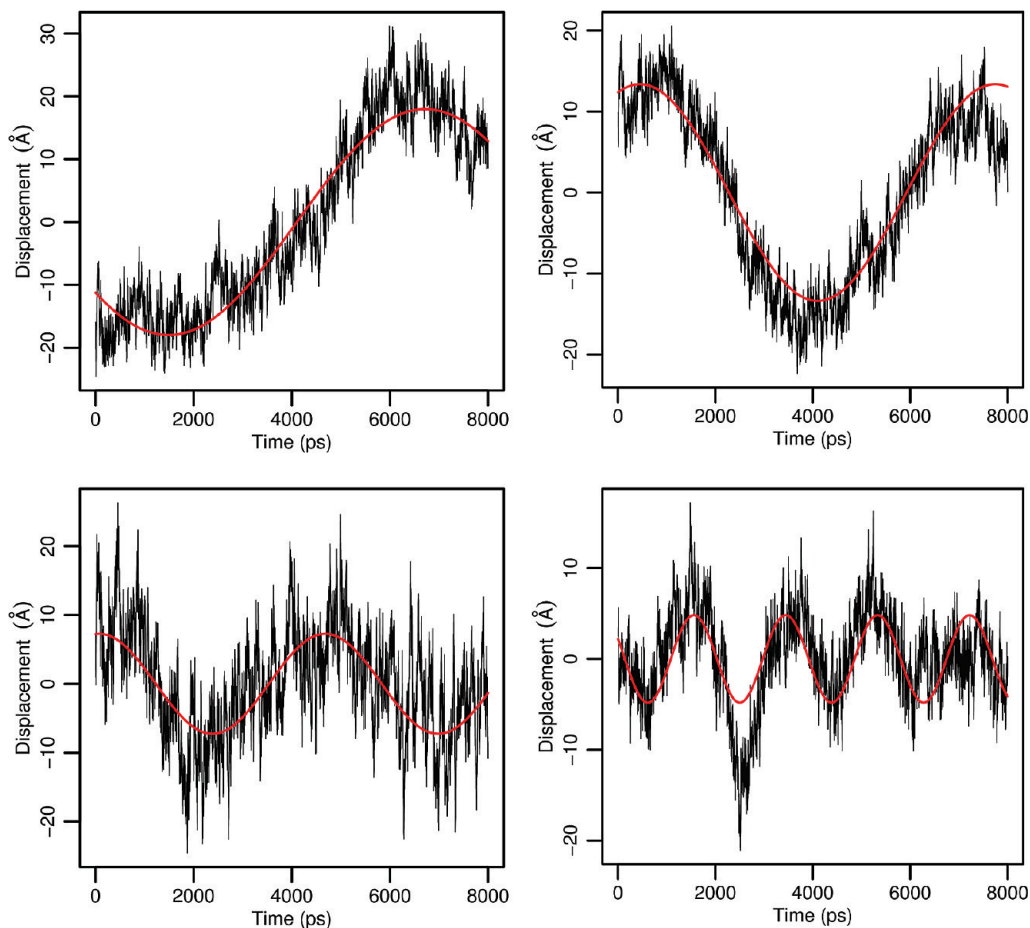


Figure 2. Sinusoidal fit of harmonic frequencies and phases (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 1 for modes number 1, 2, 4, and 10.

Table 1. Optimized Amplitudes (Å), Frequencies (rad/ns), and Phase (rad) for the Time-Dependent Quasiharmonic Mode Projection along the Molecular Dynamics Trajectory of the Protein Orotidine Monophosphate Decarboxylase As Represented by Eq 16

mode	A_i	ω_i	λ_i
1	18.0	0.60	2.25
2	13.4	0.86	5.89
3	9.8	1.55	1.05
4	7.3	1.37	6.17
5	3.0	2.71	2.50
6	1.3	7.37	4.02
7	6.5	2.12	2.97
8	5.0	2.72	1.62
9	3.9	2.66	3.32
10	4.8	3.33	1.10
11	3.2	3.37	6.00
12	3.3	2.65	1.22
13	0.4	3.79	0.69
14	2.7	4.10	5.47
15	2.6	4.17	2.66
16	0.9	5.01	6.21
17	2.1	4.72	1.33
18	2.5	4.22	5.78
19	1.7	7.59	1.71
20	1.2	7.59	1.60

property evaluations (as well as for real-time graphics animation).⁴⁵

The mode weighting factors in eq 11 are determined by minimizing the following error function:

$$\varepsilon = \frac{1}{N} \sum_{n=1}^N \sum_{ij} [S_{ij}(t_n) - \tilde{S}_{ij}(t_n)]^2 \quad (12)$$

It is straightforward to show that the minimization yields a linear equation that can be conveniently solved.

$$\sum_{k=1}^K B_{qk} W_k = D_q; \quad q = 1, \dots, K \quad (13)$$

where the matrix elements are defined as follows:

$$D_q = \frac{1}{N} \sum_{n=1}^N \sum_{ij} [\tilde{S}_{ij}(t_n) - \tilde{S}_{ij}(0)] U_{ij}^q(t_n) \quad (14)$$

$$B_{qk} = \frac{1}{N} \sum_{n=1}^N \sum_{ij} U_{ij}^q(t_n) U_{ij}^k(t_n) \quad (15)$$

$$U_{ij}^q(t_n) = \sum_{l=0}^L \sum_{m=-l}^l c_{lm}^q \cos(\omega_q t_n + \lambda_q) Y_l^m(\theta_i, \phi_j) \quad (16)$$

3. Numerical Considerations

The spherical harmonic expansion coefficients are determined by sampling surface values (coordinates) to approximate the integral of eq 3. A number of methods for optimizing surface point distribution are available including the use of a geodesic unit sphere.⁴⁶ In the present application, realizing that the

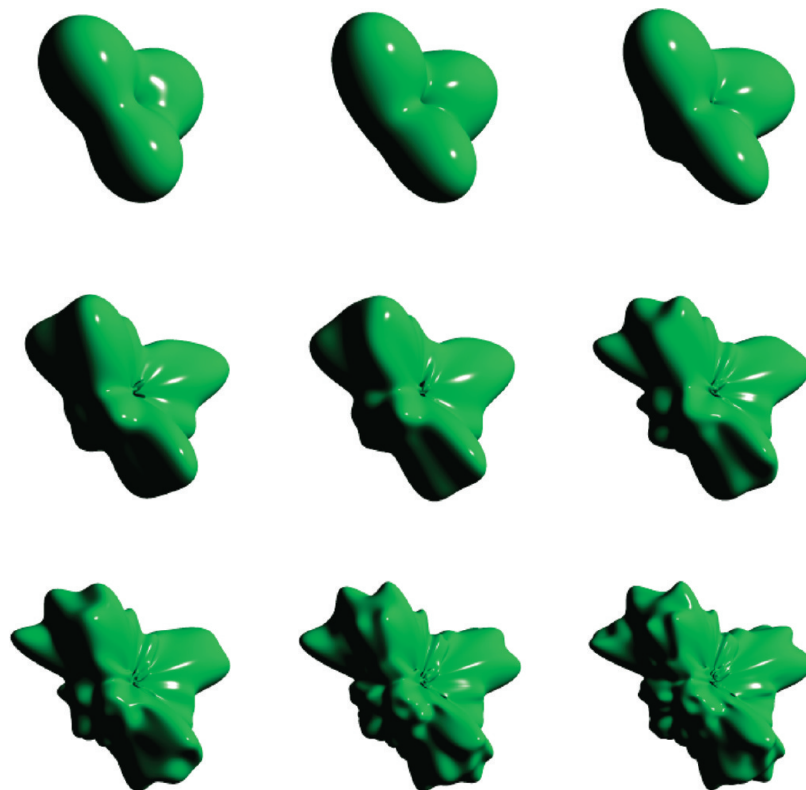


Figure 3. Spherical harmonic reconstruction of the Lee and Richards surface for the trimer complex of the capsid protein of turnip yellow mosaic virus using representation degrees of $L = 3, 4, 5, 10, 12, 15, 20, 25,$ and 30 numbered from top left to bottom right.

integral in θ is formally a Fourier transform, the numerical integration can be evaluated using M equispaced points to take advantage of fast Fourier transform (FFT) at a computing scaling of $O(M \log M)$. There are two ways of selecting points in ϕ ; the first is to use Gauss–Legendre quadrature nodes and weights, which needs only $M/2$ points, whereas a set of equally spaced points can be selected, which is equivalent to Chebychev nodes in $\cos \phi$.^{45,47} The latter is convenient to use but less efficient computationally and requires a total of M points for the same accuracy. The numerical scaling for integrating in ϕ is $O(M^3)$.

If the sampling points used in the evaluation of the integral in eq 3 are chosen to coincide with the numerical quadrature values, $\{(\theta_p, \phi_q); p = 1, \dots, M_\theta; q = 1, \dots, M_\phi\}$, the numerical procedure for property calculation can be greatly simplified. In all cases, the associated Legendre function values are precomputed along with the measure, $\sin \theta_i$, and quadrature weights (see Appendix) for a given structure and stored. The use of $(\sin \theta_i)^{1/2} P_l^m(\theta_i, \phi_j)$ preconditioning in property calculations can greatly increase numerical stability by keeping the product roughly constant.⁴⁷

All computations and illustrations are performed using a software package written in our laboratory.

4. Discussion

Figure 3 illustrates the spherical harmonic rendering of the trimeric structure of the capsid protein (2FZ2) of turnip yellow mosaic virus at various degrees of representation up to $L = 30$. By simple inspection, the domed triangular shape is not directly associated with a unit sphere, but the 3-fold symmetry of the complex is already represented at $L = 3$,

and the domed structural feature is clearly visible with $L = 4$ and 5 . As the degree of spherical harmonic basis increases, the molecular shape and detail is well described with an L above 10, while greater local features can be found using higher degrees. In principle, the spherical harmonic representation can yield any desired accuracy by increasing the value of L . However, it should be kept in mind that the protein or macromolecular structure that we model is a coarse-grained representation of a distribution over the time scale of the integration step used in Brownian dynamics simulations. Thus, there is no reason to use a very high degree of L to generate an “accurate” surface that is in fact beyond the variance of the surface amplitude fluctuation over the time interval in Brownian dynamics simulations. In fact, a certain degree of fuzziness is especially desired for these computations, a subject to be addressed in the future. We have found that a value of $L = 10–15$ is adequate to provide a compromise of quantitative shape and volume description and sufficient distinguishing details of different proteins. At this level of representation, a total of 121 to 256 terms is needed in the spherical harmonic expansion in eq 1.

To animate the dynamic fluctuation of spherical harmonics coarse-grained OMPDC, we have determined the surface radial displacement amplitudes of the ten lowest quasiharmonic modes of vibration from principal component analysis, which contribute to the overall fluctuation throughout the 8 ns molecular dynamics simulation. The optimized mode weighting factors in eq 11 for the first ten modes are listed in Table 2. The weighting factors are about 0.5 for these low frequency modes, which is a reflection that the approximate “amplitudes” used to define the surface radial

Table 2. Computed Mode Weighting Factors to Represent the Overall Complex Protein Fluctuations Using the First Ten Lowest Quasiharmonic Modes

mode	W_k
1	0.53892
2	0.48497
3	0.48692
4	0.42098
5	0.09645
6	0.05428
7	0.42884
8	0.29127
9	0.33406
10	0.37536

displacement vectors by stretching the quasiharmonic deformation to $\pm 2\sigma$ limits is used so as to obtain a large contrast in the analysis. The minimization procedure reduces the initial large variations to about $\pm 1\sigma$, further suggesting that the procedure employed in the present study is a reasonable approximation to represent the overall protein fluctuation. However, as the frequency (mode number) increases, one standard deviation is not a good measure of the dynamic contributions due to stochastic collision and coupling with fast motions. The small weighting factors for modes 5 and 6 indicate that the associated fluctuations from the principal component analysis may not be well described by quasiharmonic vibrational motions (Figure 4), perhaps due to conformational jumps, or a longer equilibration that is needed in the original MD simulation, or the fact that the explicit molecular dynamics simulation is rather short.

Using the frequencies, phases, and amplitudes optimized using the procedure outlined in section 2 by means of principal component analysis of a molecular dynamics trajectory to train the large amplitude dynamic behavior of the ACG model for OMPDC, we carried out quasiharmonic dynamics animation of the compounded motion of the ten lowest frequency quasiharmonic modes for $0.25 \mu\text{s}$ at an integration increment of 25 ps per step, which took about 1 min on a desktop workstation. Figure 5 shows three structures from the trajectory using the initial conditions listed in Tables 1 and 2. Although it is difficult to distinguish the relatively small surface variations in the static pictures, a movie that is given as Supporting Information (SMovie 1) does provide a more vivid depiction of the dynamic fluctuations of the trajectory. Figure 6 shows the computed volume histogram of the ACG protein (see below). Not surprisingly, the primary periodicity is dictated by the lowest frequency mode, which has the largest amplitude contributions to the complex motion (Tables 1 and 2), and the spectral transform of Figure 6 shows frequencies that coincide with the input listed in Table 1.

Note that although the surface radial displacement vectors were obtained by considering the corresponding quasiharmonic modes of atomic vibrations, the radial vectors do not possess an orthogonality relationship, and the least-squares fitting procedure used to optimize the displacement amplitudes also introduces contributions from other modes not specifically characterized purely by each quasiharmonic mode. Further, the amplitude for each quasiharmonic mode represents an average fluctuation sampled in the original molecular dynamics simulation; however, the maximum

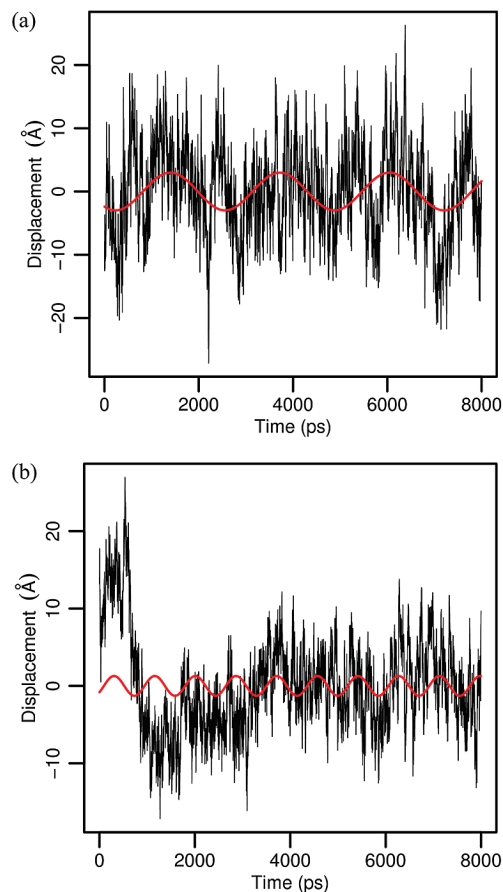


Figure 4. Sinusoidal fit of harmonic frequencies and phase (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 1 for modes number 5 (a) and 6 (b). Note that if the trajectory of the first 1 ns is discarded in mode 6 evaluation, the frequency and amplitude are both reasonable, suggesting there is either a conformational jump or change in the first 1 ns.

fluctuations can be significantly greater than the individual averages due to stochastic collisions with solvent molecules as well as mode coupling. Consequently, stochastic effects may be included in mode synthesis by randomly increasing and decreasing the amplitudes that yield the correct means over a long trajectory and satisfy the condition of the second dissipation theorem.

The fluctuation of the excluded volume defined by the protein surface for OMPDC, which can be conveniently determined by

$$\begin{aligned}
 V &= \int_0^{2\pi} \int_0^\pi \int_0^{S(\theta,\phi)} r^2 \sin \theta \, d\theta \, d\phi \, dr \\
 &= \frac{1}{3} \int_0^{2\pi} \int_0^\pi S^3(\theta,\phi) \sin \theta \, d\theta \, d\phi \quad (17) \\
 &\approx \frac{1}{3} \sum_{ij} S^3(\theta_i, \phi_j) \sin \theta_i w_i w_j
 \end{aligned}$$

is shown in Figure 6 at different degrees of approximation from $L = 5$ to $L = 20$. In eq 17, w_i and w_j are the quadrature weights and the values $\{S(\theta_i, \phi_j)\}$ are already determined during the dynamics animation. There is no

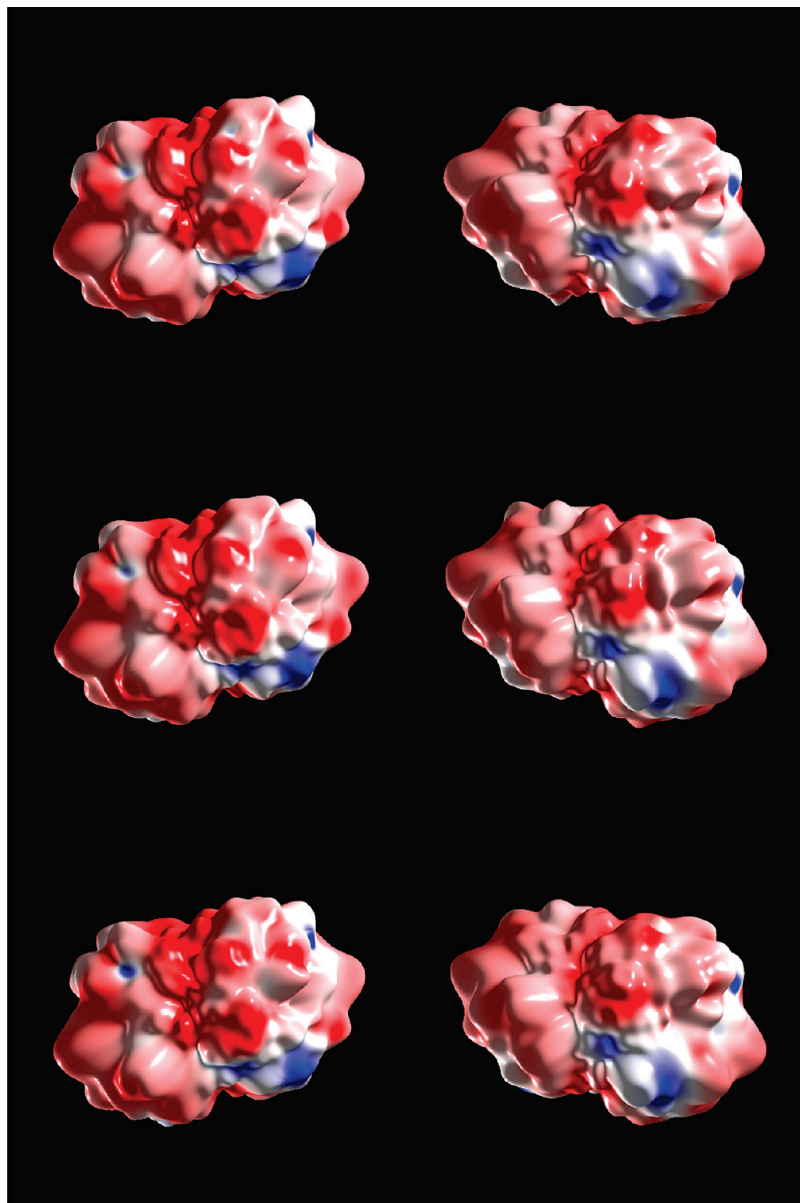


Figure 5. Snapshots of three structures of the analytically coarse-grained (ACG) protein OMPDC using spherical harmonic basis at a representation degree of $L = 15$ from the $0.25 \mu\text{s}$ composite fluctuation trajectory using the amplitudes, frequencies, and phases listed in Tables 1 and 2. The three structures on the right-hand side are the same as the corresponding ones on the left, rotated by 180° . The ACG protein surfaces are colored by the surface charge density for the illustration with red representing negative and blue positive charge densities, respectively.

major difference for the results obtained using $L = 15$ and $L = 20$, suggesting that the use of a spherical harmonic representation of degree 15 is sufficiently accurate to model the molecular volume. At $L = 10$, the average volume is about $84.3 \pm 0.2 \text{ nm}^3$, which is less than 0.4% smaller than an average value of about $84.6 \pm 0.2 \text{ nm}^3$ at higher orders. The lower order at $L = 5$ introduces an error of about 1% in volume. Note that the excluded volume determined by solvent-accessible surface is significantly larger than that encompassed by the van der Waals surface of the macromolecule. Using the initial structure in the molecular dynamics simulation, the ratio between the volumes defined by the solvent-accessible surface and the Bondi van der Waals surface (scaled by

1.20 as proposed by Luque and Orozco⁴⁸ in the calculation of solvation free energies treating the solvent as a polarizable dielectric continuum) is 1.46. If this factor is taken into account, the average molecular density of dry (without solvent molecules) OMPDC is estimated to be $1.315 \pm 0.002 \text{ g/cm}^3$, in excellent agreement with the typical protein density ($1.35\text{--}1.40 \text{ g/cm}^3$) estimated experimentally.^{49,50} If the extra volume enclosed by the solvent-accessible surface is filled with water molecules (about 850 water molecules) at the bulk density, the average density of cavity included in the spherical harmonics coarse-grained protein is estimated to be about 1.20 g/cm^3 , which may be considered as the macromolecular structure solvated by one solvent shell.

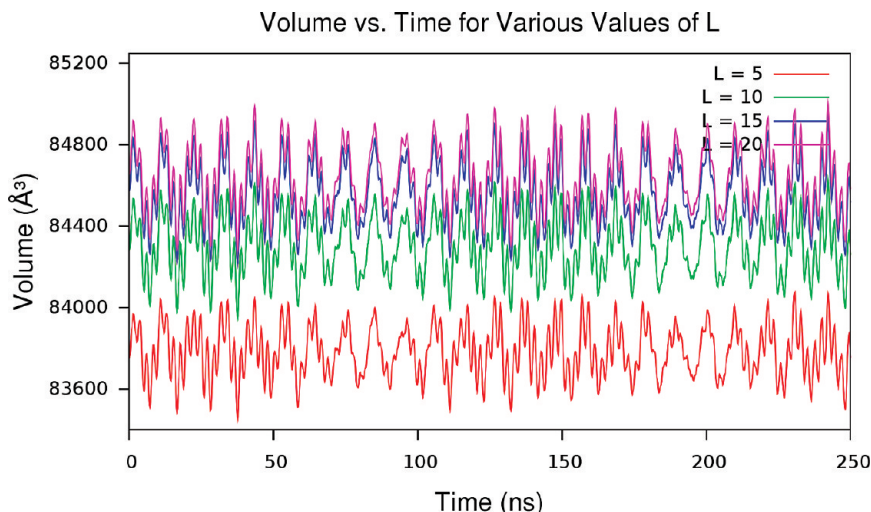


Figure 6. Histogram of the fluctuation of the excluded volume of the ACG OMPDC protein at various resolutions in L , ranging from 5 to 20 along the internal quasiharmonic fluctuation trajectory. The excluded volume illustrated in this figure is defined as the cavity enclosed by the Lee–Richards surface, which is about one solvent layer larger than the van der Waals surface.

5. Conclusion

An analytical coarse-graining (ACG) model has been introduced to represent biological macromolecules, making use of a spherical harmonic basis in the present study. In our approach, a macromolecular structure is treated as a fully coarse-grained entity with a uniform mass density without the explicit description of atomic details or “coarse-grained” interaction sites. The use of a uniform density of the ACG macromolecule is justified because the model represents an ensemble average relevant to the time series used in the dynamics simulation of cellular processes. However, the excluded volume and specific shape of the ACG macromolecule species are critical, which are explicitly treated by a spherical harmonic representation. In principle, spherical harmonic analysis can provide any desired accuracy and detail of the macromolecular surface. The present paper focuses on the first issue in a fully coarse-grained protein model, that is, the description of the internal fluctuation of the ACG macromolecule. Here, we make use of the dimeric enzyme OMPDC, consisting of 416 amino acids and 2 substrate molecules in the active site, as an illustrative example.

The internal fluctuation of the ACG protein is modeled by the superposition of a selected number of lowest frequency quasiharmonic modes of vibration, which are derived from an explicit molecular dynamics simulation of the fully solvated protein in water. A procedure for estimating the amplitudes, time scales (frequencies) of the quasiharmonic motions, and the corresponding phase is presented and used to synthesize the complex motion (note that the eigenvalues of the lowest quasiharmonic modes are close to zero and they are not quantitative for description of the time scales of the corresponding motions). In principle, all modes up to a frequency, limited by the time interval of the coarse-grained dynamics, can be included, but as numerous studies have shown, when employing principal component analysis and quasiharmonic essential dynamics, only a fraction of the lowest frequency modes are important in such a representation. The analytical description and numerical algorithm presented here can in principle provide a representation of the internal protein fluctuations as closely as needed in comparison with the atomistic molecular dynamics

simulation; however, the internal motion is restricted by the short-time nature of molecular dynamic trajectories, and the present method is not designed for the description of unfolding events unless such transitions occur during the molecular dynamics simulation.

Appendix: Algorithm for Spherical Harmonic Expansion and Evaluation

In this Appendix, we summarize the numerical procedure for spherical harmonic expansion and evaluation used in our implementation; additional details may be found in ref 13 (see also refs 47 and 51). The numerical methods are widely used in geopotential modeling and an expansion of degree, and order up to 3800 has been reported corresponding to a ground resolution of about 5 km^{52} (for the protein OMPDC considered here, it would correspond to an astonishing surface resolution of about 0.005 \AA).

It is useful to recast eq 1 for the surface function $S(\theta, \phi)$ in terms of harmonic coefficients as follows:

$$S(\theta, \phi) = \sum_{m=0}^L \sum_{l=m}^L a_{lm}^c \bar{P}_l^m(\cos \theta) \cos m\phi + a_{lm}^s \bar{P}_l^m(\cos \theta) \sin m\phi \quad (\text{A1})$$

Notice also that the order of the summations over degree and order has been switched. This is especially important in modern spherical harmonic analysis because the latitude and longitude data can now be treated independently, resulting in a two-step computational algorithm.^{13,14,47} The expansion coefficients are determined based on a set of sampling data $\{S(\theta_i, \phi_j)\}$ on a grid of equispaced $2M$ points in ϕ (for Fourier transform) and M Gauss–Legendre quadrature nodes in $\cos \theta$ (for integration):

$$a_{lm}^c = \int_0^\pi \left\{ \int_0^{2\pi} \frac{1}{\sqrt{(1 + \delta_{m0})\pi}} S(\theta, \phi) \cos(m\phi) d\phi \right\} \bar{P}_l^m(\cos \theta) \sin \theta d\theta \quad (\text{A2})$$

$$a_{lm}^s = \int_0^\pi \left\{ \int_0^{2\pi} \frac{1}{\sqrt{(1 + \delta_{m0})\pi}} S(\theta, \phi) \sin(m\phi) d\phi \right\} \bar{P}_l^m(\cos \theta) \sin \theta d\theta \quad (\text{A3})$$

The grid points are defined as follows:

$$\phi_j = j\Delta\phi = j\frac{\pi}{M}; \quad j = 0, 1, \dots, 2M - 1 \quad (\text{A4})$$

$$P_{M+1}(\cos \theta_i) = 0; \quad i = 1, 2, \dots, M + 1 \quad (\text{A5})$$

where $P_{M+1}(\cos \theta_i)$ is the Legendre polynomial. The Gauss–Legendre quadrature weights can be determined using the expression^{40,45}

$$w_i = 2 \left[\frac{\sin(\theta_i)}{(M + 1)P_M(\cos \theta_i)} \right]^2 \quad (\text{A6})$$

From this set of data, the maximum degree of the expansion coefficients that can be determined is $L = M$ because the maximum number of terms in the summation over l in eq A1, which is $L + 1$ when $m = 0$, must be less than or equal to the number of sampling points in the longitudinal direction θ . Here, we do not address the issue of aliasing,⁵³ and we typically use a larger number of sampled grid points than the maximum degree used in the expansion.

Alternatively, if equal spaced points are used in θ , which is equivalent to Chebychev nodes in $\cos \theta$, at least $2M + 1$ sampling points are needed for degree L in the expansion coefficients as opposed to a minimum of $M = L + 1$ points with the Gauss–Legendre quadrature. In this case, the Chebychev weights are obtained using the formula below:^{47,54,55}

$$w_i = \frac{\sqrt{2}}{M} \sum_{l=0}^{M-1} \frac{1}{2l + 1} \sin([2l + 1]\theta_i) \quad (\text{A7})$$

For convenience in the rest of the discussion, we use the degree of the expansion L to define the grid divisions throughout.

A.1. Spherical Harmonic Expansion. For the spherical harmonics expansion, a two-step computation algorithm is used. The first step corresponds to a Fourier transform in the inner integrals in eqs A2 and A3. For a given value m , the discretized Fourier series in the inner integrals of eqs A2 and A3 are expressed as follows:

$$U(\theta_i, \hat{m}) = \frac{1}{\sqrt{(1 + \delta_{m0})\pi}} \sum_{j=0}^{2L-1} S(\theta_i, \phi_j) \cos m\phi_j; \quad i = 1, \dots, L + 1 \quad (\text{A8})$$

$$V(\theta_i, \hat{m}) = \frac{1}{\sqrt{(1 + \delta_{m0})\pi}} \sum_{j=0}^{2L-1} S(\theta_i, \phi_j) \sin m\phi_j; \quad i = 1, \dots, L + 1 \quad (\text{A9})$$

where δ_{m0} is the Kronecker delta, and the notation \hat{m} is used to emphasize that the Fourier series can be efficiently performed by Fast Fourier Transform (FFT).

The second step involves integration by Gauss–Legendre quadrature (or equally spaced Chebychev quadrature which

requires twice as many sampling points) to yield the $2 \times (L - m + 1)$ expansion coefficients:

$$a_{lm}^c = \sum_{i=1}^{L+1} w_i \bar{P}_m^l(\cos \theta_i) U(\theta_i, \hat{m}); \quad l = m, \dots, L \quad (\text{A10})$$

$$a_{lm}^s = \sum_{i=1}^{L+1} w_i \bar{P}_m^l(\cos \theta_i) V(\theta_i, \hat{m}); \quad l = m, \dots, L \quad (\text{A11})$$

If the values $U(\theta_i, \hat{m})$ and $V(\theta_i, \hat{m})$ are arranged as column vectors $\mathbf{u}(\hat{m})$ and $\mathbf{v}(\hat{m})$, respectively, the above equations can be conveniently written in matrix form:

$$\mathbf{a}_m^c = \mathbf{P}(m)^T \mathbf{W} \mathbf{u}(\hat{m}) \quad (\text{A12})$$

$$\mathbf{a}_m^s = \mathbf{P}(m)^T \mathbf{W} \mathbf{v}(\hat{m}) \quad (\text{A13})$$

where \mathbf{a}_m^c and \mathbf{a}_m^s are the expansion coefficients of eqs A10 and A11 arranged as column vectors, $\mathbf{W} = \text{diag}\{w_i\}$ is an $(L + 1) \times (L + 1)$ diagonal matrix consisting of the quadrature weights and the matrix for the precomputed values of the normalized associated Legendre polynomial is arranged as follows:

$$\mathbf{P}(m) = \begin{pmatrix} \bar{P}_m^m(\cos \theta_1) & \cdots & \bar{P}_L^m(\cos \theta_1) \\ \cdots & \cdots & \cdots \\ \bar{P}_m^m(\cos \theta_{L+1}) & \cdots & \bar{P}_L^m(\cos \theta_{L+1}) \end{pmatrix} \quad (\text{A14})$$

The operation for the first step has a computation scale of $O(L \log L)$ using FFT, whereas the second step is of $O(L^2)$ for each order m . Thus, the overall procedure scales $O(L^2 \log L) + O(L^3)$. Obviously, the $L + 1$ parallels can be fully distributed over different processors, each having an overall computational scaling of $O(L^2)$; this is particularly suited for GPUs by choosing the number of parallels equal to that of the processors. Note that a fast spherical harmonic transform algorithm similar to that of FFT has been described.^{47,54}

A.2. Spherical Harmonic Evaluation. Evaluation (or spherical harmonic synthesis) of surface function values also involves two computational steps. For a fixed colatitude θ_i , the first step is to compute intermediate vectors $\hat{\mathbf{u}}(\theta_i)$ and $\hat{\mathbf{v}}(\theta_i)$ over l for $0 \leq m \leq L$:

$$\hat{\mathbf{u}}(\theta_i) = \mathbf{P}_i(m) \mathbf{a}_m^c \quad (\text{A15})$$

$$\hat{\mathbf{v}}(\theta_i) = \mathbf{P}_i(m) \mathbf{a}_m^s \quad (\text{A16})$$

In the second step, the $2L$ longitudinal values are computed by Fast Fourier Transform for the following discrete series:

$$S(\theta_i, \phi_j) = \sum_{m=0}^L \hat{U}(\theta_i, m) \cos m\phi_j + \hat{V}(\theta_i, m) \sin m\phi_j; \quad j = 0, \dots, 2L - 1 \quad (\text{A17})$$

The overall computational scaling is also $O(L^3)$, which can be distributed to L processors as the two computational steps are fully independent. The use of parities in the construction of the associated Legendre polynomials reduces computation by a factor of 2.¹³

Acknowledgment. This work is partially supported by the National Institutes of Health, Grant GM46736.

Supporting Information Available: Projected structural fluctuations on to the first 20 lowest frequency modes and a movie animating the trajectory of the dynamic fluctuation of the ACG protein OMPDC represented by using spherical harmonic basis functions. This material is available free of charges via the Internet at <http://pubs.acs.org>.

References

- (1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646.
- (2) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. *Structure* **2006**, *14*, 437.
- (3) Dror, R. O.; Jensen, M. O.; Borhani, D. W.; Shaw, D. E. *J. Gen. Physiol.* **2010**, *135*, 555.
- (4) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526.
- (5) Dunning, T. H.; Schulten, K.; Tromp, J.; Ostriker, J. P.; Droegemeier, K.; Xue, M.; Fussell, P. *Comput. Sci. Eng.* **2009**, *11*, 28.
- (6) Arkhipov, A.; Roos, W. H.; Wuite, G. J. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2061.
- (7) Lu, L. Y.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. *J. Chem. Theory Comput.* **2010**, *6*, 954.
- (8) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; Taylor & Francis Group: Boca Raton, FL, 2008.
- (9) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585.
- (10) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
- (11) Jekeli, C.; Lee, J. K.; Kwon, J. H. *J. Geodesy* **2007**, *81*, 603.
- (12) Wiczorek, M. A. *Treatise Geophys.* **2007**, *10*, 165.
- (13) Colombo, O. Numerical Methods for Harmonic Analysis on the Sphere, Ph.D. Thesis, Ohio State University, 1981.
- (14) Sneeuw, N. *Geophys. J. Int.* **1994**, *118*, 707.
- (15) Max, N. L.; Getzoff, E. D. *IEEE Comput. Graph. Appl.* **1988**, *8*, 42.
- (16) Duncan, B. S.; Olson, A. J. *Biopolymers* **1993**, *33*, 231.
- (17) Duncan, B. S.; Olson, A. J. *Biopolymers* **1993**, *33*, 219.
- (18) Duncan, B. S.; Olson, A. J. *J. Mol. Graph.* **1995**, *13*, 258.
- (19) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. *Biopolymers* **1996**, *38*, 305.
- (20) Ritchie, D. W.; Kemp, G. J. L. *J. Comput. Chem.* **1999**, *20*, 383.
- (21) Cai, W. S.; Shao, X. G.; Maigret, B. *J. Mol. Graph. Model.* **2002**, *20*, 313.
- (22) Morris, R. J.; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. *Bioinformatics* **2005**, *21*, 2347.
- (23) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. *J. Chem. Inf. Model.* **2007**, *47*, 1787.
- (24) Mak, L.; Grandison, S.; Morris, R. J. *J. Mol. Graph. Model.* **2008**, *26*, 1035.
- (25) Venkatraman, V.; Sael, L.; Kihara, D. *Cell Biochem. Biophys.* **2009**, *54*, 23.
- (26) Buchete, N. V.; Straub, J. E.; Thirumalai, D. *J. Mol. Graph. Model.* **2004**, *22*, 441.
- (27) Duncan, B. S.; Olson, A. J. *J. Mol. Graph.* **1995**, *13*, 250.
- (28) Brooks, B.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 4995.
- (29) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* **1993**, *17*, 412.
- (30) Brooks, B. R.; Janezic, D.; Karplus, M. *J. Comput. Chem.* **1995**, *16*, 1522.
- (31) Janezic, D.; Brooks, B. R. *J. Comput. Chem.* **1995**, *16*, 1543.
- (32) Janezic, D.; Venable, R. M.; Brooks, B. R. *J. Comput. Chem.* **1995**, *16*, 1554.
- (33) Wu, N.; Mo, Y.; Gao, J.; Pai, E. F. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2017.
- (34) Dey, T. K.; Giesen, J.; Goswami, S. *Algorithms Data Struct. Proc.* **2003**, 2748, 25.
- (35) Klees, R.; Tenzer, R.; Prutkin, I.; Wittwer, T. *J. Geodesy* **2008**, *82*, 457.
- (36) Arkhipov, A.; Freddolino, P. L.; Imada, K.; Namba, K.; Schulten, K. *Biophys. J.* **2006**, *91*, 4589.
- (37) Izvekov, S.; Swanson, J. M. J.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 4711.
- (38) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.
- (39) Connolly, M. L. *Science* **1983**, *221*, 709.
- (40) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; University of Cambridge: New York, 1992.
- (41) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D. *Nature* **2007**, *450*, 913.
- (42) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806.
- (43) Zhang, Z.; Lu, L. J.; Noid, W. G.; Krishna, V.; Plaentner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073.
- (44) Zhang, Z.; Plaentner, J.; Grafmuller, A.; Voth, G. A. *Biophys. J.* **2009**, *97*, 2327.
- (45) Stroud, A. H.; Secrest, D. *Gaussian Quadrature Formulas*; Prentice-Hall Inc.: Englewood Cliffs, NJ, 1966.
- (46) Sloan, I. H.; Womersley, R. S. *Adv. Comput. Math.* **2004**, *21*, 107.
- (47) Mohlenkamp, M. J. *J. Fourier Anal. Appl.* **1999**, *5*, 159.
- (48) Luque, F. J.; Bachs, M.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 847.
- (49) Fischer, H.; Polikarpov, I.; Craievich, A. F. *Protein Sci.* **2004**, *13*, 2825.
- (50) White, E. T.; Tan, W. H.; Ang, J. M.; Tait, S.; Lister, J. D. *Powder Technol.* **2007**, *179*, 55.
- (51) Blais, J. A. R.; Provins, D. A.; Soofi, M. A. *J. Supercomput.* **2006**, *38*, 173.
- (52) Balis, J. A. R. In *Computational Science - ICCS 2008 Part II*; Bubak, M., van Albada, G. D., Dongarra, J., Eds.; Springer-Verlag: Berlin, 2008; Vol. LNCS 5102, p 638.
- (53) Jekeli, C. *J. Geodesy* **1996**, *70*, 214.
- (54) Mohlenkamp, M. J., A Fast Transform for Spherical Harmonics. Ph.D. Thesis, Yale University, 1997.
- (55) Driscoll, J. R.; Healy, D. M. *Adv. Appl. Math.* **1994**, *15*, 202.