

## Static and Dynamical Properties of Liquid Water from First Principles by a Novel Car–Parrinello-like Approach

Thomas D. Kühne,<sup>\*,†</sup> Matthias Krack,<sup>†,‡</sup> and Michele Parrinello<sup>†</sup>

Computational Science, Department of Chemistry and Applied Biosciences, ETH Zurich, USI Campus, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

Received October 3, 2008

**Abstract:** Using the recently developed Car–Parrinello-like approach to Born–Oppenheimer molecular dynamics (Kühne, T. D.; et al. *Phys. Rev. Lett.* 2007, 98, 066401.), we assess the accuracy of ab initio molecular dynamics at the semilocal density functional level of theory to describe structural and dynamic properties of liquid water at ambient conditions. We have performed a series of large-scale simulations using a number of parameter-free exchange and correlation functionals, to minimize and investigate the influence of finite size effects as well as statistical errors. We find that finite size effects in structural properties are rather small and, given an extensive sampling, reproducible. On the other hand, the influence of finite size effects on dynamical properties are much larger than generally appreciated. So much so that the infinite size limit is practically out of reach. However, using a finite size scaling procedure, thanks to the greater effectiveness of our new method we can estimate both the thermodynamic value of the diffusion coefficient and the shear viscosity. The hydrogen bond network structure and its kinetics are consistent with the conventional view of tetrahedrally coordinated water.

### 1. Introduction

Since the first applications of molecular dynamics (MD) to realistic systems,<sup>1</sup> liquid water has been one of the most studied systems, as it is arguably the most important liquid for its role in biology, chemistry, and geophysics. This widespread interest has also sparked many controversies, the most recent being the suggestion that the average coordination of water is 2 rather than 4 as the tetrahedral coordination of ice and the nature of the hydrogen bond (HB) would suggest.<sup>2</sup> A detailed understanding of liquid water is therefore essential and at the same time demanding due to its complex behavior and unusual properties.<sup>3</sup> However, simulating water is rather challenging due to the presence of a number of difficulties in modeling the various physical effects that conspire to make water unique, such as sizable quantum corrections, large permanent dipoles and strong polarizability effects, and the cooperativity of the HB network.

Much effort has gone into developing empirical potentials for water capable of describing all of these effects and much progress has been reported.<sup>4–10</sup> However, often the transferability of these potentials to regions of the phase diagram or systems different from that in which they have been fitted is restricted. Furthermore, they are not able to simulate with sufficient predictive power chemical processes that take place in water solutions.

A first-principles based approach, like density functional theory (DFT)<sup>11</sup> based ab initio molecular dynamics (AIMD), where the forces are evaluated on the fly from accurate electronic structure calculations, is very attractive since many of these limitations can in principle be removed. However, also the ab initio approach is not without problems: the relevant energy scale is very small and an error of 0.3 kcal/mol might cause the simulated water either to freeze or to evaporate. This poses very stringent accuracy constraints. Furthermore, the computational cost associated with AIMD has forced in the past numerical approximations to extend the attainable size and length scales of the simulation, and a significant dependence on numerical details has been reported.<sup>12–19</sup>

\* Corresponding author. E-mail: tkuehne@phys.chem.ethz.ch.

<sup>†</sup> ETH Zurich.

<sup>‡</sup> Current address: Paul Scherrer Institute, CH-5232 Villigen PSI, Switzerland.

In view of great potential impact and widespread interest, it is of great value to assess its intrinsic properties as distinct from those that descend from numerical approximations, finite size effects, and insufficient sampling. This is now made possible by a new simulation method<sup>20</sup> which is highly accurate and at the same time at least 1 order of magnitude more efficient than previous ones. In this work, we focus on the Perdew–Burke–Ernzerhof (PBE) approximation to the exact exchange and correlation (XC) functional, whose parameters have been determined from first principles.<sup>21</sup> We find that ordinary PBE water at 300 K is somewhat overstructured and less fluid than real water, with a shear viscosity (here calculated for the first time) which is within a factor of 3 of the experimental one. Given the fact that here nuclear quantum effects are not included, one can conclude that PBE does provide a qualitative realistic model for water–water interactions.

The long runs (~250 ps) on relatively large systems (128 molecules) allow for the first time a number of dynamical properties to be evaluated with accuracy. We have already mentioned shear viscosity, and to this we can add the self-diffusion coefficient, which exhibits stronger than expected finite size effects, but whose asymptotic value for an infinite large system is now evaluated. We also calculate the single-particle velocity–velocity autocorrelation function and the corresponding power spectrum. In view of the current controversy, we have also examined the HB network and its dynamics. But eventually PBE water is consistent with the conventional wisdom, and this conclusion is not altered if other XC functionals are used.

However, before discussing our results we briefly summarize the principles that are at the bases of our method and is the key to our successful investigation and which has already found a number of practical applications.<sup>22,23</sup>

## 2. Second Generation Car–Parrinello MD

Contrary to the direct Born–Oppenheimer MD (BOMD),<sup>24</sup> in which during the dynamics the energy functional is fully optimized subject to the orthonormality constraint  $\langle \psi_i(r) | \psi_j(r) \rangle = \delta_{ij}$ , in the Car–Parrinello MD (CPMD)<sup>25</sup> approach this is circumvented by designing a suitable electron-ion dynamics in which, under appropriate conditions, the electrons follow adiabatically the ions very close to the instantaneous electronic ground state. An important role in this approach is played by  $\mu$ , the fictitious electronic mass, which has to be chosen small enough to ensure this adiabatic decoupling.<sup>26</sup> As a consequence, the maximal permissible integration time step  $\Delta t$  has to be considerably smaller than in BOMD.<sup>27</sup> In our novel approach the original fictitious Newtonian dynamics of the electronic degrees of freedom is replaced by another, very close to the Born–Oppenheimer (BO) dynamics, which does not require the introduction of any artificial mass. At variance with CPMD, where the electronic dynamics is defined by the introduction of a Lagrangian, the coupled electron–ion dynamics is directly specified. These are based on the always stable predictor–corrector method<sup>28,29</sup> of Kolafa, though other choices are equally possible.<sup>30,31</sup>

Let us represent the set of electronic wave functions in the form of a coefficient matrix  $\mathbf{C}$  that has the dimension of the basis set size times the number of occupied states, which is assumed to be nonorthogonal. Thus, the overlap matrix  $\mathbf{S}$  is different from unity. In terms of  $\mathbf{C}$  and  $\mathbf{S}$  we can now define the contravariant density matrix  $\mathbf{PS}$ , where  $\mathbf{P} = \mathbf{C}\mathbf{C}^T$  is the one-particle density kernel. One expects that the dynamics of  $\mathbf{PS}$  to be much smoother than that of the more widely varying wave functions even in the case of metals, where many states crowd the Fermi level. This consideration therefore suggests to propagate  $\mathbf{PS}$ , rather than the wave functions as in CPMD.

**2.1. Coupled Electron–Ion Dynamics.** Adapting Kolafa’s method to this particular case, we write the predicted wave function at time  $t_n$  in terms of  $K$  previous  $\mathbf{PS}$  matrices as

$$\mathbf{C}^p(t_n) \cong \sum_{m=1}^K (-1)^{m+1} m \frac{\binom{2K}{K-m}}{\binom{2K-2}{K-1}} \mathbf{P}(t_{n-m}) \mathbf{S}(t_{n-m}) \mathbf{C}(t_{n-1}) \quad (1)$$

The corresponding corrector involves the evaluation of only one preconditioned electronic gradient  $\text{MIN}[\mathbf{C}(t)]$  using the orbital transformation (OT) method of VandeVondele and Hutter.<sup>32</sup> This leads to the corrected  $\mathbf{C}(t_n)$ :

$$\mathbf{C}(t_n) = \omega \text{MIN}[\mathbf{C}^p(t_n)] + (1 - \omega) \mathbf{C}^p(t_n) \quad (2)$$

$$\text{with } \omega = \frac{K}{2K-1} \text{ where } K \geq 2$$

Such a predictor–corrector scheme leads to an electron dynamics which is accurate and time reversible up to  $\mathcal{O}(\Delta t^{2K-2})$ , where  $\Delta t$  is the integration time step. The efficiency of this approach is such that the ground state is very closely approached within just one such step. We thus totally avoid the self-consistency cycle and any expensive diagonalization, while remaining very close to the BO surface and allow for  $\Delta t$  to be as large as in standard MD.

Nevertheless, a small dissipative energy drift is introduced that needs to be corrected. To this effect, in ref 18 we have shown, that an excellent model for the ionic forces thus calculated is  $\mathbf{F}_I^{\text{PC}} = \mathbf{F}_I^{\text{BO}} - \gamma_D M_I \dot{\mathbf{R}}_I$ , where  $\mathbf{F}_I^{\text{BO}}$  are the correct BO forces,  $\gamma_D$  a friction coefficient,  $M_I$  the ionic masses, and  $\mathbf{R}_I$  the ionic coordinates. The presence of damping suggests a canonical sampling of the Boltzmann distribution based on a Langevin approach, rather than a microcanonical one. We therefore introduce a properly modified Langevin equation,<sup>33</sup> in which for convenience an additional friction coefficient  $\gamma_L$  is present. These damping terms are compensated by an additive white noise  $\Xi_I(t)$ , which is related to  $\gamma = \gamma_D + \gamma_L$  by the fluctuation–dissipation theorem  $\langle \Xi_I(0) \Xi_I(t) \rangle = 2\gamma M_I k_B T \delta(t)$ , thus leading to an exact sampling.

In order to determine the unknown value of  $\gamma_D$ , we perform a preliminary run in which we vary  $\gamma_D$  on the fly using a Berendsen-like algorithm<sup>34</sup> until the equipartition theorem is satisfied. This can be somewhat laborious, but once  $\gamma_D$  is fixed, very long and accurate simulation runs can be performed at a much reduced computational cost, thus unifying the best of conventional BOMD and CPMD, since

one uses large time steps, while at the same time preserving the Car–Parrinello efficiency.

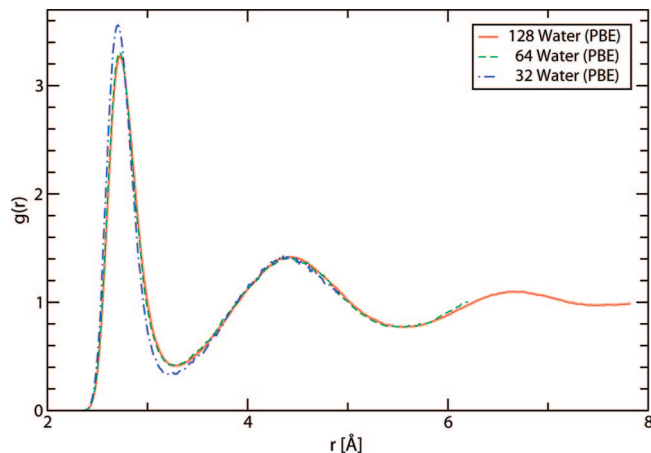
### 3. Computational Details

In the largest simulated system, we take a periodic cubic box of length  $L = 15.6627 \text{ \AA}$  consisting of 128 light water molecules that corresponds to a density that differs by only 0.3% from the experimental value. All simulations are performed at 300 K; the Langevin dynamics is integrated using the algorithm of Ricci and Ciccotti.<sup>34</sup> The discretized integration time step  $\Delta t = 0.5 \text{ fs}$ , while  $\gamma_D = 8.65 \times 10^{-5} \text{ fs}^{-1}$  and  $\gamma_L = 1.35 \times 10^{-5} \text{ fs}^{-1}$ . The simultaneous propagation of the electronic degrees of freedom proceeds with  $K = 7$ , which yields a time reversibility of  $\mathcal{O}(\Delta t^{12})$ . At each MD step the corrector is applied only once, which implies just one preconditioned gradient calculation. The deviation from the BO surface, as measured by the preconditioned mean gradient deviation is  $10^{-5} \text{ au}$ , which is only slightly larger than typical values used in fully converged BOMD simulations.

Since we are dealing with a disordered system at finite temperature that also exhibits a large band gap, the Brillouin zone is sampled at the  $\Gamma$ -point only. Furthermore, separable and norm-conserving pseudopotentials are used to describe the interactions between the valence electrons and the ionic cores.<sup>35–37</sup>

Long and well-equilibrated trajectories are necessary in order to obtain an accurate sampling. This requirement is made even more stringent by the strong dependence of the translational self-diffusion coefficient on temperature and, in the case of PBE water, on the expected low diffusivity at room temperature.<sup>16,38</sup> Unless otherwise stated, we use the PBE generalized gradient approximation to the XC energy. In each run we equilibrate carefully the system for 25 ps and accumulate statistics in the successive 250 ps. Finite size effects are studied by comparing the results of the largest system with equally long runs on 64 and 32 water molecules. For the purpose of addressing the accuracy of our simulations, we have carried out two shorter, 25 ps long, reference calculations with 128 molecules, using fully self-consistent BO forces and either Newtonian or Langevin dynamics. Otherwise, the settings for both runs were identical and started from the same well-equilibrated configuration. We also investigated the influence of the XC functional in a series of additional runs using a variety of different semilocal XC functionals, either parameter-free,<sup>39,40</sup> or empirically parameterized.<sup>41–43</sup> In each of these reference runs, statistics were accumulated for at least 30 ps after an equilibration of 20 ps, totaling more than 1 ns of AIMD simulations.

All simulations were performed using the mixed Gaussian and plane wave<sup>44</sup> code CP2K/Quickstep.<sup>45</sup> In this approach, the Kohn–Sham (KS) orbitals are expanded in Gaussians, while for the electron density a plane wave basis is used. Exploiting the efficiency of transformation methods to alternate between one representation and the other, together with advanced multigrid, sparse matrix and screening techniques, an efficient linear-scaling evaluation of the KS matrix is obtained. Efforts towards a full linear scaling algorithm are underway.<sup>47</sup> Here the orbitals are represented



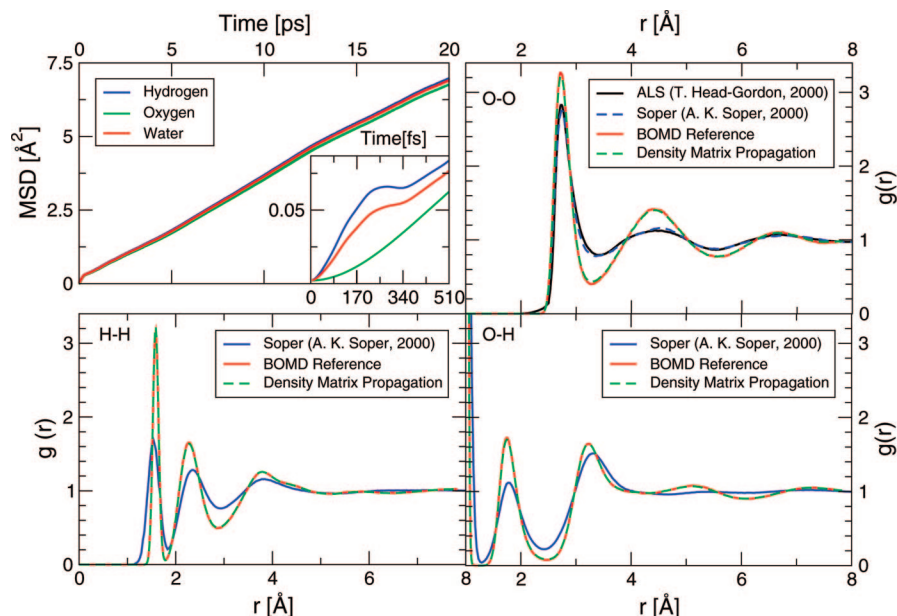
**Figure 1.** Comparison of the oxygen–oxygen PCF's, obtained from AIMD simulations, using 32, 64, or 128 water molecules.

by an accurate triple- $\zeta$  basis set with two set of polarization functions (TZV2P),<sup>48</sup> while a density cutoff of 320 Ry is used for the charge density. The use of a position-dependent basis set inevitably entails a basis set superposition error (BSSE). In our case the BSSE, as estimated by the average error in the binding energy of a single water molecule within the bulk against counterpoise corrected reference calculations at the complete basis set limit, corresponds to an approximately constant energy shift of 0.3 kcal/mol. Since the associated standard deviation is well below 0.01 kcal/mol, the nuclear forces as well as the chemical relevant energy differences are basically unaffected.

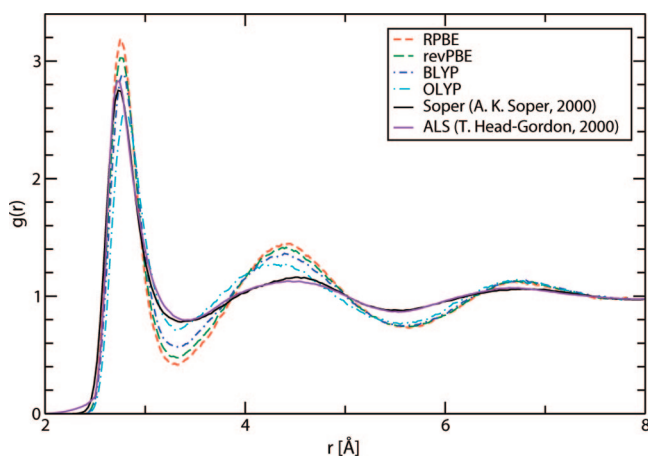
## 4. Results and Discussion

**4.1. Structural Properties.** In Figure 1 the oxygen–oxygen pair-correlation function (PCF)  $g_{OO}(r)$  evaluated on systems with different numbers of molecules are compared to assess the convergence with respect to system size. We find that, using 32 water molecules, errors due to finite size effects are not entirely negligible. However, already in the 64-molecule system the  $g_{OO}(r)$  coincides, within statistical uncertainty, with that of the larger 128-molecule calculation, providing results that are converged with respect to system size.

In Figure 2 the partial pair correlation functions are compared to recent X-ray scattering<sup>51</sup> and neutron diffraction<sup>48</sup> data, but also to BOMD reference calculations in order to assess the accuracy of our novel approach. The agreement with the reference BOMD calculation is excellent, and the results are consistent with those of others.<sup>12,14,16,38</sup> Comparison with experiments reveals a general overstructuring and an oxygen–hydrogen PCF, whose relative heights of the first two intermolecular peaks are reversed. However, the inclusion of nuclear quantum effects is expected to further improve the agreement with experiments,<sup>52</sup> though probably less than earlier calculations suggested.<sup>53</sup> In the literature, PCF's in slightly better agreement with experiment have been reported; however, such calculations have been performed either at higher temperatures<sup>12,13,15,16</sup> or using different XC functionals.<sup>17–19</sup>



**Figure 2.** Partial PCF's of liquid water at ambient conditions and its mean square displacement (top left panel). From the enclosed inset the onset of a cage effect can be observed at  $\sim 250$  fs followed by diffusion, which is in excellent agreement with Gallo et al.<sup>49</sup>



**Figure 3.**  $g_{00}(r)$ , as obtained from neutron diffraction, X-ray scattering, and AIMD simulations using a variety of different XC functionals.

**Table 1.** First Maximum and Minimum Peak Heights in the  $g_{00}(r)$ , Their Positions  $r_{00}$ , and Coordination Number  $N_c$  with Respect to the XC Functional

XC	$g_{00}^{\max}(r)$	$r_{00}^{\max}$	$g_{00}^{\min}(r)$	$r_{00}^{\min}$	$N_c$
PBE <sup>21</sup>	3.25	2.73	0.44	3.28	4.04
RPBE <sup>39</sup>	3.19	2.75	0.42	3.32	4.03
revPBE <sup>40</sup>	3.01	2.77	0.50	3.31	4.05
BLYP <sup>41-43</sup>	2.92	2.79	0.57	3.33	4.09
OLYP <sup>42-43</sup>	2.57	2.79	0.71	3.30	3.90
ALS <sup>50</sup>	2.83	2.73	0.80	3.4	4.7
Soper <sup>51</sup>	2.75	2.73	0.78	3.36	—
HASY <sup>54-55</sup>	2.58	2.76	0.83	3.40	—
PCCP <sup>50,55</sup>	2.31	2.78	0.84	3.39	—

We have also performed a series of simulations using various XC functionals, as reported in Figure 3 and detailed in Table 1, that yields a coordination number  $\sim 4.0 \pm 0.1$  and confirms a sizable functional dependence. Nevertheless, in spite of the observed variations, and given the challenge

of simulating water from first principles, all in all the performance of DFT can be judged to be encouragingly good.

The length of our simulations allows for a careful error estimation. To this end, we have randomly selected 1000 segments along the trajectory, each 25 ps long, in which we have calculated the  $g_{00}(r)$ . The length of each segment is chosen to be longer than the correlation time estimated by Grossman et al.<sup>38</sup> The fluctuations of the  $g_{00}(r)$  in each segment relative to the average of the whole trajectory are within 2 standard deviations. This shows that given a sufficient sampling the errors in our simulations are negligible, and results from AIMD calculations are reproducible.

**4.2. Hydrogen Bond Network.** The textbook picture of bonding in liquid water indicates that each water molecule sits on average in a tetrahedral cage formed by a local, but macroscopically extended, HB network that is continually deformed by the dynamic breaking and re-forming of HB's. But, as already mentioned, this view has been recently challenged by Wernet et al. and is matter of current debate.<sup>56-60</sup> Specifically, based on X-ray absorption and Raman scattering the claim is that at ambient conditions  $>80\%$  of the HB's are broken, leading to a liquid water coordination of  $\sim 2$ . This would imply, rather surprisingly, that liquid water predominantly consists of chains or rings.

To check whether their interpretation is coherent with our data, we use their HB definition. The corresponding results are summarized in Table 2. We find that all of our calculations support a tetrahedral arrangement, even in the case of the OLYP functional, which has the largest number of broken HB's. Since the definition of a HB is somewhat arbitrary we have also applied alternative HB definitions of Luzar and Chandler<sup>61</sup> as well as Kuo and Mundy.<sup>62</sup> The results are very similar and do not change if one slightly varies the cutoff radius or even introduces an additional persistence time criterion in the definition of HB.

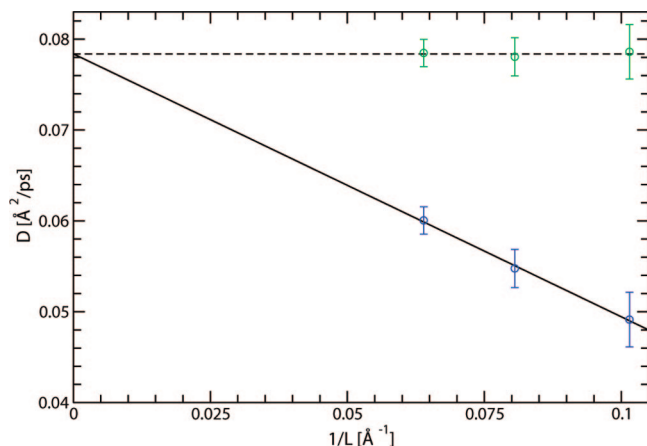
**Table 2.** Relative Occurrence of Double Donor (DD), Single Donor (SD), No Donor (ND), Donating and Free HB's, As Obtained from AIMD Simulations Averaged over Half a Million Snapshots Using Several Semilocal XC Functionals

	PBE	RPBE	revPBE	BLYP	OLYP
DD	82.8%	81.4%	76.8%	72.9%	59.0%
SD	16.6%	17.8%	22.0%	25.4%	36.3%
ND	0.7%	0.8%	1.2%	1.7%	4.7%
donated HB's	91.0%	90.3%	87.8%	85.6%	77.1%
free HB's	9.0%	9.7%	12.2%	14.4%	22.9%
mean HB's	3.642	3.613	3.513	3.423	3.085

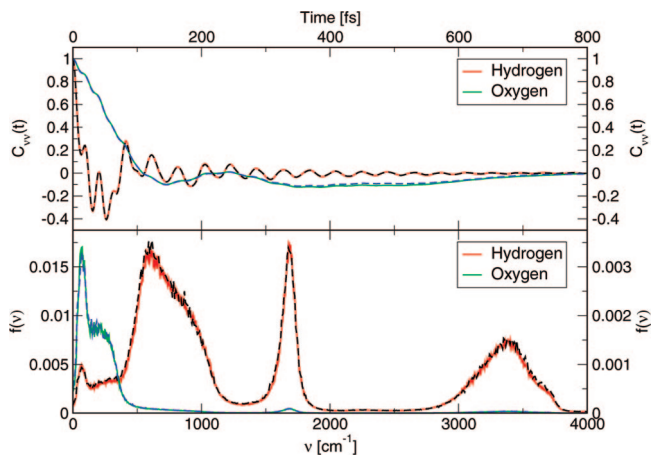
**4.3. Dynamic Properties.** Having gathered enough statistics on systems of different sizes we now address the issue of size dependence of the translational self-diffusion  $D$ . This arises from the fact that a diffusing particle sets up a hydrodynamic flow, which decays as slowly as  $1/r$ . In a periodically repeated system this leads to an interference between one particle and its periodic images. This effect has been analyzed by Dünweg and Kremer,<sup>63,64</sup> who have established the following relation for the diffusion coefficient under periodic boundary conditions as a function of simulation box length  $L$ :

$$D_{\text{PBC}}(L) = D_{\infty} - \frac{k_{\text{B}}T\zeta}{6\pi\eta L} \quad (3)$$

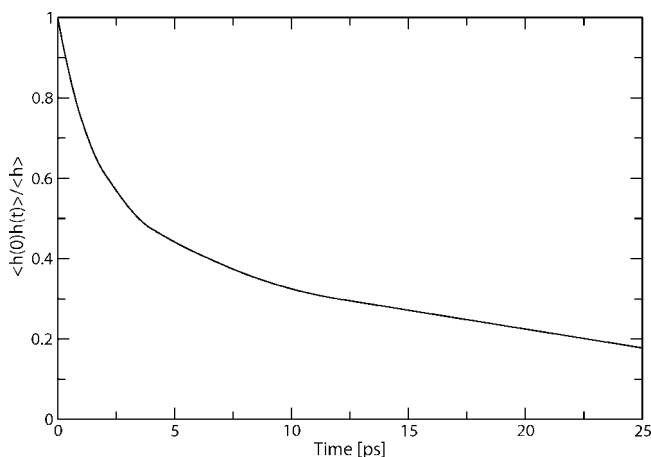
where  $D_{\infty}$  is the diffusion coefficient for an infinite system,  $\zeta = 2.837$  a numerical coefficient similar to the Madelung constants which results from an infinite summation over all replicas, and  $\eta$  the translational shear viscosity. Though this relation have been known for a while, it is not generally appreciated how large these finite size corrections might be. From Figure 4 it can be seen that, within error bars,  $D_{\text{PBC}}(L)$  obeys rather well the analytical  $1/L$  scaling. With some caution, using eq 3, we can thus extrapolate  $D_{\text{PBC}}(L)$  to  $L \rightarrow \infty$ , in order to determine  $D_{\infty} = 0.789 \times 10^{-5} \text{ cm}^2/\text{s}$ , and for the first time obtain an estimate for the translational shear viscosity  $\eta_{\infty} = 21.22 \times 10^{-4} \text{ Pa}\cdot\text{s}$ . These results have to be compared with the experimental values of  $D = 2.395 \times 10^{-5} \text{ cm}^2/\text{s}$ <sup>65</sup> and  $\eta = 8.92 \times 10^{-4} \text{ Pa}\cdot\text{s}$ .<sup>66</sup> These results confirm



**Figure 4.**  $D_{\text{PBC}}$  as a function of system size, computed by AIMD simulations using the PBE XC functional. The solid line is the linear extrapolation to an infinite system, whereas the dotted line is the mean of  $D_{\infty}$  corrected by eq 3.



**Figure 5.** Velocity–velocity autocorrelation function and the power spectrum. Full lines are obtained by Car–Parrinello-like simulations, whereas dashed lines represent BOMD reference calculations.



**Figure 6.** HB autocorrelation function as a function of simulation time.

that PBE water is less fluid than real water. However, if we take into account that we have not included nuclear quantum effects,<sup>67,68</sup> we conclude that PBE provides a good representation of the water potential energy surface. In addition, Figure 4 shows that for all of our simulations the value of  $D_{\infty}$ , as obtained by applying eq 3 together with the now determined  $\eta_{\infty}$ , is consistent with our initial estimate. This demonstrates that  $\eta_{\infty}$  is much less system size dependent than  $D_{\infty}$  and that the area of validity of the Stokes–Einstein relation, which indicates an inverse relation between these two quantities, is limited.

In Figure 5, we show the velocity–velocity autocorrelation function for the hydrogen and oxygen atoms, as well as its power spectrum that is the temporal Fourier transform. The latter is of particular interest, since, besides being in excellent agreement with our BOMD reference calculations, the shoulder due to dangling HB's in the oxygen–hydrogen high-frequency stretching band further indicates a mainly symmetric charge distribution and thus tetrahedral water coordination.

In all calculations  $D_{\text{PBC}}(L)$  is computed using Einstein's relation from the mean square displacement, and as an extra consistency check also via the Green–Kubo relation, i.e.,

by integrating the velocity–velocity autocorrelation function with respect to time.

Given the fact that our method is new and to assess the influence of the stochastic noise on the dynamics, we performed two BOMD reference calculations, one using a canonical Langevin dynamics with exactly the same damping term as before, and another one in the microcanonical NVE ensemble. The three different calculations yield results for  $D_{\text{PBC}}$  that are indistinguishable within error bars. This further strengthens our conclusion that our sampling is correct and that the use of a Langevin equation with tiny damping does not affect the dynamical properties within statistical uncertainty.

**4.4. Hydrogen Bond Kinetics.** The HB kinetics is studied via the Luzar–Chandler model<sup>69</sup> that with just two rate constants  $k$  and  $k'$  is able to describe the complex, nonexponential behavior of the following reactive flux correlation function:

$$k(t) = -\frac{dc(t)}{dt} = -\frac{\langle (dh/dt)_{t=0}[1-h(t)] \rangle}{\langle h \rangle} = kc(t) - k'n(t) \quad (4)$$

in which  $c(t) = \langle h(0)h(t) \rangle / \langle h \rangle$  is the HB autocorrelation function, and  $n(t) = \langle h(0)[1-h(t)]H(t) \rangle / \langle h \rangle$ , where  $H(t)$  is unity if a selected pair of molecules are closer than the distance  $R = 3.5 \text{ \AA}$  and zero otherwise, while  $\langle \cdot \rangle$  denotes temporal averages. Thus,  $n(t)$  denotes the number of initially bonded pairs of molecules that are broken at time  $t$ , while remaining closer than  $R$ . In the HB population operator  $h(t)$ , we use the previous mentioned HB definitions. The HB lifetime is related to  $k$  by  $\tau_{\text{HB}} = k^{-1}$ , whereas the HB relaxation time is computed as

$$\tau_r = \frac{\int dt tc(t)}{\int dt c(t)} \quad (5)$$

The reactive flux correlation function  $k(t)$  is nonexponential and monotonically decaying after a few libration periods. A least-squares fit of our data to eq 4 yields  $k = 0.143 \text{ ps}^{-1}$  and  $k' = 0.389 \text{ ps}^{-1}$ , thus  $\tau_{\text{HB}} = 6.98 \text{ ps}$  and  $\tau_r = 10.25 \text{ ps}$ . When compared to the classical results of Xu et al.,<sup>70</sup> our values for  $\tau_{\text{HB}}$  and  $\tau_r$  are both about twice as large. As a consequence, the ratio  $\tau_r/\tau_{\text{HB}} = 1.47$  is very close to the value  $\sim 1.5$  reported by others.<sup>19,70</sup>

Besides these quantitative differences, there is also a qualitative discrepancy in the sense that in our ab initio calculation  $c(t)$  decorrelates significantly slower and ceases to be exponentially decaying as displayed in Figure 6. In fact, we suspect that the decay might be biexponential, which would assume a second linear equation for  $n(t)$ . But in any case, this can be tentatively attributed to polarization effects that are possibly better described by DFT, though a final answer would require further investigation which is beyond the scope of this paper.

## 5. Conclusion

Owing to the efficiency of our novel method, to the best of our knowledge, we have presented the most extensive AIMD

simulations so far. We found that structural properties are well converged and reproducible. By contrast, dynamical properties are much less established; however, recent progress in AIMD calculations<sup>20</sup> allowed us to calculate the shear viscosity and the HB relaxation time of liquid water for the first time from ab initio. Along the way, we also reassessed the liquid water structure that has been recently questioned, but all of our calculations further supports the well-established tetrahedral coordination of liquid water.

**Acknowledgment.** We thank R. Z. Khaliullin and H. Eshet for various fruitful discussions, as well as J. A. Morrone, Y. Mantz, and L. Ojamäe for useful suggestions. The generous allocation of computer time from Swiss National Supercomputing Center (CSCS) and technical support from Neil Stringfellow is kindly acknowledged.

## References

- (1) Rahman, A.; Stillinger, F. H. *J. Chem. Phys.* **1971**, *55*, 3336–3359.
- (2) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odellius, M.; Ogasawara, H.; Näslund, L. A.; Hirsch, T. K.; Ojamäe, L.; Glatzel, P.; Petterson, L. G. M.; Nilsson, N. *Science* **2004**, *304*, 995–999.
- (3) Stillinger, F. H. *Science* **1980**, *209*, 451–457.
- (4) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331–342.
- (5) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (6) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219–260.
- (7) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896.
- (8) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.
- (9) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *Science* **2007**, *315*, 1249–1252.
- (10) Donchev, A. G.; Galkin, N. G.; Illarinov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. *Proc. Natl. Acad. U.S.A.* **2008**, *103*, 8613–8617.
- (11) Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.
- (12) Asthagiri, D.; Pratt, L. R.; Kress, J. D. *Phys. Rev. E* **2003**, *68*, 041505.
- (13) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990–12998.
- (14) Fernández-Serra, M. V.; Artacho, E. *J. Chem. Phys.* **2004**, *121*, 11136–11144.
- (15) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*, 014515.
- (16) Sit, P. H.-L.; Marzari, N. *J. Chem. Phys.* **2005**, *122*, 204510.
- (17) Mantz, Y. A.; Chen, B.; Martyna, G. J. *J. Phys. Chem. B* **2006**, *110*, 3540–3554.
- (18) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154507.

- (19) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.
- (20) Kühne, T. D.; Krack, M.; Mohamed, F.; Parrinello, M. *Phys. Rev. Lett.* **2007**, *98*, 066401.
- (21) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (22) Caravati, S.; Bernasconi, M.; Kühne, T. D.; Krack, M.; Parrinello, M. *Appl. Phys. Lett.* **2007**, *91*, 171906.
- (23) Pietrucci, F.; Caravati, S.; Bernasconi, M. *Phys. Rev. B* **2008**, *78*, 064203.
- (24) Payne, M. C.; Teter, M. P.; Allan, D. C.; Arias, T. A.; Joannopoulos, J. D. *Rev. Mod. Phys.* **1992**, *64*, 1045–1097.
- (25) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (26) Bornemann, F. A.; Schütte, C. *Numer. Math.* **1998**, *78*, 359–376.
- (27) Pastore, G.; Smargiassi, E.; Buda, F. *Phys. Rev. A* **1991**, *44*, 6334–6347.
- (28) Kolafa, J. *J. Comput. Chem.* **2004**, *25*, 335–342.
- (29) Kolafa, J. *J. Chem. Phys.* **2005**, *122*, 164105.
- (30) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1995**, *102*, 8071–8077.
- (31) Niklasson, A. M. N.; Tymczak, C. J.; Challacombe, M. *Phys. Rev. Lett.* **2006**, *97*, 123001.
- (32) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2003**, *118*, 4365–4369.
- (33) Krajewski, F. R.; Parrinello, M. *Phys. Rev. B* **2006**, *73*, 041105(R).
- (34) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (35) Ricci, A.; Ciccotti, G. *Mol. Phys.* **2003**, *101*, 1927–1931.
- (36) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703–1710.
- (37) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641–3662.
- (38) Krack, M. *Theor. Chem. Acc.* **2005**, *114*, 145–152.
- (39) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300–311.
- (40) Hammer, B.; Hansen, L. B.; Norskov, J. K. *Phys. Rev. B* **1999**, *59*, 7413–7421.
- (41) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890–890.
- (42) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (43) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.
- (44) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (45) Lippert, G.; Hutter, J.; Parrinello, M. *Mol. Phys.* **1997**, *92*, 477–488.
- (46) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.
- (47) Ceriotti, M.; Kühne, T. D.; Parrinello, M. *J. Chem. Phys.* **2008**, *129*, 024707.
- (48) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2007**, *127*, 114105.
- (49) Gallo, P.; Sciortino, F.; Tartaglia, P.; Chen, S.-H. *Phys. Rev. Lett.* **1996**, *76*, 2730–2733.
- (50) Hura, G.; Russo, D.; Glaeser, R. M.; Head-Gordon, T.; Krack, M.; Parrinello, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1981–1991.
- (51) Soper, A. K. *Chem. Phys.* **2000**, *258*, 121–137.
- (52) Morrone, J. A.; Car, R. *Phys. Rev. Lett.* **2008**, *101*, 017801.
- (53) Chen, B.; Ivanov, I.; Klein, M. L.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *91*, 215503.
- (54) Hart, R. T.; Benmore, C. J.; Neufeind, J.; Kohara, S.; Tomberli, B.; Egelstaff, P. A. *Phys. Rev. Lett.* **2005**, *94*, 047801.
- (55) Soper, A. K. *J. Phys.: Condens. Matter* **2007**, *19*, 335206.
- (56) Smith, J. D.; Cappa, C. D.; Wilson, K. R.; Messer, B. M.; Cohen, R. C.; Saykally, R. J. *Science* **2004**, *306*, 851–853.
- (57) Fernández-Serra, M. V.; Artacho, E. *Phys. Rev. Lett.* **2006**, *96*, 016404.
- (58) Prendergast, D.; Galli, G. *Phys. Rev. Lett.* **2006**, *96*, 215502.
- (59) Head-Gordon, T.; Johnson, M. E. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7973–7977.
- (60) Iannuzzi, M. *J. Chem. Phys.* **2008**, *128*, 204506.
- (61) Luzar, A.; Chandler, D. *Nature* **1996**, *379*, 55–57.
- (62) Kuo, I.-F. W.; Mundy, C. J. *Science* **2004**, *303*, 658–660.
- (63) Dünweg, B.; Kremer, K. *J. Chem. Phys.* **1993**, *99*, 6983–6997.
- (64) Yeh, I.-C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873–15879.
- (65) Hardy, E. H.; Zygari, A.; Zeidler, M. D.; Holz, M.; Sacher, F. D. *J. Chem. Phys.* **2001**, *114*, 3174–3181.
- (66) Harris, K. R.; Woolf, L. A. *J. Chem. Eng. Data* **2004**, *49*, 1064–1069.
- (67) Miller, T. F., III; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *123*, 154504.
- (68) Paesani, F.; Zhang, W.; Case, D. A.; Cheatham, T. E., III; Voth, G. A. *J. Chem. Phys.* **2006**, *125*, 184507.
- (69) Luzar, A.; Chandler, D. *Phys. Rev. Lett.* **1996**, *76*, 928–931.
- (70) Xu, H.; Stern, H. A.; Berne, B. J. *J. Phys. Chem. B* **2002**, *106*, 2054–2060.

## Striking Effects of Hydrodynamic Interactions on the Simulated Diffusion and Folding of Proteins

Tamara Frembgen-Kesner and Adrian H. Elcock\*

*Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242*

Received November 17, 2008

**Abstract:** Successful modeling of the processes of protein folding and aggregation may ultimately require accurate descriptions of proteins' diffusive characteristics, which are expected to be influenced by hydrodynamic effects; a comprehensive study of the diffusion and folding of 11 model proteins with an established simulation model extended to include hydrodynamic interactions between residues has therefore been carried out. Molecular simulations that neglect hydrodynamic interactions are incapable of simultaneously reproducing the expected experimental translational and rotational diffusion coefficients of folded proteins, drastically underestimating both when reasonable hydrodynamic radii are employed. In contrast, simulations that include hydrodynamic interactions produce diffusion coefficients that match very well with the expected experimental values for translation and rotation and also correctly capture the significant decrease in translational diffusion coefficient that accompanies protein unfolding. These effects are reflected in folding simulations of the same proteins: the inclusion of hydrodynamic interactions accelerates folding by 2–3-fold with the rate enhancement for the association of secondary structure elements exhibiting a strong sensitivity on the sequence-distance between the associating elements.

### Introduction

Efforts to develop an understanding of the mechanistic details of protein folding have been pursued for many years (for reviews, see refs 1–3) and have taken on added significance recently with the realization that a number of pathologies (the so-called 'conformational diseases'<sup>4</sup>) are caused by, or at least heavily associated with, protein misfolding and aggregation.<sup>5</sup> Theoretical and computational studies have provided a number of insights into the folding of single, isolated proteins (e.g., refs 6–8), protein–protein association (e.g., refs 9–12), oligomerization (e.g., refs 13 and 14), domain-swapping (e.g., refs 15 and 16), and aggregation processes (e.g., refs 17–20).

Up to now, very few studies have focused on how well the extant simulation models capture the diffusive motions of protein chains, either with explicit solvent<sup>21</sup> or implicit solvent methods.<sup>22</sup> This, however, is an issue that warrants special attention for the modeling of aggregation or coupled

folding-association processes since both involve an interplay (or competition) between intramolecular folding and intermolecular association events, with the latter, in particular, having a clear potential for diffusion dependence. An indication that it might be important to pay particular attention to proteins' diffusive properties is given by the fact that a number of protein aggregation processes have already been described as being controlled by kinetic rather than strictly thermodynamic factors (e.g., refs 23 and 24). Full consideration of the issue requires asking whether simulation methods are capable of accurately capturing both the translational and rotational diffusive motions of protein chains and the changes in diffusive properties that accompany their folding and/or association.

Molecular simulations of protein folding and/or association events can be roughly divided into two categories depending on whether they use explicit or implicit models for the solvent. Explicit solvent simulations have the obvious advantage of being physically more complete in the sense that individual water molecules are allowed to play specific roles during folding or association. They have the very

\* Corresponding author phone: (319)335-7894; fax: (319)335-9570; e-mail: adrian-elcock@uiowa.edu.



significant disadvantage however of being computationally demanding in comparison with corresponding implicit solvent simulations, and this expense has limited their use in the present context to a few heroic efforts in which either distributed computing resources (e.g., refs 13, 25, and 26) or highly parallelized architectures (e.g., refs 27–29) have been employed. Currently therefore, it is more common for protein folding simulations to employ implicit solvent models (e.g., refs 30 and 31). Of course, such models must attempt to incorporate in some way the energetic and dynamic influences that water exerts over the solute. The energetic effects of hydration—which are *not* the focus of the present work—can be implicitly modeled using continuum dielectric methods to account for electrostatic factors<sup>32,33</sup> and solvent-accessible surface area-based approaches to describe hydrophobic contributions.<sup>34–36</sup> Alternatively, for applications directed at studying protein folding and association processes, the issue of properly modeling the energetics is often sidestepped completely by using a native-centric ‘Gō’ model,<sup>37</sup> in which simple attractive potentials are applied to atom- or residue-pairs that form contacts in the native state, and purely repulsive potentials are applied to all other pairs. Despite the apparently drastic simplification involved in Gō-type models, the energy landscapes for folding obtained with them appear to capture a number of key features of experimental folding behavior.<sup>8,38–42</sup>

The purely dynamic effects that would be exerted by the missing water molecules in implicit solvent simulations can be accounted for by the use of Langevin dynamics (LD) (discussed in refs 43 and 44) or Brownian dynamics (BD) algorithms,<sup>45</sup> and it is with the second of these methods that the present study is concerned. In both LD and BD, water implicitly appears in the equations of motion—in the form of its viscosity—in the diffusion tensor, **D** (see Methods); in BD implementations, **D**, a  $3N \times 3N$  matrix (where  $N$  is the number of particles) determines both (a) the extent of coupling between force-induced displacements of solute particles and (b) the statistical properties of the random displacements that are applied to the solute particles.<sup>45</sup> Importantly, the diffusion tensor provides the means of including the effects of hydrodynamic interactions (HI) between solute particles (e.g., protein residues) into the equations of motion. There are several ways of conceptualizing HI; the simplest perhaps is to consider them as accounting for the fact that the displacement of one solute particle can have a ‘knock-on’ effect on a second nearby solute particle due to the displacement of intervening solvent molecules. Inclusion of HI therefore introduces correlations into the random, Brownian displacements experienced by neighboring solutes, such that they tend to be displaced in similar directions; neglect of HI (also known as the ‘free draining’ approximation<sup>44</sup>) means that the Brownian displacements applied to neighboring solute particles are completely uncorrelated.

Given the above it is not hard to imagine that the inclusion of HI into simulations of polymers such as proteins might have a significant effect on their diffusive properties. In fact, there is a considerable body of literature in the polymer physics field that has already examined the effects of HI on

the translational diffusion of high molecular weight polymers. It was shown many years ago, for example,<sup>46</sup> that the inclusion of HI—modeled via the Kirkwood-Riseman approximation<sup>47</sup>—successfully explains the  $Mw^{-1/2}$  dependence observed for polymer translational diffusion coefficients in so-called  $\theta$  conditions, whereas when HI are neglected, a much shallower, incorrect,  $Mw^{-1}$  dependence is obtained.<sup>48,49</sup> More recently it has been shown that the inclusion of HI significantly increases the collapse rate of polymers in so-called ‘bad’ solvent conditions.<sup>50–54</sup> While it is to be anticipated that these previous studies of simple polymers will provide useful clues to what might happen in proteins, very little attention has thus far been focused on the effects of HI on the folding and diffusion of actual protein chains.<sup>22,53</sup> In fact, it has been recently reported that in regard to *unfolding*, simple homopolymers and protein chains appear to exhibit distinct behaviors.<sup>55</sup> We therefore have conducted a relatively systematic comparison of the effects of HI on the BD-simulated diffusion and folding of eleven proteins using an established simulation model; the results indicate that inclusion of HI is likely to be essential for correctly capturing the full range of diffusive behavior exhibited by folding protein chains.

## Methods

**The Proteins Studied.** Eleven well-characterized proteins were chosen in order to cover a range of structural classes, subject to the restriction that they be sufficiently small (<150 residues) that complete sampling of their behavior would be computationally tractable. The proteins selected were as follows: the B1 immunoglobulin-binding domain of protein G (hereafter termed simply ‘protein G’),<sup>56</sup> B1 immunoglobulin-binding domain of protein L (protein L),<sup>57</sup> chymotrypsin inhibitor 2 (CI2),<sup>58</sup> barnase,<sup>59</sup> the fyn SH3 domain (fyn-SH3),<sup>60</sup> cold shock protein B (CSPB),<sup>61</sup> intestinal fatty acid binding protein (IFABP),<sup>62</sup> Semliki Forest viral capsid protein (SFVP),<sup>63</sup>  $\lambda$ -repressor,<sup>64</sup> colicin E9 immunity protein (Im9),<sup>65</sup> and apo-calmodulin (apoCaM).<sup>66</sup> The Protein Data Bank (PDB)<sup>67</sup> (<http://www.rcsb.org>) files used in these simulations, with the exceptions of SFVP and apoCaM, are identical with those listed in ref 68. The key characteristics of the proteins are summarized in Table 1. In addition to these complete proteins, an  $\alpha$ -helix and a  $\beta$ -hairpin, both 16 residues in length, were also simulated; the structure of the former was taken from the first helix of  $\lambda$ -repressor, the latter from strands 6 and 7 of IFABP.

**The Protein Model.** The structural and energetic model used for the proteins in all simulations is a Gō-like model,<sup>37</sup> implemented in essentially the same manner as described by Clementi, Onuchic, and others.<sup>8,39,40,42</sup> The all-atom structures of the proteins were reduced to simpler ‘bead-spring’ models and simulated at two different levels of detail. The first of these, which we refer to in the text by the shorthand expression ‘C $\alpha$ ’, represents each amino acid residue with a single bead, or pseudoatom, placed at the position of the C $\alpha$  atom. The second, finer level of detail, which we refer to by the abbreviation ‘SC’ (for Side Chain) represents residues again with a C $\alpha$  pseudoatom but supplemented by additional (up to three) pseudoatoms to model

**Table 1.** Details of the Proteins Studied

protein	PDB	fold type	number of domain(s)	number of C $\alpha$ atoms	number of SC atoms
$\alpha$ -helix		$\alpha$ -helix		16	
$\beta$ -hairpin		$\beta$ -sheet		16	
protein G	1PGA	mixed	one	56	141
protein L <sup>a</sup>	1HZ6	mixed	one	64	160
CI2	2CI2	mixed	one	65	165
barnase <sup>b</sup>	1BNI	mixed	two	108	269
fyn-SH3	1SHF	$\beta$ -sheet	one	59	151
CSPB	1CSP	$\beta$ -sheet	one	67	165
IFABP	1IFC	$\beta$ -sheet	one	131	335
SFVP	1VCP	$\beta$ -sheet	two	149	358
$\lambda$ -repressor <sup>c</sup>	1LMB	$\alpha$ -helix	one	80	202
Im9 <sup>d</sup>	1IMQ	$\alpha$ -helix	one	86	216
apo-CaM <sup>d</sup>	1CFD	$\alpha$ -helix	two	148	380

<sup>a</sup> Y47W mutant. <sup>b</sup> Added missing side chain atoms to Lys19, Glu60, Lys62, and Gln104 using WHATIF.<sup>69</sup> <sup>c</sup> Residues 6–85. <sup>d</sup> Averaged NMR structure.

the side chains. Full details of the latter model, which is conceptually similar to others previously presented in the literature,<sup>70,71</sup> are given in the Supporting Information. The total numbers of pseudoatoms that were used to represent each protein are listed in columns five and six of Table 1 for the C $\alpha$  and SC descriptions, respectively.

Following the G $\ddot{o}$  model approach, the interactions between all nonbonded pairs of pseudoatoms were assigned one of two energy functions. Pseudoatom pairs that formed a close contact in a protein's native state structure were assigned a favorable Lennard-Jones-type energy function in order to provide them with an energetic reward for forming such a contact during the simulations; pairs were considered to form a native contact if any of their constituent non-hydrogen atoms were within 5.5 Å of each other in the native state structure.<sup>39</sup> Following others,<sup>8,39,40,42</sup> the form of this energy function was chosen to be

$$E_{ij}^{\text{native}} = \varepsilon[5(\sigma_{ij}/r_{ij})^{12} - 6(\sigma_{ij}/r_{ij})^{10}] \quad (1)$$

where  $\varepsilon$  is the energy well depth of the Lennard-Jones potential,  $r_{ij}$  is the distance between pseudoatoms  $i$  and  $j$  in the simulations, and  $\sigma_{ij}$  is the distance between the two pseudoatoms in the native state structure. For simulations of proteins in their folded states and for simulations of folding events,  $\varepsilon$  values of 0.60 and 0.25 kcal/mol were used for the C $\alpha$  and SC models, respectively; the former value was obtained from our previous work matching the experimental folding free energy of CI2 at 25 °C,<sup>42</sup> the latter value was obtained here by performing similar 100  $\mu$ s-length simulations—and using histogram techniques<sup>72,73</sup>—to reproduce the folding free energy of protein L.<sup>74</sup> For simulations of proteins in unfolded states, a much smaller  $\varepsilon$  value of 0.05 kcal/mol was used for both C $\alpha$  and SC models. Pseudoatom pairs separated by four or more bonds that do not form a close contact in the native state structure were assigned a purely repulsive potential

$$E_{ij}^{\text{non-native}} = \varepsilon(\sigma/r_{ij})^{12} \quad (2)$$

with  $\sigma$  in this case being a constant value (4 Å) and  $\varepsilon$  being assigned the value of 0.60 kcal/mol.

For bonding interactions between the pseudoatoms, standard molecular mechanics terms were used,<sup>43</sup> with the total bonded energy of the protein being written

$$E_{\text{bonded}} = \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} K_\varphi^{(n)}[1-\cos(n(\varphi-\varphi_n))] \quad (3)$$

where  $r$ ,  $\theta$ , and  $\varphi$  are (pseudo)bond lengths, angles and dihedrals, respectively, and  $r_0$  and  $\theta_0$  are the corresponding native state bond lengths and angles;  $\varphi_1$  and  $\varphi_3$  are phase angles which define the energy maxima of the dihedral angles. Following others,<sup>8,39,40,42</sup> the bond and angle force constants  $K_r$  and  $K_\theta$  were set to 100 kcal/mol/Å and 20 kcal/mol/rad, respectively. For folding simulations, the force constants for the two dihedral rotations  $K_\varphi^{(1)}$  and  $K_\varphi^{(3)}$  were set to 0.50 and 0.25 kcal/mol, respectively (for C $\alpha$  models), and to 0.41 and 0.21 kcal/mol, respectively (for SC models). These values were chosen in order to maintain an appropriate balance between the nonlocal and local driving forces for folding, since this balance has been shown to affect the cooperativity of folding equilibria simulated with G $\ddot{o}$  models.<sup>8,42,75</sup> For unfolded state simulations with both models,  $K_\varphi^{(1)}$  and  $K_\varphi^{(3)}$  were set to weaker values of 0.10 and 0.05 kcal/mol, respectively.

**The Simulation Algorithm.** The time-dependent conformational behavior of the proteins was simulated using the Brownian dynamics (BD) algorithm of Ermak and McCammon<sup>45</sup>

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \sum_j \mathbf{D}_{ij} \mathbf{F}_j \Delta t / k_B T + \mathbf{R}_i \quad (4)$$

where  $\mathbf{r}_i(t)$  is the position vector of the  $i$ th pseudoatom at time  $t$ ;  $\Delta t$  is the simulation time step,  $\mathbf{D}_{ij}$  is the  $i,j$ th  $3 \times 3$  submatrix of the diffusion tensor  $\mathbf{D}$  (a  $3N \times 3N$  matrix, where  $N$  is the number of pseudoatoms in the simulated system);  $\mathbf{F}_j$  is the total force acting on the  $j$ th pseudoatom; and  $\mathbf{R}_i$  is a random displacement applied to the  $i$ th pseudoatom (see below);  $k_B$  is Boltzmann's constant, and  $T$  is the temperature in Kelvin.

The diffusion tensor is perhaps best thought of as a  $N \times N$  supermatrix of  $3 \times 3$  matrices, with each individual  $3 \times 3$  submatrix describing the coupling between the  $x$ ,  $y$ , and  $z$  components of motion for a pair of pseudoatoms  $i$  and  $j$ . For simulations that do not include hydrodynamic interactions (also known in the literature as the 'free draining' approximation<sup>44</sup>), the only nonzero elements in the entire diffusion tensor are to be found along the diagonals of the  $3 \times 3$  submatrices for those cases where pseudoatom  $i = j$ ; as is usual,<sup>45</sup> these diagonal elements were calculated from the Stokes–Einstein relation  $\mathbf{D}_{ii} = k_B T / 6\pi\eta_s a$  where  $\eta_s$  is the viscosity of the solvent (for water,  $\eta_s = 0.89$  cP at 25 °C), and  $a$  is the hydrodynamic radius of the pseudoatom. For all simulations, the hydrodynamic radii assigned to pseudoatoms were 5.3 Å and 3.5 Å for the C $\alpha$  and SC models, respectively (see the next section for the derivation of these values). For simulations that include hydrodynamic interactions, the elements of the  $3 \times 3$  matrices for interactions between pseudoatoms  $i$  and  $j$  ( $i \neq j$ ) are nonzero, which has the result that the displacement of a pseudoatom  $i$  is directly affected by the forces acting on a pseudoatom  $j$  to which it is hydrodynamically coupled (see eq 4). In the present work the elements of the  $\mathbf{D}_{ij}$  submatrices were computed using

equations due to Rotne and Prager<sup>76</sup> and Yamakawa.<sup>77</sup> The complete set of equations used to compute  $\mathbf{D}$  is therefore

$$\mathbf{D}_{ii} = (k_B T / 6\pi\eta_s a) \mathbf{I} \quad (5a)$$

$$\mathbf{D}_{ij} = (k_B T / 8\pi\eta_s) \{ (1/r_{ij}) [(1 + 2a^2/3r_{ij}^2) \mathbf{I} + (1 - 2a^2/r_{ij}^2)(\mathbf{r}_{ij}\mathbf{r}_{ij}/r_{ij}^2)] \} \text{ for } r_{ij} \geq 2a \quad (5b)$$

$$\mathbf{D}_{ij} = (k_B T / 8\pi\eta_s) \{ (1/r_{ij}) [(r_{ij}/2a)(8/3 - 3r_{ij}/4a) \mathbf{I} + (r_{ij}/4a)(\mathbf{r}_{ij}\mathbf{r}_{ij}/r_{ij}^2)] \} \text{ for } r_{ij} < 2a \quad (5c)$$

where  $\mathbf{I}$  is a unit  $3 \times 3$  matrix (1 for all diagonal elements; 0 for all nondiagonal elements),  $r_{ij}$  is the distance between pseudoatoms  $i$  and  $j$ , and  $\mathbf{r}_{ij}$  is the vector connecting them.

The fluctuation–dissipation theorem ensures that correct Boltzmann sampling is obtained by specifying a relationship between the diffusion tensor and the statistical properties required of the random displacements applied to the pseudoatoms. Formally these requirements can be written as

$$\langle \mathbf{R}_i \cdot \mathbf{R}_j \rangle = 6\mathbf{D}_{ij}\Delta t \text{ and } \langle \mathbf{R}_i \rangle = 0 \quad (6)$$

When hydrodynamic interactions are modeled, the  $3N$  vector of correlated random displacements is obtained by multiplying a  $3N$  vector of uncorrelated random numbers (with zero mean and unit variance) by the ‘square root’ matrix  $\mathbf{S}$ , which is generated from a factorization of the matrix  $\mathbf{D}$ , such that<sup>45,78</sup>

$$\mathbf{D} = \sum_l \mathbf{S}_{il} \mathbf{S}_{jl} \quad (7)$$

Thus the coupling of motion of pseudoatoms  $i$  and  $j$  specified in the diffusion tensor appears also in the random displacements. In the present study, factorization of the matrix  $\mathbf{D}$  was achieved by performing a (computationally expensive) Cholesky decomposition.<sup>78–80</sup>

Software for conducting all BD simulations was written in-house;<sup>42</sup> FORTRAN code implementing the Rotne-Prager-Yamakawa diffusion tensor calculation and the Cholesky decomposition was obtained from routines written by Allen and Tildesley.<sup>43</sup>

**Simulation Details.** For all eleven proteins studied, separate BD simulations were performed to investigate the effects of hydrodynamic interactions (HI) on the simulated diffusional properties of the proteins in their folded and unfolded states using both C $\alpha$  and SC models. Additional studies were performed to investigate the effects of hydrodynamic interactions on folding pathways and rates using C $\alpha$  models of all eleven proteins and SC models of three representative proteins, protein L, CSPB, and  $\lambda$ -repressor. In all simulations that started from the unfolded state, initial conformations were generated by randomly assigning dihedral angles such that no steric clashes were introduced.

For simulations aimed at measuring diffusive characteristics, we found that the computed translational and rotational diffusion coefficients obtained from the BD simulations were essentially insensitive to the choice of time step (see the Supporting Information); because of this, timesteps of 40 and 20 fs were used for HI and non-HI simulations, respectively. For simulations aimed at investigating the actual folding behavior of proteins more care was found to be

needed in the choice of simulation timesteps, especially in the case of simulations performed with the more detailed SC model, since simulations that included HI typically produced slightly lower energies (i.e., more stable trajectories) for a given time step than did non-HI simulations. The origins of this effect almost certainly lie in the fact that HI promote correlated motions for closely separated pseudoatoms and therefore tend to suppress abrupt changes in bonding and short-range steric interactions that would otherwise occur in corresponding non-HI simulations. This becomes an issue because the simulated folding rates of proteins are often strongly dependent on their thermodynamic stability, so artifactual differences between a protein’s stability in the presence and absence of HI due to a poor choice of timesteps could cause misleading differences in their observed rates of folding. (It is worth noting that a similar dependence of experimental folding rates on thermodynamic stability is also observed.<sup>74,81,82</sup>) To circumvent this issue, exploratory native-state simulations were performed for all eleven proteins using a range of timesteps, and a combination of timesteps that resulted in equivalent internal energies with and without HI were then selected; the timesteps chosen for HI and non-HI simulations ranged from 20 to 100 and from 10 to 100 fs, respectively. Full details of this procedure, showing the time step dependence of the proteins’ simulated energies, are provided in the Supporting Information.

In addition to finding that simulations with and without HI differed significantly in terms of their robustness to changing the timestep, we also found that they responded very differently to the inclusion of bond constraints. In order to allow longer timesteps in molecular dynamics (MD) and BD simulations it is common to constrain bonds to their equilibrium distances via application of an iterative constraint algorithm such as SHAKE<sup>83</sup> or LINCS.<sup>84</sup> In the present study, we used LINCS to constrain all pseudobonds in the non-HI simulations: in our hands this typically allowed us to increase the time step by a factor of 4. When included in HI simulations however, LINCS caused severe problems: in particular, it caused significant—and systematic—differences in the pseudobond angle energies that could not be resolved by choosing a different time step. To our knowledge this effect has not been directly reported in the literature, although the use of HI with the SHAKE constraint algorithm has been previously reported to affect transport coefficients<sup>85</sup> and the thermodynamics of  $\beta$ -hairpin folding.<sup>22</sup> Conceptually, the effect almost certainly originates from the juxtaposition of an algorithmic ‘step’ that promotes correlated displacements of bonded atoms (the inclusion of HI) with one that promotes anticorrelated displacements (the imposition of bond constraints: adjusting the positions of atoms to reach an equilibrium separation distance must either involve them being moved toward or away from each other, both of which are anticorrelated motions). Because of this problem, all simulations that included HI were performed *without* bond constraints; instead, pseudobond lengths were continually monitored during the simulations, and, if deviating by more than 0.4 Å from their equilibrium values, the simulation was backtracked by  $\sim 10$  ps and restarted (see the Supporting

Information for details). Results from control simulations that folded model proteins without HI and without using LINCS were compared to those from the two models used here (with HI, without bond constraints, and without HI, but with bond constraints). The difference in the treatment of pseudobonds between the two models had no significant effect on the folding results reported here.

In all simulations, a conventional neighbor list<sup>43</sup> was used to facilitate rapid calculation of nonbonded interactions between pseudoatoms; this list was recalculated every 200 fs. Pseudoatom pairs that did not form a contact in the native state structure were added to this list if their separation distance was 8 Å or less; pairs that did form a native contact were added to the list if their separation distance was within 9 Å of their distance in the native state structure. For simulations in which HI were included, the elements of the diffusion tensor and its Cholesky decomposition were recalculated every 1 ps; exploratory simulations indicated that updating the diffusion tensor at every time step made essentially no difference to the simulation observables (see the Supporting Information).

#### Determination of ‘Experimental’ Diffusion Coefficients.

Since experimental estimates of protein translational and rotational diffusion coefficients can vary significantly between groups and (especially) depending on the experimental technique employed,<sup>86,87</sup> we chose to estimate the diffusional properties for the eleven proteins using a single method, the hydrodynamics program HYDROPRO,<sup>88,89</sup> in previous studies the diffusion coefficients calculated with this method have been shown to be within 2 and 6% of the experimental values for translation and rotation, respectively. All ‘experimental’ diffusion coefficients referred to in this study are therefore *computed* values obtained by applying HYDROPRO to the all-atom PDB files of the eleven proteins (Table 1); in all of these calculations the recommended default hydrodynamic radius of 3.2 Å was applied to all atoms.<sup>89</sup> In addition to being used to compute ‘gold-standard’ ‘experimental’ diffusion coefficients for native state structures, HYDROPRO was also used to compute diffusion coefficients for unfolded state structures of each protein. These estimates were obtained by randomly selecting 5 snapshots each from HI and non-HI simulations of the unfolded C $\alpha$  model and submitting the 10 selected snapshots to HYDROPRO with a hydrodynamic radius of 5.3 Å assigned to each pseudoatom (see below for the derivation of this value).

For actual BD simulations of the folding and diffusion of the proteins, the pseudoatoms’ hydrodynamic radii used in eqs 5a–5c were determined in the following way. The C $\alpha$  structural models of all eleven proteins were each submitted to HYDROPRO with a range of values assigned to the hydrodynamic radius of the C $\alpha$  pseudoatom; the translational diffusion coefficients computed by HYDROPRO using these pseudoatomic models were then compared with those computed with the *fully atomic* models, and the hydrodynamic radius producing the best overall agreement was selected (note therefore that we use the same hydrodynamic radius for all residue types): the best agreement was obtained with a hydrodynamic radius of 5.3 Å (see above). Repeating the entire procedure for the SC model, an optimal hydro-

dynamic radius was found to be 3.5 Å (see the Supporting Information for further details).

**Calculation of Diffusion Coefficients from BD Simulations.** For all proteins studied, translational diffusion coefficients were calculated as an average from 10 independent BD simulations, each conducted for a production length of 100 ns following a brief (100 ps) equilibration period. From each trajectory the translational diffusion coefficient,  $D_{\text{trans}}$ , was computed using the Einstein formula<sup>43</sup>

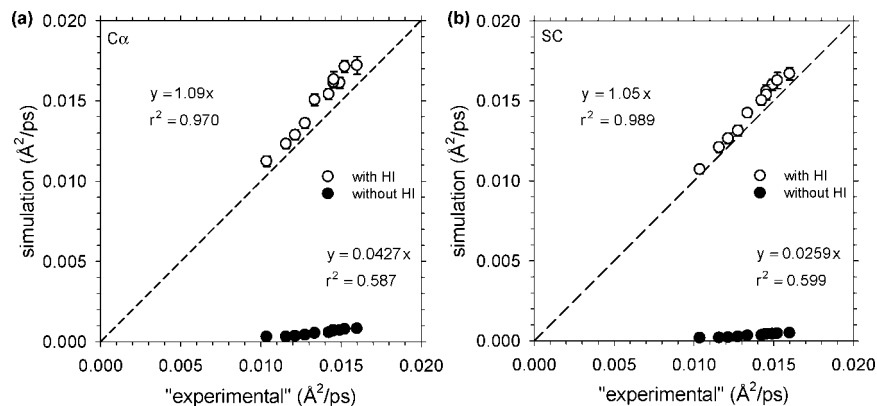
$$D_{\text{trans}} = \langle R^2 \rangle / 6\delta t \quad (8)$$

where  $\langle R^2 \rangle$  is the mean-squared distance traveled by the protein’s center of geometry in a time interval  $\delta t$ . All of the computed values reported in Results were obtained with  $\delta t = 100$  ps, but the results were insensitive to the exact choice of  $\delta t$  (see the Supporting Information).

Rotational diffusion coefficients for folded proteins were also calculated as averages from 10 independent simulations, but longer simulations were required in order to obtain robust estimates: the production length of all such simulations was 500 ns. Rotational motion was quantified by computing time-autocorrelation functions for vectors pointing along the principal axes of the proteins’ moments of inertia; these three axes were approximated by vectors connecting the two pseudoatoms closest to the principal axes of the native state structure. The three autocorrelation functions were averaged, and the first 20 ns of the averaged function were fit to a single exponential (additional exponentials were found to be unnecessary); the time constant for this decay gives the rotational relaxation time  $\tau_{\text{rot}}$ , from which the rotational diffusion coefficient  $D_{\text{rot}}$ , can be calculated as  $D_{\text{rot}} = 1/(2\tau_{\text{rot}})$ .<sup>90</sup>

In addition to calculating conventional translational and rotational diffusion coefficients for entire proteins, an ‘effective’ diffusion coefficient was also defined as a way of quantifying the *relative* diffusion of pseudoatom pairs as a function of their sequence-separation (see Results). This was achieved by analyzing ten 100 ns BD trajectories of a 108-residue peptide chain simulated (at a C $\alpha$  level of detail) with only non-native nonbonded interactions and with all dihedral angle potentials set to zero. The effective diffusion coefficient,  $D_{\text{eff}}$ , for the relative motion of two pseudoatoms was determined by applying the one-dimensional Einstein formula to the mean-squared change in the separation *distance* of the two pseudoatoms over the time-interval,  $\delta t = 100$  ps. Effective diffusion coefficients defined in this way were computed separately for pseudoatom pairs separated by 2, 4, 6, 8 residues etc. with the two pseudoatoms of each pair being symmetrically located either side of the center of the peptide chain. These calculations were carried out for simulations performed both with and without hydrodynamic interactions.

**Folding Simulations.** In order to assess the potential impact of HI on the folding kinetics of the proteins, 100 independent folding simulations were performed for all 11 proteins at a C $\alpha$  level of detail and for three proteins (protein L, CSPB, and  $\lambda$ -repressor) at the SC level of detail; each folding simulation started from a different initial conformation generated by randomly assigning dihedral angles such



**Figure 1.** Simulation-derived translational diffusion coefficients of folded proteins plotted against the corresponding 'experimental' values (see Methods for the derivation of these experimental values): (a) C $\alpha$  model proteins and (b) side chain (SC) model proteins.

that no steric interactions were included. The extent of folding during the simulations was quantified during the simulations with the conventional structural parameter 'Q',<sup>8,39</sup> defined as the number of native contacts present in the current conformation divided by the number of such contacts formed in the fully folded protein. Again, following the approach outlined by Clementi and others, a native contact was considered to have formed when the two pseudoatoms of the contact pair came within a distance  $1.2\sigma_{ij}$  where  $\sigma_{ij}$  is their separation distance in the native state structure. Proteins were considered folded when Q reached 0.90 as this approximately reflected the mean Q obtained in simulations of the folded state. Following Koga and Takada,<sup>39</sup> rates of folding were calculated as the inverse of the mean folding time.

In addition to following the folding of the protein, the rate at which the initially unfolded protein adopted a collapsed conformation was computed by monitoring the change in the radius of gyration ( $R_g$ ) versus time. The time required for  $R_g$  to drop below 115% of its native state value was considered its 'collapse' time; the remainder of the time required to fold (i.e., to reach  $Q = 0.90$ ) was considered the protein's 'search' time. Rates for these two events were calculated in the same manner as the overall folding rate. Finally, to assess any potential connection between folding rates, hydrodynamic interactions, and the relative preponderance of local and nonlocal contacts in the proteins, absolute 'contact order' values<sup>68,91</sup> for all eleven proteins were obtained, either directly from the literature<sup>68</sup> or as described therein using the program found at the Baker group's Web site (<http://depts.washington.edu/bakerpg/>).

**Simplified Descriptions of Folding Pathways.** A simplified description of the folding pathways exhibited by the proteins was obtained by monitoring the order of formation of key structural elements, with the latter being defined here not only as the conventional helices and sheets (identified using DSSP<sup>92</sup>) but also as *pairs* of contacting elements of secondary structure (i.e., helix-helix, helix-strand, and strand-strand contacts). The folding statuses of those elements possessing at least five inter-residue contacts were monitored during the simulations using a local folding coordinate,  $Q_{\text{element}}$ , defined for each element, and analogous with the overall 'Q' value defined for the entire protein (see above).

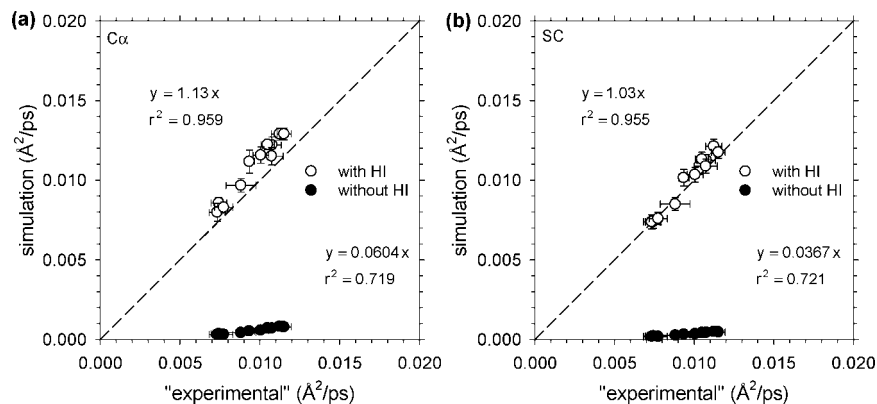
A structural element was considered folded when its  $Q_{\text{element}}$  reached 0.80, at which time its rank order in the folding pathway for that particular simulated trajectory was assigned. The mean rank order assigned to each structural element was calculated from the 100 folding trajectories and used to produce a simple vector (containing M dimensions, where M is the number of structural elements monitored during the simulations) that provides a shorthand description of the protein's folding pathway. The similarity between the folding pathways obtained from simulations performed with and without HI was then determined from the correlation coefficient of the two folding pathway vectors.

## Results

### Translational and Rotational Diffusion Coefficients.

The computed translational diffusion coefficients of folded proteins obtained from BD simulations, both with and without the inclusion of hydrodynamic interactions (HI), are plotted against the corresponding 'experimental' estimates (see Methods) in Figure 1; in (a) are shown the results obtained with a model that treats proteins at a C $\alpha$  level of representation; in (b) are shown the results obtained with a model that includes additional pseudoatoms to account for side chains (SC). With both models it can be seen that the BD simulations that include HI (open circles) reproduce the 'experimental' values very well, while the simulations that omit HI (filled circles) produce values that drastically underestimate the expected values. In fact, with the C $\alpha$  model, the translational diffusion coefficients are underestimated on average by a factor of  $\sim 23$ , whereas with the more detailed SC model, they are underestimated by a factor of  $\sim 39$ . Also displayed in Figure 1 (and subsequent figures) are the  $r^2$  values obtained from linear regressions in which an intercept of zero was enforced. Since the translational diffusion coefficients scale inversely with the overall sizes of the proteins, the much poorer quality regression fits obtained from the non-HI simulations indicates that, in addition to underestimating the absolute values of the diffusion coefficients, they also are significantly less able to capture the size-dependence of translational diffusion than are the simulations that include HI.

The above results refer to BD simulations of the proteins performed with parameters in which the native state struc-



**Figure 2.** Simulation-derived translational diffusion coefficients of unfolded proteins plotted against the corresponding 'experimental' values: (a) C $\alpha$  model proteins and (b) SC model proteins.

tures are energetically favored. A qualitatively identical picture emerges however when we conduct similar comparisons using BD simulations in which the proteins adopt only unfolded conformations (see Methods): the HI simulations again produce translational diffusion coefficient estimates that match closely with the expected values, while the non-HI simulations again dramatically underestimate them (Figure 2). Interestingly however, the magnitude of the underestimation with non-HI simulations is less than that obtained from folded state simulations: the C $\alpha$  and SC models now underestimate the 'experimental' values by factors of  $\sim 17$  and  $\sim 27$ , respectively. In addition, the quality of the regression fits for the non-HI simulation data are noticeably improved over those shown in Figure 1. It appears therefore that the omission of HI from simulations of unfolded conformations, while still being catastrophic for the simulations' ability to produce quantitatively accurate diffusion coefficients, at least does not completely destroy their ability to capture the diffusion coefficient's dependence on protein size.

The results shown up to this point have shown that the inclusion of HI in BD simulations of model proteins leads to much better descriptions of their translational diffusional behavior, regardless of whether such proteins are simulated in their folded or unfolded states. This conclusion is amplified when we consider the *ratio* of the translational diffusion coefficients for the folded and unfolded states of the proteins. Experimentally it is well-known that the folding of proteins into their globular conformations leads to significant increases in their translational diffusion coefficients.<sup>93–98</sup> This can be seen from Table 2 where we collate results from six published experimental studies in which the folded and unfolded state diffusion coefficients have been simultaneously reported for proteins of similar sizes to those studied here: the ratios of the folded to unfolded state diffusion coefficients from these studies range from 1.36 to 1.75. This behavior is conspicuously not reproduced by the BD simulations in which HI are neglected: the computed ratio for non-HI simulations is 1.00 ( $\pm 0.01$ ) for both the C $\alpha$  and SC models. The HI simulations on the other hand capture the experimental trend quite well: for the C $\alpha$  and SC models, respective ratios of  $1.38 \pm 0.08$  and  $1.47 \pm 0.11$  are obtained, indicating that the accelerated diffusion that results from

**Table 2.** Ratios of Translational Diffusion Coefficients of Folded and Unfolded States Computed from Simulation and Measured Experimentally

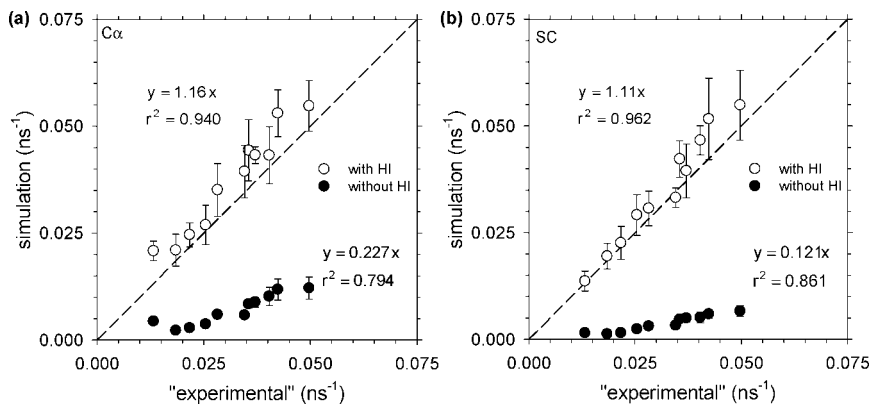
protein	C $\alpha$ w/Hi	C $\alpha$ w/o HI	SC w/Hi	SC w/o HI	expt
protein G	1.33	1.01	1.38	1.01	
protein L	1.32	1.02	1.42	0.99	
CI2	1.32	1.01	1.38	1.00	
barnase	1.41	1.00	1.55	1.00	
fyn-SH3	1.33	1.00	1.39	1.00	
CSPB	1.42	1.01	1.41	1.00	
IFABP	1.50	0.99	1.71	1.01	
SFVP	1.54	1.01	1.64	1.01	
$\lambda$ -repressor	1.33	1.00	1.45	1.01	
lm9	1.35	1.01	1.40	1.00	
apo-CaM	1.36	0.99	1.41	1.01	
BPTI <sup>a</sup>					1.36
CTL9 <sup>b</sup>					1.72
RNase A <sup>c</sup>					1.65
lysozyme <sup>d</sup>					1.69
spc-SH3 <sup>e</sup>					1.32
IFABP <sup>f</sup>					1.75

<sup>a</sup> Reference 94. Reduced disulfide bonds in unfolded state.

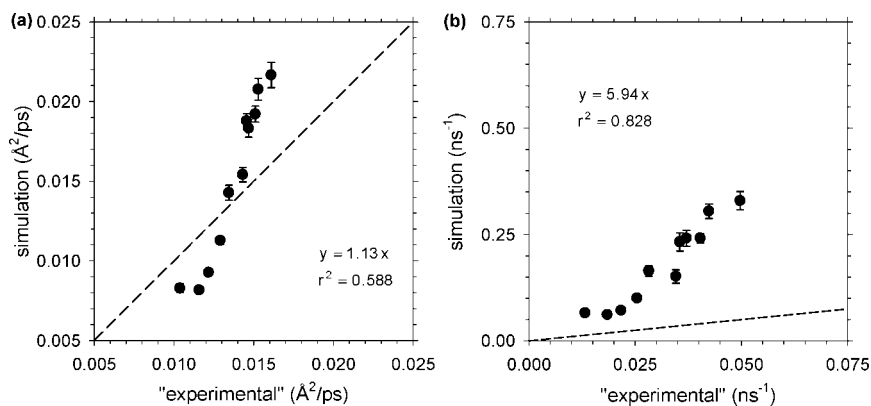
<sup>b</sup> Reference 98. <sup>c</sup> Reference 93. Reduced disulfide bonds in unfolded state. <sup>d</sup> Reference 95. Reduced disulfide bonds in unfolded state. <sup>e</sup> Reference 97. <sup>f</sup> Reference 96. With Alexa bound at V60.

folding can be successfully modeled by implicit-solvent simulations as long as HI are included.

Very similar trends are observed when the rotational diffusion coefficients of the proteins are examined. Figure 3 compares the computed rotational diffusion coefficients of the folded states of the proteins calculated from simulations, both with and without HI, with the corresponding 'experimental' estimates. Again, the values computed from the simulations with HI match very closely with the experimental estimates while those computed from simulations without HI significantly underestimate them; and again, the quality of the linear regression fits are higher for the simulations that include HI, indicating that they more faithfully reproduce the protein size-dependence of the diffusion coefficients. Interestingly however, the magnitude of the error in the non-HI simulations' rotational diffusion coefficients is significantly lower than the error obtained with the translational diffusion coefficients (e.g., compare the slopes of Figure 1a and Figure 3a). This difference suggests that the non-HI simulations are also likely to produce significant errors in the *relative* magnitudes of the translational and rotational



**Figure 3.** Simulation-derived rotational diffusion coefficients of folded proteins plotted against the corresponding 'experimental' values: (a) C $\alpha$  model proteins and (b) SC model proteins.



**Figure 4.** Simulation-derived translational and rotational diffusion coefficients of folded proteins plotted against the corresponding 'experimental' values. All simulations used a C $\alpha$  model, with a hydrodynamic radius of 0.2 Å assigned to all pseudoatoms and neglected hydrodynamic interactions between pseudoatoms: (a) translational diffusion coefficients and (b) rotational diffusion coefficients.

diffusion coefficients, i.e. in the ratio of these two diffusion coefficients. This indeed appears to be the case: the average translational-to-rotational diffusion coefficient ratio for the 11 proteins obtained from the HYDROPRO calculations is  $0.47 \pm 0.14 \times 10^3 \text{ Å}^2$ , and while the corresponding ratios obtained with HI for the C $\alpha$  and SC models are  $0.43 \pm 0.09$  and  $0.46 \pm 0.15 \times 10^3 \text{ Å}^2$ , respectively, the ratios obtained without HI for the two models are  $0.09 \pm 0.03$  and  $0.11 \pm 0.03 \times 10^3 \text{ Å}^2$ , respectively.

The results reported up to this point provide a strong indication that simulations that include HI are much better able to capture all of the investigated aspects of proteins' diffusional behavior than are simulations that neglect HI. As noted in the Discussion, a significant drawback with the HI calculations is their very considerable computational expense; it is therefore worth considering whether simulations that neglect HI (and which have the advantage that they are much faster to compute) could be optimized in some way to better reproduce the experimental behavior. The significant underestimation of both translational and rotational diffusion coefficients suggests that better results might be obtained if an artificially reduced hydrodynamic radius was assigned to the pseudoatoms in non-HI simulations. In fact, a series of simulations using the C $\alpha$  model of the proteins show that a hydrodynamic radius of 0.2 Å can give a reasonably good reproduction of the translational diffusion coefficients (Figure

4a) in the sense that the slope of the linear regression of simulation-derived diffusion coefficients with experimental estimates is now much closer to 1 than is obtained with a more 'reasonable' hydrodynamic radius. Unfortunately, two observations significantly undercut this otherwise promising result. First, the protein size-dependence of the translational diffusion coefficients is poorly reproduced: the translational diffusion coefficients of the smaller proteins (top right of Figure 4a) are consistently overestimated (by up to 35%), while those of the larger proteins (bottom left of the figure) are consistently underestimated (by as much as 30%). Second, the rotational diffusion coefficients of the proteins obtained from the same simulations are overestimated by a factor of  $\sim 6$  (see Figure 4b). These results therefore suggest that a simultaneous reproduction of all of the diffusional properties of a protein in an implicit-solvent model that does not include any modeling of HI will not be possible, even if the hydrodynamic radius of the pseudoatoms is treated as an adjustable parameter.

**Folding Simulations.** A potential consequence of the diffusional studies presented thus far is that the simulated rates of folding of the same model proteins might also be significantly affected by the inclusion of HI. To address this issue folding rates have been computed for the proteins (both with and without HI) by measuring the time taken to adopt

**Table 3.** Ratios of the Folding Rates Obtained with HI Included to the Folding Rates Obtained without HI Included<sup>a</sup>

protein	C $\alpha$	SC
$\alpha$ -helix	0.28 $\pm$ 0.04	
$\beta$ -hairpin	0.73 $\pm$ 0.05	
protein G	1.84 $\pm$ 0.25	
protein L	1.72 $\pm$ 0.40	3.15 $\pm$ 0.18
CI2	1.36 $\pm$ 0.38	
barnase	2.13 $\pm$ 0.28	
fyn-SH3	1.24 $\pm$ 0.17	
CSPB	1.52 $\pm$ 0.27	2.38 $\pm$ 0.60
IFABP	2.29 $\pm$ 0.39	
SFVP	1.86 $\pm$ 0.31	
$\lambda$ -repressor	1.65 $\pm$ 0.25	3.66 $\pm$ 0.32
Im9	1.69 $\pm$ 0.11	
apo-CaM	1.70 $\pm$ 0.27	

<sup>a</sup> Folding rates are computed from the reciprocal of the mean folding time of 100 independent folding trajectories for each protein.

the native conformation in series of 100 independent simulations that start from randomly constructed unfolded conformations; this procedure has been carried out for all 11 proteins with the C $\alpha$  representation and three of the proteins with the SC representation (owing to the significant computational expense of the latter simulations). For comparison purposes, similar simulations have also been run for the folding of a 16-residue  $\alpha$ -helix and a 16-residue  $\beta$ -hairpin. The results of all these studies are compiled in Table 3 in the form of ratios of the folding rate computed with HI to the rate computed without HI. For all *proteins* studied, the inclusion of HI causes a significant increase in the computed rate of folding: for the 11 C $\alpha$  models investigated the average ratio is 1.73  $\pm$  0.30; for the three SC models studied, the average ratio is 3.07  $\pm$  0.64, indicating that the acceleration of the folding rate is significantly greater when the more structurally detailed SC model is used.

The effects of HI on the folding rates can be explored in greater detail by examining separately their effect on first, the initial transition from an extended, unfolded conformation to a compact, collapsed conformation, and second, the transition from this collapsed state to the final, native state (see Methods). As might be expected given the large effects of HI on proteins' diffusional characteristics, the primary effect of HI on folding rates is, for the most part, exerted in the initial collapse phase: for the eleven C $\alpha$  model proteins, the ratio of the 'collapse' rates with and without HI is 2.44  $\pm$  0.75, while the ratio of the 'search' rates with and without HI is 1.34  $\pm$  0.20 (see Table 4). This qualitative picture is preserved for the three SC model proteins that we have studied, although the large error bars for the search ratio of CSPB prevent very firm conclusions being drawn.

The increased folding rates that result from the inclusion of HI when complete proteins are considered is in marked contrast to what is observed when the folding rates of the individual secondary structure elements are measured: for both the  $\alpha$ -helix and the  $\beta$ -hairpin the inclusion of HI causes a very significant *decrease* in folding rate (Table 3). In both the  $\alpha$ -helix and the  $\beta$ -hairpin the inter-residue contacts that are formed are, by construction, local since both of them are only 16 residues long; in proteins of course, the inter-

**Table 4.** Ratios of the 'Collapse' and 'Search' Rates Obtained with HI to the Rates Obtained without HI

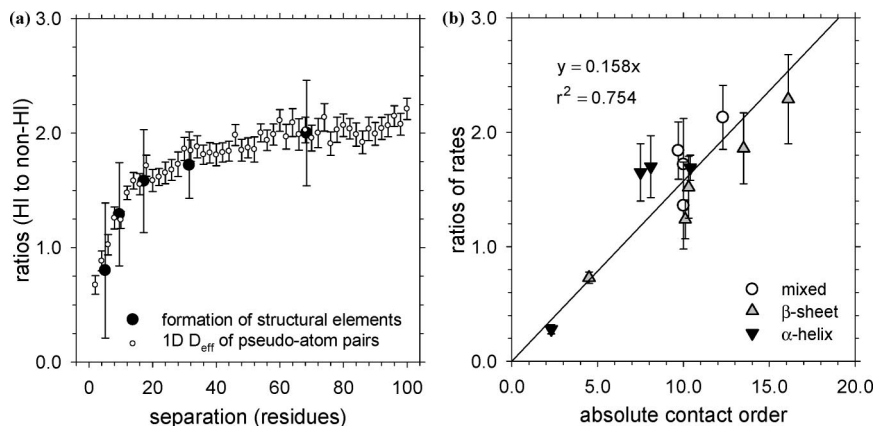
protein	collapse C $\alpha$	collapse SC	search C $\alpha$	search SC
protein G	2.35 $\pm$ 0.38		1.73 $\pm$ 0.29	
protein L	1.96 $\pm$ 0.27	3.38 $\pm$ 0.24	1.30 $\pm$ 0.32	0.86 $\pm$ 0.51
CI2	2.05 $\pm$ 0.21		1.20 $\pm$ 0.35	
barnase	2.79 $\pm$ 0.17		1.47 $\pm$ 0.47	
fyn-SH3	1.37 $\pm$ 0.23		1.13 $\pm$ 0.18	
CSPB	1.89 $\pm$ 0.18	2.35 $\pm$ 0.64	1.19 $\pm$ 0.47	2.41 $\pm$ 3.26
IFABP	2.98 $\pm$ 0.44		1.46 $\pm$ 0.44	
SFVP	2.42 $\pm$ 0.38		1.20 $\pm$ 0.55	
$\lambda$ -repressor	2.25 $\pm$ 0.78	3.81 $\pm$ 0.34	1.59 $\pm$ 0.11	1.37 $\pm$ 0.63
Im9	2.49 $\pm$ 0.34		1.09 $\pm$ 0.20	
apo-CaM	4.28 $\pm$ 0.70		1.39 $\pm$ 0.27	

residue contacts formed in the native state are of both local and nonlocal origins.<sup>91</sup> One potential explanation for the difference between the two sets of results therefore is that the inclusion of HI accelerates the formation of nonlocal interactions while decelerating the formation of local interactions. This idea has been investigated by analyzing the folding trajectories in more detail: specifically, the time taken for structural elements to fold and associate has been measured and correlated with the number of residues that separate the members of the elements along the polypeptide chain (see Methods). The combined results of such an analysis carried out on all 11 C $\alpha$  model proteins are illustrated by the closed symbols in Figure 5a where the relative rates of association of the pairs of structure elements obtained with and without HI are plotted along the y-axis. Clearly, the plot matches the expectation expressed above: the inclusion of HI decreases somewhat the rate at which closely spaced structure elements come into contact with each other but increases the rate at which more widely spaced structure elements associate.

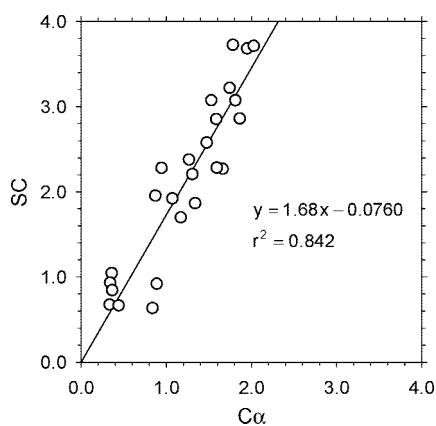
Since the association rates of structure elements in the model proteins are likely to be influenced by the folding status of intervening structural elements, it is worth considering whether a more direct connection between the association rates of structural elements and the diffusional properties of the polypeptide chain can be obtained. This has been done by performing additional BD simulations of a 108-residue C $\alpha$ -only peptide chain in which no favorable native interactions operate (i.e., one in which the only forces that operate are those acting on pseudobonds and pseudoangles and those acting to prevent steric overlap). From these simulations, an effective translational diffusion coefficient has been defined for each residue pair by applying the one-dimensional Einstein equation (see Methods) to the variation of the inter-residue distance. The ratio of the effective inter-residue diffusion coefficients,  $D_{\text{eff}}$ , obtained with and without HI is plotted as the open symbols in Figure 5a, from which it can be seen that it matches very well with the ratio of association rates for the structure elements.

The finding that the association and folding of nonlocal structural elements is accelerated more by HI than is the association and folding of more local elements would lead one to expect that the proteins for which nonlocal contacts predominate should exhibit the greatest relative increase in folding rate when HI are included. The average sequence locality of native contacts in proteins can be conveniently





**Figure 5.** Comparison of the relative rates of folding to relative effective diffusion and absolute contact order. (a) The relative rates of formation of secondary structure elements are compared to the relative  $D_{\text{eff}}$  of pseudoatoms pairs. From left to right, the first closed symbol represents helices and is offset by one-half the average length of the helices; the remaining closed symbols represent the average separation of pairs of secondary structure elements with separations (as measured by the number of residues between the midpoints of each element) of 6–11, 12–25, 26–39, and 40 and more residues, respectively. (b) The relative rates of folding of the eleven  $C\alpha$  model proteins, the  $\alpha$ -helix, and the  $\beta$ -hairpin compared to absolute contact order.



**Figure 6.** Ratios of the rates of formation of structure elements calculated from simulations using the SC model plotted against those from simulations using the  $C\alpha$  model.

represented by the ‘absolute contact order’ measure,<sup>68</sup> and when plotting the ratio of the folding rates obtained with and without HI against this measure (Figure 5b), we do indeed find that the relative folding rates tend to increase as the prevalence of nonlocal contacts increases. The relationship is especially strong for  $\beta$ -sheet proteins, for which our sample of proteins covers a quite broad range of contact order values, but is unconvincing for the purely  $\alpha$ -helical proteins unless the lone  $\alpha$ -helix is included in the correlation; this may simply reflect the fact that the range of contact orders of the chosen  $\alpha$ -helical proteins is comparatively narrow.

A final issue to note comes from comparing the folding behaviors of the three proteins for which simulations were carried out with both  $C\alpha$  and SC models. The same tendency that is observed with  $C\alpha$  models for nonlocal interactions to be more accelerated by the inclusion of HI is also seen with the SC models; in fact, a straightforward linear relationship is observed between the HI-induced accelerations of structural element association obtained with the two models (Figure 6). Interestingly, the slope of the linear regression for this plot is 1.68, which matches closely (as

expected) with the ratio of HI-induced accelerations of folding obtained with the two models ( $3.07/1.73 = 1.77$ ).

Finally, we have performed a simple comparison of the folding pathways in the presence and absence of HI by correlating the average rank-order in which the structural elements of each protein formed over 100 folding trajectories (see Methods). Perhaps surprisingly, despite HI’s effect of increasing the relative rates of formation of elements with increasing separation along the peptide chain, the average rank-order in which the elements formed was essentially unaffected: for each of the eleven proteins, the correlation coefficient of the rank-orders obtained with and without HI was at least 0.98 (details shown in the Supporting Information.)

## Discussion

Although the modeling of protein folding events has been the subject of a very large number of simulation studies (for reviews, see refs 1–3), very few have considered in detail the diffusional characteristics of the polypeptide chain.<sup>21,22</sup> The present study has investigated the simulated diffusional properties of 11 proteins and has found that the inclusion of hydrodynamic interactions (HI) between the pseudoatoms of the modeled proteins plays a critical role in allowing these properties to be correctly reproduced. Specifically, the results have indicated that the inclusion of HI dramatically improves the modeling of (a) the translational diffusion coefficient of a flexible protein model, (b) the change in the translational diffusion coefficient that accompanies folding, (c) the rotational diffusion coefficient of a flexible protein model, (d) the relative magnitudes of the translational and rotational diffusion coefficients, and (e) the protein size-dependence of the translational and rotational diffusion coefficients. The fact that all of these properties are correctly captured by simulations that include HI—regardless of the level of structural detail employed in the model—but are very poorly reproduced by simulations that omit HI, argues strongly that some kind of a HI treatment should be included in any molecular simulation that aims to address a problem in which

both diffusion and folding of a protein are likely to be important factors. In passing it is to be noted that while HI arise naturally in simulations that involve explicit solvent, this does not guarantee that a protein's actual diffusional properties will be quantitatively reproduced: in fact, a recent explicit-solvent MD simulation study found the diffusion coefficients of the 76-amino acid protein ubiquitin to be sensitive to both system size (an effect originally identified by Yeh and Hummer<sup>99</sup>) and the water model used.<sup>21</sup>

Before discussing the details and limitations of the particular simulation model used here, it is important to consider the reliability of the 'experimental' estimates for the translational and rotational diffusion coefficients used here. We have chosen to use HYDROPRO<sup>88,89</sup> as our source of 'gold standard' data here on the basis that for comparing the behavior for a number of different proteins it is important to have reference data that are obtained under identical, standardized conditions. Unfortunately, while there are many experimental estimates available for the translational and rotational diffusion coefficients of proteins, they are only very rarely reported under the same conditions. An attempt to correlate simulated properties with experimental data that are all obtained under slightly different conditions, and/or with different techniques, could be a perilous undertaking, especially given the comparatively small differences in the values for different proteins (e.g., the slowest and fastest diffusing proteins in the current data set have translational diffusion coefficients that differ only by a factor of  $\sim 1.5$ ). The use of computational estimates from HYDROPRO allows this issue of nonidentical conditions to be avoided, but, of course, it can only be justified if the computational estimates themselves can be considered reliable. Fortunately, from comparisons reported by the group of Garcia de la Torre,<sup>88,89</sup> it appears that the errors in HYDROPRO's estimates of the translational and rotational diffusion coefficients are very minor (2 and 6%, respectively). Since all of the proteins studied here are similar in size to those examined in the previous comparisons, no special difficulties or errors are likely to arise in the present HYDROPRO estimates.

There are, of course, a number of issues regarding the simulations themselves that must also be addressed. The first is the energy model used to describe inter-residue interactions in the simulations. The 'native-centric' G $\ddot{o}$  model<sup>37</sup> is unashamedly simplified: it does not consider potential favorable non-native interactions and, at least in most implementations, makes no attempt to use different kinds of energy functions to model different kinds of native interactions (e.g., hydrophobic contacts versus salt bridges). Despite these obvious limitations, the model appears surprisingly robust in its ability to describe key aspects of protein folding events:<sup>8,38–42</sup> the folding rates, for example, of a wide range of single domain proteins can be successfully reproduced by a structural and energetic model that is essentially identical to the one used here,<sup>40</sup> and, perhaps surprisingly, the same kind of model also appears able to capture changes in stability that are caused by the truncation of the polypeptide chain.<sup>42</sup> For the bulk of the work that is reported here—i.e. the exploration of the diffusional properties of simulated proteins—the exact details of the energetic

model used in the simulations are, in any case, almost certainly unimportant; we have for example been careful to ascertain that the simulated diffusional properties are unaffected by the choice of the energy well-depth,  $\epsilon$ , used in the simulations (see the Supporting Information). In fact, for simulating the diffusional properties of the folded states of proteins it is likely that *any* energetic model that retains the proteins roughly in their native shapes (e.g., a Gaussian network model<sup>100</sup>) could be used with similar success.

Of course, for modeling actual folding events the details of the energetic model are likely to be more important; however, the key point to note here is that the same energetic and structural models have been used in the simulations that include HI and those that neglect HI. As a consequence, the inherent limitations of the G $\ddot{o}$  model are also likely to have minimal impact on the central conclusion that has been reached regarding the effects of HI on folding rates: namely, that the inclusion of HI leads to a 2- to 3-fold acceleration of folding, depending on the level of structural detail employed in the model. Importantly, in the Supporting Information we show that this acceleration is not affected by the simulated stability of the protein: varying the energy well-depth,  $\epsilon$ , associated with all native contacts anywhere within the range 0.55 to 0.65 kcal/mol causes no significant change in the degree to which HI accelerates folding for protein L. Although those results indicate that changing the stability of all contacts simultaneously does not affect the contribution of HI to the folding rate, it could be interesting in the future to examine whether the inclusion of HI affects the rate of formation of different *kinds* of contacts differently, e.g. to see whether HI accelerates formation of favorable electrostatic contacts more than the rate of formation of hydrophobic contacts: given the residue separation-dependence of the HI effect (Figure 5), one might imagine that it could have a greater impact on interactions that can act over longer distances. To investigate this issue however would require the use of an energetic model that is somewhat more sophisticated than the one employed here; interestingly, just such a combined G $\ddot{o}$  + electrostatic model has already been used recently to investigate the electrostatically accelerated rates of flexible protein-DNA association events.<sup>101</sup>

A second issue connected with the technical details of the simulations concerns the specific model used to describe the hydrodynamic interactions acting between pseudoatoms. As described in Methods, the model used is one developed independently (as a modification of the Oseen tensor<sup>102</sup> by Rotne and Prager<sup>76</sup> and Yamakawa<sup>77</sup>); while not the only hydrodynamic model that is available (see for example the somewhat more sophisticated models implemented in the Stokesian dynamics methods of Brady and co-workers<sup>103</sup>), it has been chosen here owing to the fact that it is both readily implemented and is already comparatively widely used in the simulation of macromolecular conformational dynamics. The RPY model has, for example, been used (a) in BD simulations of a flexible loop that acts as a 'gate' for substrate access in the enzyme triose phosphate isomerase,<sup>104</sup> (b) in recent simulations of nucleosomal dynamics in which histone tails are treated at a subresidue level,<sup>105</sup> (c) in simulations of the shear flow-induced unfolding of N-terminus-tethered

ubiquitin,<sup>55</sup> and (d) in simulations examining the effects of HI on the folding kinetics of simple secondary structure elements<sup>22,53</sup> (discussed in detail below). The RPY description of HI therefore, while conceived many years ago, still in some respects represents the current ‘state of the art’; this is largely due to the fact that the significant computational resources needed to routinely handle HI calculations in long simulations (see below) have only recently become available. The fact that the modeling of HI with the RPY model apparently produces such a good description of the diffusional characteristics of flexible proteins argues quite strongly that it should find wider application in similar implicit-solvent simulations.

The computational expense associated with the HI calculations employed here should not however be overlooked. Although we have found that one benefit of including HI is that it can allow larger integration timesteps to be used than in simulations that neglect HI (though due to the need to match internal energies for HI and non-HI simulations, we did not always take advantage of this fact; see Methods), this speed-up is nowhere near sufficient to completely offset the additional computational burden involved in their computation. In fact, for the largest protein studied here (apocalmodulin), the computational time required for a simulation performed with HI was 5.6-fold greater than that required for the corresponding non-HI simulation when a C $\alpha$  model was used and was 40-fold greater when a SC model was used. The preceding numbers, it should be remembered, were obtained from simulations that updated the diffusion tensor only every 1 ps or every  $\sim 25$  steps (see Methods); this is an approach commonly used by others,<sup>106–108</sup> but to ensure that it is appropriate in the present setting, additional control simulations were performed to show that altering the frequency of updates of the HI tensors caused no change in any of the simulation observables (see the Supporting Information). Even with infrequent updates of the hydrodynamic tensors it is clear that for much larger systems the inclusion of HI with the current approach will result in a huge increase in computer time; the development of very fast HI methods—or at least methods that scale better with system size—will therefore likely remain an important pursuit.<sup>109–111</sup>

It is due to the computational expense of the simulations that we have not carried out a more detailed study of the effects, if any, of the inclusion of HI on the folding mechanisms of the proteins. Ideally, it would be of interest to identify a transition state ensemble (TSE) for each protein, either by finding conformations with ‘Q’ values corresponding to the free energy maximum on the folding free energy landscape<sup>8,112,113</sup> or by explicitly testing candidate conformations to ensure that they have a 50% chance of proceeding to the folded state (the ‘P<sub>fold</sub>’ approach).<sup>114,115</sup> For the present study, both methods are computationally prohibitive: the former approach requires that we first compute the folding free energy landscape as a function of Q, which, in our previous study required simulations of at least 100  $\mu$ s for the small protein, CI2;<sup>42</sup> the latter approach requires that many independent trajectories be run for each candidate conformation in order to obtain reasonable statistical esti-

mates of folding probabilities. Because of these issues, we have chosen to restrict our attention to comparing the order in which the various structural elements fold and assemble in simulations performed with and without HI. It is perhaps surprising that while the inclusion of HI accelerates folding significantly in our simulations, it does not cause any *obvious* change in the folding mechanism, at least insofar as it is reflected in the order of structural element association. Given that we find that HI exerts a stronger effect on the folding and association of elements that are more distantly separated in sequence space, it is not inconceivable that qualitative changes in folding mechanisms might be found for other proteins, especially perhaps for multidomain proteins in which widely separated domains must assemble.

Although the present study provides a reasonably broad comparison of the effects of including HI on protein diffusion and folding, it should be noted that two studies prior to this one have explored HI effects on the folding rates of simpler model systems. The first, and most directly comparable study, is that of Baumketner and Hiwatari<sup>22</sup> who investigated the effects of HI on the rates of folding of an  $\alpha$ -helix and a  $\beta$ -hairpin, each 16 residues long, using methods very similar to those employed here. In their study, it was reported that HI, modeled using the RPY equations, caused a 2-fold slowing of the folding rate for the  $\beta$ -hairpin but caused no change in the folding rate of the model  $\alpha$  helix. The former result is qualitatively similar to what we observe (Table 3); the latter result however contrasts markedly with the 4-fold deceleration of folding that we observe to be caused by HI. As is often the case when comparing simulation studies, it is not easy to determine unambiguously the reason for the discrepancy; possible sources of the difference would appear to be (a) differences in the studies’ energy functions, for both bonded and nonbonded interactions, and (b) the use of the bond constraint algorithm SHAKE in the HI simulations of Baumketner and Hiwatari.<sup>22</sup>

The only other work that we are aware of having explored the effects of HI on the folding of a freely diffusing protein is a study reported by the Yeomans’ group<sup>53</sup>—which used a HI model quite different from that employed here—and which examined the folding of a model  $\alpha$ -helix and a  $\beta$ -hairpin, each 19 residues in length, and the folding kinetics of the protein CI2 (referred to as ‘2CI2’ in their paper): in all three cases they found that HI caused a negligible change in the folding kinetics. Again, the technical differences between the present study and the previous work are significant: in particular, the model used by Kikuchi et al.<sup>53</sup> (a) models HI with the stochastic rotation dynamics model,<sup>116</sup> (b) assigns no energy parameters to pseudodihedral angles of the modeled polypeptides, and (c) uses the radius of gyration as the reaction coordinate for monitoring folding, rather than any more detailed measure of the formation of native contacts. Interestingly, in the same study, Kikuchi et al.<sup>53</sup> did find that the rate of *polymer* collapse was accelerated with the inclusion of HI, which is, of course (along with other polymer studies of the coil to globule transition<sup>50–52</sup>), in agreement with the results presented here.

The fact that so many aspects of the diffusional behavior of the 11 proteins are correctly captured by the present

simulation model gives us confidence that the basic conclusions of the folding studies are also correct. The key results that we have obtained—that folding is accelerated by 2–3-fold due to the inclusion of HI and that this results primarily from the accelerated formation of long-range native contacts—are likely to be of significance in a number of areas. First, the results may have implications for models of protein folding that emphasize the diffusional-search aspects of the process, such as the diffusion-collision model proposed by Karplus and Weaver<sup>117</sup> and successfully applied to experimental folding kinetics data by Oas and co-workers.<sup>118,119</sup> Second, the findings may be significant for very fast-folding (or ‘downhill’ folding<sup>120</sup>) proteins, as suggested by a recent dynamic Monte Carlo simulation where the kinetic transition state barrier was both shifted and increased as a result of configuration-dependent diffusion.<sup>121</sup> Third, they are also likely to be of importance for models and studies that attempt to understand the diffusion-limited kinetics of loop-closure events, an area that has already seen fruitful interplay between experiment and simulation.<sup>122,123</sup> Fourth, they may be important to keep in mind when attempts are made to directly compare folding kinetics obtained from implicit- and explicit-solvent simulations;<sup>124</sup> although such comparisons are in any case likely to be severely hampered by differences in the underlying folding free energy landscapes,<sup>125–128</sup> the present study emphasizes the fact that simulations that use a non-HI implicit solvent model also differ—in a hydrodynamic, rather than an energetic sense—from corresponding explicit-solvent simulations. Finally, it may turn out to be quite important to note that—despite the accelerated folding—proteins modeled with HI diffuse 3–4 times farther during the time it takes them to fold than do proteins modeled without HI (see the Supporting Information). Although this might not be a particularly important observation for modeling of a fundamentally intramolecular event such as single-domain protein folding (after all, we see here no obvious changes in the apparent folding mechanisms caused by HI inclusion), it would seem to have rather obvious implications for modeling the kinetics of intermolecular events such as the coupled folding and binding of proteins<sup>101</sup> or peptide- and protein-aggregation processes.<sup>129</sup> For correct modeling of such situations therefore, the continued development of fast methods for modeling hydrodynamic interactions might prove to be rather important.

**Acknowledgment.** T.F.K. is grateful for the support of an Iowa Presidential Fellowship. This work was supported in part by a grant to A.H.E. from the Carver Trust.

**Supporting Information Available:** Details regarding the conversion of all-atom protein structures to reduced pseudoatom models; control simulations for the effect of time-step choice on calculated diffusion coefficients; selection of timesteps for folding simulations and the timestep dependence of the proteins’ simulated energies; the effects of a bond constraint algorithm in simulations with HI; control simulations for the effect(s) of diffusion tensor update interval on simulation observables; finding the optimal hydrodynamic radius for the C $\alpha$  and SC models; control simulations for the effect(s) of observation interval on calculated diffusion coefficients; com-

parison of the folding pathways; control simulations for the effect of standardized energy parameters on calculated diffusion coefficients and the acceleration of folding in simulations with HI; and comparison of the average net distance traveled by the proteins prior to folding. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Daggett, V.; Fersht, A. R. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 497–502.
- (2) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.
- (3) Chen, Y.; Ding, F.; Nie, H.; Serohijos, A. W.; Sharma, S.; Wilcox, K. C.; Yin, S.; Dokhoyan, N. V. *Arch. Biochem. Biophys.* **2008**, *469*, 4–19.
- (4) Carrell, R. W.; Lomas, D. A. *Lancet* **1997**, *350*, 134–138.
- (5) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.
- (6) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694–698.
- (7) Kolinski, A.; Skolnick, J. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 353–366.
- (8) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937–953.
- (9) Northrup, S. H.; Erickson, H. P. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 3338–3342.
- (10) Gabdoulline, R. R.; Wade, R. C. *J. Mol. Biol.* **2001**, *306*, 1139–1155.
- (11) Hagan, M. F.; Chandler, D. *Biophys. J.* **2006**, *91*, 42–54.
- (12) McGuffee, S. R.; Elcock, A. H. *J. Am. Chem. Soc.* **2006**, *128*, 12098–12110.
- (13) Chong, L. T.; Snow, C. D.; Rhee, Y. M.; Pande, V. S. *J. Mol. Biol.* **2005**, *345*, 869–878.
- (14) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.
- (15) Ding, F.; Dokhoyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *J. Mol. Biol.* **2002**, *324*, 851–857.
- (16) Yang, S. C.; Cho, S. S.; Levy, Y.; Cheung, M. S.; Levine, H.; Wolynes, P. G.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13786–13791.
- (17) Smith, A. V.; Hall, C. K. *J. Mol. Biol.* **2001**, *312*, 187–2002.
- (18) Dima, R. I.; Thirumalai, D. *Protein Sci.* **2002**, *11*, 1036–1049.
- (19) Gsponer, J.; Haberthur, U.; Caffisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5154–5159.
- (20) Matysiak, S.; Clementi, C. *J. Mol. Biol.* **2006**, *363*, 297–308.
- (21) Takemura, K.; Kitao, A. *J. Phys. Chem. B* **2007**, *111*, 11870–11872.
- (22) Baumketner, A.; Hiwatari, Y. *J. Phys. Soc. Jpn.* **2002**, *71*, 3069–3079.
- (23) Taddei, N.; Capanni, C.; Chiti, F.; Stefani, M.; Dobson, C. M.; Ramponi, G. *J. Biol. Chem.* **2001**, *276*, 37149–37154.
- (24) Chiti, F.; Taddei, N.; Baroni, F.; Capanni, C.; Stefani, M.; Ramponi, G.; Dobson, C. M. *Nat. Struct. Biol.* **2002**, *9*, 137–143.

- (25) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6456–6461.
- (26) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 164902.
- (27) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740–744.
- (28) García, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898–13903.
- (29) Pitera, J. W.; Swope, W. C.; Abraham, F. F. *Biophys. J.* **2008**, *98*, 4837–4846.
- (30) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 927–937.
- (31) Jagielska, A.; Scheraga, H. A. *J. Comput. Chem.* **2007**, *28*, 1068–1082.
- (32) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (33) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.
- (34) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.
- (35) Ooi, T.; Oobatake, M.; Némethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.
- (36) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.
- (37) Taketomi, H.; Ueda, Y.; Gō, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445–459.
- (38) Hoang, T. X.; Cieplak, M. *J. Chem. Phys.* **2004**, *113*, 8319–8328.
- (39) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171–180.
- (40) Chavez, L. L.; Onuchic, J. N.; Clementi, C. *J. Am. Chem. Soc.* **2004**, *126*, 8426–8432.
- (41) Das, P.; Wilson, C. J.; Fossait, G.; Wittung-Stafshede, P.; Matthews, K. S.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14569–14574.
- (42) Elcock, A. H. *PLoS Comput. Biol.* **2006**, *2*, 824–841.
- (43) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press Inc.: New York, 1989.
- (44) Larson, R. G. *Constitutive equations for polymer melts and solutions*; Butterworths: Boston, 1988.
- (45) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352–1360.
- (46) Zimm, B. H. *J. Chem. Phys.* **1956**, *24*, 269–278.
- (47) Kirkwood, J. G.; Riseman, J. *J. Chem. Phys.* **1948**, *16*, 565–573.
- (48) Rouse, P. E. *J. Chem. Phys.* **1953**, *21*, 1272–1280.
- (49) Doi, M.; Edwards, S. F. *International Series of Monographs on Physics 73: The Theory of Polymer Dynamics*; Oxford University Press: New York, 1986.
- (50) Kuznetsov, Y. A.; Timoshenko, E. G.; Dawson, K. A. *J. Chem. Phys.* **1996**, *104*, 3338–3347.
- (51) Pitard, E. *Eur. Phys. J. B* **1999**, *7*, 665–673.
- (52) Kikuchi, N.; Gent, A.; Yeomans, J. M. *Eur. Phys. J. E* **2002**, *9*, 66–66.
- (53) Kikuchi, N.; Ryder, J. F.; Pooley, C. M.; Yeomans, J. M. *Phys. Rev. E* **2005**, *71*, 061804.
- (54) Pham, T. T.; Bajaj, M.; Prakash, J. R. *Soft Matter* **2008**, *4*, 1196–1207.
- (55) Szymczak, P.; Cieplak, M. *J. Chem. Phys.* **2007**, *127*, 155106.
- (56) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4729.
- (57) O’Neill, J. W.; Kim, D. E.; Baker, D.; Zhang, K. Y. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2001**, *57*, 480–487.
- (58) McPhalen, C. A.; James, M. N. *Biochemistry* **1987**, *26*, 261–296.
- (59) Buckle, A. M.; Henrick, K.; Fersht, A. R. *J. Mol. Biol.* **1993**, *234*, 847–860.
- (60) Noble, M. E.; Musacchio, A.; Saraste, M.; Courtneidge, S. A.; Wierenga, R. K. *EMBO J.* **1993**, *12*, 2617–2624.
- (61) Schindelin, H.; Marahiel, M. A.; Heinemann, U. *Nature* **1993**, *364*, 164–168.
- (62) Scapin, G.; Gordon, J. I.; Sacchettini, J. C. *J. Biol. Chem.* **1992**, *267*, 4253–4269.
- (63) Choi, H. K.; Lu, G.; Lee, S.; Wengler, G.; Rossmann, M. G. *Proteins* **1997**, *27*, 345–359.
- (64) Beamer, L. J.; Pabo, C. O. *J. Mol. Biol.* **1992**, *227*, 177–196.
- (65) Osborne, M. J.; Breeze, A. L.; Lian, L. Y.; Reilly, A.; James, R.; Kleanthous, C.; Moore, G. R. *Biochemistry* **1996**, *35*, 9505–9512.
- (66) Kuboniwa, H.; Tjandra, N.; Grzesiek, S.; Ren, H.; Klee, C. B.; Bax, A. *Nat. Struct. Biol.* **1995**, *2*, 768–766.
- (67) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (68) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Protein Sci.* **2003**, *12*, 2057–2062.
- (69) Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52–56.
- (70) Wallqvist, A.; Ullner, M. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 267–280.
- (71) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271–1282.
- (72) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635–2638.
- (73) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.
- (74) Plaxco, K. W.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13591–13596.
- (75) Knott, M.; Kaya, H.; Chan, H. S. *Polymer* **2004**, *45*, 623–632.
- (76) Rotne, J.; Prager, S. *J. Chem. Phys.* **1969**, *50*, 4831–4837.
- (77) Yamakawa, H. *J. Chem. Phys.* **1970**, *53*, 436–443.
- (78) Larson, R. G. *Mol. Phys.* **2004**, *102*, 341–351.
- (79) Dalquist, G.; Bjork, A. *Numerical methods*. Prentiss Hall: Englewood Cliffs: NJ, 1974.
- (80) Schlick, T.; Beard, D. A.; Huang, J.; Strahs, D. A.; Qian, X. L. *IEEE Comp. Sci. Eng.* **2000**, *2*, 38–51.
- (81) Hurler, M. R.; Michelotti, G. A.; Crisanti, M. M.; Matthews, C. R. *Proteins* **1987**, *2*, 54–63.

- (82) Jacob, M.; Schindler, T.; Balbach, J.; Schmid, F. X. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 5622–5627.
- (83) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (84) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *8*, 1463–1472.
- (85) Allison, S. A.; McCammon, J. A. *Biopolymers* **1984**, *23*, 167–187.
- (86) Price, W. S.; Tsuchiya, F.; Arata, Y. *J. Am. Chem. Soc.* **1999**, *121*, 11503–1512.
- (87) Masuda, A.; Ushida, K.; Nishimura, G.; Kinjo, M.; Tamura, M.; Koshino, H.; Yamashita, K.; Kluge, T. *J. Chem. Phys.* **2004**, *121*, 10787–10793.
- (88) Garcíde la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719–730.
- (89) Garcíde la Torre, J. *Biophys. Chem.* **2001**, *93*, 159–170.
- (90) Creighton, T. E. *Proteins: Structures and Molecular Properties*; W.H. Freeman & Co. Ltd.: New York, 1992.
- (91) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985–994.
- (92) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
- (93) Noppert, A.; Gast, K.; Muller-Frohne, M.; Zirwir, D.; Damashun, G. *FEBS Lett.* **1996**, *380*, 179–182.
- (94) Pan, H.; Barany, G.; Woodward, C. *Protein Sci.* **1997**, *6*, 1985–1991.
- (95) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. *Biochemistry* **1999**, *38*, 16424–16431.
- (96) Chattopadhyay, K.; Saffarian, S.; Elson, E. L.; Frieden, C. *Biophys. J.* **2005**, *88*, 1413–1422.
- (97) Casares, S.; Sadqi, M.; Lopez-Mayorga, O.; Conejero-Lara, F.; van Nuland, N. A. J. *Biophys. J.* **2004**, *86*, 2403–2413.
- (98) Li, Y.; Shan, B.; Raleigh, D. P. *J. Mol. Biol.* **2007**, *368*, 256–262.
- (99) Yeh, I. C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873–14879.
- (100) Haliloglu, T.; Bahar, I.; Erman, B. *Phys. Rev. Lett.* **1997**, *79*, 3090–3093.
- (101) Levy, Y.; Onuchic, J.; Wolynes, P. G. *J. Am. Chem. Soc.* **2007**, *129*, 738–739.
- (102) Oseen, C. W. *Neuere Methoden und Ergebnisse in der Hydrodynamik*; Akademische Verlagsgesellschaft: Leipzig, 1927.
- (103) Brady, J. F.; Bossis, G. *Ann. Rev. Fluid Mech.* **1988**, *20*, 111–157.
- (104) Wade, R. C.; Davis, M. E.; Luty, B. A.; Madura, J. D.; MaCammon, J. A. *Biophys. J.* **1993**, *64*, 9–15.
- (105) Arya, G.; Zhang, Q.; Schlick, T. *Biophys. J.* **2006**, *91*, 133–150.
- (106) Jian, H. M.; Schlick, T.; Vologodskii, A. *J. Mol. Biol.* **1998**, *284*, 287–296.
- (107) Huang, J.; Schlick, T.; Vologodskii, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 968–973.
- (108) Klenin, K.; Merlitz, H.; Langowski, J. *Biophys. J.* **1998**, *74*, 780–788.
- (109) Fixman, M. *Macromolecules* **1986**, *19*, 1204–1207.
- (110) Sierou, A.; Brady, J. F. *J. Fluid Mech.* **2001**, *448*, 115–146.
- (111) Geyer, T.; Winter, U. *Soft Cond. Matt.* [Online] 2008, arXiv: 0801.3212v1.
- (112) Onuchic, J. N.; Socci, N. D.; LutheySchulten, Z.; Wolynes, P. G. *Fold. Des.* **1996**, *1*, 441–450.
- (113) Nymeyer, H.; Socci, N. D.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 634–639.
- (114) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
- (115) Li, L.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13014–13018.
- (116) Malevanets, A.; Kapral, R. *J. Chem. Phys.* **1999**, *110*, 8605–8613.
- (117) Karplus, M.; Weaver, D. L. *Biopolymers* **1979**, *18*, 1421–1437.
- (118) Burton, R. E.; Myers, J. K.; Oas, T. G. *Biochemistry* **1998**, *16*, 5337–5343.
- (119) Myers, J. K.; Oas, T. G. *Nat. Struct. Biol.* **2001**, *8*, 552–558.
- (120) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191–2195.
- (121) Chahine, J.; Oliveira, R. J.; Leite, V. B. P.; Wang, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14646–14651.
- (122) Yeh, I. C.; Hummer, G. *J. Am. Chem. Soc.* **2002**, *124*, 6563–6568.
- (123) Fierz, B.; Kiefhaber, T. *J. Am. Chem. Soc.* **2007**, *129*, 672–679.
- (124) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Annu. Rev. Biomol. Struct.* **2005**, *34*, 43–69.
- (125) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.
- (126) Nymeyer, H.; García, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934–13939.
- (127) Zhou, R. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 148–161.
- (128) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.
- (129) Nguyen, H. D.; Hall, C. K. *J. Biol. Chem.* **2005**, *280*, 9074–9082.

CT800499P

# JCTC

Journal of Chemical Theory and Computation

## Coupling the Level-Set Method with Molecular Mechanics for Variational Implicit Solvation of Nonpolar Molecules

Li-Tien Cheng,<sup>†</sup> Yang Xie,<sup>‡</sup> Joachim Dzubiella,<sup>§</sup> J. Andrew McCammon,<sup>||</sup>  
Jianwei Che,<sup>⊥</sup> and Bo Li<sup>\*,#</sup>

Department of Mathematics, University of California, San Diego, 9500 La Jolla, California 92093-0112, Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, Physics Department (T37), Technical University Munich, James-Franck-Strasse, 85748 Garching, Germany, Department of Chemistry and Biochemistry, Department of Pharmacology, Howard Hughes Medical Institute, and Center for Theoretical Biological Physics, University of California, San Diego, Gilman Drive, La Jolla, California 92093-0365, The Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, and Department of Mathematics and Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, Mail Code 0112, La Jolla, California 92093-0112

Received July 25, 2008

**Abstract:** We construct a variational explicit-solute implicit-solvent model for the solvation of molecules. Central in this model is an effective solvation free-energy functional that depends solely on the position of solute–solvent interface and solute atoms. The total free energy couples altogether the volume and interface energies of solutes, the solute–solvent van der Waals interactions, and the solute–solute mechanical interactions. A curvature dependent surface tension is incorporated through the so-called Tolman length which serves as the only fitting parameter in the model. Our approach extends the original variational implicit-solvent model of Dzubiella, Swanson, and McCammon [*Phys. Rev. Lett.* **2006**, *96*, 087802 and *J. Chem. Phys.* **2006**, *124*, 084905] to include the solute molecular mechanics. We also develop a novel computational method that combines the level-set technique with optimization algorithms to determine numerically the equilibrium conformation of nonpolar molecules. Numerical results demonstrate that our new model and methods can capture essential properties of nonpolar molecules and their interactions with the solvent. In particular, with a suitable choice of the Tolman length for the curvature correction to the surface tension, we obtain the solvation free energy for a benzene molecule in a good agreement with experimental results.

### I. Introduction

The interaction of biomolecules with an aqueous environment contributes significantly to the solvation free energy, struc-

tures, and functions of biomolecular systems. Efficient descriptions of such interactions are often given by implicit-solvent (or continuum-solvent) models.<sup>1,2</sup> In such models, the solvent molecules and ions are treated implicitly, and their effects are coarse-grained. The effect of solvent is described through the continuum solute–solvent interface and related macroscopic quantities. These models are complementary to the more accurate but computationally expensive explicit-solvent models, such as molecular dynamics simulations which often provide sampled statistical information rather than direct descriptions of thermodynamics.

Most of the existing implicit-solvent models are built upon the concept of solvent-accessible surface (SAS) or solvent-excluded surface (SES) which can be defined in different ways.<sup>3–7</sup> In these models, the solvation free energy consists

\* Corresponding author e-mail: bli@math.ucsd.edu.

<sup>†</sup> Department of Mathematics, University of California.

<sup>‡</sup> Department of Mechanical and Aerospace Engineering, University of California. Current address: Woodruff School of Mechanical Engineering, 801 Ferst Drive, Georgia Institute of Technology, Atlanta, GA 30332-0405.

<sup>§</sup> Technical University Munich.

<sup>||</sup> Department of Chemistry and Biochemistry, Department of Pharmacology, Howard Hughes Medical Institute, and Center for Theoretical Biological Physics, University of California.

<sup>⊥</sup> The Genomics Institute of the Novartis Research Foundation.

<sup>#</sup> Department of Mathematics and Center for Theoretical Biological Physics, University of California.

of the surface energy which is taken to be proportional to the area of a SAS or SES and the electrostatic free energy determined by the Poisson–Boltzmann (PB)<sup>8–10</sup> or Generalized Born (GB)<sup>11–13</sup> description. While a SAS or SES based implicit-solvent approach has been extensively used and successful in many cases, its accuracy and general applicability are still questionable. One of the main issues here is the decoupling and separate descriptions of surface tension, dispersion, and the polar part of the free energy. Moreover, an ad hoc definition of SAS or SES can often lead to inaccurate free-energy estimation. It is additionally well established by now that cavitation free energies do not scale with surface area for high curvatures,<sup>14,15</sup> a fact of critical importance in the implicit-solvent modeling of hydrophobic interactions at molecular scales.<sup>16</sup>

In a recently developed *variational* implicit-solvent approach, Dzubiella, Swanson, and McCammon<sup>17,18</sup> proposed a mean-field approximation of the free energy of an underlying solvation system with an implicit solvent and fixed solute atoms as a functional of all possible solute–solvent interfaces. This free-energy functional couples both the nonpolar and polar contributions of the system. It allows for curvature correction of the surface tension to approximate the length-scale dependence of molecular hydration. Minimization of the free-energy functional determines an equilibrium solute–solvent interface and the minimum free energy of the solvation system. This stable, solute–solvent interface is an output of the theory. It results automatically from balancing the different contributions of the free energy.

Cheng, Dzubiella, McCammon, and Li<sup>19</sup> first developed a *level-set method* for numerically capturing *arbitrarily shaped* equilibrium solute–solvent interfaces that minimize the solvation free-energy functional in the variational implicit-solvent model. In such a method, a possible solute–solvent interface is represented by the zero level-set (i.e., the zero level surface) of a level-set function, and an initial surface is evolved to reduce the free energy, eventually into an equilibrium solute–solvent interface. This relaxation process is determined by solving a time-dependent equation for the level-set function. The solute–solvent interface in each time step is then located as the zero-level surface of the level-set function. Here the time is not that in the real molecular dynamics. Rather it only represents an optimization step. We note that our level-set method for the relaxation of solute–solvent interfaces is quite different from that used in ref 20 which is only for generating a SAS or SES surface.

In the present work, we extend the original variational implicit-solvent model to include the degrees of freedom of all the solute atoms. More specifically, we construct a hybrid, variational explicit-solute implicit-solvent model to couple the coarse-grained solvent with the molecular mechanics of solute atoms. Our new effective free-energy functional depends not only on a solute–solvent interface but also positions of all the solute atoms. The free energy includes both the volume and surface energies of the solutes, the solute–solvent van der Waals interaction, and the electrostatic interaction. The Tolman curvature correction to the constant surface tension is incorporated through the so-called Tolman length which serves as the only fitting parameter in

the model. The free energy also includes the van der Waals interaction as well as different kinds of molecular mechanical interactions among all the solute atoms. These mechanical interactions include the usual bond stretching, bending, and torsion. More terms can be added without difficulty. This explicit-solute implicit-solvent model is a more accurate and robust description of the structure of an underlying solvation system. Notice that we do not need to use solvation radii which are the fitting parameters in a usual SAS or SES type implicit-solvation model.

We also develop a level-set optimization method for the corresponding computer simulations. Our numerical free-energy minimization is carried out by solving two sets of time-dependent equations: one for the level-set function that determines the evolution of the solute–solvent interface and the other for the displacement of solute atoms. Our new level-set techniques include the numerical regularization for solving the level-set equation when instabilities occur and a fast algorithm for numerically evaluating integrals of radially symmetric functions in the free-energy calculation. To efficiently and accurately couple the molecular interactions with the evolution of the solute–solvent interface, we choose carefully mobilities and optimization steps in our computation. While the time here means the optimization step, our approach can be used for the further development of a theory and simulation methods for real dynamics of solvation systems in the framework of variational implicit-solvent.

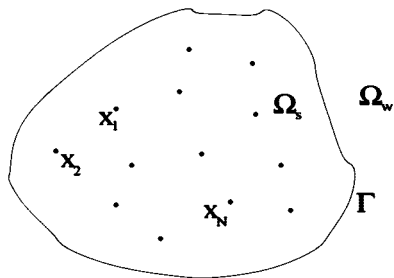
We apply our method to the solvation of the following nonpolar molecular systems: an artificial molecule of two atoms, an artificial molecule of four atoms, the ethane molecule  $C_2H_6$ , and the benzene molecule  $C_6H_6$ . Our extensive numerical results demonstrate that our approach can predict efficiently and accurately the free energy and structure of nonpolar molecules. In particular, with a suitably chosen Tolman length which is the only fitting parameter in our model and computation, we obtain a very good approximation of the experimentally measured solvation free energy for the benzene molecule. We are currently applying our theory and methods to polymers and large biomolecules. We are also working to include the electrostatics of an underlying system.

The rest of the paper is organized as follows: In Section II, we describe our hybrid explicit-solute implicit-solvent model that couples the original variational implicit-solvent with the solute molecular mechanics. In Section III, we present details of our level-set optimization algorithm. We also give formulas for the effective forces that are used as search directions in our numerical optimization. Some details of these formulas are given in the Appendix. In Section IV, we report our results of numerical computations applied to some nonpolar molecular systems. Finally, in Section V, we discuss our results and draw some conclusions.

## II. A Variational Explicit-Solute Implicit-Solvent Model

Our underlying system of molecules in a solution is divided geometrically into three parts: the solute region  $\Omega_s$ , the solvent (e.g., water) region  $\Omega_w$ , and the corresponding





**Figure 1.** The geometry of a solvation system with an implicit solvent. The free energy depends on the position of solute–solvent interface  $\Gamma$  and solute atomic positions  $\mathbf{x}_1, \dots, \mathbf{x}_N$ .

solute–solvent interface  $\Gamma$  which is the boundary of the solute region  $\Omega_s$  as well as that of solvent region  $\Omega_w$ , cf. Figure 1. Here we assume that there is a sharp interface that separates the solvent and solutes, and we treat the solvent as a continuum. We assume that there are  $N$  solute atoms in the system that are located at  $\mathbf{x}_1, \dots, \mathbf{x}_N$  inside  $\Omega_s$ . These solute atoms are treated explicitly.

Our basic assumption is that an experimentally observed equilibrium solvation system consists of a solute–solvent interface  $\Gamma$  and solute atoms located at  $\mathbf{x}_1, \dots, \mathbf{x}_N$  that together minimize an effective solvation free-energy functional. Along the line of variational implicit-solvent modeling of solvation systems,<sup>17,18</sup> we propose such an effective free-energy functional of a solute–solvent interface  $\Gamma$  and a set of solute atoms  $X = (\mathbf{x}_1, \dots, \mathbf{x}_N)$  to be

$$G[\Gamma, X] = G_{geom}[\Gamma] + G_{vdW}^{sw}[\Gamma, X] + G_{elec}[\Gamma, X] + G_{vdW}^{ss}[X] + G_{mech}[X] \quad (\text{II.1})$$

The first term  $G_{geom}[\Gamma]$  is the geometrical contribution of the solute–solvent interface  $\Gamma$ . It has the form

$$G_{geom}[\Gamma] = P \text{vol}(\Omega_s) + \int_{\Gamma} \gamma dS \quad (\text{II.2})$$

Here the term  $P \text{vol}(\Omega_s)$ , proportional to the volume of solute region  $\Omega_s$ , is the energy of creating a cavity of solute against the pressure difference  $P$  between the solvent liquid and vapor phase. This term can often be neglected for nanometer sized solutes, since the pressure difference  $P$  is usually very small. The integral term in (II.2) is the surface energy, where  $\gamma$  is the surface tension. It is known that for systems of nanometer scale, the surface tension  $\gamma$  is no longer a constant. Corrections with curvature effect must be added. For a special case of a spherical solute, Tolman<sup>21</sup> proposed that

$$\gamma = \frac{\gamma_0}{1 + 2\tau H}$$

where  $\gamma_0$  is the constant surface tension for a planar solvent liquid–vapor interface,  $\tau > 0$  is a constant with  $\tau$  often called the Tolman length,<sup>21</sup> and  $H$  is the mean curvature defined to be the average of the two principal curvatures. Since the magnitude of  $2\tau H$  is usually less than the unity, we use as in<sup>18,19</sup> the approximation (cf. also refs 22–24)

$$\gamma = \gamma_0(1 - 2\tau H) \quad (\text{II.3})$$

The second term  $G_{vdW}^{sw}[\Gamma, X]$  in the total free energy (II.1) is the nonpolar, van der Waals type interaction energy

between the solute particles  $\mathbf{x}_1, \dots, \mathbf{x}_N$  and solvent molecules that are coarse-grained. As in refs 17–19 we define it to be

$$G_{vdW}^{sw}[\Gamma, X] = \rho_0 \sum_{i=1}^N \int_{\Omega_w} U_{sw}(|\mathbf{x} - \mathbf{x}_i|) dV \quad (\text{II.4})$$

where  $\rho_0$  is the constant solvent density, and  $U_{sw} = U_{sw}(r)$  is a pairwise interaction potential. As in refs 18 and 19, here we choose  $U_{sw} = U_{sw}(r)$  to be a Lennard-Jones potential

$$U_{sw}(r) = 4\epsilon_{sw} \left[ \left( \frac{\sigma_{sw}}{r} \right)^{12} - \left( \frac{\sigma_{sw}}{r} \right)^6 \right] \quad (\text{II.5})$$

The parameters  $\epsilon_{sw}$  of energy and  $\sigma_{sw}$  of length can vary with different solute atoms as in the conventional force fields. Since the solvent is treated implicitly, the solute–solvent interaction energy (II.4) is expressed as an integral over the solvent region  $\Omega_w$ .

The third term  $G_{elec}[\Gamma, X]$  in the total free energy (II.1) is the electrostatic free energy. In the mean-field approximation, this is given by<sup>18,25</sup>

$$G_{elec}[\Gamma, X] = \int_{\Omega} \left[ \rho_f(\mathbf{x})\psi(\mathbf{x}) - \frac{\epsilon_{\Gamma}(\mathbf{x})}{8\pi} |\nabla \psi(\mathbf{x})|^2 \right] dV - \beta^{-1} \sum_{j=1}^M c_j^{\circ} \int_{\Omega_s} (e^{-\beta q_j \psi(\mathbf{x})} - 1) dV$$

Here  $\psi = \psi(\mathbf{x})$  is the electrostatic potential usually determined by the Poisson–Boltzmann equation,  $\epsilon_{\Gamma} = \epsilon_{\Gamma}(\mathbf{x})$  is the dielectric coefficient that takes one constant value in the solute region  $\Omega_s$  and a different constant value in the solvent region  $\Omega_w$ ,  $\rho_f = \rho_f(\mathbf{x})$  is the fixed charge density usually consisting of all solute point charges,  $\beta^{-1}$  is the thermal energy,  $c_j^{\circ}$  is the equilibrium concentration of the  $j$ th ionic species (a total of  $M$  is assumed), and  $q_j = ez_j$  with  $e$  the elementary charge and  $z_j$  the valence of  $j$ th ionic species in the solvent. More rigorous formulation of the electrostatic free energy can be found in ref 25 in which singularities in the potential  $\psi$  due to the point charges at solute atoms are carefully treated.

In this work, we only consider nonpolar systems and therefore set  $G_{elec}[\Gamma, X] = 0$ . This is our first step in developing our theory and methods of explicit-solute implicit-solvent modeling of biomolecular systems.

The fourth term  $G_{vdW}^{ss}[X]$  in the total free energy (II.1) is the van der Waals interaction energy among solute atoms at  $\mathbf{x}_1, \dots, \mathbf{x}_N$ . It has the form

$$G_{vdW}^{ss}[X] = \sum_{(i,j)'} U_{ss}(|\mathbf{x}_i - \mathbf{x}_j|) \quad (\text{II.6})$$

where the sum is taken over pairs of nonbonded solute atoms  $(\mathbf{x}_i, \mathbf{x}_j)$  with  $i < j$  and

$$U_{ss}(r) = 4\epsilon_{ss} \left[ \left( \frac{\sigma_{ss}}{r} \right)^{12} - \left( \frac{\sigma_{ss}}{r} \right)^6 \right] \quad (\text{II.7})$$

is a Lennard-Jones potential. The parameters  $\epsilon_{ss}$  and  $\sigma_{ss}$  can vary with solute atoms.

The last term  $G_{mech}[X]$  in (II.1) is the energy of the molecular mechanical interactions among all the solute atoms  $\mathbf{x}_1, \dots, \mathbf{x}_N$ . This includes the usual bonding, bending, and torsion energies. Specifically, we define

$$G_{\text{mech}}[X] = \sum_{(i,j)} W_{\text{bond}}(\mathbf{x}_i, \mathbf{x}_j) + \sum_{(i,j,k)} W_{\text{bend}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \sum_{(i,j,k,l)} W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \quad (\text{II.8})$$

Here the term  $\sum_{(i,j)} W_{\text{bond}}(\mathbf{x}_i, \mathbf{x}_j)$  accounts for the bonding energy of solute particles. The sum  $\sum_{(i,j)}$  is taken over all nonredundant pairs of bonded solute atoms  $(\mathbf{x}_i, \mathbf{x}_j)$ . The term  $\sum_{(i,j,k)} W_{\text{bend}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$  in (II.8) accounts for the bending energy of solute atoms. The sum  $\sum_{(i,j,k)}$  is taken over all the nonredundant triplets  $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$  such that both pairs of solute atoms  $(\mathbf{x}_i, \mathbf{x}_j)$  and  $(\mathbf{x}_j, \mathbf{x}_k)$  are bonded. The term  $\sum_{(i,j,k,l)} W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$  in (II.8) accounts for the torsion energy. The sum  $\sum_{(i,j,k,l)}$  is taken over all nonredundant quadruples  $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$  such that  $(\mathbf{x}_i, \mathbf{x}_j)$ ,  $(\mathbf{x}_j, \mathbf{x}_k)$ , and  $(\mathbf{x}_k, \mathbf{x}_l)$  are all bonded. The forms of  $W_{\text{bond}}(\mathbf{x}_i, \mathbf{x}_j)$ ,  $W_{\text{bend}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ , and  $W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$  are given in the Appendix.

In summary, our proposed free-energy functional of a nonpolar (with  $G_{\text{elec}}[\Gamma, X] = 0$ ) solvation system is given by (cf. (II.1), (II.2), (II.4), (II.6), and (A.1)–(A.3))

$$G[\Gamma, X] = P \text{vol}(\Omega_s) + \int_{\Gamma} \gamma dS + \rho_0 \sum_{i=1}^N \int_{\Omega_w} U_{\text{sw}}(|\mathbf{x} - \mathbf{x}_i|) dV + \sum_{(i,j)'} U_{\text{ss}}(|\mathbf{x}_i - \mathbf{x}_j|) + \sum_{(i,j)} \frac{1}{2} A_{ij} (r_{ij} - r_{0ij})^2 + \sum_{(i,j,k)} \frac{1}{2} B_{ijk} (\theta_{ijk} - \theta_{0ijk})^2 + \sum_{(i,j,k,l)} \left[ \frac{1}{4} (V_{ijkl}^{(1)} + 2V_{ijkl}^{(2)} + V_{ijkl}^{(3)}) + \frac{1}{4} (V_{ijkl}^{(1)} - 3V_{ijkl}^{(3)}) \Lambda_{ijkl} - \frac{1}{2} V_{ijkl}^{(2)} \Lambda_{ijkl}^2 + V_{ijkl}^{(3)} \Lambda_{ijkl}^3 \right], \quad (\text{II.9})$$

supplemented by (II.3), (II.5), and (II.7).

### III. A Level-Set Optimization Method

To find an equilibrium structure of an underlying solvation system that is a (local) minimizer of the free-energy functional (II.9), we select an initial solute–solvent interface and an initial set of solute atomic positions. We then start to move the interface and the set of solute atoms to relax the system.

To track the motion of the interface, we use the level-set method.<sup>26–28</sup> We represent the solute–solvent interface  $\Gamma = \Gamma(t)$  at time  $t$  as the zero level-set of a level-set function  $\phi = \phi(\mathbf{x}, t)$ , i.e.,  $\Gamma(t) = \{\mathbf{x}: \phi(\mathbf{x}, t) = 0\}$ . With this representation of the interface, we obtain the unit normal  $\mathbf{n} = \mathbf{n}(\mathbf{x}, t)$ , the mean curvature  $H = H(\mathbf{x}, t)$ , and the Gaussian curvature  $K = K(\mathbf{x}, t)$  of a point  $\mathbf{x}$  at the interface at time  $t$

$$\mathbf{n} = \frac{\nabla \phi}{|\nabla \phi|}, H = \frac{1}{2} \nabla \cdot \mathbf{n}, K = \mathbf{n} \cdot \text{adj}(\text{He}(\phi)) \mathbf{n} \quad (\text{III.1})$$

respectively, where  $\text{He}(\phi)$  is the  $3 \times 3$  Hessian matrix of the function  $\phi$  whose entries are all the second order partial derivatives  $\partial_{ij}^2 \phi$  of the level-set function  $\phi$ , and  $\text{adj}(\text{He}(\phi))$  is the adjoint matrix of the Hessian  $\text{He}(\phi)$ . (The Gaussian curvature is the product of the two principal curvatures.)

The level-set function is a solution to the level-set equation

$$\frac{\partial \phi}{\partial t} + v_n |\nabla \phi| = 0 \quad (\text{III.2})$$

where  $v_n = (d\mathbf{x}(t)/dt) \cdot \mathbf{n}$  is the normal velocity of the moving interface  $\Gamma(t)$  at the point  $\mathbf{x} = \mathbf{x}(t)$ . Notice that  $(d/dt)\mathbf{x}(t)$  is the velocity of the point  $\mathbf{x}(t)$  on the interface  $\Gamma(t)$ . The equation (III.2) is derived from taking the time derivative of both sides of the equation  $\phi(\mathbf{x}, t) = 0$  and using the chain rule.

To relax an underlying solvation system, we define the normal velocity  $v_n$  of the solute–solvent interface  $\Gamma(t)$  in such a way that the system moves in the steepest descent direction. Hence we define the normal velocity  $v_n$  to be the negative variation of the free energy  $G[\Gamma, X]$  with respect to the location change of the interface  $\Gamma$

$$v_n = -M_{\Gamma} \delta_{\Gamma} G[\Gamma, X] \quad (\text{III.3})$$

where  $M_{\Gamma} > 0$  is the mobility or relaxation factor which we take as a constant, and  $\delta_{\Gamma}$  denotes the first variation with respect to the location change of  $\Gamma$ . The variation  $\delta_{\Gamma} G[\Gamma, X]$  defines a function on  $\Gamma$  and is given by<sup>18,19</sup>

$$\delta_{\Gamma} G[\Gamma, X] = P + 2\gamma_0 [H(\mathbf{x}) - \tau K(\mathbf{x})] - \rho_0 \sum_{i=1}^N U_{\text{sw}}(|\mathbf{x} - \mathbf{x}_i|) \quad \forall \mathbf{x} \in \Gamma \quad (\text{III.4})$$

where  $H(\mathbf{x})$  and  $K(\mathbf{x})$  are the mean curvature and Gaussian curvature of a point  $\mathbf{x} \in \Gamma$ , respectively.

The motion of solute atoms is defined similarly to decrease the free energy. Therefore, the velocity of each of such atoms is given by

$$\frac{d\mathbf{x}_m(t)}{dt} = -M_m \nabla_{\mathbf{x}_m} G[\Gamma, X], m = 1, \dots, N \quad (\text{III.5})$$

where  $M_m > 0$  is a mobility or relaxation factor. Fix a particle  $\mathbf{x}_m$ . The gradient of  $G[\Gamma, X]$  with respect to  $\mathbf{x}_m$  is given by

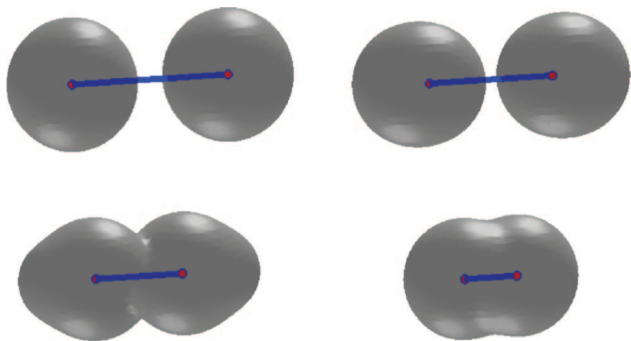
$$\nabla_{\mathbf{x}_m} G[\Gamma, X] = \rho_0 \int_{\Omega_w} U'_{\text{sw}}(|\mathbf{x}_m - \mathbf{x}|) \frac{\mathbf{x}_m - \mathbf{x}}{|\mathbf{x}_m - \mathbf{x}|} dV + \sum_{(i,j)'} \delta_{mi} U'_{\text{ss}}(|\mathbf{x}_i - \mathbf{x}_j|) \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|} + \sum_{(i,j)} \delta_{mi} A_{ij} (r_{ij} - r_{0ij}) \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|} + \sum_{(i,j,k)} \nabla_{\mathbf{x}_m} W_{\text{bend}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \sum_{(i,j,k,l)} \nabla_{\mathbf{x}_m} W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \quad (\text{III.6})$$

where  $\delta_{mi}$  is 1 if  $i = m$  and 0 if  $i \neq m$ . The derivatives  $\nabla_{\mathbf{x}_m} W_{\text{bend}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$  and  $\nabla_{\mathbf{x}_m} W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$  are given in the Appendix.

In each time step of relaxation, we solve the system of equations (III.2) and (III.5) and find the solute–solvent interface by locating all the points at which the level-set function  $\phi$  vanishes. In solving these equations, we use the first variation  $\delta_{\Gamma} G[\Gamma, X]$  and the gradients  $\nabla_{\mathbf{x}_m} G[\Gamma, X]$  that are given by (III.4), (III.6), (A.4), and (A.8), together with (A.5)–(A.7) and (A.9)–(A.15).

Notice that by a series of formal calculations we have that

$$\frac{d}{dt} G[\Gamma, X] = \delta_{\Gamma} G[\Gamma, X] v_n + \sum_{m=1}^N \nabla_{\mathbf{x}_m} G[\Gamma, X] \frac{d\mathbf{x}_m(t)}{dt} = -M_{\Gamma} [\delta_{\Gamma} G[\Gamma, X]]^2 - \sum_{m=1}^N M_m |\nabla_{\mathbf{x}_m} G[\Gamma, X]|^2 \leq 0$$



**Figure 2.** The level-set optimization of a two-atom system. The initial solute–solvent interface consists of two separated spheres. Order of snapshots: from left to right and from top to bottom.

This confirms that the free energy decays.

We have developed a numerical algorithm that combines the level-set method and optimization techniques to numerically find solutions of the system (III.2) and (III.5). Our numerical algorithm consists of the following steps:

(1) Choose a computational box, a cube in  $\mathbb{R}^3$ , and discretize the box with a uniform finite-difference grid. Initialize the level-set function  $\phi$  and position of solute atoms  $x_1, \dots, x_N$ . We place the initial solute–solvent interface a few grid points away from the boundary of the computational box, just so that we can solve the level-set equation more accurately, cf. Step (2). One choice of the initial level-set function is

$$\varphi(\mathbf{x}) = \min_{1 \leq i \leq N} (|\mathbf{x} - \mathbf{x}_i| - r_i) \quad (\text{III.7})$$

where  $r_i > 0$  ( $i = 1, \dots, N$ ) are preselected numbers. Notice that the solute atoms are not necessarily placed at grid points.

(2) Calculate and extend the normal velocity  $v_n$  using the formulas (III.3) and (III.4). The extension of the normal velocity  $v_n$  from the interface  $\Gamma(t)$  to the computational box is necessary for solving the level-set equation (III.2) on the computational box. In our current implementation, we use the level-set function  $\phi$  and the formulas in (III.1) to define the mean curvature  $H$  and Gaussian curvature  $K$  all over the computational domain. Therefore we only extend the Lennard-Jones potential part in the normal velocity, the last term in (III.4). We extend this part to a narrow band of the interface by constant in the normal direction of the interface.

(3) Solve the level-set equation (III.2). We use the forward Euler method to discretize the time derivative in the level-set equation. The normal velocity  $v_n$  consists of two parts. One is from the motion of the surface energy and is more of a parabolic type term in the equation. The other is from the solute–water interaction and only gives rise to a lower order term. We thus use the central differencing scheme to discretize the first part and an upwinding scheme for the second part. We choose our time step  $\Delta t$  to be of the order of  $(\Delta x)^2$  to satisfy the CFL stability condition. A simple linearization analysis shows that the level-set equation becomes backward parabolic if  $1 - 2\tau\kappa_1 < 0$  or  $1 - 2\tau\kappa_2 < 0$ , where  $\kappa_1$  and  $\kappa_2$  are the two principle curvatures. If this

occurs, then we numerically change the parameter  $\tau$  to regularize the interface motion.

(4) Reinitialize the level-set function. This step is necessary to keep the level-set function away from being too flat or steep. If the level-set function is  $\phi_0$  in the current step, we solve the equation

$$\frac{\partial \varphi}{\partial t} = \text{sign}(\phi_0)(1 - |\nabla \varphi|)$$

to obtain a new approximation of the level-set function, where  $\text{sign}(\phi_0)$  is the sign of  $\phi_0$ .

(5) Calculate the velocity of solute atoms using the formula (III.6). We evaluate the integral term in (III.6) as follows: For each solute atom  $\mathbf{x}_m$ , we choose a ball  $B(\mathbf{x}_m, r_m)$  centered at the  $\mathbf{x}_m$  with a radius  $r_m$ . This ball is small enough so that it is completely contained in the solute region  $\Omega_s$ . Then we compute the integral term in (III.6) using the formula

$$\int_{\Omega_w} U'_{sw}(|\mathbf{x}_m - \mathbf{x}|) \frac{\mathbf{x}_m - \mathbf{x}}{|\mathbf{x}_m - \mathbf{x}|} dV = \int_{\mathbb{R}^3 \setminus B(\mathbf{x}_m, r_m)} U'_{sw}(|\mathbf{x}_m - \mathbf{x}|) \frac{\mathbf{x}_m - \mathbf{x}}{|\mathbf{x}_m - \mathbf{x}|} dV + \int_{\Omega_s \setminus B(\mathbf{x}_m, r_m)} U'_{sw}(|\mathbf{x}_m - \mathbf{x}|) \frac{\mathbf{x}_m - \mathbf{x}}{|\mathbf{x}_m - \mathbf{x}|} dV$$

where the integral over  $\mathbb{R}^3 \setminus B(\mathbf{x}_m, r_m)$  (the region complement to the ball  $B(\mathbf{x}_m, r_m)$ ) is calculated analytically.

(6) Update the position of each solute atom  $\mathbf{x}_i$  by the formula (III.5). We use the forward Euler method for updating the atom positions. Our time step in solving the equations (III.5) for the motion of solute atoms is much smaller than that in solving the level-set equation (III.2).

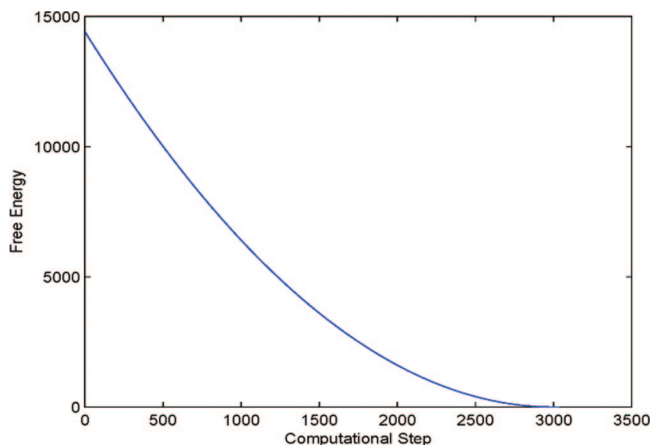
(7) Calculate the total free energy (II.9). To calculate the solute–solvent interaction term (II.4) in the total free energy (II.9), we use the same method as described in Step (5). In testing our method of calculating the total energy for a one-atom system, we found that our method is second-order accurate.

(8) Go to Step (2).

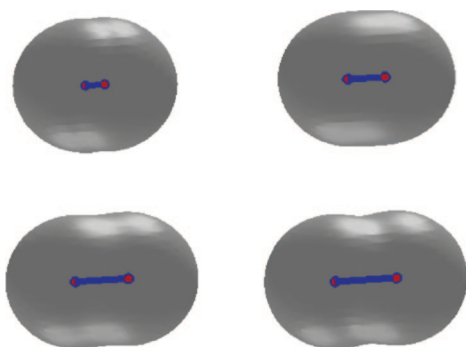
## IV. Numerical Tests and Applications

**A. A Two-Atom Molecule.** We consider an artificial molecular system of two atoms. The thermodynamic and LJ parameters we use are mostly taken from ref 19 which are for the system of two xenon atoms. These parameters are as follows: the pressure difference  $P = 0$  bar (an approximation), the constant surface tension  $\gamma_0 = 0.174 k_B T / \text{\AA}^2$ , the Tolman length  $\tau = 1.3 \text{\AA}$ , the water density  $\rho_0 = 0.033 \text{\AA}^{-3}$ , the solute–water Lennard-Jones parameters  $\sigma = 3.57 \text{\AA}$  and  $\varepsilon = 0.431 k_B T$ , the solute–solute Lennard-Jones parameters  $\sigma = 3.57 \text{\AA}$  and  $\varepsilon = 0.147 k_B T$ , and the temperature  $T = 298$  K. We additionally introduce an intrabond with the spring constant in the bond stretching energy  $A = 800 k_B T / \text{\AA}^2$ , and the equilibrium bond length  $r_0 = 3 \text{\AA}$ . To test our method, we include both the bonding and Lennard-Jones interaction between the two atoms.

We test two possible cases. In the first case, we place initially the two solute atoms far away from each other so that their distance is much larger than the equilibrium bond



**Figure 3.** The free energy (kcal/mol) vs the computational step in a level-set optimization for the two-atom system with the initial solute–solvent interface consisting of two separated spheres, cf. Figure 2.

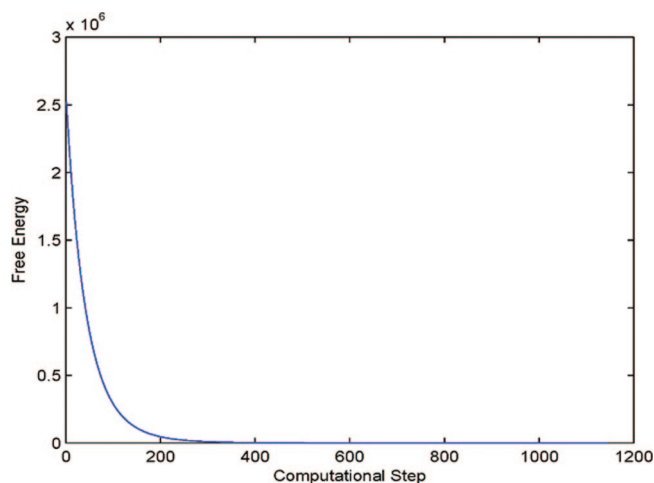


**Figure 4.** The level-set optimization of a two-atom system. The two atoms are initially close and then move apart from each other. Order of snapshots: from left to right and from top to bottom.

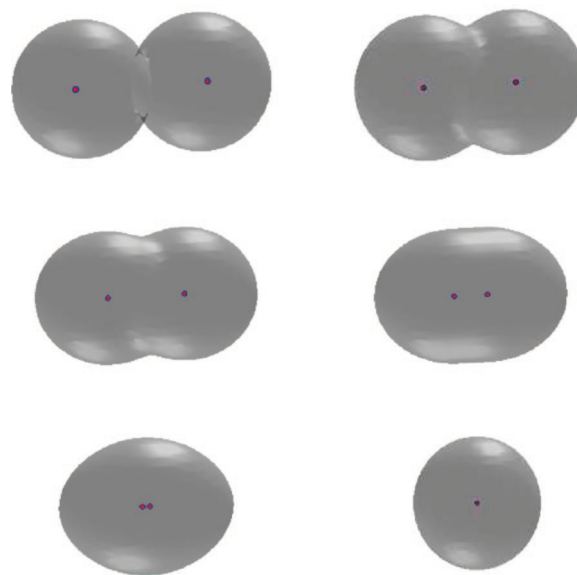
length. We also set the initial solute–solvent interface to be two separated spheres that are centered at the two solute atoms, respectively. A few snapshots of the system during the relaxation are gathered in Figure 2. We see that those two initially separated spheres merge, the distance between the two solute atoms gets closer, and then the system reaches an equilibrium state. Notice from the lower left snapshot in Figure 2 that the solute–solvent interface is not in equilibrium with respect to the two atoms. This means that the system is relaxed through the coupling of both the interface motion and atomic motion. In Figure 3 we plot the total free energy vs the computational step. It is clear that the free energy decays in each step.

In the second case, we place initially the two solute atoms very close to each other so that their distance is smaller than the equilibrium distance. We also set the initial solute–solvent interface to be a large surface that encloses both of the atoms. A few snapshots from our numerical relaxation are shown in Figure 4. The decay of the free energy in our numerical computation is shown in Figure 5. We notice that the free energy decays very fast in this case. This is due to very strong repulsion modeled by the Lennard-Jones potential.

Through this test example, we see that our level-set method works well in capturing topological changes of surfaces during relaxation. We also find that the final free energy



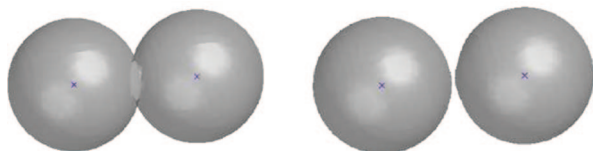
**Figure 5.** The free energy (kcal/mol) vs the computational step in a level-set optimization for the two-atom system with the initial solute–solvent interface consisting of a single surface containing the two atoms, cf. Figure 4.



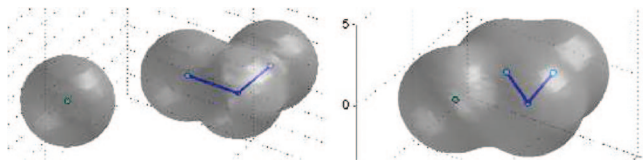
**Figure 6.** Snapshots of a relaxing system of two noninteracting particles. Order: left to right and top to bottom.

values for the two different cases are nearly the same:  $6.36397 k_B T$  and  $6.36652 k_B T$ , respectively, with an error of  $0.00255 k_B T$ .

Our results show a strong contribution of the molecular mechanical force in the relaxation of the system. To understand the solvent influence and, particularly, how the motion of solute–solvent interface affects that of solute particles, we turn off the particle–particle interaction in the two-atom system. We perform two tests. In the first test summarized in Figure 6, we set the initial center-to-center distance between the two atoms to be  $5 \text{ \AA}$ . This is in the range between 0 to about  $6 \text{ \AA}$  of attraction of the two atoms as shown in Figure 1 of our previous work.<sup>19</sup> Such attraction results from the minimization of the interfacial energy. Since there is no atom–atom interaction, the solute–solvent interaction pushes these two atoms together. The sequence of snapshots in Figure 6 demonstrates clearly that the solvent contributes significantly to the motion of solute particles.



**Figure 7.** Two noninteracting particles relax with increasing surface area.



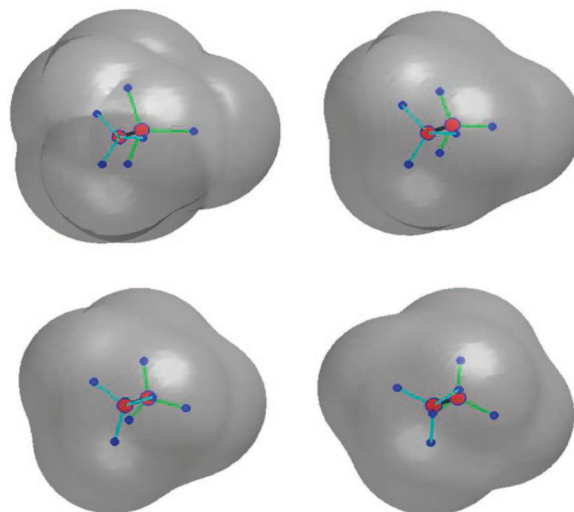
**Figure 8.** Left: Initial positions of the solute-solvent interface and solute atoms. Right: The relaxed positions of the solute-solvent interface and solute atoms.

In the second test, we set the initial center-to-center distance between the two atoms to be  $6.8 \text{ \AA}$ . This is in a range of (weak) repulsion as predicted in our previous work.<sup>19</sup> The noninteracting two atoms are then pushed away by the solute-solvent interaction, cf. Figure 7. Notice that the surface area for the initial, connected two-sphere system (Figure 7, left) increases to that of final, separated two-sphere system (Figure 7, right). Such increasing of surface area compensates the solute-solvent interaction. It obviously cannot be captured by a SAS/SES type model, in which the nonpolar part of the free energy is the surface energy, proportional to the surface area.

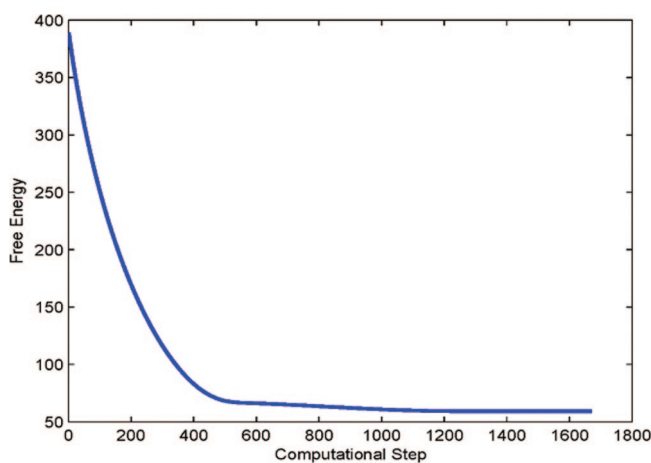
**B. A Four-Atom Molecule.** We consider an artificial molecular system of four atoms  $\mathbf{x}_1, \dots, \mathbf{x}_4$  to test how our method can handle the bending energy and the solute-solute van der Waals interaction in addition to the variational implicit-solvent. We put the pairs  $(\mathbf{x}_1, \mathbf{x}_2)$  and  $(\mathbf{x}_2, \mathbf{x}_3)$  in their equilibrium bonding position and arrange the angle between  $\mathbf{x}_2\mathbf{x}_1$  and  $\mathbf{x}_2\mathbf{x}_3$  slightly different from an equilibrium angle. We also put  $\mathbf{x}_4$  relatively far away from the first three atoms. The atom  $\mathbf{x}_4$  interacts with the other three atoms via a Lennard-Jones potential. Figure 8 (left) shows the initial conformation and Figure 8 (right) the relaxed conformation of this system. It is clear that the fourth atom moves closer to the three-atom group, while the angle between  $\mathbf{x}_2\mathbf{x}_1$  and  $\mathbf{x}_2\mathbf{x}_3$  is relaxed to its equilibrium value.

**C. An Ethane Molecule.** We consider an ethane molecule  $C_2H_6$  in water and take from<sup>29-31</sup> the solute atomic positions and force field parameters. Other parameters are as follows: the pressure differencing  $P = 0$  bar, the constant surface tension  $\gamma_0 = 0.174 k_B T / \text{\AA}^2$ , the Tolman length  $\tau = 1.3 \text{ \AA}$ , the water density  $\rho_0 = 0.033 \text{ \AA}^{-3}$ , the carbon-water Lennard-Jones parameters  $\sigma = 3.4767 \text{ \AA}$  and  $\epsilon = 0.2311 k_B T$ , and the hydrogen-water Lennard-Jones parameters  $\sigma = 3.1017 \text{ \AA}$  and  $\epsilon = 0.0989 k_B T$ .

We construct an initial conformation of the ethane molecule from its equilibrium in which three hydrogen-carbon bonds with respect to one of the carbon atoms are rotated 20 degrees. During the level-set relaxation, these three hydrogen atoms rotate back to their equilibrium positions. The solute-solvent interface also moves to its equilibrium position. Figure 9 displays a few snapshots of our numerical



**Figure 9.** The level-set relaxation of the ethane molecule. Order of snapshots: from first row to second and from left to right in each row.



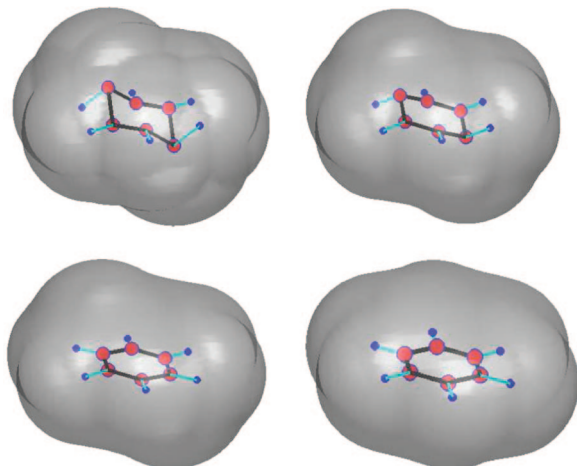
**Figure 10.** A plot of the total free energy (kcal/mol) vs computational step in the level-set optimization for the ethane molecule.

results. Figure 10 is a plot of the free energy in each step of our numerical computation.

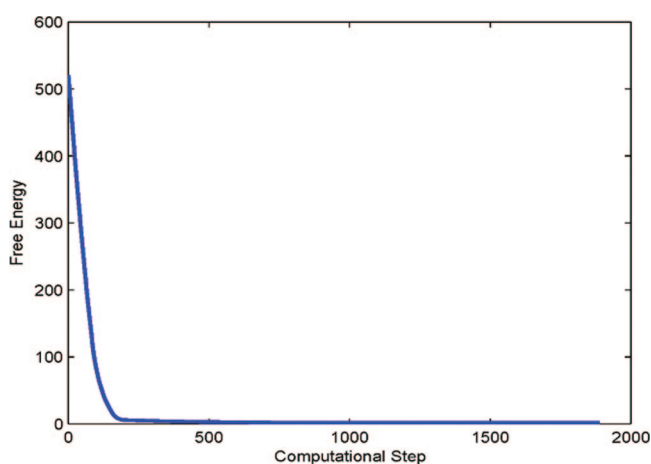
**D. A Benzene Molecule.** We consider a benzene molecule  $C_6H_6$  in water and take from<sup>29-31</sup> the solute atomic positions and force field parameters. Other parameters are the same as those for an ethane molecule.

We construct an initial conformation of the benzene molecule from its equilibrium in which all the atoms are on the same plane. We then fix a pair of C-atoms that are opposite in the hexagonal ring. We pull up one of these two C-atoms and pull down the other C-atom to form the initial positions of all the atoms. We also use (III.7) to define an initial solute-solvent interface. During the process of our level-set relaxation, those two carbon atoms move back to their equilibrium positions. The solute-solvent interface also moves to its equilibrium position. Figure 11 displays a few snapshots of our numerical results. Figure 12 is a plot of the free energy in each step of our numerical computation.

We have tried various values of the Tolman length  $\tau$ . For  $\tau = 1 \text{ \AA}$ , we find that the free energy (surface energy and the benzene-water interaction energy) is 1.98 kcal/mol. Using



**Figure 11.** The level-set relaxation of the benzene molecule. Order of snapshots: from first row to second and from left to right in each row.



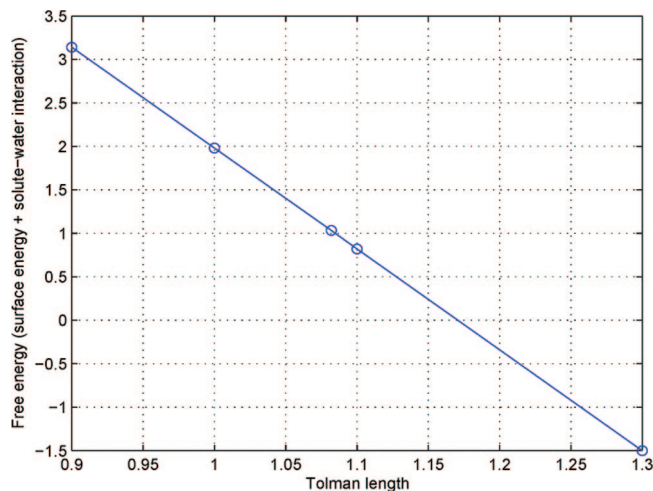
**Figure 12.** A plot of the total free energy (kcal/mol) vs computational step in the level-set optimization for the benzene molecule.

a Poisson–Boltzmann solver, we find the polar part of the solvation energy is  $-2.92$  kcal/mol. Therefore, our estimate of the solvation energy is  $-0.94$  kcal/mol. The experimental value of this solvation energy is  $-0.89$  kcal/mol.

Our PB calculation was done by impact version 50112 (released by Schrodinger, LLC) with default settings (PB grid resolution was set to high) with OPLS2005 forcefield. Since we have not implemented our own PB solver, the optimized surface was not used to define the boundary. But since the benzene molecule is small and geometrically simple, we do not expect the variational surface to be drastically different from a traditional SAS. Therefore, we believe the polar energy from traditional PB solver using a SAS is consistent with our variational implicit-solvent description of the benzene molecule.

## V. Conclusions

In this work, we construct a hybrid explicit-solute implicit-solvent model for molecular solvation. The key quantity in this model is an effective free-energy functional of positions of solute atoms and solute–solvent interface. The free energy



**Figure 13.** The sum of the surface energy and benzene–water interaction energy vs the Tolman length  $\tau$ .

couple both the polar and nonpolar contributions and also includes the molecular mechanical interactions. Minimization of this free-energy functional determines an equilibrium molecular structure and the solvation free energy. We also develop a level-set optimization method to numerically minimize the free-energy functional and to obtain equilibrium solute–solvent interface and positions of solute atoms. In our method, both a trial solute–solvent interface and a set of trial solute atoms move in the steepest descent direction of reducing the total free energy of the system. Our numerical results for the solvation of some molecules demonstrate that our model and methods can capture topological changes of the solute–solvent interface as well as the coupling between such an interface motion and molecular mechanical interactions.

We emphasize again that we model a relaxation process or free-energy minimization rather than the real dynamics of an underlying molecular system.

For each of the small, nonpolar systems that we have tested and reported here, our level-set relaxation only took a few minutes. The actual exact computational time is affected by several factors such as the number of grid points (the resolution) and the choice of initial surface. It is clear, however, that our approach is in general computationally more costly than a SAS/SES type implicit-solvent model, since we need to evolve a surface to its equilibrium state. Nevertheless, we have been developing several new level-set techniques that can speed up our computations. One such technique is the local level-set method. It can reduce the complexity of a three-dimensional problem to that of a two-dimensional one.

As noted before the only fitting parameter in our model is the Tolman length  $\tau$ . We find that the free energy calculation is sensitive to the choice of the Tolman length. Our experience is that  $\tau = 1 \text{ \AA}$  is a good value for many molecular systems. Figure 13 plots the nonpolar part of the minimum solvation energy (the surface energy plus the solute–solvent interaction energy) vs the Tolman length  $\tau$  for the benzene molecule. We can see a perfect linear dependence of the nonpolar part of the solvation energy on the Tolman length. This indicates the existence of an optimal Tolman length. It also points to the fact that the

minimized interface hardly changes with the fitting parameter  $\tau$  for this particular example and that the free energy is directly proportional to  $\tau$  as can be predicted from (II.2) and (II.3). In general, especially for larger molecules with a higher dispersion and complexity in curvature, the relation will be qualitatively different. In future, we hope a more sophisticated free-energy functional can be proposed in order to fix an optimal value of  $\tau$  once and for all, independent of the particular solute system.

In a SAS implicit-solvent model, the size effect of solvent molecules is described through the probing solvent molecule used in defining the SAS. In our variational implicit-solvent model (VISM), such size effect is reflected in the solute–solvent interaction, cf. (II.4). In general, an implicit or “continuum” solvent model is, per definition and how the name implies, not able to explicitly describe the finite size of solvent molecules. Implicitly, these effects are typically considered in fitting parameters. In a solvation system, the solvent–solute interface itself is not a sharp boundary but has a width of solvent molecule size ( $\sim 3$  Å). Thus, the “right” interface location basically does not exist within a few Angstroms. Consequently, it is hard to compare precisely a SAS/SES type implicit-solvent model to our VISM. The final goal must be to evaluate the correct free energy from any of these surface definitions, “correct” in the sense that they are quantitative in comparison to benchmarking experiments or explicit-solvent molecular dynamics simulations. The virtue of the VISM is that the interface is defined in a physically reasonable way and allows the interface—for every configuration—to respond to local solute geometry and energetic potentials and can hopefully provide accurate free energies with only very few fitting parameters (in contrast to established implicit models). If the capillary evaporation (“dewetting” or “drying”) between solutes takes places, then the VISM interface can be very different to the a priori defined SAS/SES interfaces as the latter do not capture solvent evaporation.

Our immediate next step is to add the electrostatic part of the free energy into our model and develop a corresponding level-set method. The electrostatic free energy is often described by the Poisson–Boltzmann (PB) or Generalized Born (GB) method in which the solute–solvent interface is used as the dielectric boundary. Our first step will be the development of a GB-like approach to efficiently calculate the effective electrostatic surface force (only its normal component) at each point of the evolving solute-solvent interface. This force will be used as the normal velocity in the level-set relation of the solute–solvent interface. Next we will develop a fast PB solver and couple it with our level-set code. Solving the nonlinear PB equation in each step of level-set relaxation can be very slow. To speed up our computations, we can in each step linearize the PB equation around the previous PB solution. Thus, in each step, we need only to solve a linearized PB equation. (This linearized equation is different from the Debye–Hückel equation.) Moreover, we do not need to solve very accurately the PB equation very step, since dewetting regions can be mainly captured by the surface energy term.

Besides adding the electrostatics into our models and methods, we will also apply our model and methods to larger systems of polymers and biomolecules. Further, we will apply our theory and methods to the calculation of surface forces of solute–solvent interfaces that can be used in Brownian dynamics simulations of biomolecules.

**Acknowledgment.** This work was supported by the National Science Foundation (NSF) through grant DMS-0511766 (L.-T.C.), DMS-0451466 (B.L.), and DMS-0811259 (B.L.), by the Center for Theoretical Biological Physics (B.L. and J.A.M) (NSF PHY-0822283), by the Department of Energy through grant DE-FG02-05ER25707 (B.L. and Y.X.), and by a Sloan Fellowship (L.-T.C.). J. D. thanks the Deutsche Forschungsgemeinschaft (DFG) for support within the Emmy-Noether Programme. Work in the McCammon group is supported in part by NSF, NIH, HHMI, CTBP, NBCR, and Accelrys.

## Appendix

**Molecular Mechanical Interactions and Force Calculations.** We summarize in this Appendix the molecular mechanical interaction energies and their derivatives.

For a given pair of bonded solute atoms ( $\mathbf{x}_i, \mathbf{x}_j$ ), the bonding energy is given by the harmonic approximation

$$W_{bond}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} A_{ij} (r_{ij} - r_{0ij})^2 \quad (\text{A.1})$$

where  $A_{ij}$  is a spring constant characterizing the equilibrium bonding strength,  $r_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$  is the distance between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , and  $r_{0ij}$  is the corresponding equilibrium distance.

For a triplet ( $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$ ) with ( $\mathbf{x}_i, \mathbf{x}_j$ ) and ( $\mathbf{x}_j, \mathbf{x}_k$ ) both bonded, the bending energy is given by the harmonic approximation

$$W_{bend}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \frac{1}{2} B_{ijk} (\theta_{ijk} - \theta_{0ijk})^2 \quad (\text{A.2})$$

where  $B_{ijk}$  is a constant parameter depending in general on ( $i, j, k$ ),  $\theta_{ijk}$  is the angle between the vectors  $\mathbf{r}_{ji} = \mathbf{x}_i - \mathbf{x}_j$  and  $\mathbf{r}_{jk} = \mathbf{x}_k - \mathbf{x}_j$ , and  $\theta_{0ijk}$  is the corresponding equilibrium angle constrained by  $0 \leq \theta_{0ijk} \leq \pi$  for all ( $i, j, k$ ).

For a quadruple ( $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l$ ) such that ( $\mathbf{x}_i, \mathbf{x}_j$ ), ( $\mathbf{x}_j, \mathbf{x}_k$ ), ( $\mathbf{x}_k, \mathbf{x}_l$ ) are all bonded, the torsion angle  $\phi_{ijkl}$  is the angle between the plane determined by  $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$  and that determined by  $\mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l$ . The torsion energy is<sup>32</sup>

$$W_{torsion}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = \frac{1}{2} V_{ijkl}^{(1)} [1 - \cos(\pi\phi_{ijkl})] + \frac{1}{2} V_{ijkl}^{(2)} [1 + \cos(2\pi\phi_{ijkl})] + \frac{1}{2} V_{ijkl}^{(3)} [1 - \cos(3\pi\phi_{ijkl})] \quad (\text{A.3})$$

where  $V_{ijkl}^{(1)}$ ,  $V_{ijkl}^{(2)}$ , and  $V_{ijkl}^{(3)}$  are constants.

Fix  $\mathbf{x}_i, \mathbf{x}_j$ , and  $\mathbf{x}_k$ . Denote by  $\mathbf{r}_{ji} = \mathbf{x}_i - \mathbf{x}_j$  the vector from  $\mathbf{x}_j$  to  $\mathbf{x}_i$  for any  $i$  and  $j$  and by  $r_{ji} = |\mathbf{r}_{ji}|$  the length of this vector. Routine calculations lead to

$$\nabla W_{bend}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = B_{ijk} (\theta_{ijk} - \theta_{0ijk}) \nabla \theta_{ijk} \quad (\text{A.4})$$

where

$$\nabla_{\mathbf{x}_i} \theta_{ijk} = \frac{1}{\sin \theta_{ijk}} \left( \frac{\mathbf{r}_{ji} \cdot \mathbf{r}_{jk}}{r_{ji}^3 r_{jk}} \mathbf{r}_{ji} - \frac{1}{r_{ji} r_{jk}} \mathbf{r}_{jk} \right) \quad (\text{A.5})$$

$$\nabla_{x_k} \theta_{ijk} = \frac{1}{\sin \theta_{ijk}} \left( \frac{\mathbf{r}_{ji} \cdot \mathbf{r}_{jk}}{r_{jk}^3} \mathbf{r}_{jk} - \frac{1}{r_{ji} r_{jk}} \mathbf{r}_{ji} \right) \quad (\text{A.6})$$

$$\nabla_{x_j} \theta_{ijk} = \frac{1}{\sin \theta_{ijk}} \left[ \left( \frac{1}{r_{ji} r_{jk}} - \frac{\cos \theta_{ijk}}{r_{ji}^2} \right) \mathbf{r}_{ji} + \left( \frac{1}{r_{ji} r_{jk}} - \frac{\cos \theta_{ijk}}{r_{jk}^2} \right) \mathbf{r}_{jk} \right] \quad (\text{A.7})$$

For a given quadruple  $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$ , we denote

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{r}_{ij}, \mathbf{r}_2 = \mathbf{r}_{jk}, \mathbf{r}_3 = \mathbf{r}_{kl} \\ \mathbf{u} &= \mathbf{r}_1 \times \mathbf{r}_2, \mathbf{v} = \mathbf{r}_2 \times \mathbf{r}_3 \\ \varphi &= \varphi_{ijkl}, \Lambda = \Lambda_{ijkl} = \cos \varphi = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}| |\mathbf{v}|} \\ V^{(n)} &= V_{ijkl}^{(n)}, n = 1, 2, 3 \end{aligned}$$

It follows from (A.3) and a series of elementary calculations that

$$\nabla_{r_m} W_{\text{torsion}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = \left[ -\frac{1}{2}(V^{(1)} - 3V^{(3)}) - 2V^{(2)}\Lambda + 6V^{(3)}\Lambda^2 \right] \nabla_{x_m} \Lambda \quad (\text{A.8})$$

where  $\nabla_{x_m} \Lambda = 0$  if  $m$  is not one of  $i, j, k$ , or  $l$ , and

$$\nabla_{x_i} \Lambda = -\nabla_{r_1} \Lambda \quad (\text{A.9})$$

$$\nabla_{x_j} \Lambda = \nabla_{r_1} \Lambda - \nabla_{r_2} \Lambda \quad (\text{A.10})$$

$$\nabla_{x_k} \Lambda = \nabla_{r_2} \Lambda - \nabla_{r_3} \Lambda \quad (\text{A.11})$$

$$\nabla_{x_l} \Lambda = \nabla_{r_3} \Lambda \quad (\text{A.12})$$

and

$$\nabla_{r_1} \Lambda = -\frac{(\mathbf{r}_1 \cdot \mathbf{v})|\mathbf{r}_2|^2}{|\mathbf{u}|^3 |\mathbf{v}|} \mathbf{u} \quad (\text{A.13})$$

$$\nabla_{r_2} \Lambda = \frac{\mathbf{r}_1 \cdot \mathbf{v}}{|\mathbf{u}|^3 |\mathbf{v}|^3} [(\mathbf{r}_1 \cdot \mathbf{r}_2)|\mathbf{v}|^2 \mathbf{u} + (\mathbf{r}_2 \cdot \mathbf{r}_3)|\mathbf{u}|^2 \mathbf{v}] \quad (\text{A.14})$$

$$\nabla_{r_3} \Lambda = -\frac{(\mathbf{r}_1 \cdot \mathbf{v})|\mathbf{r}_2|^2}{|\mathbf{u}|^3 |\mathbf{v}|} \mathbf{v} \quad (\text{A.15})$$

## References

- (1) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
- (2) Feig, M.; Brooks III, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
- (3) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.
- (4) Connolly, M. L. *J. Mol. Graphics* **1992**, *11*, 139–141.
- (5) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.
- (6) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
- (7) Richmond, T. J. *J. Mol. Biol.* **1984**, *178*, 63–89.
- (8) Fixman, F. *J. Chem. Phys.* **1979**, *70*, 4995–5005.
- (9) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509–521.
- (10) Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1990**, *94*, 7684–7692.
- (11) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (12) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- (13) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.
- (14) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570–4577.
- (15) Chandler, D. *Nature* **2005**, *437*, 640–647.
- (16) Chen, J.; Brooks III, C. L. *J. Am. Chem. Soc.* **2007**, *129*, 2444.
- (17) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. *Phys. Rev. Lett.* **2006**, *96*, 087802.
- (18) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. *J. Chem. Phys.* **2006**, *124*, 084905.
- (19) Cheng, L.-T.; Dzubiella, J.; McCammon, J. A.; Li, B. *J. Chem. Phys.* **2007**, *127*, 084503.
- (20) Can, T.; Chen, C.-I.; Wang, Y.-F. *J. Mol. Graphics Modell.* **2006**, *25*, 442–454.
- (21) Tolman, R. C. *J. Chem. Phys.* **1949**, *17*, 333–337.
- (22) Buff, F. P. *J. Chem. Phys.* **1951**, *19*, 1591–1594.
- (23) Stillinger, F. H. *J. Soln. Chem.* **1973**, *2*, 141–158.
- (24) Ashbaugh, H. S.; Pratt, L. R. *Rev. Mod. Phys.* **2006**, *78*, 159–178.
- (25) Che, J.; Dzubiella, J.; Li, B.; McCammon, J. A. *J. Phys. Chem. B* **2008**, *112*, 3058–3069.
- (26) Osher, S.; Fedkiw, R. *Level Set Methods and Dynamic Implicit Surfaces*; Springer: New York, 2002.
- (27) Osher, S.; Sethian, J. A. *J. Comput. Phys.* **1988**, *79*, 12–49.
- (28) Sethian, J. A. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, 2nd ed.; Cambridge University Press: 1999.
- (29) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (30) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 553–586.
- (31) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (32) Machida, K. *Principles of Molecular Mechanics*; Kodansha and Wiley: 1999.

CT800297D



# JCTC Journal of Chemical Theory and Computation

## Molecular Dynamics Simulations of Solvation and Solvent Reorganization Dynamics in CO<sub>2</sub>-Expanded Methanol and Acetone

John L. Gohres,<sup>†,§,||</sup> Alexander V. Popov,<sup>‡,||</sup> Rigoberto Hernandez,<sup>\*,‡,§,||</sup>  
Charles L. Liotta,<sup>†,‡,§</sup> and Charles A. Eckert<sup>\*,†,‡,§</sup>

*School of Chemical & Biomolecular Engineering, School of Chemistry & Biochemistry,  
Specialty Separations Center, and Center for Computational and Molecular Science  
and Technology, Georgia Institute of Technology, Atlanta, Georgia 30332-0100*

Received August 28, 2008

**Abstract:** Composition-dependent solvation dynamics around the probe coumarin 153 (C153) have been explored in CO<sub>2</sub>-expanded methanol and acetone with molecular dynamics (MD) simulations. Solvent response functions are biexponential with two distinct decay time scales: a rapid initial decay (~0.1 ps) and a long relaxation process. Solvation times in both expanded solvent classes are nearly constant at partition compositions up to 80% CO<sub>2</sub>. The extent of solvation beyond this composition has the greatest tunability and sensitivity to bulk solvent composition. Solvent rotational correlation functions (RCFs) have also been used to explore rotational relaxation. Rotations have a larger range of time scales and are dependent on a number of factors including bulk composition, solvent-solvent interactions, particularly hydrogen bonding, and proximity to C153. The establishment of the solvation structure around a solute in a GXL is clearly a complex process. With respect to the local solvent domain around C153, it was seen to be primarily affected by a nonlinear combination of the rotational and diffusive transport dynamics.

### 1. Introduction

Gas-expanded liquids (GXLs) are a leading candidate for next-generation tunable solvents and are formed by the dissolution of appreciable amounts of gas into an organic liquid. The resulting mixture is a volume-expanded liquid phase with tunable physical and solvation properties like dielectric constant and viscosity.<sup>1,2</sup> An inherent advantage of GXLs results from the increase in gas (H<sub>2</sub>, O<sub>2</sub>) solubility in the liquid phase. Relative to neat organic solvents, GXLs have improved yield and selectivity of homogeneously catalyzed oxidation reactions.<sup>3</sup> A significant amount of research has shown that GXLs are advantageous over organic solvents for a variety of reactions,<sup>4–9</sup> extractions,<sup>10</sup> and

materials processing applications.<sup>11–18</sup> Consequently a great deal of effort has focused on understanding the molecular-scale properties of GXLs so to fully exploit their unique solvent properties.

Electronic excitation of the laser dye Coumarin 153 (C153) creates an excited-state dipole moment nearly 9 Debye greater than the ground-state dipole moment.<sup>19</sup> Recent spectroscopic and molecular dynamics (MD) simulation results showed organic enrichment around C153 in CO<sub>2</sub>-expanded solvents.<sup>20–22</sup> Different solvation patterns between the ground and excited states of C153, specifically organic and CO<sub>2</sub> density enhancements relative to the ground-state cause a solvent relaxation process to solvate the excited C153. Solvent relaxation consists of electronic and nuclear rearrangements with time scales that are dependent on bulk solvent properties and molecular interactions. MD simulations provide a direct comparison to time-resolved fluorescence and give molecular-level insight into solvation mechanisms that compose solvent reorganization. Many studies

\* Corresponding author e-mail: cae@gatech.edu.

† School of Chemical & Biomolecular Engineering.

‡ Specialty Separations Center.

|| Center for Computational and Molecular Science and Technology.

§ School of Chemistry & Biochemistry.

have explored solvation dynamics in tunable solvents with experimental and computational techniques<sup>23–27</sup> because solvent dynamics affect chemical reactions.<sup>28</sup>

Solvent dynamics impact ultrafast processes like electron-transfer and free-radical reactions. Recent studies have shown that solvent polarity affects the degree of polymerization of copper-catalyzed radical polymerizations.<sup>29,30</sup> The formation of halide ions (from halide radicals) and the subsequent coordination to the copper center is directly related to the solvent medium. Solvents that can solvate the newly created halide ion have a lower coordination equilibrium constant and ultimately better molecular weight control. Electron tunneling through a donor-bridge-acceptor dyad is a direct function of solvent environment. The solvent reorganization energy impacts the ease of electron transfer and affects the optical properties and performance of the dyad.<sup>31–33</sup> GXLs are attractive solvents for dyads and other electronic materials because the solvent environment can be manipulated by CO<sub>2</sub> adjustments, allowing *in situ* control of charge transfer or a free radical polymerization.

In this manuscript, the reorganization dynamics and rotational dynamics of a solvent around a probe within two GXLs—methanol and acetone—are explored using MD simulations through a range of bulk compositions. The primary challenge in simulating these systems emerges from the difficulty of describing true dynamics in multiphase systems within cell sizes that are amenable to current computer infrastructure. Rigorous multiphase ensemble approaches<sup>34</sup> have been performed by other groups in order to obtain the phase diagram as well as to vet the quality of the underlying potentials.

While resolving these critical questions, such Monte Carlo based approaches do not provide dynamic information. An alternative approach taken by us<sup>35</sup> and Maroncelli and co-workers<sup>36</sup> has focused on the single solution phase on which a microscopic volume may be modeled using molecular dynamics. The appropriate constraints on this unit cell—such as volume, temperature, and relative composition of cosolvents—must be obtained either by the multiphase simulations or semiempirically using experimental data. In previous work, we have found that the quality of the underlying potentials employed in the simulations appears to be sufficient to be in agreement with the experimentally obtained phase diagram and hence have focused only on the use of MD simulations to reveal the structure and dynamics in a GXL phase. The present work goes further by calculating structural and rotational correlations resulting from electronic excitation of the solute. The results demonstrate the versatile nature of GXLs in providing solvent design tools for materials processing applications and free radical and electron transfer reactions.

## 2. Computational Methods

Solvation involves both electronic and nuclear rearrangements. Solute repolarization (beyond the changes in the charge distribution from the ground to excited state) was not examined in this work as it should be a higher order correction to the primary changes in the response function due to the significant charge redistribution from the ground

to excited states. Polarization effects generally slow down the solvent response in polar solvents.<sup>37</sup> As CO<sub>2</sub> is added, the solvent structure becomes increasingly nonpolar, and thereby further reduces the role of repolarization. On the other hand, nuclear motions like solvent rotation and translation have a large impact on the response function. A typical solvent response in GXLs has two distinct time scales: a fast inertial decay period that typically accounts for ~75% of the loss in correlation and a slow long-term relaxation. The inertial decay is very fast and presumably dominated by rotation. The translational diffusion does not contribute significantly because the local density autocorrelation function observed by Shukla et al.<sup>35</sup> is much slower (~10–80 ps) than the solvation time found here (~1–10 ps). The relative speed of this solvation is due to preferential solvation of the initial and final states as will be seen in the results below. C153 rotation and translation is slow relative to the solvent atoms because of its large size and does not contribute to the relaxation process.

An investigation of the mechanical response to different solvation events provides insight into the molecular interactions and solvent properties that determine the solvent response. These are surmised by the solvent response function described in Section 2.2. However, these necessarily contain a component due simply to rotational response in the neat solvent. Hence rotational correlation functions—described in Section 2.3—must also be obtained so as to resolve the dynamics effects due to a given solute.

**2.1. Model Parameterization and Methods.** All simulations have been run using the DL\_POLY v2.0<sup>38</sup> computer suite. It implements the Verlet leapfrog algorithm to integrate the equations of motion. All molecules are treated as rigid bodies interacting with each other through a Lennard-Jones plus Coulombic interactions force field. Specifically, C153 has been modeled with an OPLS-AA force field<sup>39</sup> whose partial charges are taken from Kumar and Maroncelli.<sup>40</sup> MeOH and acetone pair interactions have been treated by the 3-site J2<sup>41</sup> and 4-site OPLS<sup>42</sup> force fields, respectively, and CO<sub>2</sub> pair interactions employed the 3-site TraPPE potential.<sup>34</sup> The methyl groups in MeOH and acetone are treated as united-atom groups in these force fields. Site-site interactions between sites on a mixed pair of molecules are determined by the Lorentz–Berthelot combining rules:  $\sigma_{12} = 0.5(\sigma_1 + \sigma_2)$  and  $\epsilon_{12} = (\epsilon_1\epsilon_2)^{0.5}$ .

Simulations are performed on a unit cell with periodic boundary conditions at a density coincident with the liquid phase of the corresponding GXL following the procedure described in prior work.<sup>35</sup> As described in the Introduction, this construct is metastable in the sense that it is single-phase and presumably at higher energy than the condition which would split into two phases. However, the system is constrained such that the splitting is inaccessible during the simulations. The initial configuration for equilibration runs is a randomly distributed periodic box of 500 solvent molecules (600 for 98% CO<sub>2</sub> GXLs) and a single C153 in the ground state. The box size has been scaled to match the liquid-phase volumes as predicted by the Patel-Teja equation of state.<sup>43</sup> Starting from equilibrated structures, the ensemble of initial configurations (needed for the nonequilibrium

simulations) is obtained by sampling structures every 6 ps during long NVT trajectories of the ground-state ( $S_0$ ) electronic configuration of C153 and solvent. Representative correlation functions and structures were initially obtained from simulations performed at various timesteps and seen to converge at 3 fs; hence all the reported simulations were performed with 3 fs timesteps. The temperature has been maintained at 300 K with a Nose-Hoover thermostat whose relaxation time is 1 ps. Nonequilibrium trajectories are initialized at each of the configurations from the equilibrium ensemble, but C153 is instantaneously placed on the  $S_1$  excited-state—assuming Franck-Condon transitions—by replacing the partial charges in the molecular mechanics force field from the ground-state to excited-state values. They are usually propagated for 9 ps under NVE conditions as this was found to be sufficient to capture most of the nonzero correlation function; but they were propagated up to 20 ps when necessary. Coordinates are saved every 45 fs and analyzed using an external FORTRAN program.

**2.2. Solvent Response Function (SRF).** Solvation dynamics are explored through the solvent response function (SRF)

$$S(t) = \frac{\Delta E(t) - \Delta E(\infty)}{\Delta E(0) - \Delta E(\infty)} \quad (1)$$

where  $\Delta E(t)$  is the energy gap between the C153 electronic states and  $S(t)$  is the SRF. The SRF is a normalized function of the energy gap between the C153 electronic states and is a measure of total solvent-solute interaction energy between the electronic states. For simplicity, the C153 Lennard-Jones parameters were assumed constant in the ground and excited states. Therefore, the energy gap is composed of the electrostatic interaction between the solvent and solute and can be written as

$$\Delta E = \frac{1}{4\pi\epsilon_0} \sum_i^N \sum_\alpha \sum_\beta \frac{\Delta q_\alpha q_{i\beta}}{r_{\alpha,i\beta}} \quad (2)$$

where  $\epsilon_0$  is the relative permittivity in vacuum,  $N$  is the number of solvent molecules,  $\alpha$  denotes a solute atom, and  $i\beta$  denotes solvent atom  $\beta$  on molecule  $i$ . Terms  $q_{i\beta}$  and  $\Delta q_\alpha$  are respectively the partial charge on solvent atom  $\beta$  and difference in partial charge between ground and excited-state on C153 atom  $\alpha$ . The nonequilibrium response can be connected to equilibrium fluctuations through the equilibrium time correlation function:

$$C(t) = \frac{\langle \delta E(0) \delta E(t) \rangle}{\langle (\delta E)^2 \rangle} \quad (3)$$

A convenient assumption for nonequilibrium simulations is the linear response approximation which provides an estimate for the SRF as  $S(t) \cong C(t)$ . In this limit, the nonequilibrium response is described by fluctuations around the average,  $\delta E(t) = \Delta E(t) - [\Delta E]$ , in equilibrium or unperturbed systems. The approximation  $S(t) \cong C(t)$  is convenient and reasonable for most neat liquid solvents but breaks down in liquid mixtures with preferential solvation or highly compressible fluids systems like SCFs.<sup>25–27</sup> Preferential solvation and local density enhancements become increasingly pronounced around excited C153, and local

density enhancements become larger in magnitude than normal solvent fluctuations. Consequently the linear assumption is not applicable in GXLs because of these local density enhancements. Although the calculation and analysis of nonequilibrium simulations is more cumbersome, it is necessary to obtain the solvation dynamics in GXLs and rotational dynamics of solutes in GXLs following excitation.

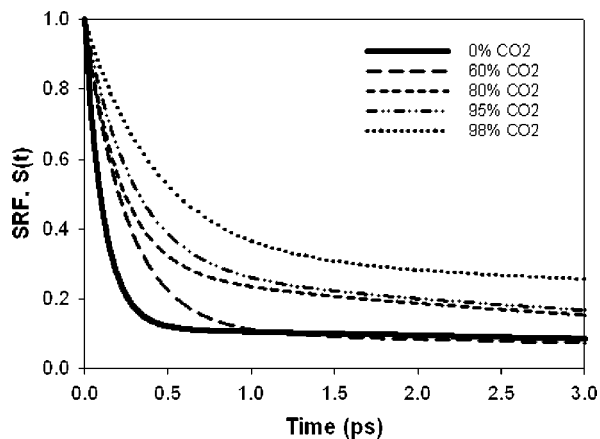
**2.3. Rotational Correlation Functions (RCFs).** Rotational dynamics of the cosolvents—CO<sub>2</sub> and the organic species—have been explored through the first- and second-order rotational correlation functions (RCF)

$$C^{(1)}(t) = \langle \vec{n}(t) \cdot \vec{n}(0) \rangle \quad (4)$$

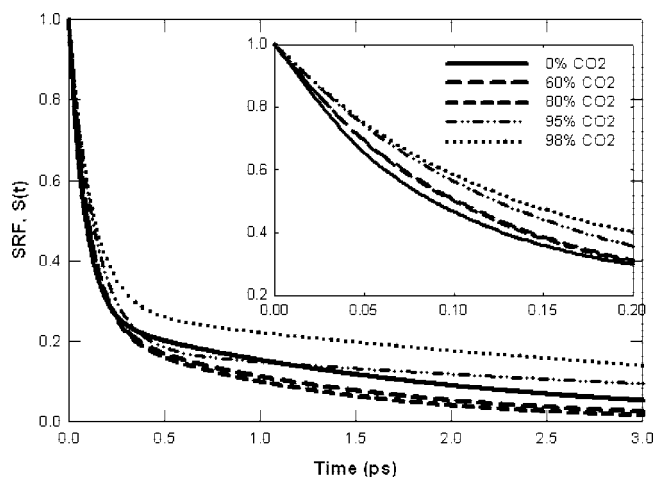
$$C^{(2)}(t) = \frac{1}{2} \langle 3[\vec{n}(t) \cdot \vec{n}(0)]^2 - 1 \rangle \quad (5)$$

where  $\vec{n}$  is a characteristic axis vector of an organic or CO<sub>2</sub> molecule, and  $C^{(1)}$  and  $C^{(2)}$  are the respective first and second ranked RCFs. CO<sub>2</sub> is a linear molecule with one characteristic vector extending from the carbon atom to an oxygen atom. A single axis of rotation is sufficient to describe MeOH rotation—the bond between the oxygen atom and the protic hydrogen. Rotation of a unit vector characterizing this bond contributes more to the solvation response than the oxygen methyl vector since the hydroxyl group is more polar and therefore more responsive to an electric field. Acetone required two axes of rotation because it is a bulky molecule with asymmetric rotations. One vector extends along the carbonyl group from the carbon to the oxygen, and the other is directed from the carbonyl carbon to a methyl group. Two orders of RCFs were used to examine solvent rotation: first-order RCFs—because their decay tends to be dominated by the collective loss of orientation relative to the initial alignment—and second-order RCFs—because their decay tends to be dominated by the loss of molecule's orientation relative to each other.

RCFs of all three solvent molecules are classified within three cases, as delineated by time relative to the excitation and the relative proximity to the C153 probe: 1) rotations of all solvent molecules immediately following C153 excitation—the SRF case; 2) rotations of molecules that are in the local or cybotactic region of C153 after excitation—the Local case; and 3) rotations that occur in the bulk fluid, i.e. no C153 molecule in the simulation—the Bulk case. This classification allows us to distinguish between bulk solvent rotations that result from normal fluctuations and solvent rotations that are affected by C153 excitation. The cybotactic region is a dynamic area that constantly changes location as the solute and solvent molecules diffuse, so assumptions were made to study this transient region. The same definition was used as was described in our previous work.<sup>21</sup> Briefly, a sphere of 7 Å was drawn outward from the C153 center of mass. Any solvent molecules that were initially in this sphere were considered part of the cybotactic region throughout the entire simulation. C153 diffusion is very slow, and most solvent rotations occur before diffusive escape from the region. Solvent molecules that entered the sphere midsimulation were not considered a part of the cybotactic region for simplification purposes.



**Figure 1.** Solvent response functions in neat acetone and CO<sub>2</sub>-expanded acetone at varying CO<sub>2</sub> compositions.



**Figure 2.** Solvent response functions in neat MeOH and CO<sub>2</sub>-expanded MeOH at varying CO<sub>2</sub> compositions. The inset magnifies the initial decay behavior.

### 3. Results and Discussion

**3.1. Solvation Dynamics.** Solvent response functions of acetone GXLs and MeOH GXLs, including neat organic liquids, are presented in Figures 1 and 2, respectively. The curves represent the average of over 1000 trajectories fit to a multiexponential decay function

$$S(t) = \sum_{i=1}^k a_i \exp(-t/\tau_i) \quad (6)$$

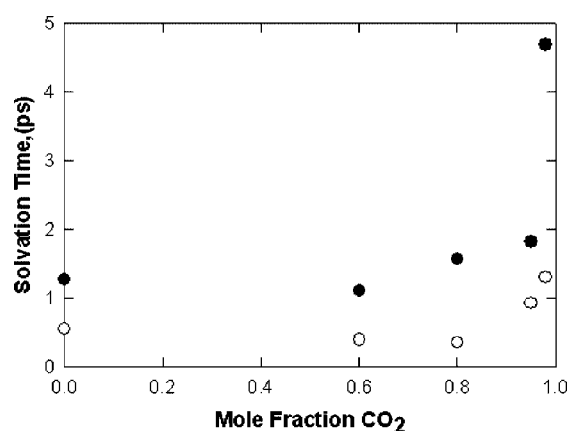
where  $a_i$  is a pre-exponential fitting parameter that gives the relative weighting of the time scale and  $\tau_i$  is a characteristic decay time that is indicative of a different solvation time scale for each of  $k$  exponentials. In the present work,  $k$  was taken to be no greater than 3, and most of the decay functions were fit well at  $k$  equal to 1 or 2. Such fits ignore the earliest ballistic part of the decay function which is nonexponential and essentially hidden at the scale displayed in, for example, Figures 1 and 2. All SRFs were fit within a standard error less than 4% using the parameters given in Table 1. Solvation time quantifies the reorganization process and is found by integrating the fit of the solvation response function as specified by eq 6, yielding

$$\tau_s = \int_0^\infty S(t) dt = \sum_{i=1}^k a_i \tau_i \quad (7)$$

where  $\tau_s$  is the effective total solvation time. A plot of solvation time versus CO<sub>2</sub> composition is shown in Figure 3 to illustrate the effects of composition on solvation that are not directly apparent in Figures 1 and 2.

The neat MeOH solvation time,  $\tau_s \approx 0.55$  ps, is much faster than the experimental result,  $\tau_s \approx 5.0$  ps, found by Horng et al.<sup>19</sup> in the supercritical regime, but it agrees reasonably well with MD simulation results from other researchers.<sup>37,40</sup> Kumar and Maroncelli<sup>40</sup> used a fixed C153 molecule and the linear time-correlation approximation of eq 3 in their analysis. Cichos et al.<sup>37</sup> used a similar approach but added a nonequilibrium case with polarizability. Polarizable force fields slow down the solvent response and provide better agreement with the experimental findings. SRFs in neat acetone ( $\tau_s \approx 1.27$  ps) are in better agreement with experimental data<sup>19</sup> ( $\tau_s \approx 0.58$  ps) than those in neat MeOH, but this result is slightly slower than the experimental data.

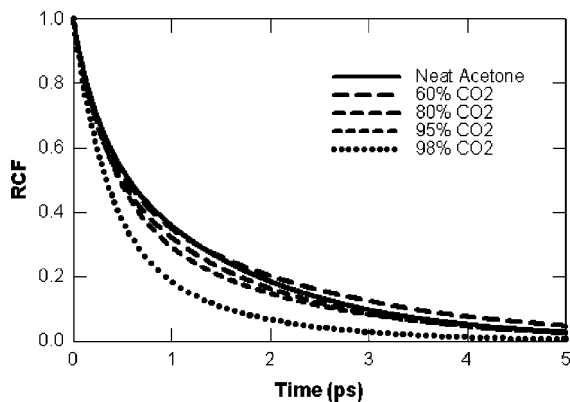
Several interesting features in the SRFs shown in Figures 1–3 suggest that solvation is solvent-dependent, and intermolecular interactions and bulk fluid properties both affect the response time. Solvation times in both GXLs remain stagnant up to  $\sim 80\%$  CO<sub>2</sub> before increasing exponentially. The effects of CO<sub>2</sub> are more pronounced in acetone GXLs where the solvent response slows down by nearly 5 ps. Similarly, CO<sub>2</sub> slows down the response at high composition in MeOH GXLs, although the time scale change is approximately 1 ps. This observed difference between the two



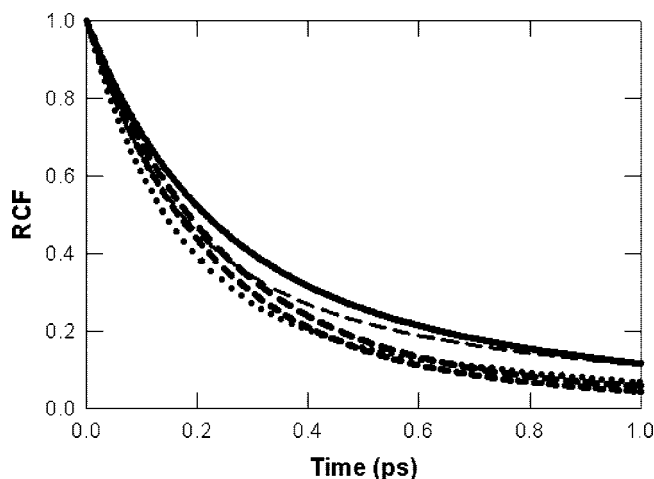
**Figure 3.** Solvation times in CO<sub>2</sub>-expanded acetone (filled circles) and MeOH (open circles).

**Table 1.** Biexponential Decay Fitting Parameters for SRF in CO<sub>2</sub>-Expanded Acetone and MeOH

% CO <sub>2</sub>	$a_1$	$\tau_1$ (ps)	$a_2$	$\tau_2$ (ps)
neat acetone	0.883	0.112	0.117	10.0
60% in acetone	0.892	0.253	0.108	8.147
80% in acetone	0.722	0.216	0.278	5.074
95% in acetone	0.715	0.288	0.285	5.66
98% in acetone	0.681	0.44	0.319	13.77
neat MeOH	0.739	0.082	0.261	1.874
60% in MeOH	0.763	0.102	0.237	1.338
80% in MeOH	0.759	0.10	0.241	1.126
95% in MeOH	0.815	0.132	0.185	4.443
98% in MeOH	0.72	0.119	0.28	4.358



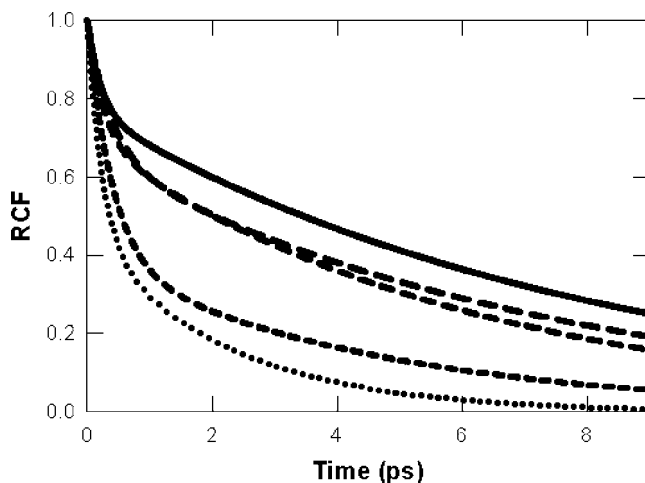
**Figure 4.** First-order acetone RCFs of the C–O bond in acetone-based GXLS for the SRF case.



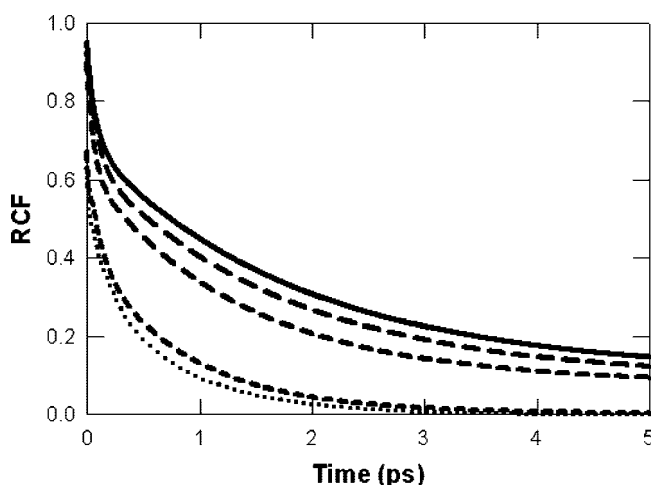
**Figure 5.** Second-order RCFs for acetone of the C–O bond around the carbonyl bond (SRF case) in neat acetone and various acetone GXLS. Lines represent the same cosolvent mixtures as used in Figure 4.

types of GXLS is directly related to the degree of cosolvent preferential solvation around C153 and its ability to respond to the excited-state dipole moment by realignment. Both acetone and MeOH preferentially solvate both ground and excited-state C153,<sup>21</sup> but there are other differences in molecular structure and solvent properties that could affect the solvent response. MeOH is a smaller molecule than acetone and could in principle diffuse faster than acetone and lead to a packing of more molecules within a local region around C153. MeOH is more polar than acetone and forms hydrogen bonds. Higher polarity causes lower solvation energy which decreases the amount of MeOH required to solvate the excited dipole; however, hydrogen bonding affects MeOH dynamics and could slow down MeOH rotations and diffusion. Thus the solvent response is a complex event that depends on the net effect of multiple factors.

**3.2. Solvent Rotation.** First- and second-order RCFs (of the C–O bond) for acetone GXLS—SRF case—are shown in Figures 4 and 5. Acetone RCFs of the C–Me bond are provided in Figures S2 and S3 in the Supporting Information. Likewise, first- and second-order RCFs for MeOH GXLS are shown in Figures 6 and 7, respectively. These figures illustrate the increase in rotational relaxation resulting from CO<sub>2</sub> addition. From the perspective of the “GXL” metaphor,

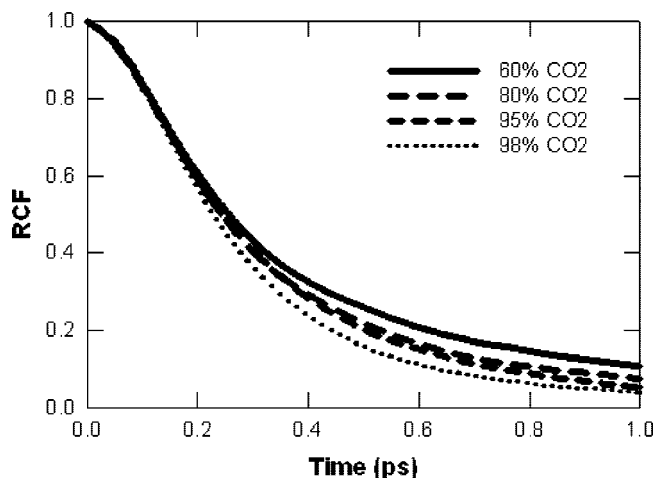


**Figure 6.** First-order MeOH RCFs (of the MeOH cylindrical symmetry axis) in MeOH-based GXLS for the SRF case. Lines represent the same cosolvent mixtures as used in Figure 4.

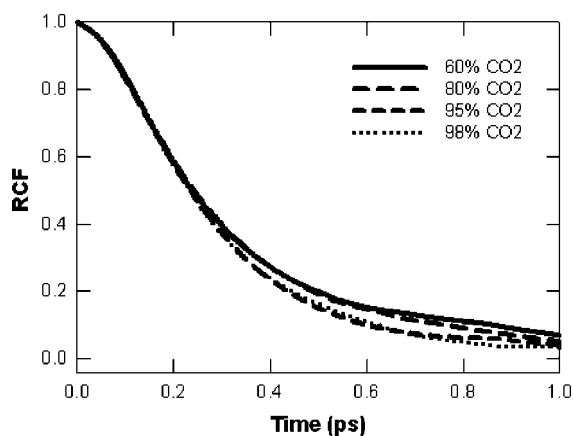


**Figure 7.** Second-order RCFs for MeOH (of the MeOH cylindrical symmetry axis) in neat MeOH and various acetone GXLS for the SRF case. Lines represent the same cosolvent mixtures as used in Figure 4.

this is not surprising because the increasing presence of CO<sub>2</sub> makes the liquid phase more gaslike and hence increases the (translational and rotational) mobility of solutes. There are several apparent features in these figures that are common throughout all the cases considered: the presence of dynamics at multiple time scales, a rapid initial decay followed by a long-term decay, and a large rate increase between 80% CO<sub>2</sub> and 95% CO<sub>2</sub>. In addition, there are two distinguishing features between acetone and MeOH RCFs: 1) initial acetone rotations are nearly identical between all GXLS and neat acetone. This is seen by the overlap of RCFs until 0.3 ps when divergence begins. 2) Acetone rotations are faster than those in MeOH. All acetone RCFs are uncorrelated within 2 ps, while MeOH RCFs indicate correlations in a range from 2 ps to 9 ps. Second-order CO<sub>2</sub> RCFs—SRF case—in acetone and MeOH GXLS are shown in Figures 8 and 9, respectively. First-order CO<sub>2</sub> RCFs can be found in Figures S1 and S4 in the Supporting Information. CO<sub>2</sub> molecules rotate faster than both organic species in the same solvent and are less sensitive



**Figure 8.** Second-order RCFs for CO<sub>2</sub> molecules in acetone GXLs.



**Figure 9.** Second-order RCFs for CO<sub>2</sub> molecules in MeOH GXLs.

to bulk composition, although rotations are faster when more CO<sub>2</sub> is present. CO<sub>2</sub> has an initial lag period during the first 0.1 ps that is not present in the organic RCFs. CO<sub>2</sub> could be initially unresponsive because weak intermolecular interactions prevent an initial thrust to start rotation. This time lag,  $\tau_{\text{lag}}$ , can be estimated from an approximation for the velocity autocorrelation function suitable for one-dimensional motion at small times<sup>44</sup>

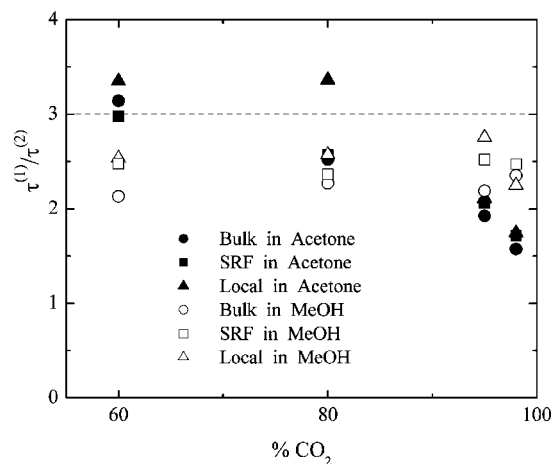
$$C(t) \approx \exp(-[\sqrt{t^2 + \tau_{\text{lag}}^2} - \tau_1]/\tau_1) \quad (8)$$

where  $\tau_1$  is the apparent monoexponential decay time. The inflection point for this function,  $t_{\text{infl}}$ , obeys the equation

$$t_{\text{infl}}^2 \sqrt{t_{\text{infl}}^2 + \tau_{\text{lag}}^2} = \tau_1 \tau_{\text{lag}}^2 \quad (9)$$

As can be seen from Figures 8, 9, S3, and S4,  $t_{\text{infl}}$  is approximately 0.2 ps. With  $\tau_1$  ranging from 0.3 to 0.9 ps (Tables S1 and S2), one obtains  $\tau_{\text{lag}}$  in a range from 0.1 ps to 0.2 ps. Note that this lag period corresponds to the average time between collisions, and, as it has already been mentioned, it is larger for CO<sub>2</sub> molecules due to their weak interaction.

All RCFs were fit by a sum of exponential decay functions per eq 6 ignoring the initial very-fast ballistic component.



**Figure 10.** Ratio of the first-order and second-order total relaxation times for CO<sub>2</sub> molecules in acetone (filled symbols) and MeOH (open symbols) GXLs. Data are taken from Tables S1 and S2.

Most organic RCFs were fit with biexponential decay functions, although second-order MeOH RCFs were fit with triexponential decay functions, and all CO<sub>2</sub> RCFs were fit with single exponential decay functions. All fits had a standard error less than 2%, but most were less than 1%. RCF fitting parameters for Bulk, SRF, and Local cases for CO<sub>2</sub> and organic RCFs are presented in the Supporting Information. As shown in Figure 10, total decay times ( $\tau^{(1)}$  in Table S1) in the first-order RCF—cf.  $C^{(1)}(t)$  in eq 4—are approximately 3 times slower than the total decay times ( $\tau^{(2)}$  in Table S2) in the second-order RCF—cf.  $C^{(2)}(t)$  in eq 5. Solvent molecules in this system range from very fast rotors like CO<sub>2</sub> to relatively slow rotors like MeOH and are thereby expected to exhibit most of the range of possible responses. The ratio of 3—about which much of the data in Figure 3 lie—is typical for liquids and emerges exactly within the Debye approximation in rotational Brownian diffusion.<sup>45</sup> Thus the fact that the computational data yields values in agreement with this limit is an indication of the quality of the results. Moreover the total relaxation decay times are seen to lie in the range of 1 ps to 5 ps in Figure 3 that includes a significant slower component relaxation decay time. The latter is in general agreement with the total relaxation times seen in C102 in acetonitrile-water mixtures and attributed to solvent reorganization and kinetic energy transfer.<sup>46</sup>

The rotational relaxation is, of course, not exactly freely diffusive, and that is also indicated by the variations in the ratios of the decay times in Figure 10. The first nontrivial exponential decay time scales are similar in all three cases; however, divergence occurs at longer time scales. This indicates that a solvent molecule becomes trapped in a stable alignment with C153 and cannot freely rotate. The fast time scales are on the order of 0.1 ps as provided in Table 1 and the Supporting Information tables but do increase with increasing CO<sub>2</sub> composition. A similar fast decay time scale was observed by Underwood and Blank<sup>47</sup> for C102 in acetonitrile. They attribute this to a dipole–dipole interaction between coumarin and the solvent. That is, it is the relaxation of the direct interaction between solvent and solute for



Solvation behavior is most tunable at higher CO<sub>2</sub> compositions, and relaxation time scales can be adjusted several ps with moderate CO<sub>2</sub> pressure changes.

The response of the GXL solvent to a repolarization of a solute—through, for example, electronic excitation—occurs at a range of time scales from 1 to 10 ps. The underlying motion—translational and rotational—induced by the excitation at these time scales are also comparable to the intrinsic (thermal) rotation of the solvent, and hence it is a nontrivial exercise to deconvolute their responses. Meanwhile, the collision time is just under 1 ps, and the response is certainly affected by solvent-solvent interactions in the local domain of the solute. First- and second-order rotational correlation functions have been obtained in the bulk and local solvent regimes to explore solute effects on rotation and rotational effects on solvation. Solvent rotational decay is determined by several factors: cosolvent polarity and interactions, proximity to the solute, and bulk composition. CO<sub>2</sub> rotations are insensitive to composition, but the two organic species depend on all factors. Acetone and MeOH rotations are faster with added CO<sub>2</sub> and distance from C153. MeOH rotations are hindered because of hydrogen bonds, while acetone rotations are viscosity dependent. Solvation and rotational dynamics in GXLs are relatively insensitive to composition until higher CO<sub>2</sub> compositions (>80%). Solvation dynamics affect reaction kinetics, and the tunability of GXLs shows potential in atom transfer radical polymerizations and electron transfer through nanodevices.

**Acknowledgment.** This work has been supported by a Department of Energy Basic Energy Sciences grant, DE-FG02-04ER15521. C.A.E. acknowledges the support of the J. Erksine Love, Jr., Institute Chair. The computational facilities at the CCMST have been supported under NSF grant CHE 0443564.

**Supporting Information Available:** Detail about the computationally observed rotational correlation functions (RCF) and their associated total and partial decay times (Figures S1–S4 and Tables S1–S5). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- Hallett, J. P.; Kitchens, C. L.; Hernandez, R.; Liotta, C. L.; Eckert, C. A. *Acc. Chem. Res.* **2006**, *39*, 531.
- Jessop, P. G.; Subramaniam, B. *Chem. Rev.* **2007**, *107*, 2666.
- Wei, M.; Musie, G. T.; Busch, D. H.; Subramaniam, B. *J. Am. Chem. Soc.* **2002**, *124*, 2513.
- Chamblee, T. S.; Weikel, R. R.; Nolen, S. A.; Liotta, C. L.; Eckert, C. A. *Green Chem.* **2004**, *6*, 382.
- Eckert, C. A.; Liotta, C. L.; Bush, D.; Brown, J. S.; Hallett, J. P. *J. Phys. Chem. B* **2004**, *108*, 18108.
- Xie, X. F.; Liotta, C. L.; Eckert, C. A. *Ind. Eng. Chem. Res.* **2004**, *43*, 7907.
- Jin, H.; Subramaniam, B.; Ghosh, A.; Tunge, J. *AICHE J.* **2006**, *52*, 2575.
- Fang, J.; Jin, H.; Ruddy, T.; Pennybaker, K.; Fahey, D.; Subramaniam, B. *Ind. Eng. Chem. Res.* **2007**, *46*, 8687.
- Nunes, R. M. D.; Arnaut, L. G.; Solntsev, K. M.; Tolbert, L. M.; Formosinho, S. J. *J. Am. Chem. Soc.* **2005**, *127*, 11890.
- Eckert, C.; Liotta, C.; Ragauskas, A.; Hallett, J.; Kitchens, C.; Hill, E.; Draucker, L. *Green Chem.* **2007**, *9*, 545.
- Myneni, S.; Hess, D. W. *J. Electrochem. Soc.* **2003**, *150*, G744.
- Levitin, G.; Myneni, S.; Hess, D. W. *J. Electrochem. Soc.* **2004**, *151*, G380.
- Anand, M.; McLeod, M. C.; Bell, P. W.; Roberts, C. B. *J. Phys. Chem. B* **2005**, *109*, 22852.
- McLeod, M. C.; Anand, M.; Kitchens, C. L.; Roberts, C. B. *Nano Lett.* **2005**, *5*, 461.
- Spuller, M. T.; Perchuk, R. S.; Hess, D. W. *J. Electrochem. Soc.* **2005**, *152*, G40.
- Kitchens, C. L.; Roberts, C. B. *Ind. Eng. Chem. Res.* **2006**, *45*, 1550.
- Song, I.; Spuller, M.; Levitin, G.; Hess, D. W. *J. Electrochem. Soc.* **2006**, *153*, G314.
- Anand, M.; You, S. S.; Hurst, K. M.; Saunders, S. R.; Kitchens, C. L.; Ashurst, W. R.; Roberts, C. B. *Ind. Eng. Chem. Res.* **2008**, *47*, 553.
- Hornig, M. L.; Gardecki, J. A.; Papazyan, A.; Maroncelli, M. *J. Phys. Chem.* **1995**, *99*, 17311.
- Li, H. P.; Arzhantsev, S.; Maroncelli, M. *J. Phys. Chem. B* **2007**, *111*, 3208.
- Gohres, J.; Kitchens, C.; Hallett, J.; Popov, A.; Hernandez, R.; Liotta, C.; Eckert, C. *J. Phys. Chem. B* **2008**, *112*, 4666.
- Gohres, J. L.; Hernandez, R.; Liotta, C. L.; Eckert, C. A. Viewing the cybotactic structure of gas-expanded liquids. In *Green Chemistry and Engineering with Gas Expanded Liquids and Near-critical Media*; ACS Symposium Series 1006; Hutchenson, K. W., Scurto, A. M., Subramaniam, B., Eds.; American Chemical Society: Washington, DC, 2008; in press.
- Parsons, D. F.; Vener, M. V.; Basilevsky, M. V. *J. Phys. Chem. A* **1999**, *103*, 1171.
- Agmon, N. *J. Phys. Chem. A* **2002**, *106*, 7256.
- Egorov, S. A. *Phys. Rev. Lett.* **2004**, *93*.
- Graf, P.; Nitzan, A. *Chem. Phys.* **1998**, *235*, 297.
- Egorov, S. A. *J. Chem. Phys.* **2004**, *121*, 6948.
- Brennecke, J. F.; Chateaufneuf, J. E. *Chem. Rev.* **1999**, *99*, 433.
- Tsarevsky, N. V.; Pintauer, T.; Matyjaszewski, K. *Macromolecules* **2004**, *37*, 9768.
- Braunecker, W. A.; Matyjaszewski, K. *J. Mol. Catal., A: Chem.* **2006**, *254*, 155.
- Weiss, E. A.; Ahrens, M. J.; Sinks, L. E.; Ratner, M. A.; Wasielewski, M. R. *J. Am. Chem. Soc.* **2004**, *126*, 9510.
- Ratera, I.; Sporer, C.; Ruiz-Molina, D.; Ventosa, N.; Baggerman, J.; Brouwer, A. M.; Rovira, C.; Veciana, J. *J. Am. Chem. Soc.* **2007**, *129*, 6117.
- Liu, M.; Waldeck, D. H.; Oliver, A. M.; Head, N. J.; Paddon-Row, M. N. *J. Am. Chem. Soc.* **2004**, *126*, 10778.
- Potoff, J. J.; Siepmann, J. I. *AICHE J.* **2001**, *47*, 1676.
- Shukla, C. L.; Hallett, J. P.; Popov, A. V.; Hernandez, R.; Liotta, C. L.; Eckert, C. A. *J. Phys. Chem. B* **2006**, *110*, 24101.



- (36) Li, H. P.; Maroncelli, M. *J. Phys. Chem. B* **2006**, *110*, 21189.
- (37) Cichos, F.; Brown, R.; Bopp, P. A. *J. Chem. Phys.* **2001**, *114*, 6834.
- (38) Smith, W.; Forester, T. R. *J. Mol. Graphics* **1996**, *14*, 136.
- (39) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810.
- (40) Kumar, P. V.; Maroncelli, M. *J. Chem. Phys.* **1995**, *103*, 3038.
- (41) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276.
- (42) Jorgensen, W. L.; Briggs, J. M.; Contreras, M. L. *J. Phys. Chem.* **1990**, *94*, 1683.
- (43) Patel, N. C.; Teja, A. S. *Chem. Eng. Sci.* **1982**, *37*, 463.
- (44) Heyes, D. M.; Powles, J. G.; Rickayzen, G. *Mol. Phys.* **2002**, *100*, 595.
- (45) Hansen, J. P.; McDonald, I. R. *Theory of simple liquids*, 2nd ed.; Academic Press: London, Orlando, 1986.
- (46) Wells, N. P.; McGrath, M. J.; Siepman, J. I.; Underwood, D. F.; Blank, D. A. *J. Phys. Chem. A* **2008**, *112*, 2511.
- (47) Underwood, D. F.; Blank, D. A. *J. Phys. Chem. A* **2005**, *109*, 3295.

CT800353S

## Simple, Efficient, and Reliable Computation of Multiple Free Energy Differences from a Single Simulation: A Reference Hamiltonian Parameter Update Scheme for Enveloping Distribution Sampling (EDS)

Clara D. Christ<sup>†</sup> and Wilfred F. van Gunsteren<sup>\*,†</sup>

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH,  
8093 Zürich, Switzerland

Received October 7, 2008

**Abstract:** We present an automatic adaptive scheme which allows fast optimization of the reference Hamiltonian parameters in enveloping distribution sampling (EDS). Six different variants of the update scheme have been tested on a condensed phase test system which included the recurrent deletion and creation of complete water molecules in water. All six schemes gave accurate free energy estimates with absolute errors of up to 1 kJ/mol for the worst scheme and up to 0.1 kJ/mol for the best scheme. Configurational sampling is focused on the regions where the end state energy difference distributions intersect, explaining the high accuracy and precision of the free energy estimates. The new update scheme makes the application of EDS to other systems, e.g. in ligand binding studies, easy as no reference state Hamiltonian parameters have to be chosen by the user. The only necessary input are the Hamiltonians of the various end states involved.

### 1. Introduction

Estimation of free energies from molecular simulation has been an active field of research for several decades and many review articles and books are devoted to the topic.<sup>1–18</sup> Although the basic equations used for free energy calculations have been proposed decades ago,<sup>19,20</sup> it still remains challenging to accurately calculate free energies. This is due to two main challenges that have to be met when estimating free energies from molecular simulation. First, the system of interest, e.g. different ligands binding to a common receptor, has to be described by an appropriate model. Classical models are often used to this end as they allow computationally cheap evaluation of the energy and the forces. For the model to be appropriate it must describe the thermodynamics of the system correctly. References 21 and 22 compare the accuracy of several classical models used for free energy prediction. In the current work, we will focus on the second challenge which consists of finding an efficient evaluation scheme for the free energy. Estimation of free

energies involves calculation of the partition function which is for most cases not accessible analytically. Fortunately, in many applications it suffices to calculate relative free energies, which reduces the problem to the estimation of partition function ratios, i.e. the estimation of relative probabilities. As an example, let us consider the free energy of folding of a peptide. Let us assume that we have a clear measure of when the peptide is in the folded state and when in the unfolded. Further assume that we have a sampling scheme (e.g., molecular dynamics) that efficiently samples configuration space. We can then estimate the free energy of folding from the ratio of number of folded configurations to unfolded configurations we have encountered. This simple example illustrates in an intuitive fashion (see section 2 for more rigorous arguments) that an efficient simultaneous sampling of the configuration space of all end states involved allows the calculation of the free energy differences. In this example, the two states were defined by the same Hamiltonian plus a set of criteria or constraints that distinguish the folded from the unfolded state. If the two or more end states correspond, however, e.g. to two chemically distinct species a combined Hamiltonian has to be constructed. This is often

\* Corresponding author. e-mail: wfvgn@igc.phys.chem.ethz.ch.

<sup>†</sup> Swiss Federal Institute of Technology.

done by the so-called coupling parameter approach<sup>23</sup> where the Hamiltonian  $H$  is now a function of the coupling parameter  $\lambda$ , and  $\lambda$  is chosen such that  $H(\lambda = 0)$  corresponds to state  $A$  and  $H(\lambda = 1)$  to state  $B$ . This basic form of the Hamiltonian is used in many methods; however, there are many variants: first, the functional dependence on  $\lambda$  differs in various methods, and second, the way the configuration space of this newly constructed combined state is sampled may differ: either  $\lambda$  is a parameter and the sampling over the whole  $\lambda$  range is achieved by performing multiple simulations at fixed  $\lambda$ -values (as e.g. in thermodynamic integration (TI),<sup>19</sup> free energy perturbation (FEP),<sup>20</sup> the Bennett acceptance ratio method (BAR),<sup>24</sup> overlap-sampling,<sup>25</sup> or also in Hamiltonian-replica-exchange),<sup>26–28</sup> or  $\lambda$  is allowed to change during the simulation either by Metropolis Monte Carlo moves like in chemical Monte-Carlo molecular dynamics (CMC/MD),<sup>29</sup> or because it is treated as a dynamic variable as in  $\lambda$ -dynamics.<sup>30</sup> If  $\lambda$  is treated dynamically, a biasing scheme<sup>31</sup> has to be used in order to ensure sampling over the whole  $\lambda$  range. Alternatively, free energy estimates can also be obtained from multiple irreversible, independent simulations in which  $\lambda$  is changed so fast that the system is driven out of equilibrium.<sup>32–34</sup>

If the important configuration space of the  $\lambda$  combined Hamiltonian is well sampled during the simulation it contains as a subset the important configuration space of the endstates  $A$  and  $B$ , allowing for an accurate estimation of the free energy difference.<sup>35</sup> However, the coupling parameter way of combining the end state Hamiltonians has two major drawbacks: first, a lot of computer time is spent on the simulation of regions of configurational space at intermediate  $\lambda$  values and, second, an extension of the approach to multiple end states rapidly makes the calculation unfeasible.

As traditional methods to calculate free energy differences are not easily extended to multiple end states, we have developed enveloping distribution sampling (EDS).<sup>36,37</sup> EDS is an implementation of the umbrella sampling method<sup>31</sup> and belongs to the class of importance sampling<sup>38</sup> methods. EDS is designed to allow for sampling of the important configuration space of multiple end states during a single simulation of a so-called reference state. This goal is also pursued in single-step perturbation.<sup>39,40</sup> Unlike the “hand-made” reference states used in single-step perturbation, EDS uses the strategy of expanding the sampled ensemble<sup>41–46</sup> using the following Hamiltonian  $H_R = -(\beta s)^{-1} \ln \sum_i^N \exp[-\beta s(H_i - E_i^R)]$ <sup>37,42,47,48</sup> (see section 2) where the energy offset parameters  $E_i^R$  ensure equal sampling of all  $N$  end states and the smoothness parameter  $s$  is chosen such that transitions between regions of phase space important to the different end states can occur, i.e. as low as necessary to overcome barriers and as high as possible in order to avoid unnecessary widening of the configuration space that has to be sampled during the simulation. As we have shown in previous work,<sup>36</sup> the accuracy of free energy differences, which can be calculated “on the fly” from the reference state simulation, strongly depends on the chosen parameters. In this work we, therefore, develop and compare schemes that allow an automatic update of the reference Hamiltonian parameters.

In section 2, the EDS working equations are derived together with the equations used for parameter update. Section 3 states the simulation protocols and the followed parameter update schemes. As a test system, we chose annihilation and creation of five water molecules in liquid water (see section 3). The system is designed such that all free energy differences are zero, i.e. the exact result is known. However, it is a highly challenging system as a water molecule is annihilated at one point in space and created at another point in space when moving from the important configuration space of one end state to that of another. In section 4 the results of the different update schemes are presented and discussed.

## 2. Theory

The free energy  $F$  of a state  $X$  is defined as

$$F_X = -\beta^{-1} \ln Q_X \quad (1)$$

where  $\beta^{-1} = k_B T$ ,  $k_B$  is Boltzmann’s constant,  $T$  is the absolute temperature, and  $Q_X$  is the partition function of state  $X$

$$Q_X = (h^{3N_p} N_p!)^{-1} \int \int \exp[-\beta H_X(\mathbf{p}, \mathbf{r})] \mathbf{p} \mathbf{r} \quad (2)$$

Here,  $h$  is Planck’s constant,  $N_p$  is the number of particles,  $\mathbf{r}$  and  $\mathbf{p}$  are the  $3N$ -dimensional vectors of the particle positions and conjugate momenta, respectively, and  $H_X$  is the Hamiltonian of state  $X$ . The factor  $(N_p!)^{-1}$  only occurs for indistinguishable particles. If the Hamiltonian can be split into a kinetic  $K_X(\mathbf{p})$  and a potential energy part  $V_X(\mathbf{r})$ ,

$$H_X(\mathbf{p}, \mathbf{r}) = K_X(\mathbf{p}) + V_X(\mathbf{r}) \quad (3)$$

a separation of the corresponding free energies is allowed,

$$F_X = -\beta^{-1} \ln \left\{ \int \exp[-\beta K_X(\mathbf{p})] \mathbf{p} \right\} - \beta^{-1} \ln \left\{ \int \exp[-\beta V_X(\mathbf{r})] \mathbf{r} \right\} + \beta^{-1} \ln(h^{3N_p} N_p!) \quad (4)$$

In molecular simulations, the kinetic part of the Hamiltonian typically reads  $K(\mathbf{p}) = \sum_{i=1}^N \mathbf{p}_i^2 / (2m_i)$  and, therefore, the first integral in eq 4 can be solved analytically. In the following we will omit the kinetic part and the constant term for simplicity. Unlike the kinetic term the potential energy part of the Hamiltonian usually involves interaction terms that cannot be separated, making it impossible to solve the second integral (the configurational integral) in eq 4 analytically. Therefore, the free energy of a multidimensional system such as e.g. a molecule in solution cannot be calculated using eq 4. Rather than calculating the absolute free energy by evaluation of the configurational integral, one can calculate the free energy difference between two states

$$\Delta F_{BA} = F_B - F_A = -\beta^{-1} \ln \frac{Q_B}{Q_A} \quad (5)$$

where one now has to calculate a ratio of partition functions rather than the partition functions themselves. Inserting the expressions for the partition functions one finds<sup>20</sup>

$$\Delta F_{BA} = -\beta^{-1} \ln \langle \exp[-\beta(V_B - V_A)] \rangle_A \quad (6)$$

where  $\langle \rangle_A$  indicates an average over an ensemble sampled at state  $A$ . Unfortunately, this expression will only yield reasonable free energy estimates if the important phase space of state  $B$  is a subset of the important phase space of state  $A$ .<sup>35</sup> Moreover, the expected error is minimal only for  $A = B$ .<sup>24,37</sup>

Introducing the energy difference distributions

$$\begin{aligned} \rho_A(\Delta V; \Delta V_{BA}) &= \langle \delta[\Delta V - (V_B - V_A)] \rangle_A \\ \rho_B(\Delta V; \Delta V_{BA}) &= \langle \delta[\Delta V - (V_B - V_A)] \rangle_B \end{aligned} \quad (7)$$

we obtain<sup>49–51</sup>

$$\rho_B(\Delta V; \Delta V_{BA}) \exp[-\beta \Delta F_{BA}] = \rho_A(\Delta V; \Delta V_{BA}) \exp[-\beta \Delta V] \quad (8)$$

or

$$\ln \frac{\rho_A(\Delta V; \Delta V_{BA})}{\rho_B(\Delta V; \Delta V_{BA})} = +\beta \Delta V - \beta \Delta F_{BA} \quad (9)$$

That is, an alternative way to calculate the free energy difference  $\Delta F_{BA}$  is to calculate the  $\rho_A(\Delta V; \Delta V_{BA})$  and  $\rho_B(\Delta V; \Delta V_{BA})$  distributions, e.g. from two simulations at states  $A$  and  $B$ . Equation 8 indicates that the free energy difference is the energy difference  $\Delta V$  where these two distributions intersect or, alternatively, one can use eq 9 to do a linear regression and obtain  $-\beta \Delta F_{BA}$  as the ordinate intercept. These equations show that in order to obtain reasonable relative free energy estimates the energy difference distributions must overlap. If the important phase space regions of states  $A$  and  $B$  lie far apart, this will, however, not be the case.

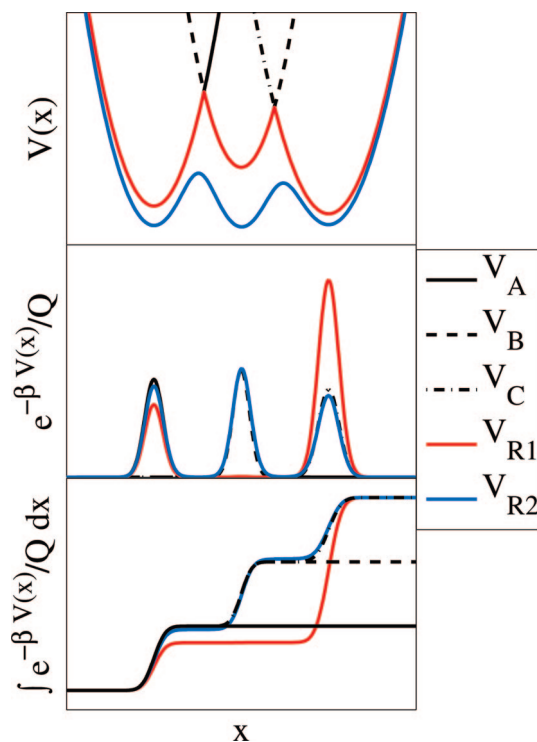
In order to sample the  $\rho_A(\Delta V; \Delta V_{BA})$  and  $\rho_B(\Delta V; \Delta V_{BA})$  distributions over the necessary range and to ensure that the important phase space of all end states is properly sampled, one can estimate the free energy difference between states  $A$  and  $B$  via the simulation of a reference state  $R$ ,

$$\Delta F_{BA} = -\beta^{-1} \ln \frac{\langle \exp[-\beta(V_B - V_R)] \rangle_R}{\langle \exp[-\beta(V_A - V_R)] \rangle_R} \quad (10)$$

A reference state Hamiltonian which does sample the important phase space of both  $A$  and  $B$  reads<sup>37,42,47,48</sup>

$$V_R(r) = -(\beta s)^{-1} \ln \left\{ \sum_{i=1}^N \exp[-\beta s(V_i(r) - E_i^R)] \right\} \quad (11)$$

where  $N$  denotes the number of EDS states, e.g. in the case of two states  $A$  and  $B$ ,  $N = 2$ .  $E_i^R$  are energy offset parameters, and  $s > 0$  is the dimensionless smoothness parameter. For  $E_i^R = F_i$  and  $s = 1$ , eq 11 is the Hamiltonian that minimizes the expected error of eq 10.<sup>37</sup> Figure 1 shows a pictorial representation of the reference state of three end states  $V_A$ ,  $V_B$ , and  $V_C$  (shown as black lines).  $V_{R1}$  corresponds to an unoptimized reference state Hamiltonian i.e.  $E_i^R = 0$ . The middle and the lower panel show that this leads to uneven sampling of the important configuration space of the three states. For example, configurations important to the  $B$  state are hardly important for  $V_{R1}$ . Equal sampling of all end states can be obtained by setting the energy offsets  $E_i^R$  to the corresponding free energies  $F_i$ . These are, however, not known in the beginning and have to be updated iteratively.



**Figure 1.** Pictorial representation of the influence of the parameters  $s$  and  $E_i^R$  of the reference Hamiltonian.  $V_{R1}$  is an unoptimized reference state Hamiltonian which does not lead to equal sampling of all three end states ( $V_A$ ,  $V_B$ , and  $V_C$ ).  $V_{R2}$  is optimized such that barriers are reduced and all states are sampled evenly. (Note that for better comparison, the end state distributions are divided by three in the central plot. In the bottom plot, the integrals over these distributions are summed up.)

Adjusting the energy offset parameters is however not sufficient as high barriers on the reference state potential energy surface between regions of configuration space important to one state and to another one may prevent efficient sampling. The smoothness parameter  $s$  can be lowered in order to decrease these barriers. Adjustment of energy offsets and smoothness parameters has been done for  $V_{R2}$  which now ensures that the important configuration space of all end states is sampled.

When using the reference state Hamiltonian eq 11 in a molecular dynamics implementation, one has to integrate the equations of motion,

$$\begin{aligned} \dot{\mathbf{r}}_k(t) &= m^{-1} \mathbf{p}_k(t) \\ \dot{\mathbf{p}}_k(t) &= \mathbf{f}_k(t) = \left( -\frac{\partial V_R(\mathbf{r})}{\partial \mathbf{r}_k} \right) \\ &= \sum_{i=1}^N \left\{ \frac{\exp[-\beta s(V_i - E_i^R)]}{\sum_{j=1}^N \exp[-\beta s(V_j - E_j^R)]} \left( -\frac{\partial V_i(\mathbf{r})}{\partial \mathbf{r}_k} \right) \right\} \\ &= \sum_{i=1}^N \left\{ \left[ \sum_{\substack{j=1 \\ j \neq i}}^N \exp[-\beta s(\Delta V_{ji} - \Delta E_{ji}^R)] + 1 \right]^{-1} \left( -\frac{\partial V_i(\mathbf{r})}{\partial \mathbf{r}_k} \right) \right\} \end{aligned} \quad (12)$$

where  $\Delta V_{ji} = V_j - V_i$ . That is, the force on an atom is a sum over force contributions of all different states. All force

contributions are multiplied with a prefactor and all these prefactors add up to one. The parameters in the prefactor determine how often a state is visited ( $E_i^R$ ) and whether transitions from configuration space important to one state to that important to another state are possible ( $s$ ).

In order to update the energy offset parameters, one has to “count” how often a state is visited. This can be done by evaluating

$$E_i^R(\text{new}) = -\beta^{-1} \ln \left\langle \left( \sum_{\substack{j=1 \\ j \neq i}}^N \exp[-\beta(\Delta V_{ji} - \Delta E_{ji}^R)] + 1 \right)^{-1} \right\rangle + E_i^R \quad (13)$$

for all states  $i$ . If a configuration is important to state  $i$  and not important to the other states  $j$  then  $\Delta V_{ji} = V_j - V_i$  is a big positive number and all exponential functions in eq 13 are approximately zero, i.e. the “counting function” (inside  $\langle \rangle$ ) returns a value close to one. If a configuration is not important to  $i$  but to another state  $k$ , then for  $j = k$  the exponential function will return a big number, i.e. the counting function returns a value close to zero. So, if a state  $i$  is insufficiently visited, the energy offset  $E_i^R$  is raised in order to increase the number of visits of state  $i$  in the next iteration. Note that the energy offset is  $E_i^R(\text{new}) = \Delta F_{iR}$  if  $s = 1$ , which can be verified by inserting the reference state Hamiltonian (eq 11) into the equation for  $\Delta F_{iR}$ .

If the important parts of phase space of two states  $i$  and  $j$  lie far apart then  $\Delta V_{ji}$  (see eq 12) is likely to be big and no transitions between these regions of phase space will occur. By lowering  $s$  ( $s > 0$ ), one can compensate for this. An optimized  $s$  parameter is obtained by solving

$$\ln \sum_{\substack{j=1 \\ j \neq i}}^N \{ [\langle \exp[-\beta(|\Delta V_{ji}| - \Delta E_{ji}^R)] \rangle_i]^s \} = \ln(N-1) - 1 \quad (14)$$

numerically for  $s$  for all states  $i$  and taking the lowest  $s$ . The reasoning behind this heuristic optimization equation is the following. Assume we are sampling configurations which are of importance to state  $i$  but unfavorable for all the other  $N-1$  states. Then all  $\Delta V_{ji}$  will be in general big positive numbers ( $V_j \gg V_i$ ). Therefore, the force prefactor in front of  $(-\partial V_i(\mathbf{r})/\partial \mathbf{r}_k)$  (see eq 12) will be approximately one whereas all the other force prefactors ( $i' \neq i$ ) will be approximately zero and no transitions from state  $i$  to any of the other states  $i'$  will be observed. However, if  $s$  ( $s > 0$ ) is lowered, transitions to the other states can occur. How much  $s$  needs to be lowered depends on the magnitude of  $\Delta V_{ji} - \Delta E_{ji}^R$  when sampling configurations that are of importance to state  $i$ . An obvious choice to estimate the average magnitude of this quantity would be to evaluate  $\langle \Delta V_{ji} - \Delta E_{ji}^R \rangle_i$ . As the distribution of  $\Delta V_{ji}$  values might, however, be very broad and have tails at very high energies, we chose to use

$$-\beta^{-1} \ln \langle \exp[-\beta(|\Delta V_{ji}| - \Delta E_{ji}^R)] \rangle_i \quad (15)$$

instead. This is an estimate of the smallest  $\Delta V_{ji} - \Delta E_{ji}^R$  values observed when sampling configurations of importance to state  $i$ , i.e. the configurations which have most overlap with state

$j$  and where a transition to this state is most likely. Taking the absolute value is done for numerical reasons. In general  $\Delta V_{ji}$  will be positive if a configuration is of importance to state  $i$  and negative if it is of importance to state  $j$ . In the latter case,  $\exp[-\beta(\Delta V_{ji} - \Delta E_{ji}^R)]$  can be very large and might contribute to the calculated ensemble average (eq 15) although the configuration has negligible importance for the  $i$  ensemble. The basic (heuristic) Ansatz is now to enforce  $\beta^{-1} = s(\Delta V_{ji} - \Delta E_{ji}^R)$  (see eq 12). Substituting this and eq 15 into the equation for the force prefactor (see eq 12), we obtain

$$\sum_{\substack{j=1 \\ j \neq i}}^N \exp[-\beta s (-\beta^{-1} \ln \langle \exp[-\beta(|\Delta V_{ji}| - \Delta E_{ji}^R)] \rangle_i)] = \sum_{\substack{j=1 \\ j \neq i}}^N \exp[-\beta \beta^{-1}] \quad (16)$$

which, after some rearrangement, leads to eq 14.

### 3. Simulation Protocols

The test system consisted of five (solute) water molecules in a cubic box of (solvent) water (1175 simple point charge (SPC)<sup>52</sup> water molecules in total, box length: 3.31 nm). In each of the five states, one of the five (solute) water molecules was interacting with the solvent water molecules while the other four were noninteracting (“dummy molecules”), i.e. their nonbonded interactions with the other water molecules were set to zero. Therefore, all states consist of one interacting solute water molecule and four noninteracting water molecules in a box of solvent water. This implies that all states have the same free energy.

All simulations were performed under  $NVT$  conditions using the weak-coupling method<sup>53</sup> ( $T = 300$  K,  $\tau_T = 0.1$  ps, and solute and solvent coupled to separate temperature baths). The translational motion of and the rotational motion around the center of mass was removed every 10 000 steps. Energies of the reference state and the end states were saved every 0.1 ps. The SHAKE algorithm<sup>54</sup> (tolerance: 0.0001) was used to constrain all bond lengths and angles to their ideal values. Nonbonded interactions were calculated using the triple range method<sup>55,56</sup> ( $R_{\text{CUTP}} = 0.8$  nm,  $R_{\text{CUTL}} = R_{\text{rf}} = 1.4$  nm,  $\epsilon_{\text{rf}} = 61$ ,  $\kappa = 0$ ). The pairlist was updated every fifth step. The leapfrog algorithm was used to integrate Newton’s equations of motion ( $\Delta t = 0.002$  ps).

In order to test the efficiency of the automatic parameter update using eqs 13 and 14, we chose initial values for the parameters of the reference state Hamiltonian which were far from the optimal ones. The five initial energy offsets were chosen to be  $E_i^R = \{0, 50, 100, 150, 200\}$  kJ/mol and the initial smoothness parameter was  $s = 1$ . Note that after each optimization of parameters the energy offsets were made relative to  $E_1^R$ . This has no influence on the trajectories or the calculated free energies as only differences of energy offset parameters occur in eq 12.

We tested six different schemes to update the parameters of the reference state Hamiltonian. These schemes, which are listed below, differ in three main points:

• Whether eqs 13 and 14 are iterated until convergence in the new parameters is obtained (denoted as “reweighting” in the listing below) or whether the calculation is stopped after one iteration (no reweighting). As eq 13 involves an ensemble average over the reference state ensemble, we calculate this average by reweighting the configurations, i.e.

$$\langle X \rangle_{R_{\text{new}}} = \langle X \exp[-\beta(V_{R_{\text{new}}} - V_R)] \rangle_R / \langle \exp[-\beta(V_{R_{\text{new}}} - V_R)] \rangle_R \quad (17)$$

Although this numerical iteration of eqs 13 and 14 has negligible cost compared to the time spent in generating the configurations of the reference state ensemble and should in principle ensure much faster convergence of the energy offset and  $s$  parameters, it might still not be useful to iterate eqs 13 and 14 at early stages of the optimization procedure where the important phase space of the end states is not well sampled yet. Therefore, we have tested both strategies.

• When the parameters should be optimized, i.e. after how much simulation time new parameters are calculated. This can be at fixed positions in time, e.g. after 150 ps,  $150 + 2 \times 150 = 450$  ps, and  $150 + 2 \times 150 + 4 \times 150 = 1050$  ps, which corresponds to an update after the 1st, 3rd, and 7th run if each run is 150 ps (denoted “update 1, 3, 7,...” below). However, this need not be the most efficient interval of parameter update. Another strategy we tested was to update once the sum of the statistical errors of the  $\Delta F_{iR}$  values calculated from the runs with the current parameters is below the sum of errors calculated from the runs that were performed with the previous set of parameters. This scheme is denoted “update when error in  $\Delta F$  is smaller than with previous parameters” below.

• Which runs should be taken into account when calculating the new parameters. If the  $E_i^R$  parameters of previous runs differ only slightly ( $< k_B T$ ) from the current ones, it would be a waste of computing effort not to take these runs also into account when calculating the new energy offsets  $E_i^R$  and smoothness parameter  $s$ . This strategy has been pursued in the last two update schemes. Combining the previously mentioned points led to the following update schemes:

1. no reweighting, update 1, 3, 7,...
2. reweighting, update 1, 3, 7,...
3. no reweighting, update when error in  $\Delta F$  is smaller than with previous parameters.
4. reweighting, update when error in  $\Delta F$  is smaller than with previous parameters.
5. reweighting, update when error in  $\Delta F$  is smaller than with previous parameters. When the  $E_i^R$  values differ less than  $k_B T$  from current parameters for the first time, take all subsequent runs into account when calculating optimized parameters. That is, compare the energy offsets  $E_i^R$  of the current run with the energy offsets of all the previous runs (starting from run 1). If all  $E_i^R$  differ less than  $k_B T$  from the current  $E_i^R$ , then take the trajectories from this previous run and from all subsequent runs into account when calculating new energy offsets and a new  $s$  parameter.
6. reweighting, update when error in  $\Delta F$  is smaller than with previous parameters or after 1, 3, 7,.... When the  $E_i^R$  differ less than  $k_B T$  from current parameters for the first time,

take all subsequent runs into account when calculating optimized parameters (see also 5). Each update scheme consisted of 127 subsequent simulations (“runs”) of 150 ps leading to a total simulation time of 19.05 ns for each of the above-mentioned schemes. When calculating  $\Delta F_{iR}$ ,  $E_i^R$ , and  $s$  values, the first 50 ps were discarded for equilibration.

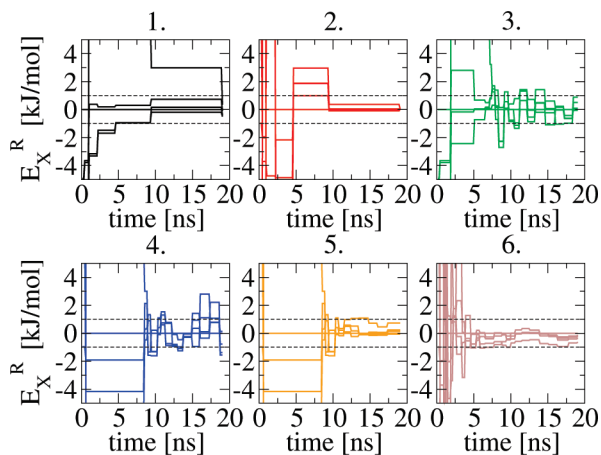
The update schemes should optimize the energy offsets to  $E_i^R = \{0, 0, 0, 0, 0\}$  kJ/mol. However, it is not clear which  $s$  parameter is the optimal one. In order to see whether the update schemes converge to the optimal  $s$  parameter, we ran 13 simulations at fixed parameters. All energy offsets were set to zero. The  $s$  values were 0.0041, 0.0082, 0.0164, 0.0328, 0.0657, 0.0821, 0.0903, 0.0985, 0.1149, 0.1313, 0.2627, 0.5254, and 1.0508. The simulation time for each of the 13 simulations was 7.5 ns. The rest of the protocol was as for the simulations with parameter update. Furthermore, a 7.5 ns non-EDS simulation of one interacting solute water molecule and four “dummy” solute water molecules in solvent water was run using the same settings as for the EDS simulations at fixed parameters.

All simulations were performed using a modified version of the GROMOS05<sup>57</sup> molecular simulation package. The method has been implemented such that only the perturbed interactions, i.e. the Hamiltonian terms that differ in the various end states, are calculated for every end state. All other interactions, i.e. in general the vast majority of the interactions, are calculated only once per time step. Moreover, the same list of atom pairs that interact with each other via nonbonded interactions is used for all end states. This results in an EDS simulation of  $N$  states being computationally much cheaper than  $N$  independent simulations.

## 4. Results and Discussion

In this study we have tested various algorithms that allow for an automatic updating of the reference Hamiltonian parameters needed for efficient EDS simulation. As we have shown in earlier work<sup>37</sup> and will show in the following, the convergence of the free energy estimates strongly depends on the chosen parameters for the reference Hamiltonian (eq 11) making it mandatory to have an efficient algorithm to automatically determine these parameters. Therefore, we have proposed several automatic update schemes (see section 3), whose ability to update the parameters efficiently is reported and discussed in this section.

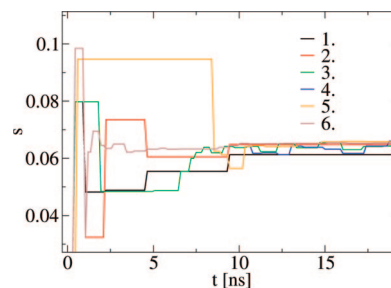
The convergence of the energy offsets is shown in Figure 2. As all states have the same free energy, the expected optimized energy offsets are  $E_i^R = \{0, 0, 0, 0, 0\}$  kJ/mol. Update scheme 6 was fastest in optimizing the energy offsets from  $E_i^R = \{0, 50, 100, 150, 200\}$  kJ/mol down to zero. Schemes 1 and 2 both calculate new parameters at fixed points during the simulation (after the 1st, 3rd, and 7th, runs). Whereas scheme 2 iterates eqs 13 and 14 until convergence, scheme 1 stops after one iteration. Although reweighting might not be reasonable at the very beginning of the simulation due to insufficient sampling, it is clear that scheme 2 which does use reweighting optimized the energy offsets faster. This shows that reweighting speeds up the convergence of the parameter estimation considerably as soon as the parameters are in a range that allows reasonable sampling



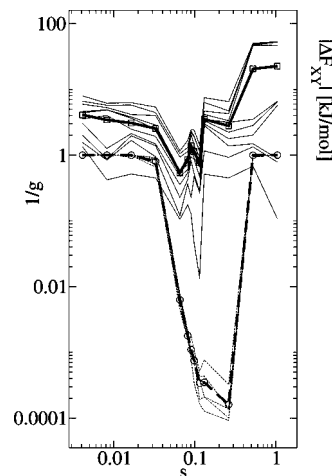
**Figure 2.** Convergence of the four ( $X = 2, \dots, 5$ ) energy offsets  $E_X^R - E_i^R$  (in kilojoules per mole) for the six parameter update schemes (see Simulation Protocols). (The dashed lines are to help guide the eye.)

of configuration space for all end states. Scheme 2 reliably optimized the energy offsets, however rather slowly. Schemes 3 and 4 were the first attempts to make convergence faster changing the criteria of when to do a calculation of new parameters. Instead of at fixed positions during simulation time the parameters were now optimized if the sum of the statistical errors of the  $\Delta F_{iR}$  values calculated from the runs with the current parameters were below the sum of errors calculated from the runs which had been performed with the previous set of parameters. Especially at the very beginning of the optimization process, this does not seem to be a good criterion as can be seen (Figure 2) from the long time it takes until the energy offsets come down to the  $\pm 1$  kJ/mol band. Furthermore, once reasonable energy offsets are found, the update criterion is fulfilled frequently leading to calculation of new energy offsets from very short pieces of trajectory. This leads to fluctuations of the energy offsets around zero. This problem has been solved in scheme 5. It differs from scheme 4 as it takes trajectories obtained with previous parameters also into account when calculating new parameters, if the previous energy offsets differ less than  $kT$ . As can be seen from Figure 2, this prevented the fluctuations around the optimal energy offset. Scheme 6 combines the findings of schemes 1–5: it uses reweighting and the information from previous runs once the energy offsets are within  $kT$  of the current energy offsets, and it updates the parameters either after runs 1, 3, 7, ... or once the statistical error in  $\Delta F_{iR}$  becomes lower. Figure 2 shows that this strategy allowed the fastest optimization of the energy offsets.

A similar discussion holds for the estimation of the smoothness parameter  $s$  (see Figure 3). Also the smoothness parameter  $s$  is optimized fastest by scheme 6. Much faster than with all other schemes the  $s$  parameter is close to the final optimized  $s$ . Whether this smoothness parameter  $s$  leads to the most accurate free energy estimate was investigated by performing 13 simulations at different  $s$  values and fixed energy offsets ( $E_i^R = \{0, 0, 0, 0, 0\}$  kJ/mol). That the optimized  $s$  parameter is the best choice for  $s$  can be seen from Figure 4. It shows all  $N(N - 1)/2 = 10$   $|\Delta F_{XY}|$  estimates (thin solid lines), where  $X$  and  $Y$  denote two end states, and



**Figure 3.** Convergence of the smoothness parameter  $s$  for the six parameter update schemes (see Simulation Protocols).



**Figure 4.** Absolute error of the ten  $\Delta F_{XY}$  estimates between the five end states ( $X, Y = 1, \dots, 5$ ) as a function of the smoothness parameter  $s$  for the 13  $s$  values mentioned in Simulation Protocols (thin solid lines). The thick solid line (with squares) denotes the mean over these 10 error estimates. The thin dashed lines indicate  $1/g$  obtained from the five  $\Delta F_{XR}$  ( $X = 1, \dots, 5$ ) estimates. Here  $g$  is the statistical inefficiency,<sup>60</sup> i.e. given a time series of  $M$  data points  $M/g$  gives the number of uncorrelated data points. The thick dashed line (with circles) indicates the mean over the five  $1/g$  estimates.

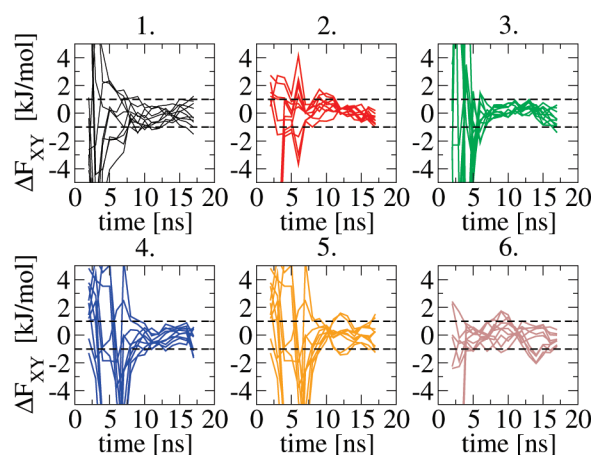
their average (thick solid line) as a function of  $s$ . As all states have the same free energy, all  $\Delta F_{XY}$  estimates should be zero and  $|\Delta F_{XY}|$  is equal to the absolute error of the free energy estimate. Minimal errors were obtained for  $s$  values between 0.06 and 0.1. The lowest error was obtained for  $s = 0.0657$  which agrees well with the final optimized  $s$  value of update scheme 6 (0.0653). Figure 4 furthermore shows that the number of uncorrelated data points  $M/g$ , where  $g$  is the statistical inefficiency,<sup>60</sup> contained within a time series of  $M$  data points strongly varies with  $s$  (dashed lines).

Figure 5 shows free energy differences  $\Delta F_{XY}$  between the end states which were calculated from a 4 ns moving window (moving in 1 ns steps). Again update scheme 6 performed best as the calculated  $\Delta F_{XY}$  values lie quickly within 1 kJ/mol from the correct result (0 kJ/mol). The calculated free energy estimates obtained from the six simulations with parameter updates and the 13 simulations at fixed parameter choices are shown in Tables 1 and 2. With an appropriate equilibration time, all six update schemes give good estimates for the free energy differences. For the simulations at fixed  $s$  values, reasonable free energy estimates were only obtained

**Table 1.** Free Energy Differences and Statistical Uncertainties<sup>37,60</sup> (in kilojoules per mole) between the Five (1–5) End States and the Reference (*R*) State Obtained from Six Simulations Performed with Six Different Update Schemes (See *Simulation Protocols*)<sup>a</sup>

	1	2	3	4	5	6
$\Delta F_{1R}$	7.8 ± 0.3	5.6 ± 0.3	5.8 ± 0.3	6.5 ± 0.3	5.4 ± 0.3	5.7 ± 0.3
$\Delta F_{2R}$	7.5 ± 0.2	5.7 ± 0.3	6.0 ± 0.3	6.8 ± 0.3	6.1 ± 0.3	5.3 ± 0.3
$\Delta F_{3R}$	7.2 ± 0.3	5.7 ± 0.3	5.9 ± 0.3	6.2 ± 0.3	6.7 ± 0.3	5.3 ± 0.3
$\Delta F_{4R}$	7.3 ± 0.3	5.8 ± 0.3	5.7 ± 0.3	6.5 ± 0.3	5.7 ± 0.3	5.5 ± 0.3
$\Delta F_{5R}$	7.5 ± 0.2	5.7 ± 0.3	5.9 ± 0.3	6.2 ± 0.3	5.9 ± 0.3	5.3 ± 0.3
$\Delta F_{21}$	−0.3 ± 0.4	0.1 ± 0.4	0.2 ± 0.4	0.3 ± 0.5	0.7 ± 0.5	−0.4 ± 0.4
$\Delta F_{31}$	−0.6 ± 0.5	0.1 ± 0.5	0.1 ± 0.5	−0.3 ± 0.5	1.2 ± 0.5	−0.4 ± 0.4
$\Delta F_{41}$	−0.5 ± 0.5	0.1 ± 0.5	−0.1 ± 0.4	−0.1 ± 0.4	0.2 ± 0.5	−0.2 ± 0.4
$\Delta F_{51}$	−0.3 ± 0.4	0.0 ± 0.4	0.1 ± 0.4	−0.4 ± 0.5	0.4 ± 0.4	−0.3 ± 0.4
$\Delta F_{32}$	−0.3 ± 0.4	0.0 ± 0.5	0.0 ± 0.5	−0.6 ± 0.5	0.6 ± 0.4	0.0 ± 0.4
$\Delta F_{42}$	−0.2 ± 0.4	0.0 ± 0.4	−0.3 ± 0.4	−0.4 ± 0.5	−0.5 ± 0.5	0.2 ± 0.4
$\Delta F_{52}$	0.0 ± 0.4	−0.1 ± 0.4	−0.1 ± 0.4	−0.6 ± 0.5	−0.3 ± 0.4	0.1 ± 0.5
$\Delta F_{43}$	0.1 ± 0.4	0.0 ± 0.5	−0.2 ± 0.4	0.2 ± 0.5	−1.0 ± 0.5	0.2 ± 0.4
$\Delta F_{53}$	0.3 ± 0.4	−0.1 ± 0.5	0.0 ± 0.4	−0.1 ± 0.5	−0.8 ± 0.5	0.1 ± 0.4
$\Delta F_{54}$	0.3 ± 0.4	−0.1 ± 0.4	0.2 ± 0.4	−0.3 ± 0.5	0.2 ± 0.5	−0.2 ± 0.5

<sup>a</sup> The averaging is performed over the last 9.6 ns of the 19.05 ns simulation time.



**Figure 5.** Ten free energy differences  $\Delta F_{XY}$  (in kilojoules per mole) between the end states ( $X, Y = 1, \dots, 5$ ) calculated from a 4 ns moving window (moving in 1 ns steps) for the six parameter update schemes (see *Simulation Protocols*). (The dashed lines are to help guide the eye.)

for a small range around the optimal  $s$ , as could already be seen from Figure 4.

The high sensitivity of the free energy estimate to the  $s$  parameter can be further explained using energy difference distributions. Figure 6 shows the energy difference distributions (eq 7)  $\rho_X(\Delta V; \Delta V_{XY})$ ,  $\rho_Y(\Delta V; \Delta V_{XY})$ , and  $\rho_R(\Delta V; \Delta V_{XY})$  for all ten  $X$ – $Y$  pairs and all 13  $s$  values. Starting at  $s = 0.0657$ , we see that the  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  are well formed and that they match the distributions obtained from an independent, non-EDS simulation. Recall that the free energy difference between two states  $X$  and  $Y$  is the energy difference where the  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  intersect (see section 2, eq 8). From Figure 6, we see that the highest probability of  $\rho_R(\Delta V; \Delta V_{XY})$  is where  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  distributions intersect, i.e. during the reference state simulation the sampling is focused on the important crossing region. The importance of focusing the sampling on the relevant energy difference space has recently been pointed out by Min and Yang.<sup>58</sup> Their approach was to improve sampling by adding a biasing potential on  $\Delta V$  in order to increase the sampled  $\Delta V$  range. A different strategy

was pursued by Wu,<sup>59</sup> who uses a modified Metropolis acceptance rule to constrain the sampling to a given  $\Delta V$  range and obtains the energy-difference distribution from overlapping umbrella windows. In EDS, sampling is automatically focused only on the crucial  $\Delta V$  range once the parameters of the reference state Hamiltonian have been optimized.

The effect of suboptimal  $s$  values can be studied in the other plots of Figure 6. Increasing the  $s$  value leads to broader sampling over the  $\Delta V$  range, leading to worse sampling of the crossing region and longer convergence times (see also Table 2). Starting with  $s = 0.1313$  not all five states are sampled equally anymore—the ten  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  distributions, respectively, start to differ. For  $s = 0.5254$  and higher  $s$  values, only the state from which the simulation started is sampled and no meaningful free energy estimates can be obtained. A decrease of  $s$  very quickly distorts the potential energy surface of the reference state Hamiltonian such that the important regions of the configuration spaces of the end states are not longer minima on this surface. This leads to very narrow  $\rho_R(\Delta V; \Delta V_{XY})$  distributions around  $\Delta V = 0$  kJ/mol and badly formed  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  distributions. This suboptimal sampling of the end states is also reflected in the rather inaccurate free energy estimates for these  $s$  values (see Table 2).

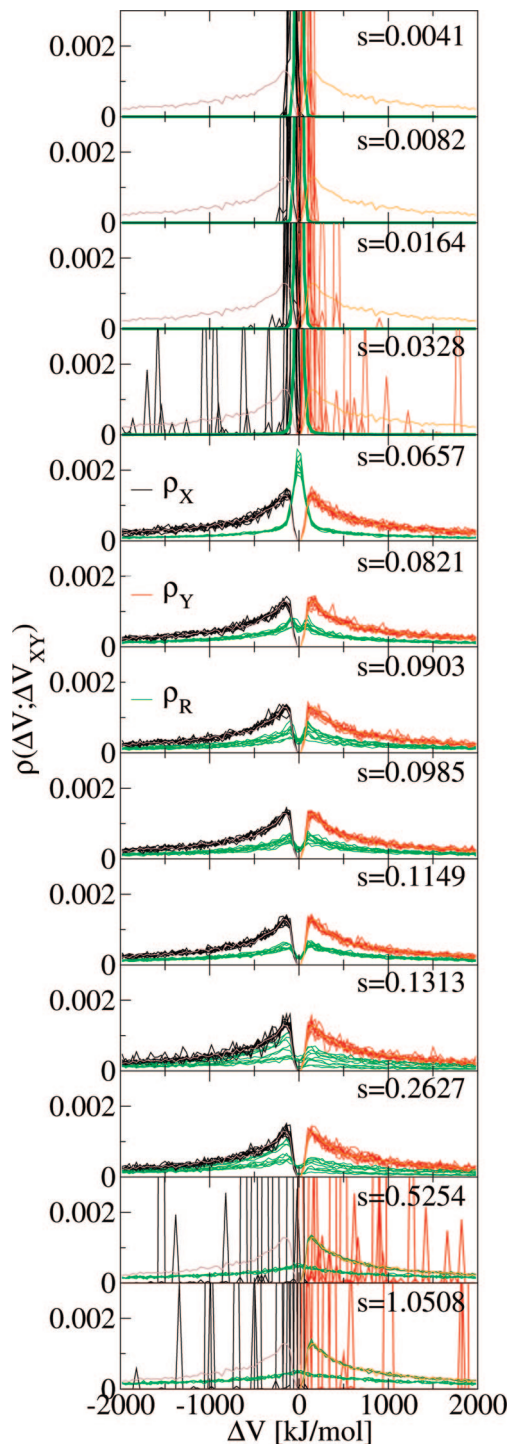
Figure 7 shows in another way how the accuracy of the free energy estimates depends on the chosen  $s$  parameter. It shows  $\ln(\rho_X(\Delta V; \Delta V_{XY})/\rho_Y(\Delta V; \Delta V_{XY}))$  as a function of  $\Delta V$  for the 13  $s$  values. From eq 9, we see that the ordinate intercept is  $-\beta\Delta F_{XY}$ . For  $s = 0.0657$ , the lines nicely cross the ordinate at  $-\beta\Delta F_{XY} = 0$ . Moving to smaller or larger  $s$  values, the variance increases. For  $s = 0.5254$  and larger, no reasonable free energy estimates can be obtained, because only a few  $\Delta V$  values near the region where the end state energy difference distributions intersect ( $\Delta V = 0$ ) are sampled in the trajectories. The average temperatures calculated from the slopes of a linear regression of the data shown in Figure 7 (see eq 9) are 330, 319, 302, 302, 303, 303, 303, 303, 303, 303, 303, 301, and 301 K for the 13  $s$  values. For  $s = 0.0657$ – $0.2627$ , where  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  are well formed (see Figure 6), the extracted



**Table 2.** Free Energy Differences and Statistical Uncertainties<sup>37,60</sup> (in kilojoules per mole) between the Five (1–5) End States and the Reference (F) State Obtained from 13 Simulations Performed with Different  $s$  Parameters<sup>a</sup>

$s =$	0.0041	0.0082	0.0164	0.0328	0.0657	0.0821	0.0903	0.0985	0.1149	0.1313	0.2627	0.5254	1.0508
$\Delta F_{1R}$	937.4	450.9 ± 1.2	205.7 ± 1.3	63.7 ± 1.6	5.9 ± 0.3	4.1 ± 0.4	4.0 ± 0.6	3.8 ± 0.6	5.0 ± 1.0	1.4 ± 0.7	1.3 ± 0.8	46.8 ± 2.0	53.2 ± 1.6
$\Delta F_{2R}$	942.0	446.9 ± 1.4	202.1 ± 1.6	62.2 ± 2.3	5.7 ± 0.4	3.6 ± 0.4	5.5 ± 0.8	4.6 ± 0.7	3.5 ± 0.8	8.9 ± 1.9	5.6 ± 2.3	50.3 ± 1.7	52.4 ± 1.1
$\Delta F_{3R}$	940.4	452.1 ± 1.3	200.0 ± 2.4	66.2 ± 1.1	4.8 ± 0.3	4.2 ± 0.5	3.3 ± 0.5	5.8 ± 0.9	3.8 ± 0.9	4.2 ± 1.3	4.7 ± 2.1	48.8 ± 2.4	46.7 ± 1.8
$\Delta F_{4R}$	944.0	446.0 ± 2.3	205.9 ± 0.9	61.8 ± 2.0	5.4 ± 0.3	5.4 ± 0.5	5.7 ± 0.6	3.3 ± 0.5	4.5 ± 1.6	6.8 ± 1.7	7.8 ± 2.4	47.5 ± 2.3	53.3 ± 1.3
$\Delta F_{5R}$	936.0	451.7 ± 1.2	203.8 ± 1.3	60.8 ± 2.1	5.9 ± 0.3	4.7 ± 0.5	3.4 ± 0.5	3.8 ± 0.9	3.8 ± 1.2	4.8 ± 1.5	5.5 ± 2.2	0.0 ± 0.0	0.0 ± 0.0
$\Delta F_{21}$	4.6	-4.0 ± 2.3	-3.6 ± 2.0	-1.5 ± 3.8	-0.2 ± 0.6	-0.5 ± 0.7	1.4 ± 1.1	0.8 ± 1.0	-1.5 ± 1.5	7.5 ± 2.3	4.3 ± 2.7	3.5 ± 3.6	-0.8 ± 2.0
$\Delta F_{31}$	3.1	1.3 ± 1.8	-5.8 ± 2.7	2.5 ± 1.9	-1.1 ± 0.5	0.2 ± 0.7	-0.7 ± 1.0	1.9 ± 1.2	-1.1 ± 1.4	2.8 ± 1.8	3.3 ± 2.5	2.0 ± 4.3	-6.5 ± 3.4
$\Delta F_{41}$	6.6	-4.9 ± 2.7	0.1 ± 1.6	-1.9 ± 2.5	-0.5 ± 0.5	1.4 ± 0.7	1.7 ± 0.9	-0.5 ± 1.0	-0.5 ± 2.2	5.4 ± 2.2	6.5 ± 2.7	0.7 ± 4.3	0.1 ± 2.1
$\Delta F_{51}$	-1.3	0.9 ± 1.9	-2.0 ± 1.8	-2.9 ± 3.7	0.1 ± 0.5	0.6 ± 0.8	-0.6 ± 0.9	-0.1 ± 1.4	-1.2 ± 1.6	3.3 ± 1.9	4.2 ± 3.1	-46.8 ± 2.0	-53.2 ± 1.6
$\Delta F_{32}$	-1.5	5.3 ± 2.0	-2.1 ± 3.0	4.0 ± 2.5	-0.9 ± 0.6	0.6 ± 0.7	-2.2 ± 1.0	1.2 ± 1.3	0.4 ± 1.3	-4.7 ± 2.4	-0.9 ± 3.9	-1.5 ± 4.1	-5.7 ± 2.1
$\Delta F_{42}$	2.0	-0.9 ± 3.7	3.8 ± 1.9	-0.4 ± 3.4	-0.3 ± 0.5	1.8 ± 0.8	0.3 ± 1.0	-1.2 ± 1.0	1.1 ± 1.9	-2.1 ± 3.6	2.2 ± 4.7	-2.8 ± 4.0	0.9 ± 2.4
$\Delta F_{52}$	-5.9	4.9 ± 2.5	1.7 ± 2.0	-1.4 ± 3.1	0.2 ± 0.6	1.1 ± 0.7	-2.0 ± 1.1	-0.8 ± 1.3	0.3 ± 1.8	-4.2 ± 3.5	-0.1 ± 3.2	-50.3 ± 1.7	-62.4 ± 1.1
$\Delta F_{43}$	3.5	-6.2 ± 2.7	5.9 ± 2.6	-4.4 ± 2.4	0.6 ± 0.5	1.2 ± 0.8	2.4 ± 1.0	-2.4 ± 1.1	0.7 ± 2.3	2.6 ± 3.0	3.2 ± 4.2	-1.3 ± 4.7	6.6 ± 2.5
$\Delta F_{53}$	-4.4	-0.4 ± 1.9	3.8 ± 2.7	-5.4 ± 2.4	1.1 ± 0.5	0.5 ± 0.7	0.1 ± 0.9	-2.0 ± 1.3	-0.1 ± 1.8	0.5 ± 2.9	0.9 ± 4.1	-48.8 ± 2.4	-46.7 ± 1.8
$\Delta F_{54}$	-7.9	5.7 ± 2.6	-2.1 ± 1.8	-0.9 ± 3.7	0.5 ± 0.5	-0.7 ± 0.8	-2.3 ± 0.9	0.4 ± 1.3	-0.8 ± 2.4	-2.1 ± 2.4	-2.3 ± 4.7	-47.5 ± 2.3	-53.3 ± 1.3

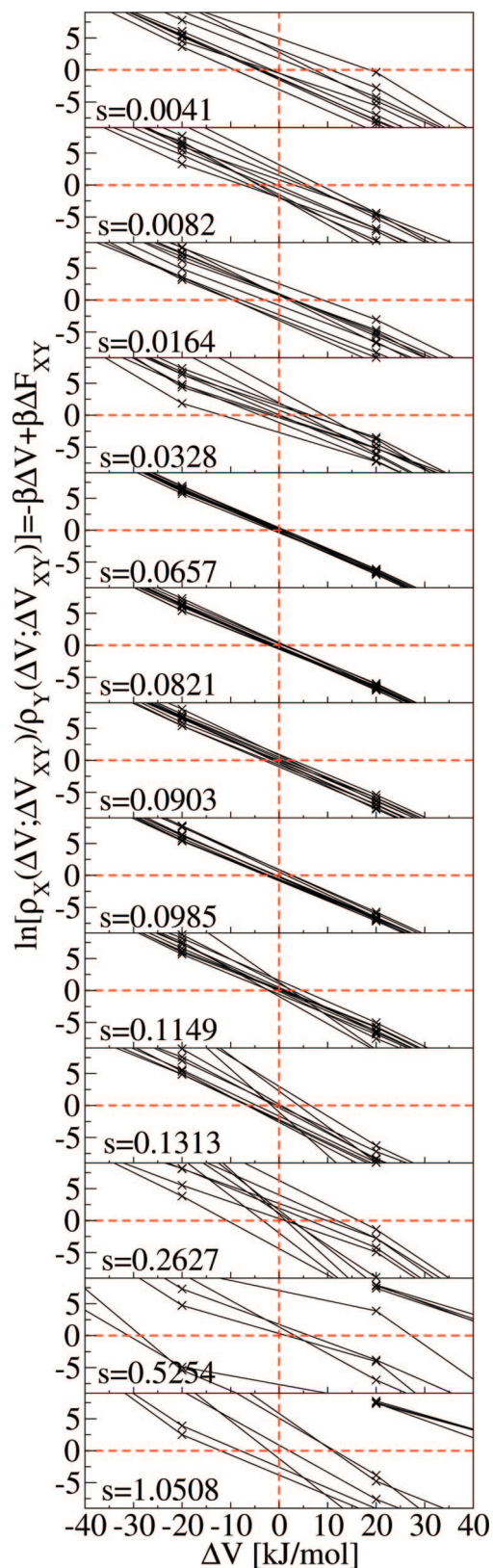
<sup>a</sup> For  $s = 0.0041$ , the variance is zero and no error estimate could be calculated. Simulation time: 7.5 ns (minus 150 ps for equilibration).



**Figure 6.** Energy difference distributions (eq 7)  $\rho_X(\Delta V; \Delta V_{XY})$  (black),  $\rho_Y(\Delta V; \Delta V_{XY})$  (red), and  $\rho_R(\Delta V; \Delta V_{XY})$  (green) for all 10  $X$ - $Y$  pairs and all 13  $s$  values (see Simulation Protocols). The distributions obtained from an independent non-EDS simulation are shown in brown ( $\rho_1(\Delta V; \Delta V_{12})$ ) and orange ( $\rho_1(\Delta V; \Delta V_{21})$ ).

temperatures agree with the average temperature which was of 303 K for all 13 simulations.

It should be stressed that for this test system  $\Delta V = 0$  kJ/mol can be observed for three different kind of configurations. First, for low  $s$  values ( $s = 0.0041$ – $0.0328$ ) regions of configuration space important to the reference state are sampled which are equally unfavorable for the end states,



**Figure 7.** Linearized representation (eq 9) of energy difference distributions for all 10  $X$ – $Y$  pairs:  $\ln(\rho_X(\Delta V; \Delta V_{XY})/\rho_Y(\Delta V; \Delta V_{XY}))$  as a function of  $\Delta V$  for all 13  $s$  values (see Simulation Protocols). (The dashed lines are to help guide the eye.)

leading to energy differences  $\Delta V_{XY}$  of zero ( $V_X \approx V_Y \gg V_R$ ). This can be observed in Figure 6 for the lowest four  $s$  values. The  $\rho_R(\Delta V; \Delta V_{XY})$  distribution is centered around zero, yet the important phase space of the end states is not well

sampled (badly formed  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  distributions) which is a necessary (but not sufficient) condition for reasonable free energy difference estimates. If parts of the  $\rho_X(\Delta V; \Delta V_{XY})$  and  $\rho_Y(\Delta V; \Delta V_{XY})$  distributions are not sampled, this will shift the intersection point and therefore the free energy estimate unless the neglected areas are of the same size and do not lie in the intersection region. Second, as in the current simulation setup the five solute water molecules are free to move, a zero energy difference can also be observed if two water molecules happen to occupy the same position ( $V_X = V_Y = V_R$ ). If this happened for longer periods during the simulation, the chosen test system would be trivial. It does indeed happen occasionally during the simulation that two water molecules occupy the same position, they however separate again after some picoseconds. In order to check whether these occasional encounters influence the presented results, we have recalculated the free energy differences using only frames where the distance between any two of the five solute water molecules is larger than 0.5 nm. The difference in the calculated  $\Delta F_{XY}$  values was found to be smaller than the statistical uncertainty of these values. The third type of configuration where a zero energy difference occurs is when during the reference state simulation a transition from the important configuration space of one end state to that of another end state occurs ( $V_X = V_Y > V_R$ ). For  $s = 0.0657$ , many of these transitions occur, which explains the focusing of  $\rho_R(\Delta V; \Delta V_{XY})$  around  $\Delta V = 0$ . With increasing  $s$  the number of these transitions decreases which is reflected in the  $\rho_R(\Delta V; \Delta V_{XY})$  distributions (i.e., the density around  $\Delta V = 0$  decreases), in the decrease of the number of uncorrelated data points (see Figure 4,  $s = 0.0657$ – $0.2627$ ), and consequently in the statistical uncertainties of the free energy estimates (see Table 2,  $s = 0.0657$ – $0.2627$ ). It is due to this decrease in the number of transitions (down to zero for  $s = 0.5254$ – $1.0508$ ) that the nominally optimal smoothness parameter  $s = 1$  is in practice not optimal. For the two smoothness parameter values  $s = 0.06568$  and  $s = 0.1149$  we have recalculated the free energy differences using the same number of uncorrelated data points, i.e. for the higher  $s$  value the complete time series was used in the analysis and for the lower  $s$  value only  $M_{s=0.06568} = M_{s=0.1149} g_{s=0.06568}/g_{s=0.1149}$  data points were used. Interestingly but not surprisingly, the obtained mean absolute error of the 10  $\Delta F_{XY}$  estimates is somewhat higher for  $s = 0.06568$  ( $\approx 2$  kJ/mol) than for  $s = 0.1149$  ( $\approx 0.8$  kJ/mol). This is in line with the expected nominally optimal smoothness parameter being  $s = 1$ . In practice, we seek the lowest  $s$  value that still ensures sampling of the complete end state energy difference distributions as this will ensure a precise and accurate free energy estimate, where the precision depends on sufficient transitions between regions of configuration space important to the end states, and the accuracy depends on complete sampling of these parts of configuration space. The presented schemes for the iterative update of the reference Hamiltonian parameters have been shown to be able to find this optimal value for the smoothness parameter.

## 5. Conclusions

We have successfully tested different schemes that allow for an automatic updating of the reference Hamiltonian parameters in enveloping distribution sampling (EDS). As a test system, we chose liquid water in which particular molecules were created and deleted. We selected five water molecules to define five states. Each state consisted of one interacting solute water molecule and four noninteracting water molecules in a box of solvent water. Starting from a set of reference Hamiltonian parameters which was far from optimal, all schemes optimized the parameters to the expected energy offset values (0 kJ/mol) and to a unique smoothness parameter. One scheme (scheme 6) was fastest in optimizing and will be used in further applications. The simulations we performed at fixed smoothness parameter  $s$  showed that the optimized  $s$  parameter gives rise to the most accurate free energy estimate (see Figure 4). The automatic update scheme is a big step toward application of EDS by nonexpert users. No parameters have to be chosen at the beginning of the simulation. The only input to the initial reference Hamiltonian are the Hamiltonians of the various end states.

Future work will imply testing of the method on flexible molecules where an optimal balance between sampling within the important configuration space of one end state and transitions between parts of configuration space important to different end states might pose a problem. A further challenge are systems where the important configuration space of some end states lie close together and those of others are far apart. For such systems a single smoothness parameter  $s$  approach might break down. Therefore, we are currently testing also multiple  $s$  approaches.

**Acknowledgment.** The authors would like to thank Chris Oostenbrink for reading the manuscript and for helpful discussions. Financial support by the National Center of Competence in Research (NCCR) Structural Biology and by Grant No. 200021-109227 of the Swiss National Science Foundation (SNSF) is gratefully acknowledged.

## References

- (1) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
- (2) van Gunsteren, W. F. Methods for calculation of free energies and binding constants: Successes and problems. In *Computer simulation of biomolecular systems: theoretical and experimental applications*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM Science publishers B. V.: Leiden, 1989.
- (3) Reynolds, C. A.; King, P. M.; Richards, W. G. *Mol. Phys.* **1992**, *76*, 251–275.
- (4) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (5) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. Computation of free energy in practice: Choice of approximations and accuracy limiting factors. In *Computer simulation of biomolecular systems: theoretical and experimental applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; ESCOM Science publishers B. V.: Leiden, 1993; Vol. 2.
- (6) Kofke, D. A.; Cummings, P. T. *Mol. Phys.* **1997**, *92*, 973–996.
- (7) Gelman, A.; Meng, X. L. *Statist. Sci.* **1998**, *13*, 163–185.
- (8) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211–243.
- (9) Chipot, C.; Pearlman, D. A. *Mol. Simul.* **2002**, *28*, 1–12.
- (10) van Gunsteren, W. F.; Daura, X.; Mark, A. E. *Helv. Chim. Acta* **2002**, *85*, 3113–3129.
- (11) Brandsdal, B. O.; Osterberg, F.; Almlof, M.; Feierberg, I.; Luzhkov, V. B.; Aqvist, J. *Adv. Protein Chem.* **2003**, *66*, 123–158.
- (12) Kofke, D. A. *Fluid Phase Equilib.* **2005**, *228*, 41–48.
- (13) Rödinger, T.; Pomes, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–170.
- (14) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181–186.
- (15) Chipot, C.; Pohorille, A. *Free energy calculations: Theory and applications in chemistry and biology*; Springer: Berlin, 2007.
- (16) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (17) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. *Annu. Rep. Comput. Chem* **2007**, *3*, 41–59.
- (18) Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.
- (19) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (20) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (21) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
- (22) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- (23) Squire, D. R.; Hoover, W. G. *J. Chem. Phys.* **1969**, *50*, 701–706.
- (24) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (25) Lu, N. D.; Wu, D.; Woolf, T. B.; Kofke, D. A. *Phys. Rev. E* **2004**, *69*, 057702.
- (26) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (27) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058–9067.
- (28) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. *J. Chem. Theory Comput.* **2006**, *2*, 217–228.
- (29) Pitera, J.; Kollman, P. *J. Am. Chem. Soc.* **1998**, *120*, 7557–7567.
- (30) Kong, X.; Brooks, C. L. *J. Chem. Phys.* **1996**, *6*, 2414–2423.
- (31) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (32) Jarzynski, C. *Phys. Rev. E* **1997**, *56*, 5018–5035.
- (33) Crooks, G. E. *Phys. Rev. E* **2000**, *61*, 2361–2366.
- (34) Jarzynski, C. *Phys. Rev. E* **2006**, *73*, 046105.
- (35) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 054103.
- (36) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2007**, *126*, 184110.
- (37) Christ, C. D.; van Gunsteren, W. F. *J. Chem. Phys.* **2008**, *128*, 174112.
- (38) Srinivasan, R. *Importance Sampling: Applications in Communications and Detection*; Springer: Berlin, Heidelberg, NY, 2002.

- (39) Liu, H.; Mark, A. E.; van Gunsteren, W. F. *J. Phys. Chem.* **1996**, *100*, 9485–9494.
- (40) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750–6754.
- (41) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.
- (42) Han, K. K. *Phys. Lett. A* **1992**, *165*, 28–32.
- (43) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776–1783.
- (44) Escobedo, F. A.; de Pablo, J. J. *J. Chem. Phys.* **1995**, *103*, 2703–2710.
- (45) Smith, G. R.; Bruce, A. D. *J. Phys. A: Math. Gen.* **1995**, *28*, 6623–6643.
- (46) Engkvist, O.; Karlstrom, G. *Chem. Phys.* **1996**, *213*, 63–76.
- (47) Han, K. K. *Phys. Rev. E* **1996**, *54*, 6906–6910.
- (48) Chen, Y. G.; Hummer, G. *J. Am. Chem. Soc.* **2007**, *129*, 2414–2415.
- (49) Shing, K. S.; Gubbins, K. E. *Mol. Phys.* **1982**, *46*, 1109–1128.
- (50) Powles, J. G.; Evans, W. A. B.; Quirke, N. *Mol. Phys.* **1982**, *46*, 1347–1370.
- (51) Jacucci, G.; Quirke, N. *Lect. Notes Phys.* **1982**, *166*, 38–57.
- (52) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, 1981; pp 331–342.
- (53) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (54) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 327–341.
- (55) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular simulation: The GROMOS96 manual and user guide*; Vdf Hochschulverlag AG an der ETH Zürich; Zürich, 1996.
- (56) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (57) 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*, 1720–1751.
- (58) Min, D. H.; Yang, W. *J. Chem. Phys.* **2008**, *128*, 191102.
- (59) Wu, D. *J. Chem. Phys.* **2008**, *128*, 224105.
- (60) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

CT800424V

# JCTC

Journal of Chemical Theory and Computation

## Fully Automated Incremental Evaluation of MP2 and CCSD(T) Energies: Application to Water Clusters

Joachim Friedrich\* and Michael Dolg

*Institute for Theoretical Chemistry, University of Cologne, Greinstrasse 4,  
50939 Cologne, Germany*

Received January 1, 2008

**Abstract:** A fully automated implementation of the incremental scheme for CCSD energies has been extended to treat MP2 and CCSD(T) energies. It is shown in applications on water clusters that the error of the incremental expansion for the total energy is below 1 kcal/mol already at second or third order. It is demonstrated that the approach saves CPU time, RAM, and disk space. Finally it is shown that the calculations can be run in parallel on up to 50 CPUs, without significant loss of computer time.

### 1. Introduction

It is well-known that a quantitative description of the electronic structure in molecules is usually not possible with the Hartree–Fock (HF) method. One way to achieve higher accuracy in electronic structure calculations is to improve the wave function, which can be routinely done by many-body perturbation theory (MBPT), configuration interaction theory (CI), or coupled-cluster theory (CC). The major drawback of these approaches is their strong dependence of the computational effort on the size of the one-particle basis set. This means that these approaches depend heavily on the system size too, if atom-centered basis functions are used.

Since for large systems the canonical HF orbitals are not necessarily the best choice for a PT, CI, or CC expansion of the wave function, many groups use a local orbital basis instead to include electron correlation. This allows one to screen out insignificant contributions to the energy, and therefore the computational cost is reduced.<sup>1–19</sup> Conceptually different approaches divide the total system into parts and then perform a perturbation expansion to obtain the total correlation energy.<sup>20–23</sup> An approach designed in this way is the incremental scheme of Stoll.<sup>24–26</sup> It is based on the Bethe–Goldstone expansion, which was introduced to quantum chemistry by Nesbet<sup>27–29</sup> more than 40 years ago. The incremental scheme was successfully applied during the past 15 years to various periodic systems<sup>30–34</sup> and mole-

cules.<sup>35–39</sup> Besides the treatment of closed-shell systems, extensions to open-shell cases have been developed too.<sup>40–42</sup>

Recently we proposed a fully automated implementation of the incremental scheme for CCSD energies,<sup>36</sup> implemented an automatic distance screening,<sup>37</sup> extended the approach for the usage of symmetry<sup>39</sup> and to the RCCSD method for open-shell calculations.<sup>42</sup> To account for the local character of the core electrons, we introduced an efficient scheme to treat the core and core–valence correlation.<sup>43</sup> In this work we now extend our fully automated implementation to second-order Møller–Plesset perturbation theory (MP2) and to the coupled-cluster ansatz with singles, doubles, and perturbative triples excitations CCSD(T). Furthermore, as recently proposed, we use a second basis set to describe the environment of the orbital domains in a computationally efficient manner.<sup>44</sup> We note that the idea of multiple basis sets is not entirely new, since Jurgens-Lutovsky and Almlöf made use of a reduced basis for the occupied space in MP2 calculations already in 1991<sup>45</sup> and Klopper et al. made use of a reduced basis for the treatment of the triples in CCSD(T) calculations in 1997.<sup>46</sup> However, we exploit this idea in the framework of the incremental scheme and check how it performs with respect to the overall accuracy and the CPU time requirements.

### 2. Theory

**2.1. Incremental Scheme.** In an incremental calculation we divide the total system into small domains consisting of groups of localized occupied orbitals according to the procedure outlined in refs 36–38. Then we calculate the

\* To whom correspondence should be addressed. Tel.: (+49) (0)221-470-6886. Fax: (+49) (0)221-470-6896. E-mail: joachim\_friedrich@gmx.de.

correlation energies for these domains. To include the nonadditivity corrections, we also calculate correction energies of pairs and triples, etc., of domains, until we reach the desired accuracy. The correlation energy is then computed according to

$$E_{\text{corr}} = \sum_i \Delta\epsilon_i + \frac{1}{2!} \sum_{ij} \Delta\epsilon_{ij} + \frac{1}{3!} \sum_{ijk} \Delta\epsilon_{ijk} + \dots \quad (1)$$

$$\Delta\epsilon_i = \epsilon_i \quad \Delta\epsilon_{ij} = \epsilon_{ij} - \Delta\epsilon_i - \Delta\epsilon_j$$

where  $\epsilon_i$  is the correlation energy of the subsystem  $i$  and  $\epsilon_{ij}$  the correlation energy of the subsystem  $i$  and  $j$  together. For the convenient treatment of higher order terms we use a notation based on simple set theory.<sup>36</sup> Now eq 1 reads

$$E_{\text{corr}} = \sum_{X \in \mathbf{P}(\mathcal{D}) \wedge |\mathcal{X}| \leq \mathbf{O}} \Delta\epsilon_X \quad (2)$$

$\mathcal{D}$  is the set of domains,  $\mathbf{P}(\mathcal{D})$  stands for the power set of the set of the domains and  $\mathbf{O}$  denotes the order of the expansion. The summation index in eq 2 runs over all increments up to the order  $\mathbf{O}$  (see refs 36 and 37 for details). When  $\epsilon_X$  represents the correlation energy of the unified subsystems of the general correlation energy increment,  $\Delta\epsilon_X$  is given as

$$\Delta\epsilon_X = \epsilon_X - \sum_{Y \in \mathbf{P}(\mathcal{X}) \wedge |Y| < |X|} \Delta\epsilon_Y \quad (3)$$

here the summation index  $Y$  runs over the power set of  $X$ .

**2.2. Introduction of a Domain-Specific Basis Set.** To reduce the computational cost, it is necessary to reduce the virtual space of the domains. As an alternative to an explicit elimination of selected orbitals from the virtual space spanned by the full basis set, we introduce a small basis set to describe the environment of an  $n$ -site domain and use the original larger basis set only in the main part of the domain.<sup>44</sup> In the current version of the code we use the distance to the center of charge of a localized orbital ( $t_{\text{main}}$ ) to determine which atoms have to be treated with the full basis set. Since a domain represents a set of occupied localized orbitals, we combine the main regions of all orbitals of the domain, to obtain the set of atoms which are treated with the full basis set.

**2.3. Localized Orbitals and Perturbation Theory.** The computationally cheap canonical representation of the MP2 energy as well as the canonical representation of the perturbative triples correction in CCSD(T) are not invariant with respect to unitary transformations within the occupied space. To account for this within the canonical representation of the theory, we construct the transformed Fock matrix  $\tilde{\mathbf{F}}(\mathbf{C}_L)$  of the local basis

$$\tilde{\mathbf{F}}(\mathbf{C}_L) = (\mathbf{S}^{-1/2})^\dagger \mathbf{F}(\mathbf{C}_L) \mathbf{S}^{-1/2} \quad (4)$$

where  $\mathbf{S}$  is the overlap matrix,  $\mathbf{S}^{-1/2}$  is constructed to orthogonalize the basis symmetrically, and  $\mathbf{C}_L$  is the MO coefficient matrix in the local basis. Next we build the matrix  $\epsilon$

$$\epsilon = \mathbf{C}'_L \tilde{\mathbf{F}}(\mathbf{C}_L) \mathbf{C}'_L \quad (5)$$

with  $\mathbf{C}'_L = \mathbf{S}^{1/2} \mathbf{C}_L$ . In the canonical basis,  $\epsilon$  is diagonal and it contains the orbital energies, but in the local basis it is not diagonal. Now we classify the occupied orbitals into four classes, the frozen core orbitals, which are not correlated in all calculations, the environment orbitals, which are frozen in a specific calculation, the domain orbitals, which have to be correlated in this specific domain and the virtual orbitals, which are unoccupied in the HF reference. Furthermore we do not distinguish between core and environment orbitals, since they are treated equally in the subsequent steps. According to these criteria, we classify the blocks of the  $\epsilon$  matrix into nine blocks:

$$\begin{array}{c} \text{core} \\ \text{domain} \\ \text{virtual} \end{array} \begin{array}{ccc} \text{core} & \text{domain} & \text{virtual} \\ \left( \begin{array}{ccc} \epsilon_{cc} & \epsilon_{cd} & \epsilon_{cv} \\ \epsilon_{dc} & \epsilon_{dd} & \epsilon_{dv} \\ \epsilon_{vc} & \epsilon_{vd} & \epsilon_{vv} \end{array} \right) = \epsilon = \mathbf{C}'^\dagger \tilde{\mathbf{F}}(\mathbf{C}_L) \mathbf{C}' \quad (6) \end{array}$$

with (frozen core + environment) = core. In our approach we diagonalize  $\epsilon$  in the subspace of the domain. In other words, we diagonalize the  $\epsilon_{dd}$  block of  $\epsilon$ . The unitary matrix  $\mathbf{U}$  which is necessary to transform the MOs has the form

$$\begin{array}{c} \text{core} \\ \text{domain} \\ \text{virtual} \end{array} \begin{array}{ccc} \text{core} & \text{domain} & \text{virtual} \\ \left( \begin{array}{ccc} \mathbf{I} & 0 & 0 \\ 0 & \tilde{\mathbf{U}} & 0 \\ 0 & 0 & \mathbf{I} \end{array} \right) = \mathbf{U} \quad (7) \end{array}$$

where  $\tilde{\mathbf{U}}$  is the matrix which diagonalizes  $\epsilon_{dd}$ . The corresponding MOs are then obtained by

$$\tilde{\mathbf{C}}_L = \mathbf{S}^{-1/2} \mathbf{C}' \mathbf{U} \quad (8)$$

This scheme includes more and more correction terms to the orbitals at higher order, and therefore it converges to the canonical treatment. In the limit of a full incremental expansion, the exact result is obtained, since the original diagonal Fock matrix is recovered.

**2.4. Distance Screening.** The second-order energy increments decay usually very rapidly with increasing distance of the domains. This can be used to introduce a distance criterion, in order to screen out the small contributions.<sup>37</sup> Since the incremental contributions decay with increasing order, we use an order-dependent distance threshold of

$$t_{\text{dist}} = \frac{f_{\text{method}}}{(\mathbf{O}-1)^2}$$

where  $\mathbf{O}$  is the order and  $f_{\text{method}}$  is an adjustable parameter for every method. A value of  $f_{\text{method}} = \infty$  means that no distance truncation is performed.

**2.5. Obtaining the Correlation Energies.** The AO basis for a domain is determined by the  $t_{\text{main}}$  parameter (vide supra). Since we change the AO basis for every domain, we have to construct MOs for every domain. In the current work this is done by HF calculations in the AO basis of the domain with a subsequent Boys localization. The orbitals of the domain  $\mathbf{K}$  are then identified by the centers of charge. This requires a unique mapping of the charge centers in the basis  $\mathbf{B}_1$  to the charge center in the basis  $\mathbf{B}_2$  (see ref 44 for details).

In the present work we apply the MP2, CCSD, and the CCSD(T) approach as implemented in MOLPRO<sup>47–49</sup> to evaluate the correlation energies. For a domain  $K$  we correlate all electrons in the orbitals of  $K$ .

### 3. Computational Details

**3.1. Incremental Calculations.** First we perform a HF calculation for the total system in a minimal basis set with a subsequent Boys localization.<sup>50,51</sup> In the next step we extract the molecular orbital coefficient matrix, the overlap matrix in AO basis, and the dipole integrals in AO basis from the MOLPRO calculation, in order to construct the  $n$ -site domains,<sup>36</sup> where the parameter  $t_{\text{con}}$  determines the connectivity in the edge-weighted graph of the correlated occupied orbitals and the parameter  $\text{dsp}$  controls the size of the domains (for details see refs 36–39). Then we determine the basis set for a domain by the distance  $t_{\text{main}}$ . Now a HF calculation with the dual basis set is performed, and the orbitals are localized with the Boys procedure of MOLPRO.<sup>47</sup> Next we classify the new MOs into the domains according to the centers of charge and construct the pseudocanonical orbitals for the domain using the procedure above. Finally we calculate the correlation energy for the domain using the MOLPRO MP2, CCSD, and CCSD(T) codes. To avoid numerical problems due to the error propagation in the incremental series, we determine the HF energy to  $10^{-11}$  hartree and use a dynamical energy threshold  $e_{\text{thres}}$  to determine the accuracy of the correlation calculations, as described in ref 43.

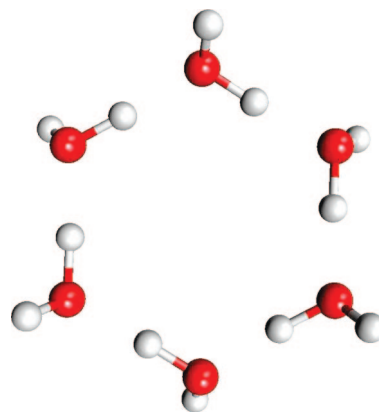
**3.2. Geometries.** If nothing else is stated, we optimized the geometries with the RI-BP86/SVP gradient-corrected density functional theory method<sup>52,53</sup> in the TURBOMOLE 5.6<sup>54</sup> quantum chemistry package. Stationary points were characterized by analyzing the Hessian matrix. Note that the goal of the current work is not to derive very accurate structural data for the compounds investigated here but rather to obtain reasonable geometrical parameters for the incremental MP2, CCSD, and CCSD(T) calculations.

**3.3. Hardware.** The calculations were performed on a cluster of Intel Core2Quad Q6600 PCs with 2.4 GHz, 4 GB random access memory (RAM), and 160 GB disk space per node. The PCs are connected with 1 Gbit ethernet.

### 4. Applications

The calculation of water clusters is an active field in quantum chemistry.<sup>22,23,55–65</sup> Since molecular clusters have natural domains, these objects were studied with a perturbation series in terms of single molecules already in the 1970 of the past century at the HF level.<sup>66</sup> Later Xantheas did a many-body analysis for these clusters in terms of water molecules at the HF and at the correlated level.<sup>56</sup> Since chemical reactions in solution are very important, we decided to study the performance of our scheme for water clusters.

**4.1.  $(\text{H}_2\text{O})_6(S_6)$ .** For intermolecular clusters such as  $(\text{H}_2\text{O})_6$  (Figure 1) we can use the full  $S_6$  symmetry of the system to reduce the number of calculations significantly. In this work we use the symmetry analysis as introduced in ref 39. The accuracy of the new scheme is tested for the correlation



**Figure 1.** RI-BP86/SVP optimized structure of  $(\text{H}_2\text{O})_6$  ( $S_6$ ).

**Table 1.** Comparison of the Incremental cc-pVDZ CCSD(T), CCSD, and MP2 Correlation Energies with the Canonical Energies for  $(\text{H}_2\text{O})_6$  ( $t_{\text{main}} = 3$  bohr;  $\text{dsp} = 4$ ,  $t_{\text{con}} = 3$  bohr;  $e_{\text{thres}} = 1 \times 10^{-6}$  au; core = 6; RAM = 800 MB; fit basis, H, O = STO-3G;  $f_{\text{CCSD(T)}} = f_{\text{CCSD}} = f_{\text{MP2}} = \infty$ ;  $S_6$ )

method	order	ith order correction (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)
CCSD(T)	1	-1.278313	-1.278313	35.06
	2	-0.056613	-1.334926	-0.47
canonical CCSD	1	-1.260129	-1.260129	31.03
	2	-0.050369	-1.310497	-0.57
canonical MP2	1	-1.210674	-1.210674	34.15
	2	-0.055719	-1.266393	-0.81
canonical			-1.265103	

consistent double- and triple- $\zeta$  basis sets of Dunning and co-workers as well as the corresponding augmented basis sets<sup>67,68</sup> (cc-pVXZ; aug-cc-pVXZ; X = D,T). Furthermore we explore the accuracy and the timing with respect to the fit basis for the environment and the energy threshold  $e_{\text{thres}}$ . The energies of the CCSD(T), CCSD, and MP2 calculations are given in Tables 1–4, the timings and the disk space requirements of the CCSD(T) calculations are given in Table 5. The  $\text{dsp}$  parameter has been set to 4 in all calculations on water, since we freeze the 1s orbitals of the oxygen atom, which means that we have to correlate 4 occupied orbitals per water molecule. Therefore a  $\text{dsp} = 4$  in combination with a  $t_{\text{con}} = 3$  forces the one-site domains to be a single water molecule.

In Table 1 we use a minimal basis set (STO-3G) as fit basis for the cc-pVDZ calculations. In this case we get chemical accuracy of 1 kcal/mol already at the second order for all applied correlation methods. If a minimal basis is used for the environment in combination with the aug-cc-pVDZ basis with diffuse functions, which are clearly necessary for the correct description of intermolecular interactions, we need a third-order expansion to obtain chemical accuracy (Table 2). The reason for this is most likely the basis set superposition error (BSSE) in the incremental energies due to the minimal basis set. If we extend the basis set of the environment and use a 6-31G basis for the oxygen atoms, the errors of the applied correlation methods are reduced significantly. For the second-order expansion we already have chemical accuracy, and at the third-order level the errors

**Table 2.** Comparison of the Incremental aug-cc-pVDZ CCSD(T), CCSD, and MP2 Correlation Energies with the Canonical Energies for (H<sub>2</sub>O)<sub>6</sub> ( $t_{\text{main}} = 3$  bohr;  $\text{dsp} = 4$ ;  $t_{\text{con}} = 3$  bohr;  $e_{\text{thres}} = 1 \times 10^{-6}$  au; core = 6; RAM = 1520 MB;  $f_{\text{CCSD(T)}} = f_{\text{CCSD}} = f_{\text{MP2}} = \infty$ ;  $S_6$ )

method	order	fit basis: H, O = STO-3G			fit basis: H = STO-3G; O = 6-31G			fit basis: aug-cc-pVDZ		
		correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)
CCSD(T)	1	-1.384842	-1.384842	38.37	-1.387153	-1.387153	36.92	-1.391714	-1.391714	34.06
	2	-0.063037	-1.447878	-1.18	-0.059193	-1.446346	-0.22	-0.053907	-1.445621	0.23
	3	0.002724	-1.445154	0.53	0.000465	-1.445881	0.07	-0.000329	-1.445950	0.03
exact			-1.445993		-1.445993			-1.445993		
CCSD	1	-1.353862	-1.353862	33.23	-1.356129	-1.356129	31.81	-1.360119	-1.360119	29.30
	2	-0.054940	-1.408802	-1.25	-0.051162	-1.407292	-0.30	-0.046544	-1.406663	0.10
	3	0.002728	-1.406074	0.46	0.000571	-1.406720	0.06	-0.000124	-1.406787	0.02
exact			-1.406814		-1.406814			-1.406814		
MP2	1	-1.313124	-1.313124	36.26	-1.314587	-1.314587	35.34	-1.318964	-1.318964	32.59
	2	-0.059355	-1.372479	-0.99	-0.056873	-1.371460	-0.35	-0.051761	-1.370726	0.11
	3	0.002330	-1.370149	0.48	0.000646	-1.370814	0.06	-0.000181	-1.370906	0.00
exact			-1.370907		-1.370907			-1.370907		

**Table 3.** Comparison of the Incremental cc-pVTZ CCSD(T), CCSD, and MP2 Correlation Energies with the Canonical Energies for (H<sub>2</sub>O)<sub>6</sub> ( $t_{\text{main}} = 3$  bohr;  $\text{dsp} = 4$ ;  $t_{\text{con}} = 3$  bohr;  $e_{\text{thres}} = 1 \times 10^{-6}$  au; core = 6; RAM = 1520 MB;  $f_{\text{CCSD(T)}} = f_{\text{CCSD}} = f_{\text{MP2}} = \infty$ ;  $S_6$ )

method	order	fit basis: H, O = STO-3G			fit basis: H = STO-3G; O = 6-31G		
		correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)
CCSD(T)	1	-1.639658	-1.639658	39.16	-1.643736	-1.643736	36.60
	2	-0.063213	-1.702870	-0.51	-0.058272	-1.702008	0.03
	3	0.001173	-1.701698	0.23			
canonical			-1.702060		-1.702060		
CCSD	1	-1.593649	-1.593649	33.77	-1.597473	-1.597473	31.37
	2	-0.054810	-1.648459	-0.63	-0.050153	-1.647626	-0.10
	3	0.001242	-1.647217	0.15			
canonical			-1.647516		-1.647516		
MP2	1	-1.564739	-1.564739	38.24	-1.568374	-1.568374	35.96
	2	-0.062153	-1.626892	-0.76	-0.057661	-1.626035	-0.23
	3	0.001530	-1.625362	0.20			
canonical			-1.625751		-1.625751		

**Table 4.** Comparison of the Incremental aug-cc-pVTZ CCSD(T), CCSD, and MP2 Correlation Energies with the Canonical Energies for (H<sub>2</sub>O)<sub>6</sub> ( $t_{\text{main}} = 3$ ;  $\text{dsp} = 4$ ;  $t_{\text{con}} = 3$ ;  $e_{\text{thres}} = 1 \times 10^{-6}$  au; core = 6; RAM = 1520 MB; fit basis, H = STO-3G and O = 6-31G;  $f_{\text{CCSD(T)}} = f_{\text{CCSD}} = f_{\text{MP2}} = \infty$ ;  $S_6$ )

method	order	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)
CCSD(T)	1	-1.678737	-1.678737	–
	2	-0.064205	-1.742942	–
CCSD	1	-1.627791	-1.627791	34.25
	2	-0.055066	-1.682857	-0.31
canonical			-1.682365	
MP2	1	-1.603516	-1.603516	38.74
	2	-0.062281	-1.665797	-0.34
canonical			-1.665256	

are negligible. To compare the error of the reduced basis set, we included a calculation with the aug-cc-pVDZ for the environment. From this we see that the usage of a smaller basis set does not change the convergence behavior for CCSD, and CCSD(T) and the errors are of the same order of magnitude compared to the 6-31G/STO-3G basis in the environment. Furthermore the calculation with the full basis set in the environment shows that the incremental expansion of the diagonalization corrections leads to the correct canonical energy in this case.

Considering the cc-pVTZ basis set, we have similar findings (Table 3). The second-order errors are larger, if a

minimal basis is used to model the environment, and they become much smaller, if we use the unpolarized 6-31G basis of double- $\zeta$  quality for oxygen. For the aug-cc-pVTZ basis we find chemical accuracy at the second order of the incremental expansion with the STO-3G/6-31G fit basis for the CCSD and the MP2 energies. In this case the canonical CCSD(T) calculation was infeasible with the 32-bit executable of MOLPRO and 2.4 GB RAM. Since the errors of CCSD and MP2 are only -0.31 and -0.34 kcal/mol, respectively, we expect a similar error for the CCSD(T) energy, too.

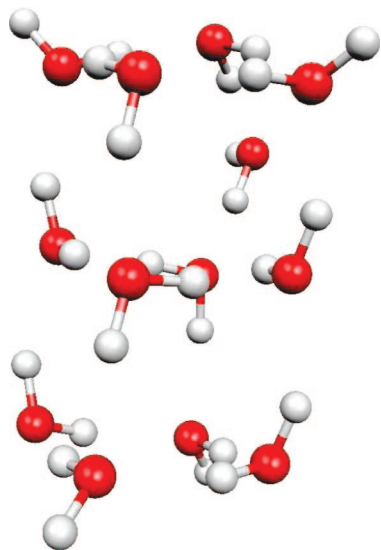
The timings of the incremental calculations are calculated by adding the real time of the client processes and the master process. Therefore they include the time for all processes required for the calculation except the time for the HF calculation in the full basis set. Since the time of the HF calculation is negligible compared to the time spent in the CCSD(T) calculation, the timings of the incremental calculations are directly comparable to the real time of the canonical calculation. Comparing the timings of the incremental calculations with the timings of the canonical calculations (Table 5), we find large improvements in the CPU time, without significant loss of accuracy. For the CCSD(T)/aug-cc-pVDZ calculations we find a reduction to less than 3.3% of the time used in the canonical case (last column), if a



**Table 5.** Timings and Disk Space Requirements of the Incremental and Canonical CCSD(T) Calculations for (H<sub>2</sub>O)<sub>6</sub>, Using the MOLPRO Quantum Chemistry Package

method	order	no. of slaves	wall time		total time		disk space	
			(s)	(%)	(s)	(%)	(GB)	(%)
cc-pVDZ								
canonical <sup>a</sup>		1	5842.6	100.0	5842.6	100.0	1.0	100.0
STO-3G fit <sup>a</sup>	2	3	107.6	1.8	260.6	4.5	0.1	8.6
aug-cc-pVDZ								
canonical <sup>d</sup>		1	60860.8	100.0	60860.8	100.0	6.0	100.0
STO-3G fit <sup>b</sup>	3	7	2279.2	3.7	10493.3	17.2	1.6	26.8
6-31G/STO-3G fit <sup>b</sup>	3	7	3041.5	5.0	14292.3	23.5	2.1	34.4
6-31G/STO-3G fit <sup>b</sup>	2	4	613.0	1.0	2022.2	3.3	0.8	13.3
aug-cc-pVDZ fit <sup>b,c</sup>	3	8	6692.1	11.0	41259.1	67.8		
aug-cc-pVDZ fit <sup>b,c</sup>	2	4	2955.2	4.9	12748.4	20.9		
cc-pVTZ								
canonical <sup>d</sup>		1	257026.5	100.0	257026.5	100.0	11.1	100.0
STO-3G fit <sup>b</sup>	3	7	8800.5	3.4	39230.6	15.3	3.5	31.5
6-31G/STO-3G fit <sup>b</sup>	2	4	1586.9	0.6	5975.1	2.3	1.7	14.8
aug-cc-pVTZ								
canonical (CCSD) <sup>d,e</sup>		1	406096.4		406096.4		115.4	
6-31G/STO-3G fit (CCSD(T)) <sup>b</sup>	2	4	8769.7		31130.3		9.7	

<sup>a</sup> 800 MB RAM. <sup>b</sup> 1520 MB RAM. <sup>c</sup> 64 bit executable of MOLPRO of the 2006.1 version. <sup>d</sup> 2400 MB RAM. <sup>e</sup> Not enough memory for the triples calculation.

**Figure 2.** Structure of the (H<sub>2</sub>O)<sub>13</sub> cluster reported by Bulusu et al.<sup>63</sup>

STO-3G/6-31G fit basis is used. Since the calculation can be run in parallel, we end up with a wall time of 1% for this calculation using four CPUs. If the incremental calculations are run in the full basis set of the environment, the required computer times are much larger compared to the reduced basis set calculations, but still stay below the time for the canonical calculation for the second and third order. For all timings of this work we included the wall time of the master process into the sum of the CPU times, even though it does not require this time completely. For the 6-31G/STO-3G fit in the cc-pVTZ basis the wall time is only 0.6% of the time, needed for the canonical calculation.

A comparison of the disk space requirements for the canonical and the incremental calculations in different basis sets is given in Table 5. As one can see from this, the disk space requirements are significantly reduced, for all incremental calculations. The disk space for the incremental calculations depend on the order of the expansion, on the

applied fit basis and on the threshold  $t_{\text{main}}$ . If the fit basis is increased, the disk space requirements increase, too. Since it was sufficient to do a second-order expansion on the water cluster to obtain chemical accuracy with the larger fit basis, the disk space requirement is lower compared to the total STO-3G fitting and a third-order expansion. Finally we conclude that the incremental scheme can be used to reduce the disk space requirements significantly.

**4.2. (H<sub>2</sub>O)<sub>13</sub>.** To check the performance of our approach for larger water clusters, we studied the (H<sub>2</sub>O)<sub>13</sub> cluster as reported by Bulusu et al.<sup>63</sup> (Figure 2). In the small 6-31G\*\* basis of Pople and co-workers,<sup>69,70</sup> we find chemical accuracy at third order for CCSD(T), CCSD, and MP2 using a STO-3G basis to fit the environment (Table 6). If we use the unpolarized 6-31G basis for oxygen and the STO-3G on hydrogen in the environment, we get chemical accuracy already at the second order of the incremental expansion. Therefore we conclude that a mixed STO-3G/6-31G basis set is a good choice to model water clusters with high accuracy at reduced cost. Since the reference calculation in the small 6-31G\*\* basis is already very expensive, we were not able to perform the CCSD(T)/CCSD calculations using the aug-cc-pVDZ/cc-pVTZ basis sets, whereas the incremental calculations were still feasible (Table 8). From the previous results we expect that the third-order calculations provide chemical accuracy for these systems. Therefore we conclude that the proposed variant of the incremental method is a useful tool to calculate high-level CCSD(T) energies for large systems which are not accessible with the standard approaches.

Figure 3 shows the error of the MP2, CCSD, and CCSD(T) third-order energies for (H<sub>2</sub>O)<sub>13</sub> with respect to the truncation parameter  $f$ . For  $f = 60$  all terms were included in the incremental series which means that larger values of  $f$  will not change the energies for third-order calculations. Since  $f = 20$  leads to small errors in this case, we use this value to perform the distance screening.

**Table 6.** Comparison of the Incremental 6-31G\*\* CCSD(T), CCSD, and MP2 Correlation Energies with the Canonical Energies for (H<sub>2</sub>O)<sub>13</sub> ( $t_{\text{main}} = 2$  bohr;  $\text{dsp} = 4$ ,  $t_{\text{con}} = 3$  bohr;  $\epsilon_{\text{thres}} = 1 \times 10^{-6}$  au; core = 13; RAM = 800 MB; fit basis, H, O = STO-3G;  $f_{\text{CCSD(T)}} = 20$  bohr;  $f_{\text{CCSD}} = 50$  bohr;  $f_{\text{MP2}} = 50$  bohr;  $C_1$ )

method	order	fit basis: H, O = STO-3G			fit basis: H = STO-3G; O = 6-31G		
		correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	error (kcal/mol)
CCSD(T)	1	-2.646002	-2.646002	67.32	-2.669007	-2.669007	52.88
	2	-0.105492	-2.751494	1.12	-0.083867	-2.752874	0.26
	3	-0.002987	-2.754480	-0.75	-0.000353	-2.753226	0.04
canonical CCSD			-2.753284			-2.753284	
	1	-2.614089	-2.614089	59.36	-2.635218	-2.635218	46.10
	2	-0.094038	-2.708126	0.35	-0.073568	-2.708786	-0.06
canonical MP2			-2.709598			-2.708836	
			-2.708687			-2.708687	
	1	-2.506036	-2.506036	65.48	-2.527991	-2.527991	51.71
canonical			-2.611444			-2.611270	
	2	-0.105408	-2.611444	-0.66	-0.083279	-2.611270	-0.55
	3	0.000399	-2.611045	-0.41	0.000858	-2.610412	-0.01
canonical			-2.610393			-2.610393	

**Table 7.** Timings of the Incremental CCSD(T) Calculations with Respect to the Canonical Ones for (H<sub>2</sub>O)<sub>13</sub>, Using the MOLPRO Quantum Chemistry Package

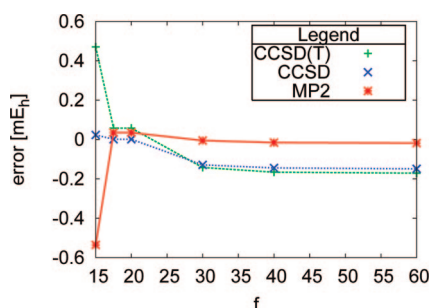
method	order	no. of slaves	wall time		total time	
			(s)	(%)	(s)	(%)
canonical <sup>a</sup>		1	1003263.6	100.0	1003263.6	100.0
STO-3G fit <sup>b</sup>	3	50	1614.0	0.2	54173.6	5.4
6-31G/STO-3G fit <sup>c</sup>	2	20	4382.8	0.4	79996.5	8.0
6-31G/STO-3G fit <sup>c</sup>	3	50	5083.2	0.5	193714.6	19.3

<sup>a</sup> 2650 MB RAM. <sup>b</sup> 800 MB RAM. <sup>c</sup> 1520 MB RAM.

**Table 8.** Incremental CCSD(T), CCSD, and MP2 Correlation Energies for (H<sub>2</sub>O)<sub>13</sub> ( $t_{\text{main}} = 2$  bohr;  $\text{dsp} = 4$ ;  $t_{\text{con}} = 3$  bohr;  $\epsilon_{\text{thres}} = 1 \times 10^{-7}$  au; core = 13; RAM = 1040 MB; fit basis, H = STO-3G and O = 6-31G;  $f_{\text{CCSD(T)}} = 30$  bohr;  $f_{\text{CCSD}} = 40$  bohr;  $f_{\text{MP2}} = 40$  bohr;  $C_1$ )

method	order	aug-cc-pVDZ <sup>a</sup>		cc-pVTZ <sup>b</sup>	
		correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)	correction( <i>i</i> ) (au)	$E_{\text{corr}}(i)$ (au)
CCSD(T)	1	-2.974964	-2.974964	-3.529315	-3.529315
	2	-0.138585	-3.113549	-0.138913	-3.668228
	3	0.005169	-3.108381	-0.000497	-3.668725
CCSD	1	-2.911434	-2.911434	-3.433496	-3.433496
	2	-0.119177	-3.030611	-0.119799	-3.553295
	3	0.004951	-3.025661	0.000601	-3.552694
MP2	1	-2.815292	-2.815292	-3.364553	-3.364553
	2	-0.128732	-2.944024	-0.136900	-3.501453
	3	0.004889	-2.939135	0.003631	-3.497821

<sup>a</sup> The canonical reference calculations with 533 contracted Gaussian basis functions, 52 correlated occupied orbitals, and 468 virtual orbitals were not feasible. <sup>b</sup> The canonical reference calculations with 754 contracted Gaussian basis functions, 52 correlated occupied orbitals, and 689 virtual orbitals were not feasible.

**Figure 3.** Error of the MP2, CCSD, and CCSD(T) energy with respect to the truncation parameter  $f$ .

Considering the timing for the (H<sub>2</sub>O)<sub>13</sub> cluster in Table 7, we find a reduction of the total CPU time of 94.6% compared to the canonical calculation for the STO-3G fit. The wall time is 0.2% on 50 slave PCs. The parallel efficiency as

defined by Pulay and co-workers<sup>71</sup> (total CPU time divided by the product of the wall time and the number of CPUs) is in this case 65.8%, which is similar to the efficiency reported by other groups.<sup>64,71,72</sup> If we calculate the efficiency with respect to the canonical calculation:

$$\text{total efficiency} = \frac{\text{time of the canonical calculation}}{\text{product of wall time and no. of CPUs}} \quad (9)$$

we get 1218.9%, due to the local approximations. Note that the master process was included in the total CPU time as well in the denominator of the efficiency. We want to point out that the accuracy is still within 1 kcal/mol in this case. Considering the results of the (H<sub>2</sub>O)<sub>6</sub> cluster, we expect that we will not meet the high accuracy for calculations with larger basis sets. If we use the more accurate STO-3G/6-31G fit, we end up with

a total CPU time of 8.0% at second order. The wall time is 0.4% on 20 slave PCs and the parallel efficiency 86.9%. The reason for the larger parallel efficiency is the better ratio between the number of calculations (91) and the number of slaves (20). In general the efficiency of our approach will decrease, if the number of CPUs gets comparable to the number of calculations. The reason for this is that the time for the calculations of domains with different size may vary largely. Due to the nature of the incremental scheme the domain sizes vary largely for different orders of the incremental expansion; e.g., for one-site domains of the same size, the size is doubled at second order and tripled at third order. However, since we have no idle processes in our approach, it is not very important to optimize the efficiency defined in this way. The reason why the efficiency as defined by Pulay may become small in our scheme is just because the denominator of eq 9 counts contributions of CPUs which may already compute another job.

Going to the third-order level, we have a total CPU time of 19.3% and a wall time of 0.5% using 50 slave PCs. This corresponds to a parallel efficiency of 74.7% and a total efficiency of 387.0%.

The timing of the STO-3G/6-31G fit is somewhat worse compared to the timing of the STO-3G fit. However, due to the higher accuracy and the robustness of the STO-3G/6-31G fit for larger basis sets we conclude that it is superior to the STO-3G fit. As judged from the timings and the high accuracy, we conclude that the proposed approach is a useful tool to calculate the correlation energies of large systems at reduced cost.

Since the calculations in the domains can be run independently, we conclude that the incremental scheme is inherently parallel. Therefore it is not necessary to have all slaves available at the beginning of the calculation, and no slave has to wait for the other ones. We do not waste CPU time by idle CPUs; therefore the total CPU efficiency as defined by Pulay<sup>71</sup> is not that important for our incremental approach. The only process which collects the total data is the master process. Since our master and slave structure is build upon the socket++ library, the master process does not consume much CPU time. On the basis of the improved timings compared to canonical calculations and the high accuracy (within 1 kcal/mol) achieved in all cases, we conclude that our modified variant of the incremental scheme is probably competitive with the currently proposed parallel CCSD(T) approaches of Olson et al.<sup>64</sup> and Auer et al.<sup>72</sup> Although the current pilot implementation is not generally applicable, we demonstrated for an important class of chemical systems that the approach considerably improves the speed of the calculations. In our future research we plan to use a projection technique to remove the mapping step for the identification of the domains, in order to make the approach generally applicable.

If we compare the proposed incremental method with the approach of Werner and Schütz<sup>73</sup> or Head-Gordon and Subotnik,<sup>74</sup> we obtain more accurate correlation energies if the incremental series is truncated in a proper way. On the other hand the approach of Werner and Schütz as well as the approach of Subotnik and Head-Gordon scale linearly with the system size. Therefore we conclude that the

proposed incremental scheme is somewhat in the middle between the efficiency of the local coupled cluster methods and the accuracy of the parallel CCSD(T) implementations.

## 5. Conclusion

We introduced a new efficient variant of the incremental scheme to calculate CCSD(T), CCSD, and MP2 correlation energies of large systems. We demonstrated for CCSD(T), CCSD, and MP2 energies that chemical accuracy can be reached at reduced computational cost, if a dual basis set approach is followed and the incremental expansion is truncated in a proper way. Furthermore we have shown that the disk space and memory requirements are reduced significantly. We have shown that our scheme is systematically improvable by including higher orders of the expansion and by applying better basis sets for the description of the environment orbitals. In addition the approach is inherently parallel with essentially no loss in CPU time due to dependencies of the individual processes. Therefore we conclude that our modified incremental scheme is an alternative way to calculate high-level correlation energies for large systems. Although in the current work we only applied the approach to water clusters, it can also be applied to hydrocarbons and the glycine tetramer as demonstrated elsewhere.<sup>44</sup>

**Acknowledgment.** We gratefully acknowledge financial support by the German Research Foundation (DFG) through the priority program 1145 and the SFB 624.

## References

- (1) Pulay, P.; Saebø, S. *Theor. Chim. Acta* **1986**, *69*, 357.
- (2) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86* (2), 914.
- (3) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.
- (4) Boughton, J. W.; Pulay, P. *J. Comput. Chem.* **1993**, *14*, 736.
- (5) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286.
- (6) Schütz, M.; Hetzer, G.; Werner, H. J. *J. Chem. Phys.* **1999**, *111* (13), 5691.
- (7) Maslen, P. E.; Head-Gordon, M. *Chem. Phys. Lett.* **1998**, *283*, 102.
- (8) Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 7093.
- (9) Maslen, P. E.; Lee, M. S.; Head-Gordon, M. *Chem. Phys. Lett.* **2000**, *319*, 205.
- (10) Lee, M. S.; Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **2000**, *112*, 3592.
- (11) Schütz, M. *J. Chem. Phys.* **2000**, *113* (22), 9986.
- (12) Schütz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661.
- (13) Flocke, N.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *121*, 10935.
- (14) Subotnik, J. E.; Head-Gordon, M. *J. Chem. Phys.* **2005**, *123*, 64108.
- (15) Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 74116.
- (16) Auer, A.; Nooijen, M. *J. Chem. Phys.* **2006**, *125*, 24104.
- (17) Weijs, V.; Manninen, P.; Jørgenson, P.; Christiansen, O.; Olsen, J. *J. Chem. Phys.* **2007**, *127*, 074106.

- (18) Doser, B.; Lambrecht, D. S.; Ochsenfeld, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3335.
- (19) Hughes, T. F.; Flocke, N.; Bartlett, R. J. *J. Phys. Chem. A* **2008**, *112*, 5994.
- (20) Li, W.; Li, S. *J. Chem. Phys.* **2004**, *121*, 6649.
- (21) Federov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.
- (22) Kamiya, M.; Hirata, S.; Valiev, M. *J. Chem. Phys.* **2008**, *128*, 074103.
- (23) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
- (24) Stoll, H. *Chem. Phys. Lett.* **1992**, *191*, 548.
- (25) Stoll, H. *Phys. Rev. B* **1992**, *46*, 6700.
- (26) Stoll, H. *J. Chem. Phys.* **1992**, *97*, 8449.
- (27) Nesbet, R. K. *Phys. Rev.* **1967**, *155*, 51.
- (28) Nesbet, R. K. *Phys. Rev.* **1968**, *175*, 2.
- (29) Nesbet, R. K. *Adv. Chem. Phys.* **1969**, *14*, 1.
- (30) Doll, K.; Dolg, M.; Fulde, P.; Stoll, H. *Phys. Rev. B* **1997**, *55*, 10282.
- (31) Rosciszewski, K.; Paulus, B.; Fulde, P.; Stoll, H. *Phys. Rev. B* **1999**, *60*, 7905.
- (32) Shukla, A.; Dolg, M.; Fulde, P.; Stoll, H. *Phys. Rev. B* **1999**, *60*, 5211.
- (33) Stoll, H.; Paulus, B.; Fulde, P. *J. Chem. Phys.* **2005**, *123*, 144108.
- (34) Voloshina, E.; Paulus, B. *J. Phys. Chem.* **2006**, *124*, 234711.
- (35) Paulus, B. *Int. J. Quantum Chem.* **2004**, *100*, 1026.
- (36) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 154110.
- (37) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Phys. Chem. A* **2007**, *111*, 9830.
- (38) Friedrich, J.; Hanrath, M.; Dolg, M. *Chem. Phys.* **2007**, *338*, 33.
- (39) Friedrich, J.; Hanrath, M.; Dolg, M. *Chem. Phys.* **2008**, *346*, 266.
- (40) Mödl, M.; Dolg, M.; Fulde, P.; Stoll, H. *J. Chem. Phys.* **1997**, *106*, 1836.
- (41) Bezugly, V.; Birkenheuer, U. *Chem. Phys. Lett.* **2004**, *399*, 57.
- (42) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Phys. Chem. A* **2008**, *112*, 8762.
- (43) Friedrich, J.; Walczak, K.; Dolg, M. *Chem. Phys.* **2008**, doi: 10.1016/j.chemphys.2008.10.030.
- (44) Friedrich, J.; Dolg, M. *J. Chem. Phys.* **2009**, *129*, 244105.
- (45) Jurgens-Lutovsky, R.; Almlöf, J. *Chem. Phys. Lett.* **1991**, *178*, 452.
- (46) Kloppe, W.; Noga, J.; Koch, H.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 164.
- (47) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D. Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, a package of ab initio programs designed* version 2002; Werner H.-J., Knowles, P. J., Eds.; Technical Report; University of Birmingham: Birmingham, England, 2002.
- (48) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219.
- (49) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106.
- (50) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300.
- (51) Edmiston, C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457.
- (52) Becke, A. D. *Phys. Rev. A* **1988**, *38* (6), 3098.
- (53) Perdew, J. P. *Phys. Rev. B* **1986**, *33* (12), 8822.
- (54) Ahlrichs, R.; Bär, M.; Baron, H.-P.; Bauernschmitt, R.; Böcker, S.; Ehrig, M.; Eichkorn, K.; Elliott, S.; Furche, F.; Haase, F.; Häser, M.; Horn, H.; Huber, C.; Huniar, U.; Kölmel, C.; Kollwitz, M.; Ochsenfeld, C.; Öhm, H.; Schäfer, A.; Schneider, U.; Treutler, O.; von Arnim, M.; Weigend, F.; Weis, P.; Weiss, H. *Turbomole 5*; Institut für Physikalische Chemie, Universität Karlsruhe: Karlsruhe, Germany, 2002.
- (55) Laasonen, K.; Parrinello, M.; Car, R.; Lee, C.; Vanderbilt, D. *Chem. Phys. Lett.* **1993**, *207*, 208.
- (56) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.
- (57) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. *J. Phys. Chem. A* **1997**, *101*, 9163.
- (58) Kim, J.; Majumdar, D.; Lee, H. M.; Kim, K. S. *J. Chem. Phys.* **1999**, *110*, 9128.
- (59) Kozmutza, C.; Kryachko, E. S.; Tfirst, E. *THEOCHEM* **2000**, *501*, 435.
- (60) Day, P.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys.* **2000**, *112*, 2063.
- (61) Upadhyay, D. M.; Shukla, M. K.; Mishra, P. C. *Int. J. Quantum Chem.* 2001.
- (62) Tschumper, G. S. *Chem. Phys. Lett.* **2006**, *427*, 185.
- (63) Bulusu, S.; Yoo, S.; Apra, E.; Xantheas, S.; Zeng, X. C. *J. Phys. Chem. A* **2006**, *110*, 11781.
- (64) Olson, R. M.; Bentz, J. L. R. A. K.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Theory Comput.* **2007**, *3*, 1312.
- (65) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.
- (66) Hankins, D.; Moskowitz, J. W.; Stillinger, F. H. *J. Chem. Phys.* **1970**, *53*, 4544.
- (67) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (68) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (69) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (70) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
- (71) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368.
- (72) Harding, M. E.; Metzroth, T.; Gauss, J.; Auer, A. A. *J. Chem. Theory Comput.* **2008**, *4*, 64.
- (73) Schütz, M.; Werner, H.-J. *Chem. Phys. Lett.* **2000**, *318*, 370.
- (74) Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 034103.

## pK<sub>a</sub> Calculation of Some Biologically Important Carbon Acids - An Assessment of Contemporary Theoretical Procedures

Junming Ho and Michelle L. Coote\*

ARC Centre of Excellence for Free-Radical Chemistry and Biotechnology, Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia

Received August 13, 2008

**Abstract:** In this study, the aqueous pK<sub>a</sub> values for 13 neutral, 10 cationic, and 5 anionic carbon acids, including amino acids, peptides, and related species have been calculated using the high level ab initio composite procedure, G3MP2+//BMK, combined with solvation energies that were calculated using the CPCM-(UAKS/UAHF), COSMO-RS, and SM6 continuum models. The pK<sub>a</sub>s were further calculated using three schemes, namely the direct method and the proton exchange method as well as the inclusion of an explicit solvent water molecule. The results of this study indicate that the direct method is unsuitable for computing the pK<sub>a</sub> of carbon acids, whereas the other two schemes perform significantly better with varying degrees of success, depending on the charge of the carbon acid. Specifically, the combination of the proton exchange scheme and CPCM-UAKS model performed particularly well for neutral species, with mean absolute deviations (MADs) of ~1 pK<sub>a</sub> unit. The ionic species were more problematic, though the combination of the proton exchange scheme and the SM6 and CPCM-UAKS models performed reasonably well for the cationic and anionic acids, respectively. The inclusion of an explicit water molecule generally improved the calculated values for anionic carbon acids.

### 1. Introduction

Carbon acids are ubiquitous in nature and in the chemistry laboratory. Amino acids and peptides are prominent examples of carbon acids that occur naturally. It is well-known that all living organisms synthesize peptides and proteins that are composed entirely of amino acids in the L-configuration. However, spontaneous racemization can result in the generation of D-residues during the life span of the protein. For example, the accumulation of D-aspartic acid in the brain,<sup>1</sup> tooth enamel,<sup>2,3</sup> bones,<sup>4</sup> and lens proteins<sup>5</sup> has been associated with aging and can contribute to loss of tissue functions. The relative rates of racemization of different amino acid residues at proteins have been reported,<sup>6,7</sup> and this relates to the acidity of the  $\alpha$ -C-H protons, which depends on the nature of the amino acid side chain, the amino acid sequence, and the peptide conformation.

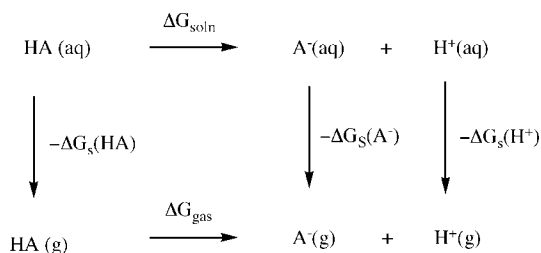
D-Amino acids are also widely found in prokaryotic (bacteria) peptides and less commonly in eukaryotic peptides,

which are invariably components of venoms.<sup>8–15</sup> Contrary to spontaneous racemization as a result of aging, these D-amino acids are critical for the biological activity of these peptides, and specialized enzymes such as racemases and enolases are used to catalyze the heterolytic cleavage of these very stable  $\alpha$ -C-H bonds. The mechanism as to how enzymes activate these bonds for proton transfer, i.e. by increasing the kinetic acidity of these protons, has been a subject of intense research.<sup>16–23</sup>

The acidity of  $\alpha$ -C-H protons also plays an important role in chemical synthesis. For example, the carbanions generated from deprotonation of the  $\alpha$ -carbonyl protons of ketones and aldehydes are routinely used in nucleophilic substitution reactions for forming carbon–carbon bonds. Cyclic dipeptides or diketopiperazines have also been explored as chiral auxiliaries for the stereoselective synthesis of amino acids, where regioselective deprotonation of  $\alpha$ -C-H protons is critical.<sup>24–26</sup>

Clearly, an understanding of substituent effects on the acidity of the  $\alpha$ -C-H protons has important implications in

\* Corresponding author e-mail: mcoote@rsc.anu.edu.au.

**Scheme 1.**  $pK_a$  Calculation via the Direct Method

both biological systems as well as in chemical synthesis. The acid dissociation constant,  $pK_a$ , is the most common measure of the thermodynamic acidity. However, accurate measurements of  $pK_a$  values are further complicated by the extremely weak acidity of these carbon protons, with  $pK_a$  values typically in the range of 20 or higher.<sup>27,28</sup> Thus, very sensitive methods are required to determine the dissociation constants of these acids.

In recent years, Richard et al. have pioneered the use of NMR methods for measuring the  $pK_a$  values of a wide variety of carbon acids with different functionalities, viz. carboxylic acids, esters, amides, derivatives of amino acids, and heterocycles as well as small peptides in aqueous solution.<sup>17,18,22,29–36</sup> Bordwell et al. have also compiled a large database of  $pK_a$  values for various organic acids, including carbon acids, in DMSO.<sup>27,28</sup> This has led to a better understanding of the substituent effects and the mechanisms employed by enzymes to accelerate the deprotonation of the  $\alpha$ -protons. For example, pyridoxal 5'-phosphate (PLP) is known to catalyze carbon deprotonation of  $\alpha$ -amino acids at enzyme active sites by formation of an imine. Through quantitative measurements of  $pK_a$ , Richard et al. have found that a dramatic increase in acidity at the  $\alpha$ -carbons, by  $\sim 7$   $pK_a$  units, may be achieved through the formation of such an intermediate.<sup>22,37</sup>

In light of the recent reports on acidity constants for carbon acids of amino acid derivatives and peptides and their associated exciting insights into enzymatic mechanisms, the theoretical calculation of  $pK_a$  of such carbon acids has become even more appealing. There has been significant effort targeted at making reliable predictions of  $pK_a$  values using quantum chemical methods. Liptak and Shields utilized the thermodynamic cycle shown in Scheme 1 in which gas-phase free energies obtained via high level *ab initio* methods (e.g., CBS-QB3 and G-*n* models) are combined with the solvation free energies obtained from continuum solvation models.<sup>38,39</sup> Despite the intrinsically larger errors in continuum solvation calculations, the authors showed that it was possible to predict  $pK_a$  values of carboxylic acids to within 0.5  $pK_a$  units, presumably due to systematic error cancelation. More generally, they suggested that accurate  $pK_a$  values could be obtained by means of a proton exchange scheme that allowed for further error cancelation.<sup>40</sup> Similar approaches have been used for calculating the  $pK_a$  values of carboxylic acid derivatives,<sup>41</sup> alcohols,<sup>42</sup> carbenes,<sup>43</sup> amines,<sup>44</sup> phosphoranes,<sup>45</sup> substituted phenols,<sup>46</sup> and pharmaceutically important compounds.<sup>47</sup> Various assessment and methodology-development studies employing large data sets of various neutral and charged organic and inorganic acids have also been reported.<sup>47–52</sup> In general, these studies have found that accurate  $pK_a$  values

can be obtained through the combination of high-level *ab initio* methods with continuum solvation models, particularly when proton exchange reactions are used.

However, because accurate experimental  $pK_a$  values for carbon acids are relatively scarce, far fewer studies have been carried out to examine the performance of computational methods on these systems. Brinck et al. have studied the performance of the polarizable continuum model (PCM) for the solvation of small aliphatic carbanions in organic solvents DMF and THF.<sup>53</sup> The deviation with experimental values was found to be quite large ( $\sim 20$ – $30$  kJ/mol), and this has been attributed to the neglect of short-range solvent effects in continuum models. Fu et al. have made use of continuum solvent model combined with the proton exchange scheme for the  $pK_a$  calculation of a large data set of organic molecules, including carbon acids, in DMSO.<sup>51</sup> In their test set, the carbon acids include ketones and substituted aliphatic systems, and the calculated values are generally in good agreement with experiment. Gao et al. have also examined the use of continuum methods and QM/MM-Ewald simulations to calculate the aqueous  $pK_a$  of the acetate anion using both direct method and the proton exchange scheme.<sup>54</sup> The authors found that the QM/MM-ewald protocol yields the best result using the latter scheme, with calculated values within 2  $pK_a$  units of experiment. However, further testing of this method is necessary to establish its general applicability to other carbon acids.

In this light, it is thus important to establish whether such procedures are suitable for the calculation of  $pK_a$  values for carbon acids, particularly those in biological systems. In this study, we wish to examine the performance of several popular procedures for calculating the  $pK_a$  of a range of neutral, cationic, and anionic carbon acids in aqueous solution. It is also worth highlighting that most of these molecules are not included in the training sets for the parametrization of various continuum solvation models and should therefore provide an objective and rigorous test of these methods.

## 2. $pK_a$ Calculation Methods

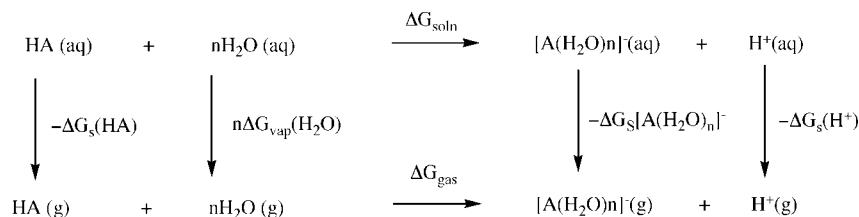
Invariably, most  $pK_a$  calculations involve the representation of the acid dissociation process as a sum of several intermediate steps such as in the thermodynamic cycle shown below in Scheme 1. By virtue of Hess's law, the free energy of acid dissociation in solution,  $\Delta G_{\text{soln}}$ , is equal to

$$\Delta G_{\text{soln}}^* = \Delta G_{\text{gas}}^* + \Delta \Delta G_{\text{solv}}^* \quad (1)$$

where  $\Delta \Delta G_{\text{solv}}^* = \sum \Delta G_{\text{solv,products}} - \sum \Delta G_{\text{solv,reactants}}$ . The “\*” symbol is used for a standard state of 1 mol/L in any phase. The  $K_a$  and  $pK_a$  may be obtained through the thermodynamic relationship

$$\Delta G_{\text{soln}}^* = -RT \ln K_a \quad (2)$$

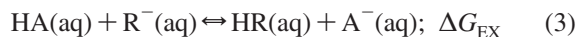
Eq 1 allows us to decompose the errors in the acidity constant into a gaseous component,  $\Delta G_{\text{gas}}$ , and a solvation component,  $\Delta \Delta G_{\text{solv}}$ . Gas-phase energies and hence acidities can now be calculated with chemical accuracy provided that electron correlation effects are included through appropriate post-Hartree–Fock or density functional methods.<sup>55–68</sup> Liptak et al. have found that, for gas-phase energies, the CBS-QB3

**Scheme 2.**  $pK_a$  Calculation with Explicit Water Molecules

method gives the most accurate results.<sup>38,39</sup> In this study, we examine the use of various levels of theories, such as CBS-QB3 and *Gn* composite methods for calculating  $\Delta G_{\text{gas}}$ .

However, the situation is much less satisfactory in solution, mostly due to the difficulty of treating the solvent–solute interactions rigorously. In particular, the acid dissociation involves the formation of charged species starting from neutral molecules. Short-range intermolecular interactions (e.g., ion-dipole and hydrogen bonding) are considered to be particularly important in solvation of charged species. Since these effects are not explicitly taken into account by continuum solvation models,<sup>50,69–71</sup> the direct calculation of  $pK_a$  values via the above thermodynamic cycle is likely to incur significant errors. Two general procedures are used to remedy this deficiency in continuum solvation models.

First, an isodesmic reaction has been shown to yield very accurate  $pK_a$  values ( $\pm 1$   $pK_a$  unit) for moderately strong acids.<sup>41,42,72,73</sup> In this study, a proton exchange reaction between the acid and a reference acid (HR) with known  $pK_a$  is considered.



Since the number of charged species is conserved on both sides of the equation, one can expect some cancelation of the errors due to the neglect of short-range solvent effects. Accordingly, the equilibrium constant,  $K_{\text{EX}}$ , for reaction 3 can be calculated from eqs 1 and 2. The acid dissociation constant,  $K_a(\text{HA})$ , can then be determined by the product of  $K_a(\text{HR})$  and  $K_{\text{EX}}$ . Unfortunately, the success of this approach depends heavily on the choice of reference acid, with best results expected if HR is structurally similar to HA. The accuracy of the calculated value also depends very much on the accuracy of the experimental  $K_a(\text{HR})$ .

The second approach involves inclusion of explicit solvent molecules. There are several variants of this approach, including the cluster-continuum model<sup>51,74</sup> and the implicit-explicit solvent approach.<sup>48</sup> In the latter approach, the  $pK_a$  of an acid is calculated via the thermodynamic cycle in Scheme 2. Based on eq 1, the free energy of acid dissociation in solution is therefore given by eq 4:

$$\Delta G_{\text{soln}} = \Delta G_{\text{gas}} + \Delta G_s(\text{H}^+) + \Delta G_s[\text{A(H}_2\text{O)}_n\text{]}^- - \Delta G_s(\text{HA}) + n\Delta G_{\text{vap}}(\text{H}_2\text{O}) \quad (4)$$

Kelly et al. have used this thermodynamic cycle to calculate  $\Delta G_{\text{solv}}$  and  $pK_a$  in aqueous solution for a variety of organic acids and found that the agreement with experiment is significantly improved when one explicit water molecule was included in the thermodynamic cycle.<sup>48</sup>

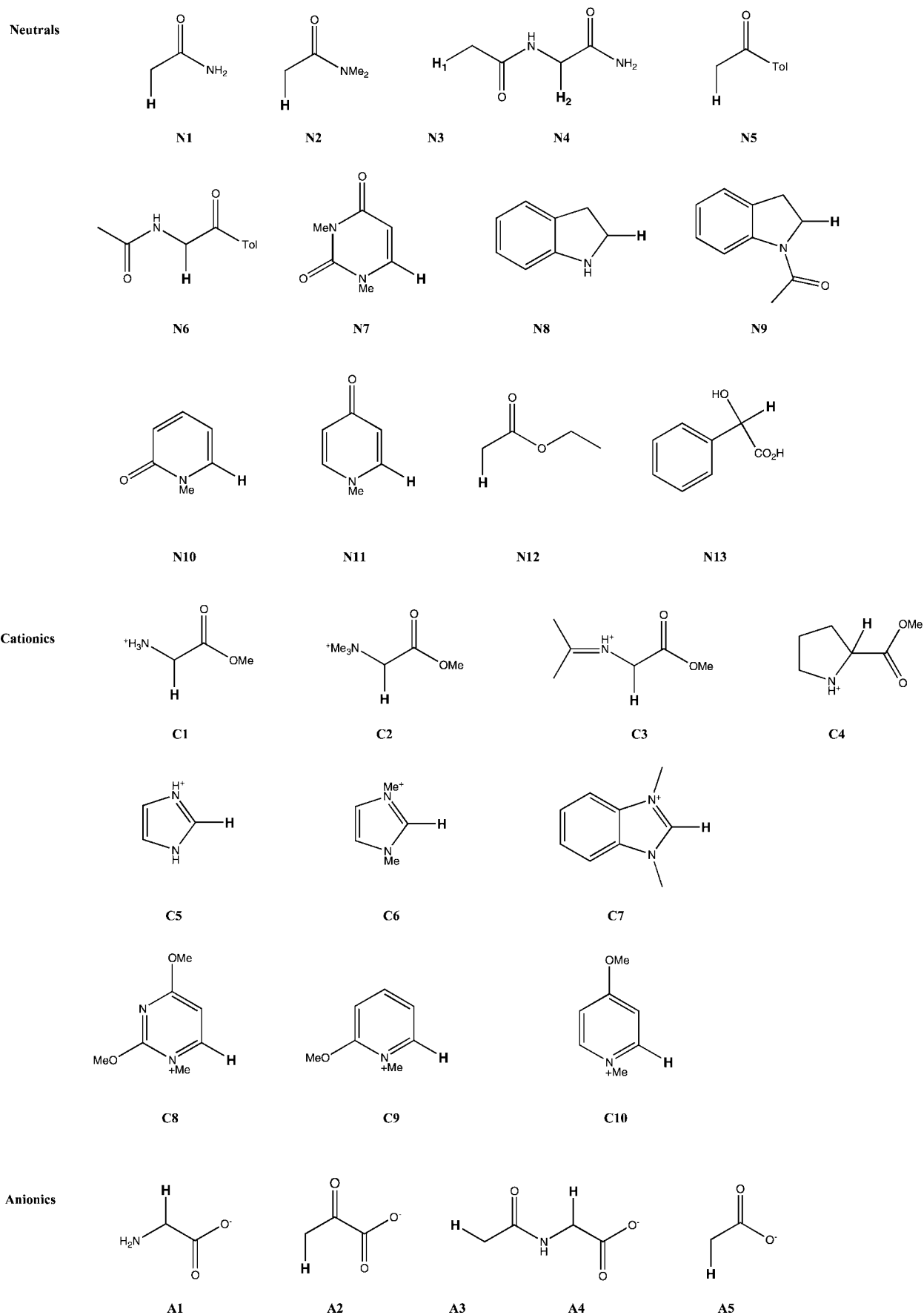
In this study, we examine the three approaches highlighted above, namely the direct method, the proton exchange

method, and the implicit-explicit solvent method, for calculating  $pK_a$ s. The test set of molecules, mainly amino acids and derivatives, has been compiled from a list of studies that were mostly conducted by the research group of Richard's.<sup>20,22,29,30,33–37,75–79</sup> These molecules are further categorized as neutral, cationic, and anionic as shown in Figure 1.

**3. Theoretical Procedures**

Gaussian 03 software<sup>80</sup> has been used for all gas-phase ab initio molecular orbital theory<sup>81</sup> and density functional theory calculations.<sup>82</sup> The gas-phase acid dissociation free energy of several carbon acids has been calculated at the G3MP2,<sup>83</sup> G3,<sup>62</sup> and CBS-QB3<sup>84</sup> levels of theory. The Gaussian methods (G3, G3MP2, and various other modified versions) approximate QCISD(T) energies with a large triple- $\zeta$  basis using cheaper QCISD(T)/6–31G(d) calculations in conjunction with additivity corrections, obtained at the MP2, MP3, and/or MP4 levels of theory.<sup>62,83</sup> The complete basis set methods (CBS) are a model chemistry that makes use of a complete basis set extrapolation of the correlation energy, which is performed at the MP2 level of theory and then corrected to the CCSD(T) level via additivity corrections.<sup>84</sup> These high-level composite procedures have been designed particularly for the prediction of reliable energies of molecules in the gas phase and have been demonstrated to provide an accuracy of 1–2 kcal/mol when assessed against large test sets of thermochemical data.<sup>62,84</sup> In addition to these standard procedures, calculations were also performed using the modified procedures, G3+ and G3MP2+, in which calculations with the 6–31G(d) basis set have been replaced with the 6–31+G(d) basis set, so as to allow for an improved description of anionic species.

In their original forms, the G3MP2 and G3 methods both employ the geometries which are optimized at the MP2(Full)/6–31G(d) level; CBS-QB3 employs B3-LYP/CBSB7 optimized geometries. However, in the present work we sought to identify an optimal procedure for the calculation of geometries that balanced accuracy and computational expense. To this end, proton affinities of a selection of carbon acids were first calculated at a consistent level of theory, MP2(Full)/6–311+G(d, p), using geometries that had been optimized at various levels of theory. On the basis of this initial study, the BMK/6–31+G(d) level of theory was selected for geometry optimizations and frequency calculations for the remainder of the study. Scale factors for the BMK/6–31+G(d,p) vibrational frequencies have been used for the free energy calculations.<sup>85</sup>



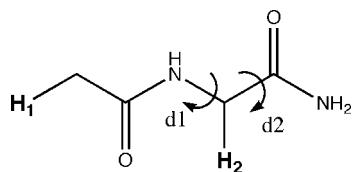
**Figure 1.** Test set of studied carbon acids.

For some of the larger carbon acids such as **N3/N4**, several conformations are possible. Because the acids examined in

this study are relatively small, we have performed a grid search on these molecules. This involves optimization of all



possible conformers generated from combinations of rotations about certain chemical bonds. The rotations were examined at 120° and 180° resolution for sp<sup>2</sup> and sp<sup>3</sup> hybridized centers, respectively. For instance, the rotations examined for the acid **N3/N4** are illustrated where d1 and d2 were examined at 120° and 180° resolution, respectively. The amide bond was fixed in its thermodynamically preferred trans-configuration.



In an earlier study by Liptak and Shields, the authors showed that calculations using continuum models on the lowest energy gas-phase conformer and the conformationally averaged structure gave comparable results.<sup>40</sup> Accordingly, in this study we only consider the solvation free energies on the lowest energy gas-phase conformer. The lowest energy gas-phase optimized geometries and their associated conformers are provided in the Supporting Information.

The free energies of solvation were evaluated using several popular procedures as recommended by several earlier studies of other types of acid.<sup>38,39,43,46</sup> The conductor-polarizable continuum model (CPCM)<sup>86,87</sup> was used to compute solvation free energies at the HF and B3LYP levels of theory in conjunction with various basis sets. These calculations were carried out using GAUSSIAN 03 software<sup>80</sup> using the radii of the united atom topological model, optimized for the Hartree–Fock and DFT methods (UAHF and UAKS), and default values for the other parameters. All geometries of the studied species have been optimized fully in the presence of solvent using different basis sets at the level of HF and B3LYP.

In addition, the free energies of solvation were also computed using the COSMO-RS<sup>88–90</sup> and SM6<sup>91</sup> models. The COSMO-RS model is a variant of the CPCM model (conductor-like screening model for real solvents) that describes the interactions in a fluid as local interaction of molecular surfaces, the interaction energies being quantified by the values of the two screening charge densities that form a molecule contact.<sup>88,89</sup> The resulting energies are presumably more accurate than a typical PCM calculation because the real character of the solvent is taken into account and not a simple homogeneous continuum. The ADF package<sup>92</sup> was used to compute the COSMO-RS solvation free energies on the gas-phase geometries at the BP/TZP level of theory, with the rest of the parameters (e.g., atomic cavity radii, radius of the probing sphere, and cavity construction) kept as default values.<sup>92</sup> We have computed the ADF COSMO-RS solvation free energies for a selection of molecules and compared them with values obtained from the original paper,<sup>90</sup> where the COSMO-RS model was parametrized slightly differently. The agreement is generally very good (within 0.5 kcal/mol), and the data are tabulated in Table S3 in the Supporting Information.

The SM6 model is based on a generalized born approach which uses a dielectric continuum to treat bulk electrostatics effects combined with atomic surface tensions to account for

**Table 1.** Absolute Deviations of Proton Affinities<sup>a</sup> (kJ/mol) Calculated on Geometry Optimized at Various Levels of Theory Relative to MP2(full)/6-311+G(d,p)

geometry	N1	N3	N4	C4	A4
RHF/6–31+G(d)	0.58	0.31	1.78	2.03	5.04
RHF/6–31+G(d,p)	0.26	0.00	1.75	1.78	4.88
RHF/6–311+G(d,p)	0.86	0.70	0.72	1.40	4.74
B3LYP/6–31+G(d)	2.31	0.55	0.31	0.27	1.16
B3LYP/6–31+G(d,p)	0.70	0.49	0.13	0.20	1.35
B3LYP/6–311+G(d,p)	0.75	0.60	0.08	0.11	1.20
BMK/6–31+G(d)	1.04	0.85	0.32	0.26	0.46
BMK/6–31+G(d,p)	0.84	0.72	0.08	0.02	0.51
BMK/6–311+G(d,p)	1.05	0.84	0.02	0.07	0.28
MP2(full)/6–31+G(d)	0.13	0.15	0.05	0.03	0.24
MP2(full)/6–31+G(d,p)	0.23	0.17	0.42	0.04	0.21
MP2(full)/6–311+G(d,p)	0.00	0.00	0.00	0.00	0.00

<sup>a</sup> Proton affinities are calculated as  $E_a(\text{conjugate base}) - E_a(\text{conjugate acid})$ .

first shell solvent effects, and it has been shown to give aqueous solvation free energies accurate to within ~2 kJ/mol for neutral species.<sup>91</sup> The SM6 free energy of solvation is calculated on the gas-phase geometries at the B3LYP level using the GAMESSPLUS program.<sup>93</sup>

As noted above, the pK<sub>a</sub> values for the carbon acids in this assessment study have been calculated using the direct method and the proton exchange method. In the direct method, we have used the most recent experimental-theoretical values of –26.3 kJ/mol<sup>38</sup> and –1112.5 kJ/mol<sup>94</sup> for the gas-phase Gibbs free energy of H<sup>+</sup>, G<sup>o</sup>(g, H<sup>+</sup>), and solvation energy of H<sup>+</sup> in water, ΔG<sub>s</sub><sup>o</sup>(H<sup>+</sup>). Calculation of the gas-phase energies are for a standard state of 1 atm, but solvation energies use a standard state of 1 mol/L; therefore, the value of 7.9 kJ/mol which corresponds to RTln(24.46) has been added to gas-phase energies in Scheme 1.

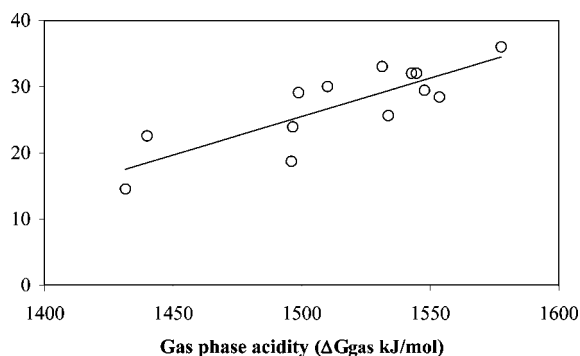
## 4. Results

**Gas-Phase Proton Affinities.** Although the largest source of error in pK<sub>a</sub> calculations is likely to be the treatment of solvation effects (and how this error is mitigated through the pK<sub>a</sub> calculation method), it is nonetheless necessary to ensure that the chosen electronic-structure procedures used are capable of delivering accurate gas-phase proton affinities in a cost-effective manner. To this end, we first investigated the effect of the level of theory used in the geometry optimization on the proton affinities for a small selection of neutral, cationic, and anionic carbon acids. Table 1 shows the proton affinities as calculated at a consistent level of theory, using geometries obtained using HF, B3LYP, BMK, and MP2(full) in conjunction with a variety of basis sets. As shown in Table 1, the proton affinities calculated on these geometries appear to be relatively insensitive to the level of theory since significant error cancellation is expected in the reaction energies. Even at the lowest level of theory studied, RHF/6–31+G(d) geometries are generally within 2 kJ/mol of the corresponding MP2(full)/6–311+G(d,p) level, though there are some exceptions to this. The lowest cost procedure that consistently delivered proton affinities within 1 kJ/mol of the corresponding benchmark level was BMK/6–31+G(d), hence this procedure has been adopted for the remainder of the study.

**Table 2.** Experimental and Calculated<sup>a</sup> Gas-Phase Free Energies (kJ/mol) of the Deprotonation Reaction at 298 K

level of theory	acetamide (N1)	N-methylpyrazole	N-methylimidazole	1,3-dimethyluracil <sup>b</sup>
G3MP2//BMK	1551.8 (24.6)	1574.8 (2.1)	1583.5 (7.2)	1540.3 (3.6)
G3MP2+//BMK	1553.7 (26.5)	1576.5 (0.4)	1585.0 (5.7)	1540.3 (3.6)
G3//BMK	1554.1 (26.9)	1577.2 (0.3)	1586.6 (4.1)	1541.7 (2.2)
G3+//BMK	1555.3 (28.1)	1577.6 (0.7)	1586.3 (4.4)	1542.1 (1.8)
CBS-QBMK	1554.9 (27.7)	1577.9 (1.0)	1585.6 (5.1)	1539.6 (4.3)
W1/BMK	1554.0 (26.8)	—	—	—
experimental	1527.2 ± 12.6 <sup>95</sup>	1576.9 ± 2.9 <sup>100</sup>	1590.7 ± 4.2 <sup>100</sup>	1543.9 ± 8.4 <sup>101</sup>

<sup>a</sup> Based on BMK/6-31+G(d) optimized geometries and BMK/6-31+G(d,p) scaled frequencies. Absolute deviation from experiment in parentheses. <sup>b</sup>  $\Delta H_{\text{acid}}$ .



**Figure 2.** The correlation between gas-phase (G3MP2+//BMK/6-31+G\*) and aqueous acidities of neutral carbon acids. Least-squares correlation  $pK_a = 0.12\Delta G_{\text{gas}} - 148.55$  ( $r = 0.68$ ).

Using the BMK/6-31+G(d) geometries, we then investigated the accuracy of the various ab initio procedures for the calculation of the gas-phase acidities. The calculated and experimental gas-phase proton affinities (as free energies) are provided in Table 2. The free energies computed using the various composite methods are all in good agreement with one another. Furthermore, with the exception of **N1**, the calculated values are generally within the error margins of the experimental values. It has been acknowledged by Hare et al.<sup>95</sup> that the true value for **N1** is likely to be on the high side of their experimental value, which would be in accordance with our calculated values. To verify this, a more accurate W1<sup>96</sup> calculation has been performed on **N1**, and there is good agreement with the value calculated from the other composite methods. Thus, on the basis of accuracy and computational cost, we have chosen the G3MP2+//BMK method for calculating the gas-phase free energies of the carbon acids in this study and estimate that the gas-phase errors associated with this level of theory are of the order of 5 kJ mol<sup>-1</sup> (i.e., chemical accuracy).

We were also interested in whether there is any correlation between the gas-phase and solution acidities. As shown in Figure 2, there is a relatively strong linear correlation between gas-phase and aqueous acidities for the neutral carbon acids ( $r^2 = 0.68$  and  $pK_a = 0.12 \cdot \Delta G_{\text{gas}} - 148.6$ ). Of greater interest is that substituent effects on aqueous acidities are also manifested in the gas-phase data. Specifically, the activating effect of the N-acetyl group is clearly seen in **N9** (cf. **N8**) and **N6** (cf. **N5**), where the electron-withdrawing group is expected to stabilize the adjacent carbanionic charge via inductive effects.<sup>36</sup> The experimental  $pK_a$  of **N9**, **N8**, **N6**, and **N5** are 33, 36, 14.5,

and 18.7, and their corresponding gas-phase acidities ( $\Delta G_{\text{gas}}$ ) are 1531, 1577, 1431, and 1496 kJ/mol, respectively. The weaker aqueous acidity of an  $\alpha$ -carbonyl proton in an amide compared to an ester is also mirrored in the gas-phase data. A similarly strong correlation was observed in the cationic systems ( $r^2 = 0.71$ ), although a somewhat weaker correlation was observed in the anionic systems ( $r^2 = 0.33$ ). Presumably, this is due to the divalent anionic conjugate bases of these acids where the solvation energies are expected to dominate the trends in aqueous acidities.

**Solvation Energies and  $pK_a$  Values.** Having identified suitable electronic structure methods for calculating the gas-phase proton affinities, these procedures were then used in conjunction with a variety of solvation models to calculate the corresponding  $pK_a$  values. In calculating the solvation energies and  $pK_a$  values, there are a number of additional variables to consider, including the choice of solvation model, the level of theory at which it is applied, whether or not explicit solvent molecules are included in the calculation, and whether the  $pK_a$  value is calculated via a direct or proton exchange approach. To simplify the experimental design, in this section we consider only the CPCM model and study the effect of level of theory on the (directly calculated)  $pK_a$  values. Then, having selected an appropriate level of theory for the solvation energy calculations, in subsequent sections we explore the effect of solvation model,  $pK_a$  calculation method on the accuracy of the results.

Table 3 shows the  $pK_a$  values for a selection of neutral, anionic, and cationic carbon acids, in which the gas-phase energies were calculated at a consistent high level of theory (G3MP2+//BMK), and solvation energies were calculated using CPCM at various levels of theory. The geometry was fully optimized in the presence of solvent at each of the studied levels of theory. As shown in Table 3, increasing the basis set has a minimal effect on the accuracy of the  $pK_a$  for the neutral, cationic, and anionic carbon acids, and the smaller basis set 6-31+G(d) is sufficient for the solvation free energy calculations. However, there are significant differences between the CPCM results at the HF and B3LYP levels of theory in a number of acids (such as **N6** and **A4**), and thus both methods are retained for the remainder of the study.

The COSMO-RS model in ADF has been parametrized to some extent, and the BP/TZP level of theory was used as recommended.<sup>92</sup> The SM6 is a density functional theory continuum solvation model and can be used in conjunction with any good density functional, including the mPW0, B3LYP, and B3PW91 functionals.<sup>91</sup> As such, the SM6 solvation free energies have been computed at the B3LYP/

**Table 3.** Aqueous  $pK_a$  Values for Selected Neutral, Cationic, and Anionic Carbon Acids Calculated Using CPCM Solvent Models with Various Basis Sets at 298 K<sup>a</sup>

method	level of theory	N1	N2	N6	C1	C2	A1	A3	A4
CPCM	B3LYP/6-31+G(d)	35.6	37.0	21.4	24.6	24.0	43.6	38.9	42.0
CPCM	B3LYP/6-31+G(d,p)	35.7	37.1	21.5	24.8	24.1	43.7	39.1	42.2
CPCM	B3LYP/6-311+G(d,p)	35.4	36.9	21.4	23.5	24.0	42.8	38.5	41.8
CPCM	HF/6-31+G(d)	35.3	36.3	16.9	25.0	24.4	42.5	40.5	39.8
CPCM	HF/6-31+G(d,p)	35.3	36.4	17.1	25.2	24.6	42.0	40.5	39.9
CPCM	HF/6-311+G(d,p)	35.2	36.4	17.2	25.2	24.4	41.7	40.3	39.7

<sup>a</sup> All associated gas-phase calculations performed using the G3MP2+//BMK/6-31+G(d) level of theory.  $pK_a$  values calculated via the direct method. B3LYP calculations used UAKS radii; HF calculations used UAHF radii.

**Table 4.** Calculated<sup>a</sup> and Experimental Aqueous Acid Dissociation Constants at 298 K for Neutral Carbon Acids<sup>b</sup>

carbon acid	direct method				proton exchange method				expt
	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	
N1	35.6 (7.2)	35.3 (6.9)	28.6 (0.2)	28.1 (-0.3)	reference	reference	reference	reference	28.4 ± 0.5 <sup>33</sup>
N2	37.0 (7.6)	36.3 (6.9)	31.7 (2.3)	28.5 (-0.9)	29.8 (0.4)	29.4 (0.0)	31.5 (2.1)	28.8 (-0.6)	29.4 ± 0.5 <sup>33</sup>
N3	35.9 (6.8)	39.3 (10.2)	31.2 (2.1)	29.5 (0.4)	28.6 (-0.5)	32.3 (3.2)	31.0 (1.9)	29.8 (0.7)	29.1 <sup>36</sup>
N4	32.1 (8.2)	35.9 (12.0)	24.7 (0.8)	25.8 (1.9)	24.9 (1.0)	29.0 (5.1)	24.5 (0.6)	26.1 (2.2)	23.9 <sup>36</sup>
N5 <sup>e</sup>	25.6 (6.9)	21.8 (3.1)	22.8 (4.1)	20.5 (1.8)	18.4 (-0.3)	14.9 (-3.8)	22.7 (4.0)	20.8 (2.1)	18.7 <sup>36</sup>
N6	21.4 (6.9)	16.9 (2.4)	16.7 (2.2)	17.7 (3.2)	14.2 (-0.3)	10.0 (-4.5)	16.5 (2.0)	18.0 (3.5)	14.5 <sup>36</sup>
N7	38.4 (8.4)	38.2 (8.2)	35.6 (5.6)	32.5 (2.5)	31.1 (1.1)	31.2 (1.2)	35.5 (5.5)	32.8 (2.8)	~30 <sup>77</sup>
N8	44.3 (8.3)	42.8 (6.8)	48.0 (12.0)	39.9 (3.9)	37.1 (1.1)	35.9 (-0.1)	47.9 (11.9)	40.2 (4.2)	~36 <sup>78</sup>
N9	41.3 (8.3)	40.6 (7.6)	41.3 (8.3)	36.1 (3.1)	34.1 (1.1)	33.7 (0.7)	41.2 (8.2)	36.4 (3.4)	~33 <sup>78</sup>
N10 <sup>e</sup>	41.7 (9.7)	41.2 (9.2)	40.2 (8.2)	36.2 (4.2)	34.4 (2.4)	34.3 (2.3)	40.1 (8.1)	36.5 (4.5)	32 ± 2 <sup>77</sup>
N11 <sup>e</sup>	38.1 (6.1)	35.7 (3.7)	39.1 (7.1)	36.0 (4.0)	30.8 (-1.2)	28.8 (-3.2)	38.9 (6.9)	36.3 (4.3)	32 ± 2 <sup>77</sup>
N12	33.9 (8.3)	33.4 (7.8)	28.3 (2.7)	24.7 (-0.9)	26.6 (1.0)	26.4 (0.8)	28.2 (2.6)	25.0 (-0.6)	25.6 ± 0.5 <sup>30</sup>
N13	29.7 (7.1)	27.9 (5.3)	23.4 (0.8)	18.8 (-3.8)	22.4 (-0.2)	21.0 (-1.6)	23.2 (0.6)	19.1 (-3.5)	22.6 <sup>79</sup>
AD <sub>max</sub>	9.7	12.0	12.0	4.2	2.4	5.1	11.9	4.5	-
MAD	7.7	6.9	4.3	2.4	0.9	2.2	4.5	2.7	-

<sup>a</sup> Signed errors are shown in brackets. <sup>b</sup> All associated gas-phase calculations performed using the G3MP2+//BMK/6-31+G(d) level of theory. All solvation energy calculations performed at the B3LYP/6-31+G(d) level of theory unless noted otherwise. <sup>c</sup> Solvation energy calculations performed at the HF/6-31+G(d) level of theory. <sup>d</sup> Solvation energy calculations performed at the BP/TZP level of theory. <sup>e</sup> MP2/6-31+G(d) geometries, frequencies, and corresponding scale factors were used due to convergence problems during geometry optimization.

**Table 5.** Calculated<sup>a</sup> and Experimental Aqueous Acid Dissociation Constants at 298 K for Cationic Carbon Acids<sup>b</sup>

carbon acid	direct method				proton exchange method				expt
	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	
C1	24.6 (3.6)	25.0 (4.0)	27.2 (6.2)	26.2 (5.2)	reference	reference	25.3 (4.3)	reference	21 ± 1 <sup>34</sup>
C2	24.0 (6.0)	24.4 (6.4)	19.9 (1.9)	24.5 (6.5)	20.4 (2.4)	20.5 (2.5)	reference	19.3 (1.3)	18 ± 1 <sup>34</sup>
C3	18.9 (4.9)	17.4 (3.4)	13.5 (-0.5)	17.2 (3.2)	15.3 (1.3)	13.5 (-0.5)	11.6 (-2.4)	12.0 (-2.0)	14 ± 1 <sup>22</sup>
C4	23.7 (2.7)	24.5 (3.5)	25.5 (4.5)	28.4 (7.4)	20.1 (-0.9)	20.6 (-0.4)	23.6 (2.6)	23.3 (2.3)	21 ± 1 <sup>75</sup>
C5	24.4 (0.6)	22.0 (-1.8)	26.6 (2.8)	30.9 (7.1)	20.8 (-3.0)	18.0 (-5.8)	24.7 (0.9)	25.7 (1.9)	23.8 ± 0.5 <sup>29</sup>
C6	24.4 (1.1)	24.4 (1.4)	24.1 (1.1)	29.7 (6.7)	20.5 (-2.5)	20.5 (-2.5)	22.2 (-0.8)	24.6 (1.6)	23.0 ± 0.5 <sup>29</sup>
C7	21.6 (0.0)	22.3 (0.7)	21.4 (-0.2)	24.9 (3.3)	18.0 (-3.6)	18.4 (-3.2)	19.5 (-2.1)	19.8 (-1.8)	21.6 ± 0.5 <sup>29</sup>
C8	28.8 (-4.2)	28.6 (-4.4)	29.0 (-4.0)	32.3 (-0.7)	25.2 (-7.8)	24.6 (-8.4)	27.1 (-5.9)	27.2 (-5.8)	33 ± 2 <sup>76</sup>
C9	30.8 (-3.2)	30.8 (-3.2)	32.1 (-1.9)	36.9 (2.9)	27.2 (-6.8)	26.9 (-7.1)	30.2 (-3.8)	31.7 (-2.3)	34 ± 2 <sup>76</sup>
C10	29.8 (-3.2)	29.6 (-3.4)	31.3 (-1.7)	36.7 (3.7)	26.2 (-6.8)	25.6 (-7.4)	29.4 (-3.6)	31.5 (-1.5)	33 ± 2 <sup>76</sup>
AD <sub>max</sub>	6.0	6.4	6.2	7.4	7.8	8.4	5.9	5.8	-
MAD	2.9	3.2	2.5	4.7	3.9	4.2	2.9	2.3	-

<sup>a</sup> Signed errors are shown in brackets. <sup>b</sup> All associated gas-phase calculations performed using the G3MP2+//BMK/6-31+G(d) level of theory. All solvation energy calculations performed at the B3LYP/6-31+G(d) level of theory unless noted otherwise. <sup>c</sup> Solvation energy calculations performed at the HF/6-31+G(d) level of theory. <sup>d</sup> Solvation energy calculations performed at the BP/TZP level of theory.

6-31+G(d) level of theory. Having selected appropriate levels of theory for the gas- and solution-phase calculations, we now examine the effects of solvation model and  $pK_a$  calculation method on the accuracy of the results for the neutral, cationic, and anionic acids. These results, including mean absolute deviation (MAD), maximum absolute deviations (AD<sub>max</sub>), and signed errors (in parentheses), are presented in Tables 4, 5 and 7, respectively. It should be noted that the MAD for the proton-exchange  $pK_a$  values also

provides a measure of the accuracy of the *relative* values of  $pK_a$ , as obtained by the direct method.

## 5. Discussion

**Neutral Carbon Acids.** Shown in Table 4 are the calculated acid dissociation constants for the neutral carbon acids using the direct and proton exchange methods. In the latter approach, acetamide was chosen as the reference acid.

**Table 6.** Multipole Derived Atomic Charges Computed at the BP/TZP Level of Theory

carbon acid	atomic charge <sup>a</sup>	formal charge	carbon acid	atomic charge <sup>b</sup>	formal charge
N1	+0.33	-1	C1	+0.6	+1
N2	+0.28	-1	C2	-0.53	+1
N3	+0.32	-1	C3	-0.15	+1
N4	+0.33	-1	C4	+0.25	+1
N5	+0.30	-1	C5	-0.10	+1
N6	+0.27	-1	C6	-0.42	+1
N7	-0.27	-1	C7	-0.46	+1
N8	-0.36	-1	C8	-0.45	+1
N9	-0.32	-1	C9	-0.42	+1
N10	-0.29	-1	C10	-0.42	+1
N11	-0.34	-1	-	-	-
N12	+0.36	-1	-	-	-
N13	+0.27	-1	-	-	-

<sup>a</sup> Atomic charge on the  $\alpha$  carbon of the conjugate base of the neutral carbon acid. <sup>b</sup> Atomic charge on the nitrogen adjacent to the  $\alpha$  carbon in the cationic acid.

Using the direct method, the mean absolute deviation (MAD) from experiment for CPCM-UAKS and CPCM-UAHF are in excess of 7  $pK_a$  units, which is unacceptably large. On the other hand, the COSMO-RS and, in particular, the SM6 model perform significantly better with MADs of 4.3 and 2.4, respectively.

As mentioned earlier, gas-phase free energies calculated using high-level composite methods have an intrinsic error of about 5 kJ/mol. On the other hand, a recent assessment study conducted by Houk et al. also showed that the MADs in the CPCM free energy of solvation are  $\sim 5$  kJ/mol and  $\sim 16$  kJ/mol for neutral and anionic species, respectively.<sup>52</sup> Assuming that these errors are additive, a crude estimate of the error in  $\Delta G_{\text{soln}}$  for the overall reaction, as calculated via the direct method, would be about 25 kJ/mol energy or approximately 5  $pK_a$  units. On this basis, we note that the MADs in Table 4 are in the right range.

It is interesting to note that COSMO-RS performed quite poorly for carbon acids **N7** to **N11** ( $AD > 6$ ) but otherwise performs rather well, with calculated values within  $\sim 2$  units of experiment. The conjugate bases of these acids are somewhat different due to the absence of an adjacent carbonyl group resulting in a highly localized anionic charge at the  $\alpha$  carbon. Very recently, Eckert and co-workers have attempted to use the COSMO-RS model for calculating the  $pK_a$  of carbon acids in acetonitrile.<sup>97</sup> The authors noted that the COSMO-RS model performed better on acids that produced anions with delocalized charges as they are less affected by solvation. Conversely, the short-range interactions in the solvation of anions with localized charges are not fully accounted for in the COSMO-RS model.<sup>97</sup> In this light, we examined the atomic charge on the anionic carbon, which has a formal charge  $-1$ , using multipole derived charge analysis<sup>98</sup> (as implemented in ADF) to assess the degree of charge delocalization in these species. These charges are tabulated in Table 6. As shown, the atomic charges on the anionic carbon in the conjugate bases of **N7** to **N11** are negative, whereas the corresponding charges in the other acids are  $\sim +0.3$ , indicating a high degree of charge delocalization in the latter and further supports the argument by Eckert and co-workers.

In the proton exchange method, marked improvement in the MADs was observed for all of the solvation models, except COSMO-RS and SM6, where the MAD is similar in both approaches. The performance of the CPCM-UAKS model is most noteworthy, with an MAD and  $AD_{\text{max}}$  of about 1 and 2.5  $pK_a$  units, respectively. Moreover, the neutral carbon acids studied here included both cyclic and acyclic systems with a range of functionalities (amides, amines, and ketones), and the performance of the CPCM-UAKS model was relatively insensitive to these structural variations. Thus, when used in conjunction with a proton exchange scheme, this model should provide a useful strategy for accurate  $pK_a$  calculation of a wide range of neutral carbon acids in biological systems as well as in chemical synthesis.

The success of the proton exchange scheme relies on several factors. First, the accuracy of the experimental  $pK_a$  of the reference acid is critical since any errors in the experimental data would propagate into the calculations. Acetamide was chosen as the reference acid for the present study as it is the smallest neutral carbon acid and has a relatively small experimental uncertainty of  $\pm 0.5$   $pK_a$  units. Second, the errors in the solvation model must be systematic, i.e. it needs to consistently over- or underestimate the experimental values so as to allow for optimal error cancellation. In practice, this would involve choosing a reference acid that is structurally similar to the carbon acid of interest. For example, one would not choose a cationic acid as a reference for a neutral carbon acid since the magnitude and sign of the errors incurred by the continuum solvation model is likely to be very different for the two species. As shown in Figure 3, there is a very strong correlation ( $r^2 = 0.98$ ) between the experimental  $pK_a$  and those calculated using the direct approach (CPCM-UAKS). The unsigned errors in Table 4 also indicate that this approach systematically overestimates the  $pK_a$ s of the neutral carbon acids. As such, a marked improvement is obtained using the proton exchange scheme. On the other hand, using the COSMO-RS and SM6 models, the errors in the direct method are less systematic, and, depending on the choice of reference, the use of a proton exchange reaction reduces the errors in only some of the species of the test set.

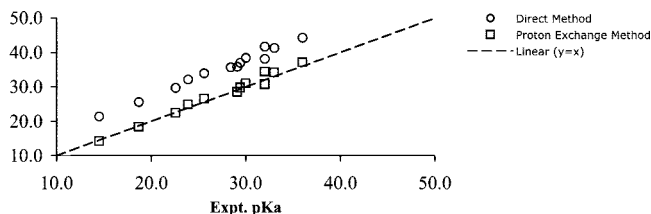
**Cationic Carbon Acids.** Shown in Table 5 are the calculated acid dissociation constants for the cationic acids using the direct and proton exchange methods, respectively. In contrast to the neutral acids, inspection of the MAD values reveal that the direct method performs better for cationic systems, with MADs generally under 3  $pK_a$  units across the various solvation models. Unfortunately, the method still fails in several cases, such as in carbon acid **C2**, which accounts for the  $AD_{\text{max}}$  of  $> 5$   $pK_a$  units in the CPCM models. Closer inspection of the data reveals that the smaller MAD observed in these systems is partly due to the good agreement between the calculated and experimental values for carbon acids **C5**, **C6**, and **C7**, where the agreement is generally within 1  $pK_a$  unit. It is also interesting to note that the good performance of the SM6 model on neutral systems is not reflected in the cationic carbon acids.

The COSMO-RS model, which performed best for these species, has an MAD and  $AD_{\text{max}}$  of 2.5 and 6.2, respectively.

**Table 7.** Calculated<sup>a</sup> and Experimental Aqueous Acid Dissociation Constants at 298 K for Anionic Carbon Acids<sup>b</sup>

carbon acid	direct method				proton exchange method				expt
	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	CPCM/UAKS	CPCM/UAHF <sup>c</sup>	COSMO-RS <sup>d</sup>	SM6	
A1	43.6 (9.6)	42.5 (8.5)	35.4 (1.4)	26.1 (-7.9)	35.0 (1.0)	32.3 (-1.7)	38.8 (4.8)	31.7 (-2.3)	~34 <sup>20</sup>
A2	26.4 (9.4)	23.5 (6.5)	11.8 (-5.2)	12.9 (-4.1)	17.8 (0.8)	13.3 (-3.7)	15.1 (-1.9)	18.6 (-1.6)	17 <sup>102</sup>
A3	38.9 (8.6)	40.5 (10.2)	26.9 (-3.4)	24.6 (-5.7)	reference	reference	reference	reference	30.3 <sup>36</sup>
A4	42.0 (11.2)	39.8 (9.0)	30.7 (-0.1)	30.5 (-0.3)	33.4 (2.6)	29.6 (-1.2)	34.1 (3.3)	36.1 (5.3)	30.8 <sup>36</sup>
A5	44.9 (11.4)	42.3 (8.8)	36.7 (3.2)	20.6 (-12.9)	36.3 (2.8)	32.1 (-1.4)	40.1 (-6.6)	26.3 (-7.2)	33.5 <sup>33</sup>
AD <sub>max</sub>	11.4	10.2	5.2	12.9	2.8	3.7	6.6	7.2	-
MAD	10.1	8.6	2.7	6.2	1.8	2.0	4.2	4.1	-

<sup>a</sup> Signed errors are shown in brackets. <sup>b</sup> All associated gas-phase calculations performed using the G3MP2+//BMK/6-31+G(d) level of theory. All solvation energy calculations performed at the B3LYP/6-31+G(d) level of theory unless noted otherwise. <sup>c</sup> Solvation energy calculations performed at the HF/6-31+G(d) level of theory. <sup>d</sup> Solvation energy calculations performed at the BP/TZP level of theory.



**Figure 3.** The correlation between experimental and calculated (direct and proton exchange methods using CPCM/UAKS model) aqueous acidities of neutral carbon acids at 298 K. Least-squares correlation for (a) direct method:  $pK_a(\text{Calc}) = 1.06pK_a(\text{Expt}) + 6.03$ ;  $r^2 = 0.98$  and (b) proton exchange method:  $pK_a(\text{Calc}) = 1.06pK_a(\text{Expt}) - 1.19$ ;  $r^2 = 0.98$ .

Furthermore, we observed that the method performed rather well for most of the carbon acids with the exception of **C1** and **C4**, which might be related to charge distribution in these systems. The atomic charges for the nitrogen with a formal +1 charge are tabulated in Table 6 where it is seen that the cationic charge is strongly localized in **C1** and **C4**; all the other systems exhibit negative charges on the nitrogen adjacent to the  $\alpha$  carbon. The poorer performance of the COSMO-RS in these two carbon acids is thus consistent with our earlier observations in the neutral systems. Accordingly, it appears that neutral and cationic carbon acids are amenable to moderately accurate  $pK_a$  calculations ( $\sim 2$   $pK_a$  unit of experiment) using the COSMO-RS solvation model provided the charges on these systems are ‘sufficiently delocalized’. Admittedly, a consistent and direct measurement of charge delocalization is required for the appropriate application of the COSMO-RS model. In this work, we have used the *sign* of the atomic charge for this purpose; however, more extensive studies need to be carried out to establish the general applicability of this approach.

It should also be brought to the readers’ attention that the discrimination between localized and delocalized systems is not observed in the other solvation models. This may be attributed to the different approaches adopted by the various models to account for explicit intermolecular interactions. For example, the CPCM-UAHF model indirectly accounts for these effects by utilizing cavity radii optimized at the HF/6-31G(d) level of theory,<sup>99</sup> whereas the SM6 model uses different parameters such as atomic surface tensions and a different set of atomic radii that are fitted against a much larger data set of experimental solvation free energies.<sup>91</sup> In this light, we have chosen glycine methyl ester (**C1**) as the

reference acid for the CPCM and SM6 models and betaine methyl ester (**C2**) for the COSMO-RS model in the proton exchange approach.

As mentioned earlier, the success of the proton exchange scheme depends on the nature of the errors incurred in the solvation models. Inspection of the signed errors in direct approach suggest that these errors are specific to the functionality of the carbon acid; the CPCM model overestimates the acidity constants for **C1** to **C4** (amino acids) and underestimates the values for **C8** to **C10** (pyrimidiums). This observation is in contrast to the neutral systems. Accordingly, the proton exchange method has resulted in an increase in the MAD of the CPCM solvation models. Specifically, the choice of glycine methyl ester, **C1**, as the reference acid resulted in reasonably accurate results for carbon acids **C2**, **C3**, and **C4** but led to deviations that are much larger than those encountered using the direct method for the remaining carbon acids **C5** to **C10**.

On the other hand, the SM6 model appears to perform better in the proton exchange scheme. In the direct method, the SM6 method consistently overestimates the experimental  $pK_a$ s, and the proton exchange reaction ameliorates some of this error, reducing the MAD from 4.7 to  $\sim 2$   $pK_a$  unit. The performance of the COSMO-RS model in the exchange scheme is less satisfactory because of its sensitivity to the charge distribution in these acids. On the basis of these results, the combination of the SM6 model and a proton exchange scheme is most likely to give the best results.

**Anionic Carbon Acids.** The data for the anionic acids are summarized in Table 7. The performance of the direct method is clearly unsatisfactory for these systems. In particular, the CPCM model consistently overestimates the  $pK_a$  by  $\sim 10$ , suggesting that the solvation free energies of divalent anions are significantly underestimated. Interestingly, despite the high charge density of the divalent anions, the COSMO-RS model fared reasonably well, with MAD of  $\sim 3$ , although its performance was less consistent, as was the SM6 model.

In the proton exchange scheme, **A3** was used as the reference acid. Not surprisingly, there is an enormous improvement in the CPCM values, where the MADs were reduced to 2 or less. This result is encouraging in view of the much larger errors incurred by continuum models for multi-valent ions. For reasons explained earlier, the proton exchange scheme was less effective when used in combination with the COSMO-RS and SM6 models.

**Table 8.** Effect of Adding an Explicit Water Molecule<sup>a</sup> on Accuracy of Calculated Acid Dissociation Constants of Anionic Carbon Acids<sup>b</sup>

carbon acid	$n = 0$	$n = 1$	expt
A1	26.1 (7.9)	33.3 (0.7)	34 <sup>20</sup>
A2	12.9 (4.1)	16.5 (0.5)	17 <sup>102</sup>
A3	24.6 (5.7)	28.6 (1.7)	30.3 <sup>36</sup>
A4	30.5 (0.3)	33.9 (3.1)	30.8 <sup>36</sup>
A5	20.6 (12.9)	28.6 (4.9)	33.5 <sup>33,79</sup>
AD <sub>max</sub>	12.9	4.9	–
MAD	6.2	2.2	–

<sup>a</sup>  $n$  = number of water molecules. <sup>b</sup> All associated gas-phase calculations performed using the G3MP2+//BMK/6–31+G(d) level of theory.  $pK_a$  values calculated via the direct method. All solvation energy calculations performed using SM6 at the B3LYP/6–31+G(d) level of theory.

Given these problems, we examined whether the results could be improved through the inclusion of an explicit water molecule as shown in Scheme 2. Previously, Kelly and co-workers have reported some success using this approach for the divalent carbonate anion.<sup>48</sup> The configuration of the aqua-complexes has been chosen such that the water molecule is hydrogen bonded to the carbonyl oxygen and/or the  $\alpha$  carbon where the anionic charge is likely to reside based on resonance structures. Where several configurations are possible, the lowest energy gas-phase conformer was used.

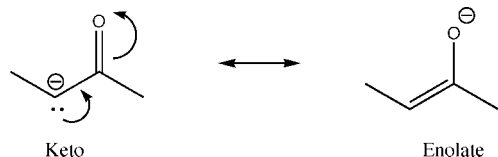


Table 8 shows the results obtained for the anionic carbon acids using the combined SM6 implicit-explicit solvent approach (Scheme 2). As shown, adding a single water molecule to the anion significantly improves the calculated values, where the MAD is reduced to 2.2 as compared to 6.2 for the direct method. Despite the promise of this approach, there are other issues relating to the number of water molecules to add and the conformational sampling problems associated with larger ion-clusters. In particular, we note that this approach led to a larger deviation for carbon acids **A4** and **A5**, and the agreement is likely to worsen with the addition of more water molecules.

## 6. Conclusions

There are a wide variety of carbon acids in biological systems, including amino acids, peptides, esters, and ketones, which exist in neutral, charged, and zwitterionic forms. As such, the computation of accurate  $pK_a$  values of these acids poses a serious challenge for continuum models since their solvation patterns are very different. We find that gas-phase acidities can be accurately obtained using G3MP2+//BMK/6–31+G(d); however, solvation energies are subject to much larger errors. In particular, the direct approach (Scheme 1) yields unacceptably large errors for all three categories of carbon acids in this study.

Nevertheless, the  $pK_a$  values of *neutral* carbon acids can be accurately obtained to within  $\sim 1$  unit of experiment via

the combination of a proton exchange scheme with the CPCM-UAKS solvation model. Furthermore, the accuracy of this approach is not sensitive to the structure of the (neutral) reference acid and should therefore be useful for the  $pK_a$  calculation of a wide range of neutral carbon acids. Alternatively, moderately accurate results may be obtained through the direct approach using the SM6 and COSMO-RS solvent models.

Ionic carbon acids are more problematic, where the success of the proton exchange scheme is highly sensitive to the choice of reference acid. This limits the applicability of this approach for studying charged carbon acids. For cationic systems, the SM6 model combined with a proton exchange scheme delivered the best results. For anionic acids, the combination of the proton exchange scheme with the CPCM-UAKS model also gave reasonably good results, and the addition of an explicit water molecule using the SM6 model significantly improved the computed  $pK_a$ s for anionic acids ( $\sim 3$  fold reduction in MAD) compared to using the direct method.

Admittedly, the  $pK_a$  calculation strategies that have emerged from this work are somewhat *ad hoc*, as they do not directly address the problems associated with the solvation of ionic species. Nevertheless, it is also intended that this work helps to identify limitations of present continuum solvation models and to spur further research aimed at improving the presented results. In this regard we note that the COSMO-RS model provided the best overall performance in the direct method  $pK_a$  calculations in all three classes of carbon acids, although there are problems associated with ionic species having highly localized charges. As noted before, the COSMO-RS model is a more sophisticated variant of the CPCM model and takes the real character of the solvent (rather than a simple continuum) into account. Thus, its success could possibly indicate the importance of explicit consideration of real character of the solvent in the future development of solvation models beyond the continuum approximation.

**Acknowledgment.** We gratefully acknowledge support from the Australian Research Council under their Centres of Excellence program and generous allocations of computing time on the National Facility of the Australian Partnership for Advanced Computing.

**Supporting Information Available:** Complete BMK/6–31+G(d) optimized gas-phase geometries of all species and their associated conformers (Table S1), corresponding electronic energies and solvation energies (Table S2), and COSMO-RS solvation energies computed using ADF and corresponding values from ref 90 is also available (Table S3). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Man, E. H.; Sandhouse, M. E.; Burg, J.; Fisher, G. H. *Science* **1983**, *220*, 1407.
- (2) Helfman, P. M.; Bada, J. L. *Nature* **1976**, *262*, 279.
- (3) Helfman, P. M.; Bada, J. L. *Proc. Natl. Acad. Sci., U.S.A.* **1975**, *72*, 297.

- (4) Bada, J. L.; Kvenvolden, K. A.; Peterson, E. *Nature* **1973**, *245*, 308.
- (5) Masters, P. M.; Bada, J. L.; Zigler, J. S. *Nature* **1977**, *262*, 71.
- (6) Liardon, R.; Friedman, M. *J. Agric. Food Chem.* **1987**, *35*, 661.
- (7) Liardon, R.; Ledermann, S. *J. Agric. Food Chem.* **1986**, *34*, 557.
- (8) Auvynet, C.; Seddiki, N.; Dunia, I.; Nicolas, P.; Amiche, M.; Lacombe, C. *Eur. J. Cell Biol.* **2006**, *85*, 25.
- (9) Jilek, A.; Mollay, C.; Tippelt, C.; Grassi, J.; Mignogna, G.; Muellegger, J.; Sander, V.; Fehrer, C.; Barra, D.; Kreil, G. *Proc. Natl. Acad. Sci., U.S.A.* **2005**, *102*, 4235.
- (10) Kreil, G. *Annu. Rev. Biochem.* **1997**, *66*, 337.
- (11) Kreil, G. *Science* **1994**, *266*, 996.
- (12) Montecucchi, P. C.; de Castiglione, R.; Piani, S.; Gozzini, L.; Erspamer, V. *Int. J. Pept. Protein Res.* **1981**, *17*, 275.
- (13) Torres, A. M.; Menz, I.; Alewood, P. F.; Bansal, P.; Lahnstein, J.; Gallagher, C. H.; Kuchel, P. W. *FEBS Lett.* **2002**, *524*, 172.
- (14) Torres, A. M.; Tsampazi, C.; Geraghty, D. P.; Bansal, P. S.; Alewood, P. F.; Kuchel, P. W. *Biochem. J.* **2005**, *391*, 215.
- (15) Torres, A. M.; Tsampazi, M.; Tsampazi, C.; Kennett, E. C.; Belov, K.; Geraghty, D. P.; Bansal, P. S.; Alewood, P. F.; Kuchel, P. W. *FEBS Lett.* **2006**, *580*, 1587.
- (16) Tanner, M. E. *Acc. Chem. Res.* **2002**, *35*, 237.
- (17) Richard, J. P. *J. Am. Chem. Soc.* **1984**, *106*, 4926.
- (18) Richard, J. P. *Biochemistry* **1998**, *37*, 4305.
- (19) Richard, J. P.; Amyes, T. L. *Bioorg. Chem.* **2004**, *32*, 354.
- (20) Richard, J. P.; Amyes, T. L. *Curr. Opin. Chem. Biol.* **2001**, *5*, 626.
- (21) Richard, J. P.; Amyes, T. L.; Toteva, M. M. *Acc. Chem. Res.* **2001**, *34*, 981.
- (22) Rios, A.; Crueiras, J.; Amyes, T. L.; Richard, J. P. *J. Am. Chem. Soc.* **2001**, *123*, 7949.
- (23) Sievers, A.; Wolfenden, R. *J. Am. Chem. Soc.* **2002**, *124*, 13986.
- (24) Bull, S. D.; Davies, S. G.; Garner, A. C.; Parkes, A. L.; Roberts, P. M.; Sellers, T. G. R.; Smith, A. D.; Tamayo, J. A.; Thomson, J. E.; Vickers, R. J. *New J. Chem.* **2007**, *31*, 486.
- (25) Davies, S. G.; Garner, C. A.; Ouzman, J. V. A.; Roberts, P. M.; Smith, A. D.; Snow, E. J.; Thomson, J. E.; Tamayo, J. A.; Vickers, R. J. *Org. Biomol. Chem.* **2007**, *5*, 2138.
- (26) Davies, S. G.; Rodriguez-Solla, H.; Tamayo, J. A.; Garner, A. C.; Smith, A. D. *Chem. Commun.* **2004**, 2502.
- (27) Matthews, W. S.; Bares, J. E.; Bartmess, J. E.; Bordwell, F. G.; Cornforth, F. J.; Drucker, G. E.; Margolin, Z.; McCallum, R. J.; McCollum, G. J.; Vanier, N. R. *J. Am. Chem. Soc.* **1975**, *97*, 7006.
- (28) Taft, R. W.; Bordwell, F. G. *Acc. Chem. Res.* **1988**, *21*, 463.
- (29) Amyes, T. L.; Diver, S. T.; Richard, J. P.; Rivas, F. M.; Toth, K. *J. Am. Chem. Soc.* **2004**, *126*, 4366.
- (30) Amyes, T. L.; Richard, J. P. *J. Am. Chem. Soc.* **1996**, *118*, 3129.
- (31) Amyes, T. L.; Richard, J. P. *J. Am. Chem. Soc.* **1992**, *114*, 10297.
- (32) Richard, J. P.; Williams, G.; Gao, J. *J. Am. Chem. Soc.* **1999**, *121*, 715.
- (33) Richard, J. P.; Williams, G.; O'Donoghue, A. C.; Amyes, T. L. *J. Am. Chem. Soc.* **2002**, *124*, 2957.
- (34) Rios, A.; Amyes, T. L.; Richard, J. P. *J. Am. Chem. Soc.* **2000**, *122*, 9373.
- (35) Rios, A.; Richard, J. P. *J. Am. Chem. Soc.* **1997**, *119*, 8375.
- (36) Rios, A.; Richard, J. P.; Amyes, T. L. *J. Am. Chem. Soc.* **2002**, *124*, 8251.
- (37) Crueiras, J.; Rios, A.; Riveiros, E.; Amyes, T. L.; Richard, J. P. *J. Am. Chem. Soc.* **2008**, *130*, 2041.
- (38) Liptak, M. D.; Shields, G. C. *J. Am. Chem. Soc.* **2001**, *123*, 7314.
- (39) Liptak, M. D.; Shields, G. C. *Int. J. Quantum Chem.* **2001**, *85*, 727.
- (40) Toth, A. M.; Liptak, M. D.; Phillips, D. L.; Shields, G. C. *J. Chem. Phys.* **2001**, *114*, 4595.
- (41) Namazian, M.; Kalantary-Fotooh, F.; Noorbala, M. R.; Searles, D. J.; Coote, M. L. *THEOCHEM* **2006**, *758*, 275.
- (42) Namazian, M.; Heidary, H. *THEOCHEM* **2003**, *620*, 257.
- (43) Magill, A. M.; Cavell, K. J.; Yates, B. F. *J. Am. Chem. Soc.* **2004**, *126*, 8717.
- (44) Kallies, B.; Mitzner, R. *J. Phys. Chem. B* **1997**, *101*, 2959.
- (45) Lopez, X.; Schaefer, M.; Dejaegere, A.; Karplus, M. *J. Am. Chem. Soc.* **2002**, *124*, 5010.
- (46) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. *J. Am. Chem. Soc.* **2002**, *124*, 6421.
- (47) Klicic, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. *J. Phys. Chem. A* **2002**, *106*, 1327.
- (48) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 2493.
- (49) Almerindo, G. I.; Tondo, D. W.; Pliego, J. R., Jr. *J. Phys. Chem. A* **2004**, *108*, 166.
- (50) Chipman, D. M. *J. Phys. Chem. A* **2002**, *106*, 7413.
- (51) Fu, Y.; Liu, L.; Li, R.-Q.; Liu, R.; Guo, Q.-X. *J. Am. Chem. Soc.* **2004**, *126*, 814.
- (52) Takano, Y.; Houk, K. N. *J. Chem. Theory Comput.* **2005**, *1*, 70.
- (53) Brinck, T.; Larsen, A. G.; Madsen, K. M.; Daasbjerg, K. *J. Phys. Chem. B* **2000**, *104*, 9887.
- (54) Gao, D.; Wong, P. K.; Maddalena, D.; Hwang, J.; Walker, H. *J. Phys. Chem. A* **2005**, *109*, 10776.
- (55) Smith, B. J.; Radom, L. *J. Phys. Chem.* **1991**, *95*, 10549.
- (56) Smith, B. J.; Radom, L. *J. Am. Chem. Soc.* **1993**, *115*, 4885.
- (57) Smith, B. J.; Radom, L. *Chem. Phys. Lett.* **1994**, *231*, 345.
- (58) Martin, J. M.; Lee, T. J. *Chem. Phys. Lett.* **1996**, *258*, 136.
- (59) Ochterski, J. W.; Petersson, G. A.; Wiberg, K. B. *J. Am. Chem. Soc.* **1995**, *117*, 11299.
- (60) Wiberg, K. B. *J. Org. Chem.* **2002**, *67*, 4787.
- (61) Peterson, K. A.; Xantheas, S. S.; Dixon, D. A.; Dunning, T. H. *J. Phys. Chem. A* **1998**, *102*, 2449.

- (62) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- (63) Remko, M. *J. Phys. Chem. A* **2002**, *106*, 5005.
- (64) Pokin, E. K.; Liptak, M. D.; Feldgus, S.; Shields, G. C. *J. Phys. Chem. A* **2001**, *105*, 10483.
- (65) Seo, Y.; Kim, Y.; Kim, Y. *Chem. Phys. Lett.* **2001**, *340*, 186.
- (66) Burk, P.; Koppel, I. A.; Koppel, I.; Leito, I.; Travnikova, O. *Chem. Phys. Lett.* **2000**, *323*, 482.
- (67) Ervin, K. M.; DeTuri, V. F. *J. Phys. Chem. A* **2002**, *106*, 9947.
- (68) Hammerum, S. *Chem. Phys. Lett.* **1999**, *300*, 529.
- (69) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.
- (70) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.
- (71) Chipman, D. M. *J. Chem. Phys.* **2003**, *118*, 9937.
- (72) Namazian, M.; Halvani, S. *J. Chem. Thermodyn.* **2006**, *38*, 1495.
- (73) Namazian, M.; Halvani, S.; Noorbala, M. R. *THEOCHEM* **2004**, *711*, 13.
- (74) Pliego, J. R., Jr.; Riveros, J. M. *J. Phys. Chem. A* **2002**, *106*, 7434.
- (75) Williams, G.; Maziarz, E. P.; Amyes, T. L.; Wood, T. D.; Richard, J. P. *Biochemistry* **2003**, *42*, 8354.
- (76) Wong, F. M.; Capule, C.; Chen, D. X.; Gronert, S.; Wu, W. *Org. Lett.* **2008**, *2008*, 2757.
- (77) Wong, F. M.; Capule, C.; Wu, W. *Org. Lett.* **2006**, *8*, 6019.
- (78) Reutov, O. A.; Beletskaya, I. P.; Butin, K. P. A guide to all existing problems of CH-acidity with new experimental methods and data, including indirect electrochemical, kinetic and thermodynamic studies; CH acids: A guide; Crompton, T. R., Ed.; Pergamon Press Ltd.: New York, 1978.
- (79) Chiang, Y.; Kresge, A. J.; Popik, V. V.; Schepp, N. P. *J. Am. Chem. Soc.* **1997**, *119*, 10203.
- (80) Frisch, M. J. T.; G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.
- (81) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (82) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (83) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.
- (84) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822.
- (85) Merrick, J. P.; Moran, D.; Radom, L. *J. Phys. Chem. A* **2007**, *111*, 11683.
- (86) Klamt, A.; Schueuermann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.
- (87) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.
- (88) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224.
- (89) Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier Science Ltd.: Amsterdam, The Netherlands, 2005.
- (90) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074.
- (91) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.
- (92) (a) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *Chemistry with ADF. J. Comput. Chem.* **2001**, *22*, 931. (b) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391. (c) Pye, C.; Louwen, J. N.; van Lenthe, E. Manuscript in preparation. (d) ADF 2008.01, COSMO-RS, SCM. Available via the Internet at [www.scm.com](http://www.scm.com), accessed October 2008.
- (93) Higtashi, M.; Marenich, A. V.; Olson, R. M.; Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Zhu, T.; Hawkins, G. D.; Chuang, Y.-Y.; Fast, P. L.; Lynch, B. J.; Liotard, D. A.; Rinaldi, D.; Gao, J.; Cramer, C. J.; Truhlar, D. G. GAMESSPLUS - version 2008-2; University of Minnesota, MN, 2008, based on the General Atomic and Molecular Electronic Structure System (GAMESS) as described in Schmidt, M. W.; Baidridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (94) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066.
- (95) Hare, M. C.; Marimanikkuppam, S. S.; Kass, S. R. *Int. J. Mass Spectrom.* **2001**, *210/211*, 153.
- (96) Martin, J. M.; Parthiban, S. In *Quantum Mechanical Prediction of Thermochemical Data*; Cioslowski, J., Ed.; Kluwer-Academic: Dordrecht, The Netherlands, 2001; p 31.
- (97) Eckert, F.; Leito, I.; Kaljurand, I.; Kütt, A.; Klamt, A.; Diedenhofen, M. *J. Comput. Chem.* **2008**, . in press.
- (98) Swart, M.; Duijnen, P. T. v.; Snijders, J. G. *J. Comput. Chem.* **2001**, *22*, 79.
- (99) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.
- (100) Villano, S. M.; Gianola, A. J.; Eyet, N.; Ichino, T.; kato, S.; Bierbaum, V. M.; Lineberger, W. C. *J. Phys. Chem. A* **2007**, *111*, 8579.
- (101) Kurinovich, M. A.; Lee, J. K. *J. Am. Soc. Mass. Spectrom* **2002**, *13*, 985.
- (102) Chiang, Y.; Kresge, A. J.; Pruszynski, P. *J. Am. Chem. Soc.* **1992**, *114*, 3103.



## Evaluation of Electronic Coupling in Transition-Metal Systems Using DFT: Application to the Hexa-Aquo Ferric–Ferrous Redox Couple

Agostino Migliore,\* Patrick H.-L. Sit,\* and Michael L. Klein

*Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104-6323*

Received August 16, 2008

**Abstract:** We present a density-functional theory (DFT) approach, with fractionally occupied orbitals, for studying the prototypical ferric–ferrous electron-transfer (ET) process in liquid water. We use a recently developed ab initio method to calculate the transfer integral (also named electronic-coupling or ET matrix element) between the solvated ions. The computed transfer integral is combined with previous ab initio values of the reorganization energy, within the framework of Marcus' theory, to estimate the rate of the electron self-exchange reaction. The self-interaction correction incorporated (through an appropriate treatment of the electronic correlation effects) into a Hubbard  $U$  extension to the DFT scheme leads to a theoretical value of the ET rate relatively close to an experimental estimate from kinetic measurements. The use of fractional occupation numbers (FON) turned out to be crucial for achieving convergence in most self-consistent calculations because of the open-shell d-multiplet electronic structure of each iron ion and the near degeneracy of the redox groups involved. We provide a theoretical justification for the FON approach, which allows a description of the chemical potential and orbital relaxation, and possible extension to other transition-metal redox systems. Accordingly, the methodology developed in this paper, which rests on a suitable combination of Hubbard  $U$  correction and a FON approach to DFT, seems to offer a fruitful approach for the quantitative description of ET reactions in biochemical systems.

### 1. Introduction

ET reactions play an essential role in inorganic and organic redox chemistry. In particular, a wide variety of reactions relevant to chemistry and molecular electronics involve nonadiabatic ET processes.<sup>1</sup> The  $\text{Fe}^{2+}$ – $\text{Fe}^{3+}$  redox couple is an archetypal system for the theoretical analysis of homogeneous ET reactions<sup>2–6</sup> and is relevant in many practical contexts, such as corrosion studies and environmental remediation strategies.<sup>7–9</sup>

ET reactions are characterized by means of their rate constants (conductance in molecular electronics applications<sup>10</sup>) and the relevant electron-tunneling pathways. Within the general context of Marcus' ET theory,<sup>11</sup> the rate constant

of nonadiabatic ET reactions is essentially controlled by three key quantities: the reorganization energy (i.e., the free energy change due to the nuclear rearrangement that follows the ET process), the nuclear frequency factor (the frequency of the crossover through the transition-state barrier), and the electronic transmission coefficient or electronic factor. The latter can be strongly dependent on the transfer integral,<sup>12</sup> which is the effective electronic coupling between the donor and acceptor redox groups. In particular, according to the Landau–Zener model<sup>13,14</sup> the electronic factor is proportional to the square modulus of the transfer integral for ET reactions in the nonadiabatic limit.

The resultant expression of the nonadiabatic electron-transfer rate at a given temperature depends on the reorganization energy and the transfer integral. The latter provides a compact link between the ET rates and the electronic

\* Corresponding author phone: 215-898-7058; fax: 215-573-6233; e-mail: amigliore@cmm.upenn.edu.

properties of the interacting redox groups. Since electronic structure plays a pivotal role in determining the kinetics of ET reactions,<sup>15</sup> considerable effort has been devoted to computing electronic couplings by means of several quantum chemical methods.<sup>16–23</sup> Nevertheless, transfer integrals are often very small and thus difficult to compute with high accuracy. As to the system under study, an accurate estimation of the electronic coupling is further complicated by the multiplet nature of the system. Moreover, it is expected<sup>5</sup> that the transfer integral, and the ET rate, is sensitive to the configuration of the solvation water molecules around the two metal ions. In fact, water molecules can have a dramatic effect on the ET kinetics, resulting from the interplay<sup>24,25</sup> between the water–ion electrostatic interactions (which can either promote or oppose the electron transfer as a consequence of the water arrangement) and the effectiveness of the water molecules between the two redox centers in mediating ET coupling pathways. (For any redox system, the latter amounts to a favorable contribution, lowering the barrier for electron tunneling relative to the vacuum.<sup>26</sup>)

Two noteworthy estimates of the transfer integral for the  $\text{Fe}^{2+}$ – $\text{Fe}^{3+}$  redox couple (refs 3 and 5) provide benchmark evaluations of the transfer integral. A more recent approach<sup>27</sup> uses a quasi-experimental derivation of the electronic coupling from experimental ET rate data. This method avoids direct calculation of the transfer integral by assuming the validity of Marcus' ET rate equation and suitably modeling the water medium.

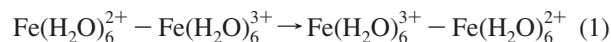
In the present work, we perform a direct many-electron calculation of the transfer integral, within the DFT scheme, using several solvent configurations sampled from ab initio molecular dynamics (MD) simulations. This allows exploration of the non-Condon effects related to the dynamical nature of the water medium and a reliable estimate of the root-mean-square transfer integral to be inserted into Marcus' equation for the ET rate constant. The DFT scheme treats the electronic correlation effects in the presence of metal ions and offers the best compromise between accuracy and feasibility in the study of complicated systems involving the redox couple  $\text{Fe}^{2+}$ – $\text{Fe}^{3+}$  (e.g., in the presence of mineral surfaces and contaminants). The formula<sup>23</sup> adopted to compute the electronic couplings does not resort to empirical parameters, require knowledge of the exact transition-state coordinate, and make use of excited-state quantities. However, the approximate character of any currently available exchange-correlation functional (leading to spurious self-interaction terms in the energy) along with the complicated electronic structure of the transition-metal system under consideration requires the use of the abovementioned formula in a suitable FON-DFT scheme and a remedy for self-interaction errors. In the present context, such errors are corrected to the level allowed by DFT +  $U$ <sup>28</sup> electronic structure calculations, while the effects related to the nondynamical electron correlation<sup>29</sup> are coped with via a suitable FON approach.

The work is organized as follows. Section 2 deals with the theory. In section 2.1 we introduce the model system on which the ab initio calculations are performed. Section 2.2 briefly reviews Marcus' equation for the ET rate constant and the adopted formula<sup>23</sup> for transfer integral evaluation. The general issue of the problematic self-consistent field calculations in

transition-metal systems is addressed in section 2.3. This is followed by analysis (sections 2.4 and 2.5, Appendix B, and Supporting Information) of the FON-DFT scheme for values of the smearing parameter well beyond those studied in previous works. We identify and rationalize a linear regime (and a wider range of approximately linear behavior) of energy eigenvalues, chemical potential, and FON entropic term. Our theoretical analysis provides a recipe for an appropriate choice of the broadening parameter in calculating transfer integrals. The main computational results and their analysis are presented in section 3: Computational details are reported in section 3.1. The ab initio transfer integral calculations, with and without the Hubbard  $U$  correction scheme (introduced to improve the description of the electronic charge distribution), for selected nuclear configurations are presented in section 3.2. They are compared with the couplings obtained by the pathway model<sup>16</sup> in section 3.3 and with the experimental expectations in section 3.4. In particular, the ab initio root-mean-square electronic coupling is combined with prior theoretical estimates of the reorganization energy<sup>30,31</sup> to get a fully ab initio value of the electron self-exchange rate. Finally, the calculated and observed<sup>32</sup> rates are compared, and Marcus' equation for the concerned ET rate is assessed. Analytical details of the theoretical development on FON-DFT achieved in this work and used to compute the transfer integrals are presented in Appendix B (and Supporting Information) after an overview of the standard FON-DFT Scheme with Gaussian Broadening in Appendix A. The Hubbard  $U$  correction to DFT is detailed in Appendix C.

## 2. Theory

**2.1. Model Redox System.** The redox system under consideration is  $\text{Fe}_{\text{aq}}^{2+}$ – $\text{Fe}_{\text{aq}}^{3+}$ . In this paper attention is focused on the solvation cages since the effects of the outside water on the electronic coupling can be considered relatively minor<sup>3,15</sup> (also based on general considerations about the tunneling nature of the electronic coupling between redox partners<sup>21</sup>). The nonadiabatic electron self-exchange reaction under study is



where the electron-transfer process leads to the formation of the successor complex (right side) from the precursor complex (left side). The two groups are the localized donor (D) and acceptor (A) species, whose structural identity is maintained throughout the (outer-sphere) reaction.<sup>15</sup> The D and A groups are assumed to be in contact around the transition state, at the ET optimal interionic distance of 5.5 Å,<sup>3–5,33</sup> also adopted in the MD simulations of ref 30.

**2.2. ET Rate and Transfer Integrals.** In the present work we deal with the rate of the ET process (once the transition state is reached) as controlled by the magnitude of the transfer integral. For nonadiabatic ET reactions, which are characterized by a weak electronic coupling between the D and A groups, the rate constant is approximately given by the high-temperature expression

$$k_{\text{ET}} = \sqrt{\frac{\pi}{\lambda k_{\text{B}} T}} \frac{\langle V_{\text{IF}}^2 \rangle}{\hbar} \exp \left[ -\frac{(\Delta G^\circ + \lambda)^2}{4\lambda k_{\text{B}} T} \right] \quad (2)$$

where  $\lambda$  is the reorganization energy,  $\Delta G^\circ$  is the reaction free energy (in particular, it is zero for a self-exchange reaction),  $k_B$  is Boltzmann's constant,  $T$  is the temperature,  $\langle V_{IF}^2 \rangle$  is the mean-square value of the transfer integral, which measures the coupling between the initial state I and the final state F of the ET reaction. The average in eq 2 expresses a "relaxed" Condon approximation, which holds in the limit of slow modulation of the ET rate by the nuclear motion<sup>34</sup> and captures the average effects of the changes in the water configuration. In fact (see Supporting Information), even if the Condon approximation does not hold (because of a significant dependence of the transfer integral on the arrangement of the water molecules), its "relaxed" form appropriately accounts for the effects of the electronic coupling on the ET rate due to the actual uncoupling of nuclear and electronic motions.

The averaging on the transfer integral in eq 2, as performed in our calculations, is strictly related to the distinction between accepting modes and inducing modes.<sup>34</sup> The first ones support the energy exchanges necessary both to make the relevant donor and acceptor levels nearly degenerate in energy and to relax the nuclear structure after the electron transition occurs. Along these modes ET can happen only for a regime of configurations near the transition state, so that the Condon approximation can be applied. Indeed, the reaction coordinate depends on the overall set of accepting modes. The inducing modes are weakly coupled to the D and A states, so that the electron transfer is not limited to a small range of configurations along them. In fact, the effects of the disordered water motion can be negligible on the energies of the localized (*diabatic*) D and A electronic states while considerable and probably fluctuating on the coupling between such states. Our implementation of eq 2 can be depicted in terms of a two-dimensional space of the water nuclear configurations, spanned by a reaction coordinate and an orthogonal inducing coordinate.<sup>35</sup> The MD simulations from ref 30 exploited in this work allow a mapping onto the aforementioned space, and the transfer integral is computed on sampled configurations around the reaction coordinate of the transition state. Hence, variable magnitudes necessarily arise from distinct points along the inducing coordinate.

The ab initio computation of the transfer integrals is performed by means of the formula<sup>23</sup>

$$V_{IF} = \frac{|\Delta E_{IF} ab|}{a^2 - b^2} \quad (3)$$

where  $\Delta E_{IF}$  is the energy difference between the ET diabatic states I and F, and  $a$  and  $b$  are their respective overlaps with the ground state of the system. The initial state vector is defined as  $|\psi_I\rangle = |D\rangle|A\rangle$  and the final state vector as  $|\psi_F\rangle = |D^+\rangle|A^-\rangle$ , where  $|D\rangle$  ( $|D^+\rangle$ ) denotes the reduced (oxidized) ground state of the isolated donor site and  $|A^-\rangle$  ( $|A\rangle$ ) the reduced (oxidized) ground state of the isolated acceptor site. Therefore, in the two-state model the ground-state vector of the system is given by  $|\psi\rangle = a|\psi_I\rangle + b|\psi_F\rangle$ . The approximations involved in eq 3 and the feasibility of appropriately using DFT wave functions are detailed in ref 23, where it is also stressed that it yields a dependence of the electronic coupling on the distance between the redox centers in good

agreement with the empirical average packing density model.<sup>36</sup> Moreover, within the theoretical framework of ref 23 eq 3 gives the best estimate of the transfer integral also when the two-state approximation is not satisfied. The quantity  $\Delta E_{IF}$  is given by

$$\Delta E_{IF} = (E_D + E_A) - (E_{D^+} + E_{A^-}) + W_{D-A} - W_{D^+-A^-} \quad (4)$$

where  $E_D$  ( $E_{D^+}$ ) is the ground-state energy of the isolated donor group in its reduced (oxidized) state of charge,  $E_A$  ( $E_{A^-}$ ) is the same for the acceptor group in its oxidized (reduced) state,  $W_{D-A}$  and  $W_{D^+-A^-}$  are the energies of (essentially electrostatic) interaction between the D and A groups in the initial and final diabatic states, respectively.

**2.3. Problematic SCF Convergence.** Equation 3 has been implemented into a spin-polarized DFT scheme. In this section we show that self-consistent field (SCF) calculations, in the absence of fractional occupations of the Kohn–Sham (KS) spin orbitals, hardly manage to converge. Moreover, the convergence, if it is achieved, generally leads to a dramatic failure in the description of the electronic ground state with the transferring electron charge abnormally shared between the two iron centers.<sup>30,37</sup>

The issue of the troublesome SCF convergence can be well appreciated starting from the effective one-particle KS equations. In atomic units they are written as<sup>38</sup>

$$\begin{aligned} H([n_s], \mathbf{r})\psi_i(\mathbf{r}) &\equiv \left[ -\frac{1}{2}\nabla^2 + V_H([n_s], \mathbf{r}) + v_{xc}([n_s], \mathbf{r}) \right] \psi_i(\mathbf{r}) \\ &= \varepsilon_{si} \psi_i(\mathbf{r}) \quad (5) \end{aligned}$$

where  $i = 1$  to  $N$ ,  $N$  is the total number of electrons,  $H$  denotes the KS Hamiltonian operator,  $n_s(\mathbf{r})$  is the ground-state density of the auxiliary system of noninteracting electrons (which equals the exact density of the interacting system),  $V_H$  is the Hartree potential,  $v_{xc}$  is the exchange-correlation (XC) potential,  $\psi_i$  is the  $i$ th spin orbital, and  $\varepsilon_{si}$  is the corresponding energy eigenvalue.<sup>39</sup> Equation 5 is to be solved in a self-consistent manner under the constraint  $n_s(\mathbf{r}) = \sum_{i=1}^N |\psi_i(\mathbf{r})|^2$ , which expresses the ground-state density in terms of the  $N$  lowest KS spin orbitals. Due to the approximate character of any currently available XC functional the iterative solution of eq 5 leads to incorrect wave functions ( $\varphi_i$ ), energy eigenvalues ( $\varepsilon_i$ ), and density ( $n$ ), which satisfy the approximate equations

$$\begin{aligned} H^{\text{approx}}([n], \mathbf{r})\varphi_i(\mathbf{r}) &\equiv \left[ -\frac{1}{2}\nabla^2 + V_H([n], \mathbf{r}) + \right. \\ &\quad \left. v_{xc}^{\text{approx}}([n], \mathbf{r}) \right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}) \quad (6) \end{aligned}$$

where  $n(\mathbf{r}) = \sum_{i=1}^N |\varphi_i(\mathbf{r})|^2$  is the electron density and  $v_{xc}^{\text{approx}}$  the approximate XC potential. By considering the operators into eqs 5 and 6 after the respective self-consistency is achieved we recast eq 6 in the form

$$(H([n_s], \mathbf{r}) + W([n], [n_s], \mathbf{r}))\varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}) \quad (7)$$

with

$$\begin{aligned} W([n], [n_s], \mathbf{r}) &= (V_H + v_{xc})([n], \mathbf{r}) - (V_H + v_{xc})([n_s], \mathbf{r}) + \\ &\quad v_{xc}^{\text{approx}}([n], \mathbf{r}) - v_{xc}([n], \mathbf{r}) \quad (8) \end{aligned}$$

Equation 8 displays two different, although related, contributions to  $W$ . The former comes from the difference in the

total potential, with the correct functional form, evaluated on the approximate and exact charge densities. The latter is due to the incorrect functional form of  $v_{xc}$ . It persists even if  $n$  happens to be so close to  $n_s$  that the first contribution can be neglected.

By applying the stationary perturbation theory<sup>40</sup> to eq 7 we get the following connection between the exact and the approximate KS spin orbitals in terms of the “perturbation”  $W$  (i.e., the deviation from the exact one-particle Hamiltonian)

$$\varphi_i(\mathbf{r}) = \psi_i(\mathbf{r}) + \sum_{n \neq i} \frac{\langle \psi_n | W | \psi_i \rangle}{\varepsilon_{si} - \varepsilon_{sn}} \psi_n(\mathbf{r}) + O(W^2) \quad (9)$$

According to eq 9, the departure of  $\varphi_i$  from the corresponding wave function  $\psi_i$  (i.e., the spin orbital coming from the use of the exact XC potential) is determined by the mixing of  $\psi_i$  with the other orbitals  $\psi_n$  ( $n \neq i$ ) through  $W$ . The closer the KS energy levels  $\varepsilon_{si}$  and  $\varepsilon_{sn}$  and the stronger the coupling  $\langle \psi_n | W | \psi_i \rangle$ , the larger the mixing between  $\psi_i$  and  $\psi_n$ . Therefore, the approximate character of the XC potential can affect (sometimes to a great extent) the shapes of the wave functions and thus the values of the overlap parameters  $a$  and  $b$  entering on eq 3. In this respect, a crucial source of errors is the unphysical self-interaction of the relevant electronic charge with itself arising from the fact that the approximations to the XC energy are independent of the electrostatic repulsion term.

In the presence of open-shell atoms a further complication stems from the multiplet problem since the currently available XC functionals do not have the appropriate spin and spatial symmetry behavior to correctly describe multiplet systems.<sup>41</sup> Such a problem is exacerbated in the case that molecular orbitals centered on different redox sites come to be virtually degenerate. One way out consists of the use of a linear combination of determinants to describe the wave function of the system.<sup>41</sup>

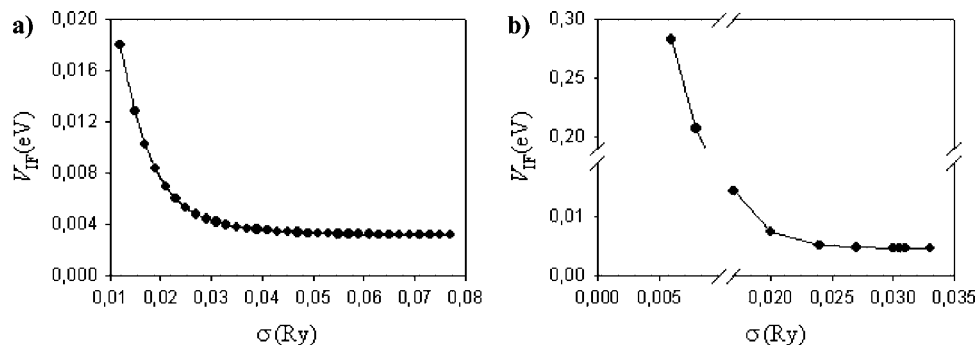
Finally, a computational issue (possibly related to the previous ones in a DFT scheme) comes out by considering eq 9 (and the equations which express the second-order correction to the energy as well as higher terms in both the energy and the spin-orbital perturbation expansions) over a SCF calculation. In fact, as the potential (initially corresponding to the wrong guess density) is changed between subsequent iterations to move toward the self-consistency, a computational contribution to  $W$  intervenes to mix the spin orbitals. On one hand, this is an intrinsic feature of the iterative procedure, exploitable to obtain quadratic SCF convergence.<sup>42</sup> On the other hand, it cannot be used effectively when the gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) is particularly small. Rather, in such a case, the convergence is generally problematic and can lead to unstable solutions, all the more that a small HOMO–LUMO gap is often accompanied by a high density of states near the Fermi energy, defined as the average of the HOMO and LUMO energies for finite systems with integral occupation numbers.<sup>42</sup> Indeed, it has been formally shown, within the HF scheme, that SCF convergence is slowed by a small gap at the Fermi energy.<sup>43</sup> In general,

oscillations in the iterative solution of one-electron equations arise whenever some orbitals close to the Fermi level are alternatively occupied and unoccupied from one iteration to the next. One way to damp these oscillations consists in allowing fractional occupations.<sup>44</sup>

In the ferrous–ferric system the generally asymmetric arrangement of the aqueous environment around the two ions moves the ground state out of degeneracy. Therefore, a single-determinant picture of the ground-state wave function is allowed. However, the levels corresponding to some d-like molecular orbitals preserve a multiplet structure. This can be envisaged by taking jointly into account the near degeneracy of the atomic d orbitals and the fact that close enough to the transition state one-half of the gap between the HOMO and LUMO levels are comparable with the expected value of the transfer integral.<sup>22,45</sup> In particular, the HOMO and the LUMO, each expected to be essentially localized on a different ion (although with a tail onto the other ion, owing to the electronic coupling between the two redox sites), correspond to very similar energies for all of the sampled nuclear configurations. Hence, due to the abovementioned computational issues, DFT calculations without fractional occupation numbers, when they manage to convergence, almost always lead to a quite inaccurate HOMO, which is a linear combination of the correct HOMO and LUMO exceedingly spread over the two sites. Thus, the ground-state wave function, obtained as a Slater determinant of the lowest occupied orbitals, is unduly delocalized over the D and A groups and the overlaps  $a$  and  $b$  entering on eq 3 are correspondingly similar, leading to anomalously large values of the transfer integral. This behavior is illustrated in Figure 1. It shows that the value of the transfer integral diverges when the electron smearing becomes so small that all the molecular orbitals have integral occupation numbers. Moreover, for several other nuclear configurations SCF convergence without FON has been not achieved.

It is worth noting that, while the interplay between the computational issues and the approximations incidental to the DFT scheme is in general responsible for an almost equal spread of the HOMO (which describes the ET system in the one-electron picture and is the crucial orbital in determining the transfer integral also in the multielectron picture<sup>12,15</sup>) over the two redox centers, the spurious electronic self-interaction is, by itself, sufficient for a considerable overestimation of the electronic coupling. In fact, the latter is crucially dependent on the small tail of the HOMO, which can be drastically affected by self-interaction errors, even if they cause a negligible relative change in the correct charge density around the ion where the spin orbital is prominently localized.

In this work, the electron-smearing technique proposed by ref 46 (where a broadening parameter is used to get the SCF convergence and then gradually reduced to zero) has been successfully used for some calculations on the isolated redox groups, necessary to derive the localized wave functions  $\psi_I$  and  $\psi_F$ , while it turned out not to work in the presence of both ions. For the electronic structure calculations on the overall system we use an alternative approach. Fractional occupations are suitably exploited to get the convergence. Then the spin orbitals up to the nominal HOMO are used to build the necessary wave



**Figure 1.** Transfer integrals by eq 3 vs spreading parameter  $\sigma$  for the Gaussian broadening of the orbital occupation numbers (see below). The panels correspond to different nuclear configurations taken from a Car–Parrinello molecular dynamics (CPMD) run in ref 30 after (a) 9000 and (b) 24 000 steps. Each time step is 5 au. The overall production run lasts 60 000 steps.

functions, while the virtual spin orbitals with “fake” occupation are disregarded.<sup>47</sup> With reference to eq 3, where only the ground-state wave function of the overall system is employed, the latter procedure requires that the orbital relaxation exclusively due to the smearing of the electron charge (once the abnormal spreading of the HOMO is avoided) is negligible. This is in analogy with the approximation on relaxation pertaining to Koopmans’ theorem (in both the HF<sup>48</sup> and DFT<sup>49</sup> schemes), although that theorem involves an integer change of the LUMO occupation with a corresponding increase in the net electronic charge. On the other hand, the analogy cannot be pushed far enough to justify the employed approach, which requires a specific theoretical basis. The pertinent analysis (sections 2.4 and 2.5, Appendix B, and Supporting Information) is also essential to identify the appropriate values of the smearing parameter to be used in transfer integral evaluation.

From an analytical point of view we note that in the presence of fractional occupations  $f_i$  of the orbitals eq 6 turns into

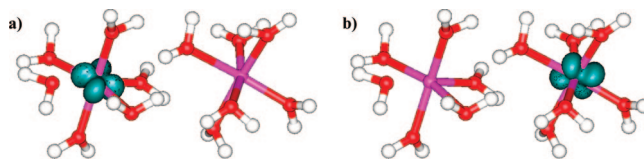
$$\left[-\frac{1}{2}\nabla^2 + v_{\text{eff}}([n_\sigma], \mathbf{r})\right]\varphi_i(\mathbf{r}; \sigma) = \varepsilon_i(\sigma)\varphi_i(\mathbf{r}; \sigma) \quad (10)$$

where  $v_{\text{eff}}([n_\sigma], \mathbf{r}) \equiv v_{\text{eff}}(\mathbf{r}; \sigma) = V_{\text{H}}([n_\sigma], \mathbf{r}) + v_{\text{xc}}^{\text{approx}}([n_\sigma], \mathbf{r})$  is the effective single-particle potential and  $n_\sigma$  is the electronic charge density given by

$$n_\sigma(\mathbf{r}) \equiv n(\mathbf{r}; \sigma) = \sum_{i=1}^N f_i(\sigma) |\varphi_i(\mathbf{r}; \sigma)|^2 \quad (11)$$

The Hamiltonian operators in eqs 6 and 10 differ by the term  $v_{\text{eff}}([n_\sigma], \mathbf{r}) - v_{\text{eff}}([n], \mathbf{r})$ . In section 2.5 we show that this difference, stemming from the FON approach, can only lead to minor changes in the orbitals when a suitable SCF convergence can be achieved also for a vanishing electron smearing. Thus, the orbitals derived by means of the FON-DFT approach can only suffer significantly from the intrinsic approximations of  $v_{\text{eff}}$ , that is, mainly from self-interaction errors. We will take care of the latter by the method illustrated in Appendix C, which also allows calculation of the transfer integral without fractional occupations for some of the considered nuclear configurations.

**2.4. Single-Particle Energies, Chemical Potential, and Energy Terms in the Gaussian FON-DFT Scheme.** For all the sampled configurations of the water medium the computed electronic structure of the aqueous ferrous–ferric



**Figure 2.** Minority-spin (a) HOMO and (b) LUMO for the configuration after 24 000 simulation steps. The small tail of each orbital on the other redox center (for the HOMO it is due to the nonzero transfer integral) is not visible at the represented isovalue of 0.02.

system (in its high-spin state) turns out to be characterized by a multiplet of levels  $\varepsilon_j$  ( $j = 1, \dots, N_d$ ), which correspond to minority-spin orbitals with a predominant d-like character. For instance, Figure 2 displays the two lowest d-like MOs for one of the selected configurations. The remaining MOs in the multiplet have a similar shape (i.e., they are essentially d type orbitals of  $t_{2g}$  symmetry), dictated by the electrostatic field of the solvation water. In fact, we always obtain  $N_d = 6$ . Instead, the atomic 4s orbital and the other atomic 3d orbitals give rise to higher virtual levels well separated from the multiplet.

In fact, the half-width of the d-like multiplet turns out to be within ca. 0.25 eV for all of the selected nuclear configurations, whereas the separation between the mean energy of the multiplet and the higher lying levels is always above 1 eV and the lower lying levels are at least 2 eV far apart. Such a level structure allows a relatively wide range of  $\sigma$  values larger than the spreading of the d-like multiplet and smaller than the separation from the remaining levels, which are denoted by  $\varepsilon_k$ . In the present section we exploit this feature, common to many other open-shell transition-metal systems, and develop the formalism of the FON-DFT approach with Gaussian broadening in order to obtain useful connections between the KS single-particle energies, the chemical potential, and the entropy. The resultant analysis provides the theoretical justification for the derivation of the ground-state wave function from the FON-DFT approach, whereas the corresponding electronic density is not used.

For the aforementioned values of  $\sigma$ , by analytical elaboration of eq 27 of Appendix A and exploitation of Cardano’s formula,<sup>50</sup> we obtain (see Supporting Information)

$$\varepsilon_j(\sigma) - \mu(\sigma) = [A(N_d) + B(N_d)\omega_j(\sigma)]\sigma \quad (12)$$

where the deviation numbers

$$\omega_j(\sigma) = f_j(\sigma) - \frac{1}{N_d} \quad (13)$$

measure the departure from an even occupation of the MOs in the multiplet. For  $N_d = 6$ , the numerical values of  $A$  and  $B$  are 0.57 and  $-1.57$ , respectively. Moreover, the dependence of  $B$  on  $N_d$  is negligible. The addition of eq 12 corresponding to the different  $\varepsilon_j$  and the requirement that the deviation numbers add up to zero (as long as the fractional occupations  $f_j$  add up to unity) yield the following useful connection between the chemical potential and the mean energy  $\langle \varepsilon(\sigma) \rangle$  of the multiplet

$$\mu(\sigma) = \langle \varepsilon(\sigma) \rangle - A(N_d)\sigma \quad (14)$$

If  $\langle \varepsilon(\sigma) \rangle$  is approximately independent of  $\sigma$ , as the levels  $\varepsilon_j$  mix among them,<sup>51</sup> and  $\sigma$  is sufficiently large for the validity of eq 14, then the chemical potential has a linear dependence on  $\sigma$ . As exemplified by Figure 3, this is the case in a wide range of  $\sigma$  values. Moreover, the slope of the dashed line in Figure 3b is 0.58, which is very close to the value 0.57 predicted by eq 14 for  $N_d = 6$ .

When  $\sigma$  becomes comparable with the energy gap between the d-like multiplet and the other levels,  $\langle \varepsilon(\sigma) \rangle$  changes appreciably (e.g., see Figure 3a) and  $\mu$  is no longer a linear function of  $\sigma$ . The energy gap below the multiplet is wider than the one above it. However, the spreading of the transferring electron over the multiplet determines a decrease of the chemical potential, which gets closer to the lower lying levels and increasingly smaller than the mean energy of the multiplet. On the whole, at the upper edge of any explored  $\sigma$  range the deviations of the lower lying levels from unit occupation are small (less than 0.05) and comparable with the deviations of the higher lying levels from zero occupation. Therefore,  $\mu$  levels off beyond the linear regime (see Figure 3b). On the other hand, the smaller energy gap above the multiplet determines a stronger mixing between the latter and the higher virtual levels. This causes the increase of  $\langle \varepsilon \rangle$  illustrated by Figure 3a. We conclude that although the chemical potential is an even function of  $\sigma$  when the electron density does,<sup>52</sup> its behavior can be linearized in a suitable (wide) range of the broadening parameter.

When  $\sigma$  is much larger than the spreading of the d-like multiplet, while still small relative to the gap with the higher KS levels (i.e., for  $\sigma$  well inside the range of values where the chemical potential has a linear behavior, referred to as the *linear regime*), the levels  $\varepsilon_j$  get almost equally occupied

so that  $f_j \approx 1/N_d$  and  $\omega_j \approx 0$ . This is to say that in the asymptotic expansion of  $\omega_j$  around any of such  $\sigma$  values

$$\omega_j(\sigma) = \omega_{0j} + \frac{\omega_{1j}}{\sigma} + \dots \quad (15)$$

the zero-order term is quite small and  $\omega_{1j} \ll \sigma$ . More generally, in the  $\sigma$  range where the expansions of eq 15 can be truncated to the first order in  $1/\sigma$ , their insertion into eq 12 gives

$$\varepsilon_j(\sigma) - \mu(\sigma) = [x_j(\sigma) + A(N_d)]\sigma = B\omega_{1j} + [B\omega_{0j} + A(N_d)]\sigma \quad (16)$$

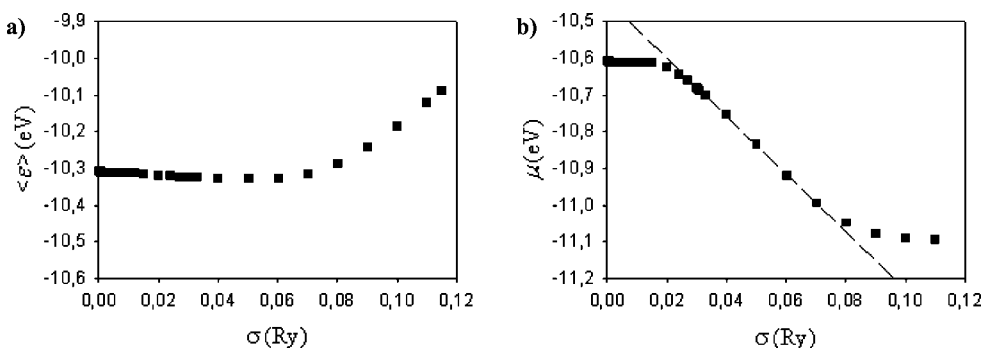
Equation 16 predicts the behavior of the relative energies  $\varepsilon_j - \mu$  in the linear regime and is closely confirmed by the computational results (see Supporting Information). It is also worth noting that the slopes of the curves represented by eq 16 are quite similar since  $A$  is much larger than  $B\omega_{0j}$ .

For values of the Gaussian broadening beyond the linear regime the orbitals corresponding to out-of-multiplet levels ( $\varepsilon_k$ ) begin to have appreciable fractional occupations. Since all the levels  $\varepsilon_j$  correspond to d-like MOs and are very close in energy relative to the remaining levels  $\varepsilon_k$ , they mix to a similar amount with the latter (as can be seen by means of the stationary perturbation theory, accomplished up to second-order corrections<sup>53</sup>). As a consequence, beyond the linear regime the levels  $\varepsilon_j$  do not branch off. Rather, they continue to experience an almost identical shift as functions of  $\sigma$ . More specifically, in the Supporting Information we describe the behavior of the KS relative energies for values of  $\sigma$  beyond the linear regime, although still corresponding to the plateau of the transfer integral (that is, according to eq 3, corresponding to a negligible orbital relaxation).

As shown in the Supporting Information, the linear regime is also characteristic of the behavior of the entropic contribution  $-\sigma S(\sigma)$  to the fictitious free energy. In fact, it is

$$S(\sigma) = \frac{N_d}{2\sqrt{\pi}} \exp(-A^2) \left[ 1 + \frac{2A^2 - 1}{N_d} \sum_j \left( B^2 \omega_{0j}^2 + \frac{2B^2 \omega_{0j} \omega_{1j}}{\sigma} + \dots \right) \right] \cong S_0 - \frac{S_1}{\sigma} \quad (17)$$

where the positive coefficients  $S_0$  and  $S_1$  are introduced, the sum is restricted to the levels  $\varepsilon_j$  (which is appropriate in the linear regime, where the levels  $\varepsilon_k$  are still empty), and the terms up to the first order in  $1/\sigma$  are retained in the last



**Figure 3.** (a)  $\langle \varepsilon \rangle$  vs  $\sigma$  and (b)  $\mu$  vs  $\sigma$  for the nuclear configuration after 24 000 simulation steps. The vertical axis is translated upward in the b panel. The dashed line fits to the data points in the linear region of the entropic term (see below).

approximate expression. By considering the relation<sup>52</sup>  $dE/d\sigma = \sigma(dS/d\sigma)$ , eq 17 yields the following expression for the energy

$$E(\sigma) = E(\sigma_0) + S_1 \ln \frac{\sigma}{\sigma_0} \quad (18)$$

Finally, from eqs 17 and 18 we obtain the fictitious free energy

$$F(\sigma) = E(\sigma_0) + S_1 \ln \left( e \frac{\sigma}{\sigma_0} \right) - S_0 \sigma \quad (19)$$

Equations 17, 18, and 19 describe the behavior of the FON-DFT entropy, energy, and free energy in a wide range of  $\sigma$ , starting from a value  $\sigma_0$  at the onset of the linear regime. Therefore, they are complementary to the power series (up to the second order in  $\sigma$ ) in eqs B4 and B5 of ref 52, which are valid for small enough  $\sigma$ . Equation 11 of ref 52 shows that  $1/2[F(\sigma) + E(\sigma)] = E(0) + O(\sigma^n)$  with  $n > 2$ , thus providing a method to calculate the correct energy (without electron smearing) from the results at a suitably small  $\sigma$ . In this work the value of  $n$  is not found, but eqs 18 and 19 clearly show that such a method cannot be used in the linear regime. Furthermore, as shown in next section, eq 18 yields a sufficient condition for suitably picking the value of  $\sigma$  in the calculation of the transfer integral.

**2.5. Single-Particle Quantities and Orbital Relaxation Analysis: An Alternative Approach to FON-DFT.** In section 2.4 we showed that the KS levels experience almost identical shifts in a large  $\sigma$  range, with approximately equal slopes. In particular, the value of the transfer integral is always picked in correspondence to a  $\sigma$  in the linear regime, although the relative changes of the transfer integral are negligible also beyond such a regime. Indeed, there is a strict connection between the regime of nearly uniform behavior of the single-particle energies and the orbital relaxation. In the HF scheme the shape and the energy of an orbital are independent of its occupation number if the remaining orbitals do not relax.<sup>54</sup> This circumstance does not occur in any DFT calculations due to the spurious electron self-interaction. However, as stressed in ref 55 within the context of Janak's theorem, when orbital relaxation is negligible a substantial linear response (i.e., the orbital energy is a linear function of the occupation number) can be expected in the presence of the electron self-interaction. Moreover, in refs 49 and 56 (with particular reference to the DFT Koopmans' theorem in large molecular systems) it is noted that orbital relaxation can be appraised by the nonuniformity of the KS level shift.

In this section (see also Appendix B) we present a theoretical analysis which provides a general connection between the Gaussian broadening and the orbital relaxation as mediated by the KS effective potential and is able to circumscribe the errors that can arise when the orbital relaxation is completely ignored. The latter point is particularly important when dealing with ET between small complexes. It is also worth noting that the provided formalism can be directly applied to other electron-transfer systems (e.g., electron self-exchange reactions, where the HOMO and the LUMO are characterized by the same quantum numbers). The two main goals of the approach, as for the evaluation

of the transfer integrals, are as follows: (i) theoretical proof that the relaxation of the spin orbitals is negligible in the entire large  $\sigma$  range, so that a stable value of the transfer integral can be derived from them; (ii) analytical assessment that such orbitals, denoted by  $\bar{\varphi}_i(\mathbf{r};\sigma)$ , are an adequate approximation to the computationally inaccessible orbitals in the absence of the occupation Gaussian broadening, denoted by  $\varphi_i(\mathbf{r};0)$ , so that the corresponding value of the transfer integral is reliable.

A general equation, which transparently discloses the connection between  $\bar{\varphi}_i(\mathbf{r};\sigma)$  and  $\varphi_i(\mathbf{r};0)$ , owing to the dependence of the KS effective potential  $v_{\text{eff}}$  on  $\sigma$ , can be derived by exploiting the formalism of the quantum theory of scattering by a potential. In fact, the Kohn–Sham equations, after self-consistency is achieved, can be written as

$$\left[ -\frac{1}{2}\nabla^2 - \varepsilon_i(0) + v_{\text{eff}}(\mathbf{r};0) \right] \bar{\varphi}_i(\mathbf{r};\sigma) = [\Delta\varepsilon_i(\sigma) - \Delta v_{\text{eff}}(\mathbf{r};\sigma)] \bar{\varphi}_i(\mathbf{r};\sigma) \quad (20)$$

where

$$\Delta\varepsilon_i(\sigma) = \varepsilon_i(\sigma) - \varepsilon_i(0) \quad (21a)$$

$$\Delta v_{\text{eff}}(\mathbf{r};\sigma) = v_{\text{eff}}(\mathbf{r};\sigma) - v_{\text{eff}}(\mathbf{r};0) \quad (21b)$$

and  $\Delta\varepsilon_i - \Delta v_{\text{eff}}$  works as a “scattering potential”. The solution of the homogeneous equation associated to eq 20 is just  $\varphi_i(\mathbf{r};0)$ , while the Green's function of the pertaining operator, that is the solution of the equation

$$\left[ -\frac{1}{2}\nabla^2 - \varepsilon_i(0) + v_{\text{eff}}(\mathbf{r};0) \right] G_i(\mathbf{r}) = \delta(\mathbf{r}) \quad (22)$$

is

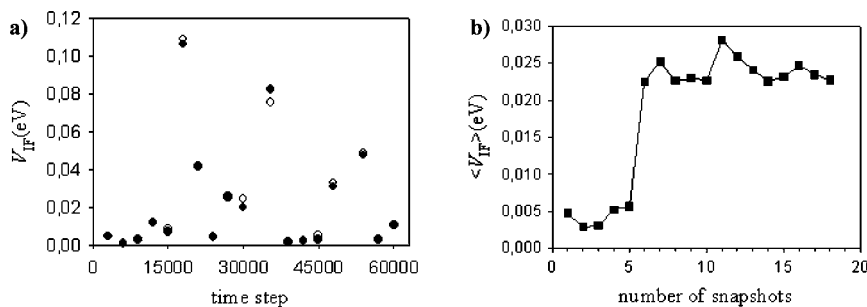
$$G_i(\mathbf{r}) = \frac{1}{2\pi r} \frac{\varphi_i(\mathbf{r};0)}{\varphi_i(\mathbf{0};0)} \quad (23)$$

The origin of the coordinate reference system can be arbitrarily chosen wherever the denominator of eq 23 is not null. From eqs 20, 21a, 21b, 22, and 23 it is seen that the wave function  $\bar{\varphi}_i(\mathbf{r};\sigma)$  fulfills the integral “scattering” equation

$$\bar{\varphi}_i(\mathbf{r};\sigma) = \varphi_i(\mathbf{r};0) + \frac{1}{2\pi\varphi_i(\mathbf{0};0)} \int d^3r' \frac{\varphi_i(\mathbf{r}-\mathbf{r}';0)}{|\mathbf{r}-\mathbf{r}'|} \times [\Delta\varepsilon_i(\sigma) - \Delta v_{\text{eff}}(\mathbf{r}';\sigma)] \bar{\varphi}_i(\mathbf{r}';\sigma) \quad (24)$$

Equation 24 illustrates the role played by  $\Delta v_{\text{eff}}(\mathbf{r};\sigma)$  (i.e., the change in the KS effective potential due to the Gaussian broadening) in the rearrangement of the orbitals. The theoretical analysis of eq 24, illustrated in Appendix B, shows that the integral term, owing to the “scattering potential” and mixing the correct spin orbitals, can be neglected at any  $\sigma$  within a suitable range, which includes all the linear regime. In particular, this holds for the HOMO, which is the crucial spin orbital in the calculation of the parameters  $a$  and  $b$  in eq 3, and thus of the effective electronic coupling.

Once demonstrating the reliability of the transfer integral evaluation in the plateau region, the sufficient condition represented by the fulfillment of eq 18 provides a useful computational tool, as the exploration of the  $\sigma$  range can be



**Figure 4.** (a) Effective electronic coupling vs CPMD step number for the PW91-GGA (○) and PBE-GGA (●) calculations. (b) Dependence of the PBE-GGA mean transfer integral on the number of configurations used to compute the average.

stopped when the onset of the linear regime is detected. Ultimately, we also note that the nearly even occupation of the d-like MOs well inside the plateau region (where the coupling is obtained) can be regarded as an extension of the Slater transition-state notion<sup>57</sup> (but see the discussion in p 58 of ref 38 about the definition of the transition-state density).

Note that the theoretical analysis in this section and Appendix B adequately delimits the errors arising from the Gaussian spreading of the electron charge, which is responsible for a change  $\Delta v_{\text{eff}}(\mathbf{r};\sigma)$  in the KS effective potential, whereas it does not envisage the errors in both  $v_{\text{eff}}(\mathbf{r};\sigma)$  and  $v_{\text{eff}}(\mathbf{r};0)$  resulting from the spurious electronic self-interaction. The computational achievement of a plateau of the transfer integral over a wide range of  $\sigma$ , where occupation of the HOMO changes by a significant fraction of the electron charge  $e$  (e.g., from  $0.2e$  to more than  $0.3e$  for the cases depicted in Figure 1), indicates that  $v_{\text{eff}}(\mathbf{r};\sigma)$  and  $v_{\text{eff}}(\mathbf{r};0)$  are affected by similar self-interaction errors. The recipe used for self-interaction correction (SIC), thus yielding a FON-DFT +  $U$  approach, is described in Appendix C.

### 3. Computation

**3.1. Computational Details.** The transfer integrals are computed through eqs 3 and 4, on the system of eq 1, for selected nuclear configurations along a Car–Parrinello MD<sup>58</sup> (CPMD) run taken from ref 30. The CPMD time step is 5 au, and the selected snapshots are 3000 steps apart ( $\sim 0.36$  ps). The required electronic properties are computed by means of the PWscf code,<sup>59</sup> in the repeated supercell approach, by using the plane-wave spin-polarized DFT scheme in both the Perdew–Wang<sup>60</sup> (PW91) and the Perdew–Burke–Ernzerhof<sup>61</sup> (PBE) generalized gradient approximation (GGA). The DFT +  $U$  method is applied with the PBE exchange–correlation functional. The wave function and charge density cutoffs are 25 and 200 Ry, respectively. The atomic cores are represented by ultrasoft pseudopotentials from the standard QUANTUM-ESPRESSO distribution<sup>59</sup> (specifically, H.pw91-van\_ak.UPF, O.pw91-van\_ak.UPF, and Fe.pw91-sp-van\_ak.UPF for the PW91 exchange–correlation functional; H.pbe-rrkjus.UPF, O.pbe-rrkjus.UPF, and Fe.pbe-sp-van\_mit.UPF for the PBE functional). The Fe pseudopotential corresponds to 16 valence electrons. The size of the repeating cell is  $15 \times 15 \times 20 \text{ \AA}^3$  ( $20 \text{ \AA}$  along the direction of the irons). Calculations with large values of  $\sigma$  are performed by constraining the total magnetization in such

a way that the total numbers of the majority-spin and minority-spin electrons are fixed, and the system is strictly preserved in the correct high-spin state (which is used in ref 30; see also refs 3 and 62).

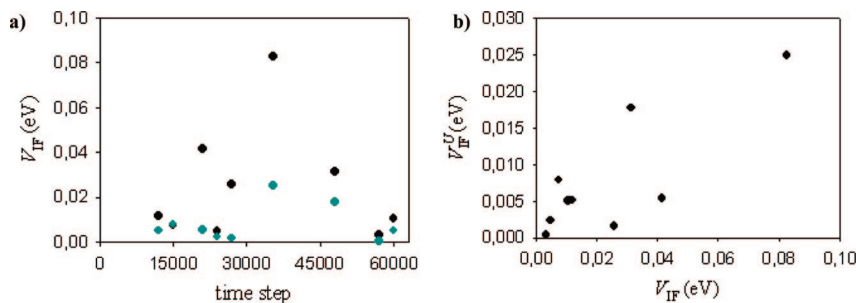
The relevant wave functions are constructed as single Slater determinants of the lowest lying KS spin orbitals up to the nominal HOMO (which is a true HOMO for the single-site wave functions, obtained without fractional occupations). The feasibility of using DFT wave functions is argued in ref 23 and references therein. The overlaps between the spin orbitals, necessary to derive the overlap integrals  $a$  and  $b$  into eq 3, are computed through the DTI program.<sup>63</sup> The interactions in eq 3 ( $W_{D-A}$  and  $W_{D^+-A^-}$ ) are obtained by full electrostatics calculations exploiting the Poisson equation and the electrostatic potential provided by the PWscf program.

In the (FON-)DFT +  $U$  approach the  $U$  parameter is computed through the PWscf code. The same  $U$  is used for obtaining the electronic structures of the isolated donor and acceptor groups (in the oxidation states corresponding to the initial and final ET states) and of the whole system, which are required by eq 3. The dependence of the electronic charge distribution and thus of the electronic correlation on nuclear coordinates is taken into account by separately evaluating the  $U$  interaction parameter for each selected configuration.

**3.2. Transfer Integrals.** The couplings computed through the spin-polarized FON-DFT approach in the PW91-GGA and the PBE-GGA are represented in Figure 4a. For each point a plateau of the transfer integral (as in Figure 1) has been obtained. The changes along the plateaus are generally nonmonotonic, amount to less than 5%, and are mainly attributable to the shortcomings of the two-state model. For any nuclear configuration the choice of the  $\sigma$  value (within the respective plateau) for evaluation of the electronic coupling is essentially arbitrary due to the small relative error. Anyhow, the adopted criterion lies in the use of the  $\sigma$  value which yields the maximum localization of the nominal HOMO on one of the ET sites and thus the smallest electronic coupling. This choice opposes the overestimation of the coupling by any DFT scheme without full self-interaction correction and corresponds to the maximum (nominal) HOMO–LUMO gap. The latter circumstance matches the fact that any SIC approach tends to enlarge that gap (e.g., this can be seen from the expressions of the single-particle energies in the Perdew–Zunger SIC scheme<sup>64</sup>).

As shown in Figure 4a, PW91-GGA and PBE-GGA give very similar results. This matching is generally expected,





**Figure 5.** (a) Transfer integral vs CPMD step number. The dark circles represent the transfer integrals by the FOND-DFT approach with the PBE-GGA. The cyan circles are obtained with introduction of the  $U$  correction term. (b) Mapping between the two sets of couplings in the left panel.

although not trivial, as the two approximations are not equivalent (e.g., see ref 65). At each estimate  $V_{IF}$  of the coupling can be associated an error  $cV_{IF}/(|a| + |b|)$ ,<sup>23</sup> which measures the maximal uncertainty (in fact, an upper bound for the uncertainty) due to departure from the two-state condition. Hence, a maximal error can be associated to the mean transfer integral by exploiting the well-known rules for independent error propagation. The estimates corresponding to the configurations after 33 000 and 51 000 simulation steps have been rejected because of a strong failure of the two-state model (i.e.,  $a^2 + b^2 < 0.75$ , from which  $c > 0.5$ ), resulting from localization of the transferring electron on different d-like MOs in the ET diabatic states and in the ground state of the system. Within the accuracy determined by the maximal error the PW91 and PBE exchange-correlation functionals yield the same mean value and root-mean-square (rms) value for the electronic coupling, that is  $\langle V_{IF} \rangle_{PW91} = \langle V_{IF} \rangle_{PBE} = (23 \pm 3) \times 10^{-3}$  eV and  $\sqrt{\langle V_{IF}^2 \rangle_{PW91}} \equiv (V_{IF})_{rmsPW91} = (V_{IF})_{rmsPBE} = (37 \pm 7) \times 10^{-3}$  eV, respectively. The number of snapshots used to compute the averages is more than enough, as shown by the cumulative average in Figure 4b, according to which the mean value plateaus after six data points.

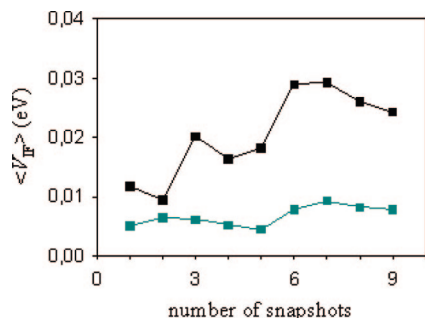
The difference between the two kinds of average is a consequence of the spreading of the results due to the dynamical nature of the water medium and clearly indicates that the inducing coordinate has been explored. In fact, water can affect the electronic coupling both by mediation of ET coupling pathways (mainly, through hydrogen bonds)<sup>66,67</sup> and by the electrostatic field determined around the two ions.<sup>25</sup> The magnitude of both effects can be strongly altered by the water displacements still allowed within the solvation cages, thus leading to the wide range of the transfer integral shown in Figure 4a. This range can be characterized by the standard deviation  $\sigma_V = \sqrt{\langle V_{IF}^2 \rangle - \langle V_{IF} \rangle^2} = 0.029$  eV, which is comparable with the mean values given above, as expected for flexible systems.<sup>25,68</sup> Indeed, the values of  $V_{IF}$  corresponding to different nuclear configurations can be viewed as “independent estimates” of the quantity  $\langle V_{IF} \rangle$  (or  $(V_{IF})_{rms}$ ). Thus, as an estimate of the mean value obtained after an infinite number of calculations,  $\langle V_{IF} \rangle$  or  $(V_{IF})_{rms}$  can be endowed with a statistical error  $\sigma_{(V)} = \sigma_V/\sqrt{18} = 0.007$  eV. Note that the two errors attributed to the mean transfer integral cannot be directly added.

As shown in the next section, the above value of the root-mean-square transfer integral is significantly larger than the

experimental estimate derived from the ET Marcus’ equation. This overestimation arises from the too fast asymptotic decay of the XC potential associated with the spurious self-interaction<sup>69,70</sup> because the d-like HOMO, essentially localized around one metal ion, has a too large tail on the other redox center. The correction of the charge distribution around the ions through the  $U$  term (and FON, whenever required) leads to the electronic couplings  $V_{IF}^U$  shown in Figure 5a. Relying on the matching between the results by PW91-GGA and PBE-GGA we employed only the PBE functional (used in the CPMD run of ref 30).

We note that for four of the selected MD configurations the DFT +  $U$  scheme yields the electronic coupling without the need for FON. However, as a consequence of the strong interdependence between energies and occupation numbers of the relevant orbitals in the employed Hubbard  $U$  correction (see eq 39 in Appendix B), for some MD configurations the plateau regime of the electronic coupling is reached only at very large values of  $\sigma$ , where a level scheme analogous to the one without the  $U$  correction term is recovered. In such cases, the maximal errors incidental to the smearing of the electronic charge, which are proportional to  $\sigma$  (see Appendix B, eqs 35 and 37), can be correspondingly large. In fact, the change in the density ensuing from the unduly large electron smearing can yield a significant orbital relaxation through the corresponding change in the KS effective potential (see eq 24). Thus, the transfer integrals corresponding to those plateaus are rejected, although the HOMO can be accidentally correct (as shown in Appendix B and the Supporting Information). Ultimately, we retain only the nine nuclear configurations whose transfer integrals are obtained either in the absence of FON or by plateaus that extend over  $\sigma$  ranges similar to the ones in Figure 1.

Although the amount of change in the transfer integral varies with the nuclear configuration, the sets of values  $V_{IF}$  and  $V_{IF}^U$  are correlated, as displayed by the mapping in Figure 5b. In fact, their correlation coefficient is  $r_{V,V^U} = 0.81$ , and the probability of finding an at least equal value of the coefficient if the two sets of data (each including nine points) are uncorrelated is  $P_9(r \geq r_{V,V^U}) = 0.9\%$ . This probability indicates a highly significant correlation<sup>71</sup> between the two sets of electron transfer integrals. Therefore, the relative values of the effective coupling obtained by means of eq 3 in the “bare” (without SIC) FOND-DFT scheme are meaningful, on average, for the system under consideration. For example, the above conclusion supports the use of the bare



**Figure 6.** Dependence of the PBE-GGA mean transfer integral on the number of configurations used to compute the average. Cyan and dark squares are obtained by the FON-DFT approach with and without the  $U$  correction, respectively.

scheme in order to find the decay constant of the transfer integral with the distance between the two irons. Further analysis is required to ascertain the point. However, we wish to stress that eq 3, implemented within a bare DFT scheme, yielded a decay constant in good agreement with the empirical average packing density model<sup>36</sup> for the system studied in ref 23.

Using the  $U$  correction we obtain  $\langle V_{IF}^U \rangle = 7.8 \times 10^{-3}$  eV and  $(V_{IF}^U)_{\text{rms}} = 11.0 \times 10^{-3}$  eV. Both averages are endowed with the statistical error  $\sigma_{\langle V \rangle}^U = 2.6 \times 10^{-3}$  eV. By comparison, the corresponding quantities obtained without the  $U$  correction by averaging on the reduced set of nuclear configurations are  $\langle V_{IF} \rangle = 24 \times 10^{-3}$  eV,  $(V_{IF})_{\text{rms}} = 34 \times 10^{-3}$  eV, and  $\sigma_{\langle V \rangle} = 8 \times 10^{-3}$  eV. Their differences from the respective estimates based on 18 configurations are unimportant. The best estimate of the transfer integral and its standard error are considerably decreased using the  $U$  approach. However, we note that  $(V_{IF}^U)_{\text{rms}}$  and  $(V_{IF})_{\text{rms}}$  are of the same order of magnitude, which is also attributable to the fact that eq 3 does not resort to any excited-state quantity, thus limiting the shortcomings of DFT.<sup>23,69</sup>

The significantly smaller standard deviation resulting from the  $U$ -corrected scheme is illustrated by the reduced spreading of the pertinent electronic coupling values (represented by cyan circles in Figure 5a) and results in a more rapid plateauing of the cumulative average transfer integral, as displayed by Figure 6. However, the statistical error continues to be relatively high. Actually, this is a common feature of high-level quantum chemical methods.<sup>25,72</sup> In fact, they generally reduce the systematic errors of the average quantities coming from the approximate description of the electronic structure, but hardly address the statistical errors since the number of data points used in the averaging is limited by their computational cost. Moreover, for a given number of points, the spreading of the results and thus the statistical error in the estimate of the mean transfer integral depend on the specific non-Condon effects characterizing the system under consideration. Ultimately, the higher accuracy of the FON-DFT +  $U$  computational scheme, relative to the FON-DFT scheme, pushes the spreading of the transfer integrals toward the correct one, exclusively determined by the failure of the Condon approximation.

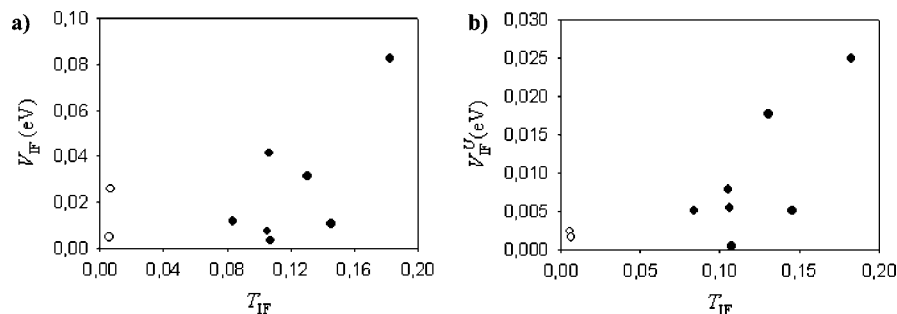
The following points are worthy of note. (i)  $(V_{IF}^U)_{\text{rms}}$  has a larger relative statistical error than  $(V_{IF})_{\text{rms}}$ , although its

absolute statistical error is smaller. This is due to the physical limit in the spreading of the electronic couplings (expressing the departure from the Condon approximation) in conjunction with the overestimate of the root-mean-square (or mean) transfer integral without  $U$  correction. (ii) An exact (linear) correlation between the FON-DFT and FON-DFT +  $U$  results should imply the proportionality of the respective spreads. The achieved high correlation indicates a random contribution of the shortcomings in the electronic structure calculations employing the less accurate FON-DFT approach. The possible use of eq 3 in the bare DFT scheme, discussed above, essentially rests on such a random feature.

**3.3. Coherence Parameter and Comparison with the Pathways Model.** A quantity strictly related to  $\sigma_V$  is the coherence parameter  $R_c = \langle V_{IF} \rangle^2 / \langle V_{IF}^2 \rangle$ ,<sup>66</sup> which gives a measure of the fluctuations in the transfer integral. The effect of the fluctuations is negligible when  $R_c$  is close to unity (so that the Condon approximation holds), while it is at a maximum when  $R_c$  approaches zero (breakdown of the Condon approximation). The latter case is typical of extremely flexible systems (although it can occur also in symmetry forbidden processes at the equilibrium nuclear configuration<sup>34</sup>). We obtain  $R_c = 0.4$  (without the  $U$  correction) and  $R_c^U = 0.5$ . Both estimates reflect the fact that the solvation cages are far from the free-motion condition while still relatively far from a tight binding.

The value of the coherence parameter can be interpreted in terms of ET coupling pathways.<sup>66,68</sup> In fact, it is expected to be very small when the interaction between the donor and acceptor groups is mediated by several interfering coupling pathways, whereas it is close to unity in the case of a dominant-coupling pathway. The estimates of  $R_c$  for the system under consideration are relatively high, thus suggesting the presence of a dominant pathway. On the other hand, they are sufficiently far from unity so that a more complicated picture is necessary to describe some features of the ET system. To analyze the point we compare the ab initio electronic couplings  $V_{IF}$  and  $V_{IF}^U$  with the corresponding pathways products<sup>73</sup>  $T_{IF}$ , derived<sup>74</sup> using the pathway model.<sup>16</sup> In fact, according to this semiempirical model the electronic coupling is proportional to the product  $T_{IF}$  of the (coupling) decay factors for each step in the dominant pathway tube connecting the D and A groups. Each decay factor describes the exponential decay of the electronic coupling along a step. The pertaining decay constant depends on the through-bond, through-hydrogen-bond, or through-space character of the step. As the decay constant weighs the length of the corresponding step, a (unit-less) effective distance is associated to the dominant pathway and  $T_{IF}$  corresponds to the shortest effective distance between the redox centers.

In the solvated ferrous–ferric system with the face-to-face conformation the region between the ions comprises six waters (see Figure 2). Two of such water molecules (one from each ion complex) are always involved in the best pathway. The value of  $T_{IF}$  strongly increases when they are connected by a hydrogen bond (so that the through-space decay factor is replaced by the through-hydrogen-bond decay factor for one step along the best pathway). The mappings



**Figure 7.** (a) FON-DFT electronic couplings ( $V_{IF}$ ) and (b) FON-DFT +  $U$  electronic couplings ( $V_{IF}^U$ ) vs pathway products ( $T_{IF}$ ): (●) nuclear configurations including a hydrogen bond in the best ET pathway; (○) remaining configurations.

between the ab initio electronic couplings  $V_{IF}$  and  $V_{IF}^U$  and the products  $T_{IF}$  are shown in Figure 7a and 7b, respectively. The relative values of  $V_{IF}$  and  $T_{IF}$  have a similar spreading. In addition, the pathway couplings show a significant gap between the two nuclear configurations not including the hydrogen bond in the best ET pathway (empty circles) and the remaining ones (full circles). These two configurations yield a small coupling also in the ab initio  $U$ -corrected approach. The fact that other configurations, including the hydrogen bond in the best pathway, lead up to similarly small ab initio transfer integrals is attributable to the interplay between the bridging and electrostatic effects of the water environment, which is not grasped by the single-pathway picture.<sup>25</sup>

According to standard statistics<sup>71</sup> the correlation between the  $V_{IF}$  and  $T_{IF}$  sets is not significant, although appreciable, whereas the correlation between the  $U$ -corrected electronic couplings  $V_{IF}^U$  and the pathway products  $T_{IF}$  is significant.<sup>75</sup> In spite of the appreciable correlation between ab initio transfer integrals and pathway products, the dominant pathway can provide only an approximate picture of the electron-transfer process and cannot capture important features of the electronic structure. In fact, the scattering of the data points in Figure 7, the not-high level of correlation between the ab initio method and the semiempirical model, and the above argument about the ion–water electrostatic interactions (which can also comprise water molecules not directly involved in any tunneling pathway) point to the ab initio calculations with  $U$  correction as a reliable method for obtaining electronic couplings to be compared with experiment. Moreover, the ab initio method does not require the use of empirically adapted prefactors resting on the maximum ET rate constant as is the case when using the quantities  $T_{IF}$ .<sup>76</sup> Ultimately, let us stress that from the comparison between Figure 7a and 7b (where the solvent configurations characterized by an ET relevant hydrogen bond between the ions are in evidence) emerges the ability of the  $U$  correction to DFT in describing the effects of relevant hydrogen bonds on the transfer integral between the redox centers.

**3.4. Comparison with Experiment.** According to Marcus' ET theory the experimental estimate of the rms electronic coupling is obtained from eq 2 once the experimental value of the reorganization energy is inserted. The best estimates of ET rate constant and reorganization energy from the experimental data are<sup>31,32</sup>  $k_{ET}^{(exp)} = 7.9 \times 10^2 \text{ s}^{-1}$  and  $\lambda^{(exp)} = 2.1 \text{ eV}$ , respectively. The resulting transfer integral is  $(V_{IF}^{(exp)})_{rms} = 7.1 \times 10^{-3} \text{ eV}$ . The slightly larger value

$(V_{IF}^{(exp)})_{rms} = 7.3 \times 10^{-3} \text{ eV}$  is obtained from the following refined expression for the rate constant<sup>77</sup>

$$k_{ET} = \nu_n \frac{1 - \exp(-\nu_{el}/2\nu_n)}{1 - \frac{1}{2}\exp(-\nu_{el}/2\nu_n)} \exp\left(-\frac{\lambda}{4k_B T}\right) \quad (25)$$

where

$$\nu_{el} = \frac{\langle V_{IF}^2 \rangle}{\hbar} \sqrt{\frac{\pi}{\lambda k_B T}} \quad (26)$$

is an electronic frequency for the electron transfer within the activated complex and  $\nu_n$  is an effective nuclear frequency for the motion along the reaction coordinate. In eq 25 it is explicitly considered that the reaction free energy  $\Delta G^\circ$  is zero for an electron self-exchange reaction. Our ab initio best value for the rms transfer integral, provided with the statistical error (i.e., the uncertainty due to the finite number of data points used in the averaging), is  $(V_{IF}^U)_{rms} = (11.0 \pm 2.6) \times 10^{-3} \text{ eV}$ . It can be seen that the discrepancy between the theoretical and the experimental best estimates is not statistically significant.<sup>78</sup> At any rate, the agreement can be considered good by taking into account the various steps connecting the observed rate constant of the bimolecular reaction and the electron self-exchange rate constant (e.g., the valuation of the pair distribution function). Moreover, the effective distance between the two ions is not necessarily a useful guide for assessing the magnitude of the transfer integral<sup>15</sup> due to the significant non-Condon effects here unraveled for the face-to-face conformation of the reactants.

As shown in Table 1, the ab initio estimate of the rms transfer integral obtained in the present work lies between the experimental value and two benchmark theoretical estimates in the literature.<sup>3,5</sup> The single-configuration value in ref 3,  $V_{IF} = 98 \text{ cm}^{-1} = 12.2 \text{ eV}$ , refers to a reduced model system, which includes all ligands lying between the two metal ions, with a common value for the Fe–O distances in the ET complex. The estimate in eq 6 of ref 5,  $V_{IF} = 124 \text{ cm}^{-1} = 15.3 \text{ eV}$ , rests on a simple model with one electron in the pseudopotential field of two ferric ions. The discrepancy between the latter and our theoretical estimate is beyond the assumed significance threshold, while we find a considerable agreement with the value in ref 3. On the other hand, the rms transfer integral obtained in the present work more closely approaches the experimental estimate. Moreover, it allows a statistical analysis of the remaining gap, also addressing the issue of the Condon approximation

**Table 1.** Comparison between Transfer Integral Best Estimates from Experimental Rates and from Different Theoretical Approaches

method	(model) system	$V_{IF}$ ( $10^{-3}$ eV)	ref
from eq 2, $\lambda^{(\text{exp})}$	experimental, rms value		
from eq 25, $\lambda^{(\text{exp})}$	$(\text{Fe}_{\text{aq}}-\text{Fe}_{\text{aq}})^{5+}$	7.1	31, 32
	$(\text{Fe}_{\text{aq}}-\text{Fe}_{\text{aq}})^{5+}$	7.3	31, 32
	theoretical		
FON-DFT + $U$ , rms value	$[\text{Fe}(\text{H}_2\text{O})_6-\text{Fe}(\text{H}_2\text{O})_6]^{5+}$	$11.0 \pm 2.6$	present work
FON-DFT, rms value	$[\text{Fe}(\text{H}_2\text{O})_6-\text{Fe}(\text{H}_2\text{O})_6]^{5+}$	$34 \pm 8$	present work
ROHF <sup>a</sup>	$[\text{Fe}(\text{H}_2\text{O})_3-\text{Fe}(\text{H}_2\text{O})_3]^{5+}$	12.2	3
fitting to Schrödinger equation solution	$\text{Fe}^{3+}-\text{Fe}^{3+} + e$	15.3	5

<sup>a</sup> Spin-restricted open-shell HF.

through the theoretical analysis on CPMD water configurations<sup>79</sup> (note that the CPMD simulations, performed at a suitably high ionic temperature, allowed reproducing the expected structure of the solvation shells and led to the right free-energy profile for the electron self-exchange process<sup>30</sup>).

By inserting the theoretical estimate of the reorganization energy in ref 31,  $\lambda = 2.11$  eV, into eq 25 we can obtain a fully ab initio estimate of the self-exchange rate constant, that is  $k_{\text{ET}}^{(\text{theo})} = 15.5 \times 10^2 \text{ s}^{-1}$ , with a confidence interval  $(9.4-22.8) \times 10^2 \text{ s}^{-1}$ , corresponding to the uncertainty interval for the electronic coupling. As a matter of fact, the electronic coupling affects the self-exchange rate approximately in a quadratic way (as is the case for eq 2) and the reorganization energy has a crucial role because of its presence in the (exponential) nuclear factor. As a consequence of these two circumstances, the absolute change in the ab initio rate constant, resulting from a difference in the employed set ( $\lambda, V_{IF}$ ), is magnified relative to the latter (e.g., the best value of the rate constant resulting from the ab initio reorganization energy in ref 30,  $\lambda' = 2.0$  eV, is  $k_{\text{ET}}^{(\text{theo})'} = 46.5 \times 10^2 \text{ s}^{-1}$ ). Therefore, the agreement between  $k_{\text{ET}}^{(\text{theo})}$  and  $k_{\text{ET}}^{(\text{exp})}$  can be considered quite good.

#### 4. Conclusions

The main achievements of the present work are (i) a methodology to treat the thorny transition-metal redox system, (ii) theoretical analysis to justify a method that can be extended to other FON schemes, and (iii) an ab initio value of the root-mean-square transfer integral.

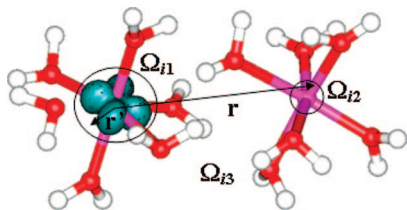
The use of a FON approach is crucial for computational treatment of the open-shell system under consideration. We provided the theoretical basis for the adopted computational methodology by suitable elaboration on the FON-DFT scheme with Gaussian broadening and subsequent exploitation of the formalism pertaining to the quantum theory of scattering by a potential. The resultant theoretical development transparently discloses the connections between the relevant quantities of the Kohn–Sham scheme with FON (namely, single-particle effective potential, energy eigenvalues, chemical potential, energy, entropy, and free energy associated to the fractional occupation numbers). Moreover, it describes the behavior of these quantities as a function of the Gaussian broadening over a wide range of the latter. In particular, a linear regime, characterizing the behavior of single-particle energies, chemical potential, and entropic contribution to the (variational) free energy, is identified and

rationalized. The established connection between Gaussian broadening of the orbital occupations and orbital relaxation also provides a useful sufficient condition (corresponding to the linear regime of the KS levels) for future applications of the proposed FON method.

The linear regime of the KS single-particle energies is expected in the absence of orbital relaxation within the context of Janak's theorem.<sup>55,80</sup> In the present work it is formulated and extended to the behavior of other relevant KS quantities, within the context of the FON-DFT scheme with a Gaussian broadening, applied to open-shell transition-metal systems. Moreover, it is shown (Appendix B) that the relaxation of the orbitals is negligible beyond the linear response regime of the spin–orbital energies. It is worth noting that the provided theoretical analysis is amenable to extensions to different FON schemes with particular concern for the Fermi–Dirac broadening.<sup>81</sup>

Besides the FON approach, the ab initio  $U$  correction to DFT of ref 28 was essential for a reliable description of the electron charge distribution around the two redox centers, which in turn leads to correspondingly reliable estimates of the effective electronic couplings. Moreover, through the analysis of our results we can assert and quantify the limited adequacy of a single-tunneling pathway picture of the ET reaction. We also infer its limitations in describing the ET system under consideration because of the interplay between bridging and electrostatic effects of the water medium. Such limitations are, indeed, strictly related to the failure of the Condon approximation. In fact, the computed coherence parameter offers a common measure for both features. The analyzed departure from the Condon approximation supports two important points: (i) despite their lack of low-lying vacant orbitals, the water molecules can play a relevant role in ET to the extent that some ligand-to-metal charge transfer is present in the ground states of the reactants.<sup>3</sup> (ii) The nominal Fe-to-Fe distance (fixed in the present work) is not necessarily a useful guide for assessing (single) coupling magnitudes.<sup>15</sup>

The use of a recently proposed ab initio method for transfer integral valuation within the proposed FON-DFT approach along with the ab initio  $U$  correction<sup>28,82</sup> yields an estimate of the transfer integral for the solvated ferrous–ferric system that is in fair agreement with the experimental estimate derived from the observed kinetic data and reorganization energy.



**Figure 8.** Partition of the space around the irons into three regions  $\Omega_{ik}$  ( $k = 1, 2, 3$ ).  $\Omega_{11}$  and  $\Omega_{12}$  denote suitable regions around the two metals where the  $i$ th MO is appreciable, and  $\Omega_{13}$  comprises all the space around. The HOMO of Figure 2a is represented. In the calculation of  $I_i(\mathbf{r};\sigma)$ ,  $\mathbf{r}'$  spans  $\Omega_{11}$ .

Finally, we wish to stress that the present methodology offers a pragmatic approach for the quantitative investigation of electron self-exchange reactions in transition-metal systems.

### Appendix A. Overview of the FON-DFT Scheme with Gaussian Broadening

The formal introduction of the fractional occupation numbers rests on the thermal DFT, founded by Mermin<sup>83</sup> through the extension of the Hohenberg–Kohn theorem to nonzero temperatures, within the grand canonical ensemble. In fact, at any finite temperature the Fermi–Dirac distribution leads to fractional occupations around the chemical potential  $\mu$  (or the Fermi energy,  $\varepsilon_F$ , in the low-temperature limit, i.e.,  $k_B T \ll \varepsilon_F$ ). Then, the variational energy functional appropriate to FON-DFT is formally identical to the grand potential  $\Omega = E - TS - \mu N$  of the finite-temperature DFT once an “entropy” is associated with the occupation numbers of the KS spin orbitals and is defined as<sup>81</sup>

$$S = - \sum_i [f_i \ln f_i + (1 - f_i) \ln(1 - f_i)] \quad (27)$$

although no physical meaning needs to be associated to the  $T$  (or  $k_B T$ ) parameter. The same formalism is allowed by the other broadening functions<sup>81,84,85</sup> once the corresponding spreading parameter  $\sigma$  and occupation numbers are introduced, although  $\sigma$  has no simple physical interpretation.

If the total number of electrons  $N$  is fixed (i.e., in the canonical ensemble), the fictitious “free energy”  $F = E - \sigma S$  is the suitable variational functional.<sup>52</sup> The stationary condition for  $\Omega$  (or  $F$ ) with respect to the occupation numbers  $f_i$  is written as

$$\frac{\partial \Omega}{\partial n_i} = \frac{\partial}{\partial n_i} (F - \mu N) = 0 \quad (28)$$

where the chemical potential is introduced as a Lagrange multiplier when using the free energy  $F$ . Note that by exploiting the property of the entropy<sup>52</sup>  $\partial S / \partial f_i = (\varepsilon_i - \mu) / \sigma$  eq 28 yields Janak’s theorem, that is  $\partial E / \partial f_i = \varepsilon_i$ .<sup>80</sup> Moreover, at zero temperature eq 28 leads to the classical FON solution,<sup>38</sup> according to which the levels with fractional occupation are degenerate and equal to  $\varepsilon_F$ .

In this work we use the Gaussian broadening scheme, so that the fractional occupations are given by

$$f_i(\sigma; \varepsilon_i - \mu) = 1 - \int_{-\infty}^{\varepsilon_i - \mu} g(\sigma; x) dx = \frac{1}{2} \operatorname{erfc} \left( \frac{\varepsilon_i - \mu}{\sigma} \right) \quad (29)$$

where  $\operatorname{erfc}$  denotes the complementary error function. It is obtained from the Fermi occupation function at zero temperature by replacing the Dirac delta with its Gaussian broadening

$$g(\sigma; x) = \frac{1}{\sigma \sqrt{\pi}} \exp \left( - \frac{x^2}{\sigma^2} \right) \quad (30)$$

### Appendix B. Fractional Occupations and Orbital Relaxation

In this appendix we elaborate on eq 24 in order to demonstrate that the ground-state wave function, used for evaluation of the transfer integral by means of eq 3, can be reliably obtained in a wide range of  $\sigma$  values, though the corresponding density is not physically meaningful.

First, it is worth noting that if  $v_{\text{eff}}$  undergoes an essentially homogeneous change as a consequence of the occupation broadening, the Hamiltonian operator correspondingly changes by an additive constant dependent only on  $\sigma$ . Therefore, the solutions  $\tilde{\varphi}_i(\mathbf{r};\sigma) = \varphi_i(\mathbf{r};0)$  are directly obtained from eq 20 and the KS eigenvalues  $\varepsilon_i(\sigma)$  are uniformly shifted relative to  $\varepsilon_i(0)$ . In the general case the change in  $v_{\text{eff}}$  due to smearing of the electron charge depends on the space coordinate and is related to the kinetic energy and the chemical potential. To provide physical insight on this point let us consider a small variation  $\delta n(\mathbf{r};\sigma)$  around the minimum density  $n(\mathbf{r};\sigma)$ , involving arbitrary changes in the shapes of the orbitals, while the occupation numbers are set at the values pertaining to the given equilibrium density. The corresponding functional derivative of the kinetic energy satisfies the equation<sup>38,86</sup>

$$\frac{\delta T[n(\sigma)]}{\delta n(\mathbf{r};\sigma)} = -v_{\text{eff}}(\mathbf{r};\sigma) + \mu(\sigma) \quad (31)$$

By writing eq 31 for zero smearing and subtracting term by term we get

$$\frac{\delta T[n(\sigma)]}{\delta n(\mathbf{r};\sigma)} - \frac{\delta T[n(0)]}{\delta n(\mathbf{r};0)} = -\Delta v_{\text{eff}}(\mathbf{r};\sigma) + \mu(\sigma) - \mu(0) \quad (32)$$

Equation 32 shows that the change in  $v_{\text{eff}}$  is independent of  $\mathbf{r}$  if the functional derivative of the kinetic energy does or, at any rate, if  $\delta T / \delta n$  is independent of  $\sigma$ . By considering the functional form of  $T[n(\sigma)]$  it can be seen<sup>56</sup> that the term in the left-hand side of eq 32 is negligible compared to  $\Delta v_{\text{eff}}$  (which includes the change in the Coulomb potential) for small and delocalized density changes (involving also a variation of the total charge in ref 56). For the system under study, the changes in the electron density due to large enough values of  $\sigma$  (within the linear regime and beyond) are relatively dispersed around the two irons. Moreover, they do not entail a change in the total electronic charge and essentially involve d-like MOs (coming from atomic orbitals with the same quantum numbers), which are characterized by very similar orbital angular momenta, spatial extents, and thus kinetic energies (so that  $T[n(\sigma)] \cong T[n(0)]$  at every  $\sigma$ ). Consequently, the difference in the left-hand side of eq 32 is expected to be quite small over the entire  $\sigma$  range (especially at large enough  $\sigma$ ). This working hypothesis (resting on the above physical arguments and corroborated by Figures 1 and 3) amounts to saying that although  $\Delta v_{\text{eff}}$

generally depends on  $\mathbf{r}$ , it continues to be of the order of magnitude of  $\Delta\mu(\sigma) = \mu(\sigma) - \mu(0)$ , that is a fraction of eV (e.g., at most 0.5 eV in the case of Figure 3b). The consequences of such a consideration on the integral of eq 24 can be well appreciated by partitioning the space around the redox centers as in Figure 8 and recasting eq 24 in the form

$$\bar{\varphi}_i(\mathbf{r};\sigma) = \varphi_i(\mathbf{r};0) + I_{i1}(\mathbf{r};\sigma) + I_{i2}(\mathbf{r};\sigma) + I_{i3}(\mathbf{r};\sigma) \quad (33)$$

with

$$I_{ik}(\mathbf{r};\sigma) = \frac{1}{2\pi\varphi_i(\mathbf{0};0)} \int_{\Omega_{ik}} d^3r' \frac{\varphi_i(\mathbf{r}-\mathbf{r}';0)}{|\mathbf{r}-\mathbf{r}'|} [\Delta\varepsilon_i(\sigma) - \Delta\nu_{\text{eff}}(\mathbf{r}';\sigma)] \bar{\varphi}_i(\mathbf{r}';\sigma) \quad (34)$$

$\Omega_{ik}$  ( $k = 1, 2$ ) denote the regions where the  $i$ th MO is appreciably nonzero, while  $\Omega_{i3}$  comprises all the space around. Such a partition rests on the consideration that with the ions 5.5 Å apart less than 1% of the relevant electron charge resides in the tunneling region between the two ions.<sup>5</sup> In accordance, a calculation based on the Schrödinger equation for the two ions without water molecules puts 98% of the electronic charge within 1.65 Å of one of the two ions.<sup>5</sup> In particular, as shown by Figure 8, the partition is meaningful for the HOMO,  $\bar{\varphi}_H(\mathbf{r};\sigma)$ , which is the crucial molecular orbital in evaluating the transfer integral. In fact, both  $\bar{\varphi}_H(\mathbf{r};\sigma)$  and  $\varphi_H(\mathbf{r};0)$  are essentially localized on the same metal ion (named “ion 1” and depending on the sign of the energy difference  $\Delta E_{\text{IF}}$  between the diabatic states) with a small tail on the other ion (“ion 2”). The accuracy in the determination of such a tail is the crucial factor in the calculation of the electronic coupling. However, even an excessive delocalization of the HOMO onto ion 2, responsible for a drastic overestimate of the transfer integral, corresponds to a negligible relative change of the orbital distribution around ion 1 once the localization of the orbital on the right ion (indicated by  $\Delta E_{\text{IF}}$ ) has been achieved. An analogous argument can be reported to the other orbitals. Consequently, it is a reliable approximation to replace  $\bar{\varphi}_i(\mathbf{r};\sigma)$  with  $\varphi_i(\mathbf{r};0)$  in the region  $\Omega_{i1}$ , that is into the expression of  $I_{i1}(\mathbf{r};\sigma)$ , which amounts to applying the Born approximation<sup>40,87</sup> to eq 24. By focusing attention on the tail of the HOMO (so that  $\mathbf{r}$  points toward the ferric ion 2 and  $|\mathbf{r}-\mathbf{r}'|$  is on the order of the distance  $R$  between the two ions)<sup>88</sup> and dropping the explicit indication of the spin orbital in the quantities  $I_{ik}$  and  $\Omega_{ik}$  we can write

$$I_1(\mathbf{r};\sigma) \lesssim \frac{1}{2\pi|\varphi_H(\mathbf{0};0)|} \frac{|\varphi_H(\mathbf{R};0)|}{R} |\Delta\varepsilon_H(\sigma) - \Delta\mu(\sigma)| \frac{4}{3} \pi r_d^3 \langle |\varphi_H(\mathbf{r}';0)| \rangle_{\Omega_1} \approx A(N_d) \frac{2r_d^3}{3R} |\varphi_H(\mathbf{r};0)| \sigma. \quad (35)$$

In eq 35 it is considered that  $\langle |\varphi_H(\mathbf{r}';0)| \rangle_{\Omega_1}$  is of the order of  $|\varphi_H(\mathbf{0};0)|$  and  $\Omega_1 \approx 4/3 \pi r_d^3$ , where  $r_d$  is the Fe d-state radius. Moreover, the order of magnitude of  $|\Delta\varepsilon_H - \Delta\mu|$ , i.e.,  $A(N_d)\sigma$ , can be derived from eq 16 by considering that  $\varepsilon_H - \mu$  is an approximately linear function of  $\sigma$  and that  $A(N_d)$  is much larger than  $B\omega_{0j}$ . As a matter of fact, the order

of magnitude of  $|\Delta\varepsilon_H - \Delta\mu|$  remains the same within the  $\sigma$  range of nonlinear behavior described in the Supporting Information as well as when the quantities  $B\omega_{0j}$  are appreciable, so that the differences among the slopes of the relative energies  $\varepsilon_j - \mu$  cannot be disregarded. Therefore, eq 35 establishes a more general relation between the level shift and the orbital relaxation relative to the case of the uniform level shift. Using the value  $r_d = 0.744$  Å, provided by the atomic-surface method,<sup>89</sup> and  $N_d = 6$ , the corrective factor for the HOMO ensuing from eq 35 is numerically equal to  $\sigma$ , as expressed in Ry. Therefore, the relative error due to neglecting  $I_1(\mathbf{r};\sigma)$  in eq 33 is within a few percent, all the more that eq 35 actually supplies an upper bound for such an error. For example, since  $R \gg \Omega_1^{1/3}$ , by considering the case  $\Delta\nu_{\text{eff}}(\mathbf{r}';\sigma) \cong \Delta\mu(\sigma) \forall \mathbf{r}' \in \Omega_1$ , we find

$$I_1(\mathbf{r};\sigma) \cong \frac{1}{2\pi\varphi_H(\mathbf{0};0)} \frac{\varphi_H(\mathbf{R};0)}{R} |\Delta\varepsilon_H(\sigma) - \Delta\mu(\sigma)| \int_{\Omega_1} \bar{\varphi}_H(\mathbf{r}';0) d^3r' \cong 0 \quad (36)$$

since the positive and negative lobes of the d-like MO approximately cancel out in the integration. In general, the contribution of  $I_1(\mathbf{r};\sigma)$  to the HOMO lies between the boundary provided by eqs 35 and 36 and can also be a nonmonotonic function of  $\sigma$ . The analog of eq 35 for the region  $\Omega_2$  is

$$I_2(\mathbf{r};\sigma) \lesssim A(N_d) \frac{r_i^2}{2\pi} |\varphi_H(\mathbf{r};0)| \sigma \quad (37)$$

where  $r_i$  is some length, much smaller than  $r_d$ , measuring the size of  $\Omega_2$ . Note that the approximations leading to eq 35 (in particular, the fact that  $\Delta\nu_{\text{eff}}$  and  $\Delta\mu$  have the same order of magnitude) apply even better to the small region  $\Omega_2$ , also considered that the Coulomb potential gives the major contribution to  $\Delta\nu_{\text{eff}}$ . By taking into account that  $I_3(\mathbf{r};\sigma)$  is negligible by construction, from eqs 35, 36, and 37 we deduce that the HOMO and thus the transfer integral can be reliably evaluated at any  $\sigma$  since the onset of the linear regime. Note that the existence of a plateau of the transfer integral remains demonstrated by the fact that  $\bar{\varphi}_H(\mathbf{r};\sigma) \cong \varphi_H(\mathbf{r};0) \cong \bar{\varphi}_H(\mathbf{r};\sigma')$ , for conveniently large  $\sigma$  and  $\sigma'$ . However, according to eqs 35, 36, and 37, the effects of the quantities  $I_1(\mathbf{r};\sigma)$  and  $I_2(\mathbf{r};\sigma)$  can become important for unduly large broadenings, where eq 16 also drastically fails. Then,  $\bar{\varphi}_H(\mathbf{r};\sigma)$  is no longer a good approximation to  $\varphi_H(\mathbf{r};0)$  and the ground-state wave function,  $\psi$ , is correspondingly wrong. Such a circumstance can be easily recognized by means of a considerable departure from the two-state condition  $|\psi\rangle = a|\psi_1\rangle + b|\psi_2\rangle$ , as measured by  $c = \sqrt{1-a^2-b^2}$ .

## Appendix C. DFT + Hubbard $U$ Approach to SIC

As mentioned in section 2.3, the approximate character of any currently available XC functional brings to the spurious electron self-interaction. In particular, the exchange potential does not have the correct inverse-distance asymptotic behavior; rather, it decreases exponentially with the distance

between the redox centers,<sup>29</sup> thus yielding a wrongly delocalized HOMO and a correspondingly overestimated electronic coupling. Self-interaction errors (whose importance can generally depend on the nature and configuration of the system under study) are not accounted for by eqs 35, 36, and 37, which delimit only the errors arising from electron smearing.

A substantial SIC rests on the appropriate treatment of the electronic correlation effects. Such effects can be expected to cancel, to a large extent, when the energy difference  $\Delta E_{\text{IF}}$  between the diabatic states is computed. On the other hand, since the expected electronic coupling is much smaller than  $\Delta E_{\text{IF}}$ , the delocalization of the transferring electron and thus the overlaps  $a$  and  $b$  of eq 3 can still be considerably affected.

In order to improve the description of the valence electron charge distribution (in particular, by taking properly into account the strong electron correlation effects) an orbital-dependent correction functional  $E_U$ , adapted from the Hubbard model<sup>90</sup> (dealing with strongly correlated electron systems), is added to the standard DFT functional  $E_{\text{DFT}}$ . According to the DFT +  $U$  approach, first introduced by Anisimov and co-workers,<sup>91,92</sup> a few localized orbitals (the  $d$  orbitals for the transition metals) are selected and the corresponding correlation is treated in a special way. The magnitude of  $E_U$  and thus the amount of the corrective electron correlation is controlled by the Hubbard  $U$  parameter, which measures the screened on-site Coulomb interaction. The total energy functional is written as

$$E_{\text{DFT}+U}[n] = E_U[\{n_{mm}^{I,s}\}] + E_{\text{DFT}}[n] \quad (38)$$

where  $I$  identifies the atomic site experiencing the Hubbard-like interaction (i.e., the ferric and ferrous species in the system under study),  $s$  denotes the electron spin,<sup>93</sup>  $m$  is the magnetic quantum number, and  $\mathbf{n}^{I,s}$  is the atomic orbital occupation matrix, which describes the degrees of freedom associated to the strongly correlated electrons on which the Hubbard  $U$  acts. In this work we adopt a rotationally invariant DFT +  $U$  scheme,<sup>28,82</sup> where the  $U$  parameter is obtained from first-principles, using a linear response approach internally consistent with the chosen definition for the occupation matrix.<sup>28</sup> The expression for  $E_U$  is

$$E_U[\{n_{mm}^{I,s}\}] = \frac{U}{2} \sum_{I,s} \text{Tr}[\mathbf{n}^{I,s}(1 - \mathbf{n}^{I,s})] \quad (39)$$

which is described and detailed in ref 28. It is worth stressing that the Hubbard  $U$  computed within the scheme of ref 28 is not an empirical fitting parameter. It is a truly ab initio quantity, derived from the bare and screened linear responses of the system to a change in the occupation numbers.

The orbital-dependent correction potential changes the level structure because of the strict connection between level energy and occupation number embedded into eq 39. In the  $\sigma$  range corresponding to the linear regime the minority-spin (nominal) HOMO and LUMO are the only  $d$ -like MOs close in energy (0.01–0.3 eV) and with appreciable fractional occupation. They are lowered relative to the other levels corresponding to  $d$ -like MOs by an amount dependent on the nuclear configuration and the value of  $\sigma$  but generally

of the order of 1–2 eV. This is also the order of magnitude of the separation from the remaining lower lying and higher lying levels. Hence, the value  $N_d = 2$  has to be used in the equations of section 2.4.

**Acknowledgment.** We thank Stefano Corni, Clotilde Cucinotta, Grace Brannigan, Giacomo Fiorin, Matteo Cococcioni, Nicola Marzari, and Paolo Giannozzi for helpful discussions. This work was supported by the NIH, grant no. GM 067689.

**Supporting Information Available:** Relaxation of the Condon approximation in the expression for the electron-transfer rate constant; derivation of the equations in section 2.5, further analytical development, and computational tests; transfer integral vs  $\sigma$  in the presence of the Hubbard  $U$  correction; table with the values of the electronic couplings and pathway products for the individual nuclear configurations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Kuznetsov, A. M.; Ulstrup, J. *Electron Transfer in Chemistry and Biology*; John Wiley & Sons: New York, 1999.
- (2) Brunschwig, B. S.; Logan, J.; Newton, M. D.; Sutin, N. *J. Am. Chem. Soc.* **1980**, *102*, 5798–5809.
- (3) Logan, J.; Newton, M. D. *J. Chem. Phys.* **1983**, *78*, 4086–4091.
- (4) Tembe, B. L.; Friedman, H. L.; Newton, M. D. *J. Chem. Phys.* **1982**, *76*, 1490–1507.
- (5) Kuharski, R. A.; Bader, J. S.; Chandler, D.; Sprik, M.; Klein, M. L.; Impey, R. W. *J. Chem. Phys.* **1988**, *89*, 3248–3257.
- (6) Rustad, J. R.; Rosso, K. M.; Felmy, A. R. *J. Chem. Phys.* **2004**, *120*, 7607–7615.
- (7) Cui, D. Q.; Eriksen, T. E. *Environ. Sci. Technol.* **1996**, *30*, 2259–2262.
- (8) Buerge, I. J.; Hug, S. J. *Environ. Sci. Technol.* **1998**, *32*, 2092–2099.
- (9) Amonette, J. E.; Workman, D. J.; Kennedy, D. W.; Fruchter, J. S.; Gorby, Y. A. *Environ. Sci. Technol.* **2000**, *34*, 4606–4613.
- (10) Nitzan, A. *Annu. Rev. Phys. Chem.* **2001**, *52*, 681–750.
- (11) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265–322.
- (12) Newton, M. D. *Chem. Rev.* **1991**, *91*, 767–792.
- (13) Landau, L. D. *Phys. Z. Sowjetunion* **1932**, *1*, 88–98. *ibid.* **1932**, *2*, 46–51.
- (14) Zener, C. *Proc. R. Soc. London A* **1932**, *137*, 696–702. *ibid.* **1933**, *140*, 660–668.
- (15) Newton, M. D. *J. Phys. Chem.* **1988**, *92*, 3049–3056.
- (16) Beratan, D. N.; Onuchic, J. N.; Hopfield, J. J. *J. Chem. Phys.* **1987**, *86*, 4488–4498.
- (17) Farazdel, A.; Dupuis, M.; Clementi, E.; Aviram, A. *J. Am. Chem. Soc.* **1990**, *112*, 4206–4214.
- (18) Cave, R. J.; Newton, M. D. *Chem. Phys. Lett.* **1996**, *249*, 15–19.
- (19) Prezhdov, O. V.; Kindt, J. T.; Tully, J. C. *J. Chem. Phys.* **1999**, *111*, 7818–7827.

- (20) Voityuk, A. A.; Rösch, N.; Bixon, M.; Jortner, J. *J. Phys. Chem. B* **2000**, *104*, 9740–9745.
- (21) Stuchebrukhov, A. A. *Theor. Chem. Acc.* **2003**, *110*, 291–306.
- (22) Prytkova, T. R.; Kurnikov, I. V.; Beratan, D. N. *J. Phys. Chem. B* **2005**, *109*, 1618–1625.
- (23) Migliore, A.; Corni, S.; Di Flice, R.; Molinari, E. *J. Chem. Phys.* **2006**, *124*, 064501.
- (24) Migliore, A.; Corni, S.; Di Felice, R.; Molinari, E. *J. Phys. Chem. B* **2006**, *110*, 23796–23800.
- (25) Migliore, A.; Corni, S.; Di Felice, R.; Molinari, E. *J. Phys. Chem. B* **2007**, *111*, 3774–3781.
- (26) Troisi, A.; Ratner, M. A.; Zimmt, M. B. *J. Am. Chem. Soc.* **2004**, *126*, 2215–2224.
- (27) Bu, Y.; Wang, Y.; Xu, F.; Deng, C. *J. Mol. Struct. (Theochem)* **1998**, *453*, 43–48.
- (28) Cococcioni, M.; de Gironcoli, S. *Phys. Rev. B* **2005**, *71*, 035105.
- (29) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley-VCH Verlag GmbH: New York, 2000.
- (30) Sit, P. H.-L.; Cococcioni, M.; Marzari, N. *Phys. Rev. Lett.* **2006**, *97*, 028303.
- (31) Rosso, K. M.; Rustad, J. R. *J. Phys. Chem. A* **2000**, *104*, 6718–6725.
- (32) Sutin, N. Theory of electron transfer reactions. In *Electron-transfer and electrochemical reactions. Photochemical reactions and other energized reactions*; Zuckermann, J. J., Ed.; VCH: New York, 1986; Vol. 15, p 16.
- (33) Newton, M. D.; Sutin, N. *Annu. Rev. Phys. Chem.* **1984**, *35*, 437–480.
- (34) Troisi, A.; Nitzan, A.; Ratner, M. A. *J. Chem. Phys.* **2003**, *119*, 5782–5788.
- (35) Okuno, Y. *J. Chem. Phys.* **1999**, *111*, 8034–8038.
- (36) Page, C. C.; Moser, C. C.; Chen, X.; Dutton, P. L. *Nature* **1999**, *402*, 47–52.
- (37) Indeed, the poor convergence is a common feature for self-consistent field calculations on open-shell transition-metal systems both within the unrestricted or open-shell restricted Hartree–Fock (HF) and the unrestricted DFT approaches, see: Neese, F. *Chem. Phys. Lett.* **2000**, *325*, 93–98.
- (38) Dreizler, R. M.; Gross, E. K. U. *Density functional theory*; Springer-Verlag: Berlin Heidelberg, 1990.
- (39) For brevity, the spin dependence is not explicitly shown; however, the index  $i$  can be defined in such a way to distinguish the spin state.
- (40) Cohen-Tannoudji, C.; Diu, B.; Laloë, F. *Quantum Mechanics*; Hermann: Paris, 1977; Vol. 2.
- (41) Jamorski, C.; Martinez, A.; Castro, M.; Salahub, D. R. *Phys. Rev. B* **1997**, *55*, 10905–10921.
- (42) Dunlap, B. I.; Rösch, N. *Adv. Quantum Chem.* **1990**, *21*, 317–339.
- (43) Natiello, M. A.; Scuseria, G. E. *Int. J. Quantum Chem.* **1984**, *26*, 1039–1049.
- (44) Springborg, M.; Albers, R. C.; Schmidt, K. *Phys. Rev. B* **1998**, *57*, 1427–1435.
- (45) Görling, A. *Phys. Rev. A* **1996**, *54*, 3912–3915.
- (46) Michelini, M. C.; Pis Diez, R.; Jubert, A. H. *Int. J. Quantum Chem.* **1998**, *70*, 693–701.
- (47) Barcaro, G.; Fortunelli, A. *Faraday Discuss.* **2008**, DOI: 10.1039/b705105k. In the above article the HOMO–LUMO gap is derived from the nominal (fractionally occupied) HOMO and LUMO.
- (48) Koopmans, T. *Physica* **1934**, *1*, 104–113.
- (49) Luo, J.; Xue, Z. Q.; Liu, W. M.; Wu, J. L.; Yang, Z. Q. *J. Phys. Chem. A* **2006**, *110*, 12005–12009.
- (50) Cardano, G. *Artis magna, sive de regulis algebraicis*; Petrius: Nuremberg, 1545.
- (51) The validity of this consideration when  $\sigma$  approaches zero is clearly not affected by the fact that the HOMO can result from a wrong linear combination of d-like orbitals.
- (52) Grotheer, O.; Fähnle, M. *Phys. Rev. B* **1998**, *58*, 13459–13464.
- (53) According to the second-order energy correction in the stationary perturbation theory the mixing between two levels is determined not only by the difference between their unperturbed energies, but also by the matrix element of the perturbation Hamiltonian term between the corresponding orbital states. The latter depends also on the orbital localization. On the other hand, the KS level structure and the shapes of the orbitals reflect the approximate degeneracy of the two metal sites. Therefore, each level in the multiplet can mix in a similar way with couples of very close levels  $\epsilon_k$  corresponding to symmetrically arranged orbitals.
- (54) Kryachko, E. S.; Ludeña, E. V. *Energy density functional theory of many-electron systems*; Kluwer: Dordrecht, 1990.
- (55) Beste, A.; Harrison, R. J.; Yanai, T. *J. Chem. Phys.* **2006**, *125*, 074101.
- (56) Luo, J.; Yang, Z. Q.; Xue, Z. Q.; Liu, W. M.; Wu, J. L. *J. Chem. Phys.* **2006**, *125*, 094702.
- (57) Slater, J. C. *The self-consistent field for molecules and solids*; McGraw-Hill: New York, 1974; Vol. 4.
- (58) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- (59) Baroni, S.; Dal Corso, A.; de Gironcoli, S.; Giannozzi, P.; Cavazzoni, C.; Ballabio, G.; Scandolo, S.; Chiarotti, G.; Focher, P.; Pasquarello, A.; Laasonen, K.; Trave, A.; Car, R.; Marzari, N.; Kokalj, A. <http://www.pwscf.org/>.
- (60) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (61) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (62) Knops-Gerrits, P. P.; Jacobs, P. A.; Fukuoka, A.; Ichikawa, M.; Faglioni, F.; Goddard, W. A., III *J. Mol. Catal. A: Chem.* **2001**, *166*, 3–13.
- (63) (a) Ferretti, A.; Ruini, A.; Bussi, G.; Molinari, E.; Caldas, M. J. *Phys. Rev. B* **2004**, *69*, 205205. (b) Ferretti, A. DTI program, 2005; available from <http://www.s3.infm.it/dti> on request. Contact information: ferretti.andrea@unimore.it.
- (64) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.
- (65) Mattsson, A. E.; Armiento, R.; Schultz, P. A.; Mattsson, T. R. *Phys. Rev. B* **2006**, *73*, 195123.
- (66) Balabin, I. A.; Onuchic, J. N. *Science* **2000**, *290*, 114–117.
- (67) Lin, J.; Balabin, I. A.; Beratan, D. N. *Science* **2005**, *310*, 1311–1313.



- (68) Prytkova, T. R.; Kurnikov, I. V.; Beratan, D. N. *Science* **2007**, *315*, 622–625.
- (69) Ruiz, E.; Salahub, D. R.; Vela, A. *J. Phys. Chem.* **1996**, *100*, 12265–12276.
- (70) Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *125*, 164105.
- (71) Taylor, J. R. *Introduzione all'analisi degli errori. Lo studio delle incertezze nelle misure fisiche*, 2nd ed.; Zanichelli: Bologna, 1999.
- (72) Skourtis, S. S.; Balabin, I. A.; Kawatsu, T.; Beratan, D. N. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3552–3557.
- (73) Jones, M. L.; Kurnikov, I. V.; Beratan, D. N. *J. Phys. Chem. A* **2002**, *106*, 2002–2006.
- (74) The pathway products are computed by using Kurnikov's HARLEM program, which is available from <http://www-w.kurnikov.org/>.
- (75) The correlation between ab initio transfer integrals and pathway products can be measured by the correlation coefficients, which are  $r_{T,V} = 0.51$  and  $r_{T,V^*} = 0.69$ . The probabilities of finding at least equal values of those coefficients, if the corresponding data sets are uncorrelated, are  $P_9(r \geq r_{T,V}) = 16\%$  and  $P_9(r \geq r_{T,V^*}) = 4\%$ , respectively. These values can be compared with the commonly accepted threshold of 5% for delimiting significant correlations. The two probabilities get closer to each other by excluding the two nuclear configurations not including the hydrogen bond in the best ET pathway. In fact, in this event we obtain  $r_{T,V} = 0.73$  and  $r_{T,V^*} = 0.76$ , from which  $P_7(r \geq r_{T,V}) = 6\%$  and  $P_7(r \geq r_{T,V^*}) = 5\%$ , respectively.
- (76) Miyashita, O.; Okamura, M. Y.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3558–3563.
- (77) Sutin, N. *Acc. Chem. Res.* **1982**, *15*, 275–282.
- (78) The two estimates differ by about 1.4 standard deviations. The one-tail probability of obtaining a discrepancy which is at least 1.4 standard deviations is 8%. In other words, by assuming that our average value complies with a normal distribution centered on the expected (i.e., experimental) transfer integral, the probability that our single valuation of the rms electronic coupling gives a result at least as large as  $11.0 \times 10^{-3}$  eV is 8%. Therefore, according to the usual 5% criterion the discrepancy between the two values is not significant.
- (79) In ref 3 the Condon approximation is tested on the apex-to-apex conformation by translating the reactants along the metal–metal direction.
- (80) Janak, J. F. Proof that  $\partial E/\partial n_i = \epsilon_i$  in density-functional theory. *Phys. Rev. B* **1978**, *18*, 7165–7168.
- (81) Weinert, M.; Davenport, J. W. *Phys. Rev. B* **1992**, *45*, 13709–13712.
- (82) Cococcioni, M.; Dal Corso, A.; de Gironcoli, S. *Phys. Rev. B* **2003**, *67*, 094106.
- (83) Mermin, N. D. *Phys. Rev.* **1965**, *137A*, 1441–1443.
- (84) De Vita, A Ph.D. thesis, University of Keele, 1992.
- (85) Elsässer, C.; Fähnle, M.; Chan, C. T.; Ho, K. M. *Phys. Rev. B* **1994**, *49*, 13975–13978.
- (86) Takeda, R.; Yamanaka, S.; Yamaguchi, K. *Int. J. Quantum Chem.* **2003**, *93*, 317–323.
- (87) Born, M. *Fisica Atomica*; Bollati Boringhieri: Torino, 1993.
- (88) The argument can be suitably extended to the generic orbital. However, the analysis provided in the main text is appropriate for evaluation of the transfer integral.
- (89) Harrison, W. A. *Elementary Electronic Structure*; World Scientific: Singapore, 1999.
- (90) Hubbard, J. *Proc. R. Soc. A* **1963**, *276*, 238–257.
- (91) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. *Phys. Rev. B* **1991**, *44*, 943–954.
- (92) Anisimov, V. I.; Solovyev, I. V.; Korotin, M. A.; Czyżyk, M. T.; Sawatzky, G. A. *Phys. Rev. B* **1993**, *48*, 16929–16934.
- (93) The symbol  $s$  is here adopted, in place of the commonly used  $\sigma$ , to avoid confusion with the notation for the Gaussian broadening parameter.

CT800340V

## Benchmark Energetic Data in a Model System for Grubbs II Metathesis Catalysis and Their Use for the Development, Assessment, and Validation of Electronic Structure Methods

Yan Zhao and Donald G. Truhlar\*

Department of Chemistry and Supercomputing Institute, University of Minnesota,  
Minneapolis, Minnesota 55455-0431

Received September 15, 2008

**Abstract:** We present benchmark relative energetics in the catalytic cycle of a model system for Grubbs second-generation olefin metathesis catalysts. The benchmark data were determined by a composite approach based on CCSD(T) calculations, and they were used as a training set to develop a new spin-component-scaled MP2 method optimized for catalysis, which is called SCSC-MP2. The SCSC-MP2 method has improved performance for modeling Grubbs II olefin metathesis catalysts as compared to canonical MP2 or SCS-MP2. We also employed the benchmark data to test 17 WFT methods and 39 density functionals. Among the tested density functionals, M06 is the best performing functional. M06/TZQS gives an MUE of only 1.06 kcal/mol, and it is a much more affordable method than the SCSC-MP2 method or any other correlated WFT methods. The best performing meta-GGA is M06-L, and M06-L/DZQ gives an MUE of 1.77 kcal/mol. PBEh is the best performing hybrid GGA, with an MUE of 3.01 kcal/mol; however, it does not perform well for the larger, real Grubbs II catalyst. B3LYP and many other functionals containing the LYP correlation functional perform poorly, and B3LYP underestimates the stability of stationary points for the *cis*-pathway of the model system by a large margin. From the assessments, we recommend the M06, M06-L, and MPW1B95 functionals for modeling Grubbs II olefin metathesis catalysts. The local M06-L method is especially efficient for calculations on large systems.

### 1. Introduction

The ground-breaking advances in catalytic olefin metathesis<sup>1–3</sup> have revolutionized organic synthesis and greatly broadened the scope of its applicability to medicine, biology, and materials science as well as promoting green chemistry. As a result of its impact, the Nobel Prize in Chemistry 2005 was awarded to Chauvin,<sup>4</sup> Grubbs,<sup>5</sup> and Schrock<sup>6</sup> “for the development of the metathesis method in organic synthesis”.<sup>7</sup> Schrock’s Mo-based olefin metathesis catalysts are air sensitive but generally more active than air-stable Grubbs’ Ru-based catalysts, and they are complementary in reactivities and other properties.<sup>2,8</sup> Grubbs second-generation (Grubbs II) Ru metathesis catalysts<sup>9–16</sup> are a hundred to a thousand

times more active than first-generation Ru metathesis catalysts, and they also exhibit greater thermal and chemical stability with significant functional group tolerance.<sup>9–11</sup> The difference between the Grubbs I and II catalysts is the substitution of one of the phosphine ligands, usually tricyclohexylphosphine, PCy<sub>3</sub>, of the bisphosphine first-generation precatalyst, (PCy<sub>3</sub>)<sub>2</sub>Cl<sub>2</sub>Ru=CHPh, by a *N*-heterocyclic carbene (NHC), usually 1,3-dimesityl-4,5-dihydro-2-ylidene, which is abbreviated as H<sub>2</sub>IMes.

Together with experimental studies,<sup>1–3,9–16</sup> density functional theory (DFT) has been used in the past decade to model the mechanisms in Grubbs catalysts;<sup>17–32</sup> most of the computational studies employed the BP86 or B3LYP functionals. BP86 was chosen due to its early success in describing metal–carbonyl compounds,<sup>33–35</sup> whereas B3LYP

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

**Table 1.** Basis Sets Employed in the Present Study

basis sets	Ru		H	C,N,P,Cl	$N^a$
	ECP	valence basis			
DZQ	CEP <sup>b</sup>	CEP <sup>b</sup>	6-31+G(d,p) <sup>c</sup>	6-31+G(d,p) <sup>c</sup>	260
TZQS	CEP <sup>b</sup>	CEP+d3f <sup>b,d</sup>	MG3S <sup>e</sup>	MG3S <sup>e</sup>	522
ATZQ	aug-cc-pVTZ-PP <sup>f</sup>	aug-cc-pVTZ-PP <sup>f</sup>	cc-pVTZ <sup>g</sup>	aug-cc-pVTZ <sup>g</sup>	668
AQZQ	aug-cc-pVQZ-PP <sup>f</sup>	aug-cc-pVQZ-PP <sup>f</sup>	cc-pVQZ <sup>g</sup>	aug-cc-pVQZ <sup>g</sup>	1197
AQZQ+d	aug-cc-pVQZ-PP <sup>f</sup>	aug-cc-pVQZ-PP <sup>f</sup>	cc-pVQZ <sup>g</sup>	aug-cc-pV(Q+d)Z <sup>h</sup>	1212

<sup>a</sup>  $N$  is the number of contracted basis functions for 1, and “valence basis” denotes the basis set used for the 16 electrons of Ru that are treated explicitly. The other 28 electrons (a [Ar]3d<sup>10</sup> core) of Ru are replaced by a relativistic ECP. <sup>b</sup> Reference 58. <sup>c</sup> Reference 57. <sup>d</sup> References 55 and 56. <sup>e</sup> Reference 59. <sup>f</sup> Reference 62. <sup>g</sup> Reference 61. <sup>h</sup> The aug-cc-pVQZ is used for C and N atoms, and the aug-cc-pV(Q+d)Z basis set<sup>63</sup> is used for P and Cl atoms.

is the most popular DFT method; many users have used B3LYP as a “reliable” black-box computational tool. However, evidence delineating poor performance of popular density functionals in several areas in chemistry has been presented by many research groups.<sup>36–52</sup> Indeed, BP86 and B3LYP are not accurate for the description of Grubbs metathesis catalysis, as shown by Tsipis et al.<sup>23</sup> and us;<sup>30,53</sup> both functionals fail to predict the trend of the phosphine binding energies between the first- and second-generation Grubbs’ ruthenium precatalysts for olefin metathesis.<sup>30,31,53</sup> Moreover, Piacenza et al.<sup>31</sup> assessed the performance of five density functionals against benchmark energetic data for RuCl<sub>2</sub>(PH<sub>3</sub>)<sub>2</sub>CH<sub>2</sub>, a small model system for the first-generation Grubbs catalysts. They found that B3LYP gives the worst performance with a maximum error of 17.8 kcal/mol. Thus, we think that it is now important to scrutinize the strength and limitation of popular and new-generation density functionals for the description of Grubbs olefin metathesis catalysis.

In order to assess the performance of density functionals for olefin metathesis, one needs to use accurate benchmark data. Unfortunately, there are very few experimental data<sup>16</sup> that one can directly compare to. Therefore we generate high-quality data by using the most reliable available levels of wave function theory (WFT). For real catalysts having the size of the Grubbs olefin catalysts, state-of-the-art correlated WFT methods (for example, CCSD(T)<sup>54</sup>) are prohibitively expensive. Alternatively, one can use CCSD(T) on small model systems for the Grubbs olefin metathesis catalysts. One objective of the present study is to use high-level CCSD(T) theory to develop benchmark energetic data for ethene metathesis reactions catalyzed by a model system [(PH<sub>3</sub>)(C<sub>3</sub>H<sub>6</sub>N<sub>2</sub>)Cl<sub>2</sub>Ru=CH<sub>2</sub>] (**1**) that mimics the coordinate covalent bonding in Grubbs second-generation catalysts.

Another goal of our study is to validate a number of low-cost density functional theory (DFT) methods and to determine if there are DFT methods that can describe the energetics of coordinate covalent bonding in Grubbs second-generation olefin metathesis catalysis sufficiently well for practical simulations.

This paper is organized as follows. The computational details and DFT methods are described in Section 2, and results and discussion are in Section 3. Section 4 presents concluding remarks.

## 2. Computational Methods

**2.1. Basis Sets.** In the present study we employed five basis sets, and they are listed in Table 1. The DZQ basis set was defined elsewhere,<sup>55,56</sup> it uses the 6-31+G(d,p)<sup>57</sup> basis set for main-group elements and uses the relativistic effective core potential and valence basis set of Stevens et al.<sup>58</sup> for Ru. The TZQS basis set is slightly different from the TZQ basis used in our previous studies,<sup>55,56</sup> TZQS uses the MG3S<sup>59</sup> basis set (for comparison, TZQ uses MG3<sup>59</sup>) for main-group elements and the same basis set for Ru as in the TZQ basis set. The MG3<sup>59,60</sup> and MG3S<sup>59</sup> basis sets are triple- $\zeta$  quality basis sets, and they have been defined in previous studies. The ATZQ basis set is also of a triple- $\zeta$  quality, and it employs the cc-pVTZ<sup>61</sup> basis for H, the aug-cc-pVTZ<sup>61</sup> basis for C, N, P, and Cl, and the aug-cc-pVTZ-PP<sup>62</sup> relativistic effective core potential and basis set for Ru. The AQZQ basis set is of a quadruple- $\zeta$  quality; it employs the cc-pVQZ<sup>61</sup> basis for H, the aug-cc-pVQZ<sup>61</sup> basis for C, N, P and Cl, and the aug-cc-pVQZ-PP<sup>62</sup> relativistic effective core potential and basis set for Ru. The fifth basis set, labeled AQZQ+d, differs from the fourth by the use of aug-cc-pV(Q+d)Z<sup>63</sup> for P and Cl. Note that the MG3S and aug-cc-pV(Q+d) Z basis sets for P and Cl include tight  $d$  functions, but the other basis sets used for P and Cl do not.

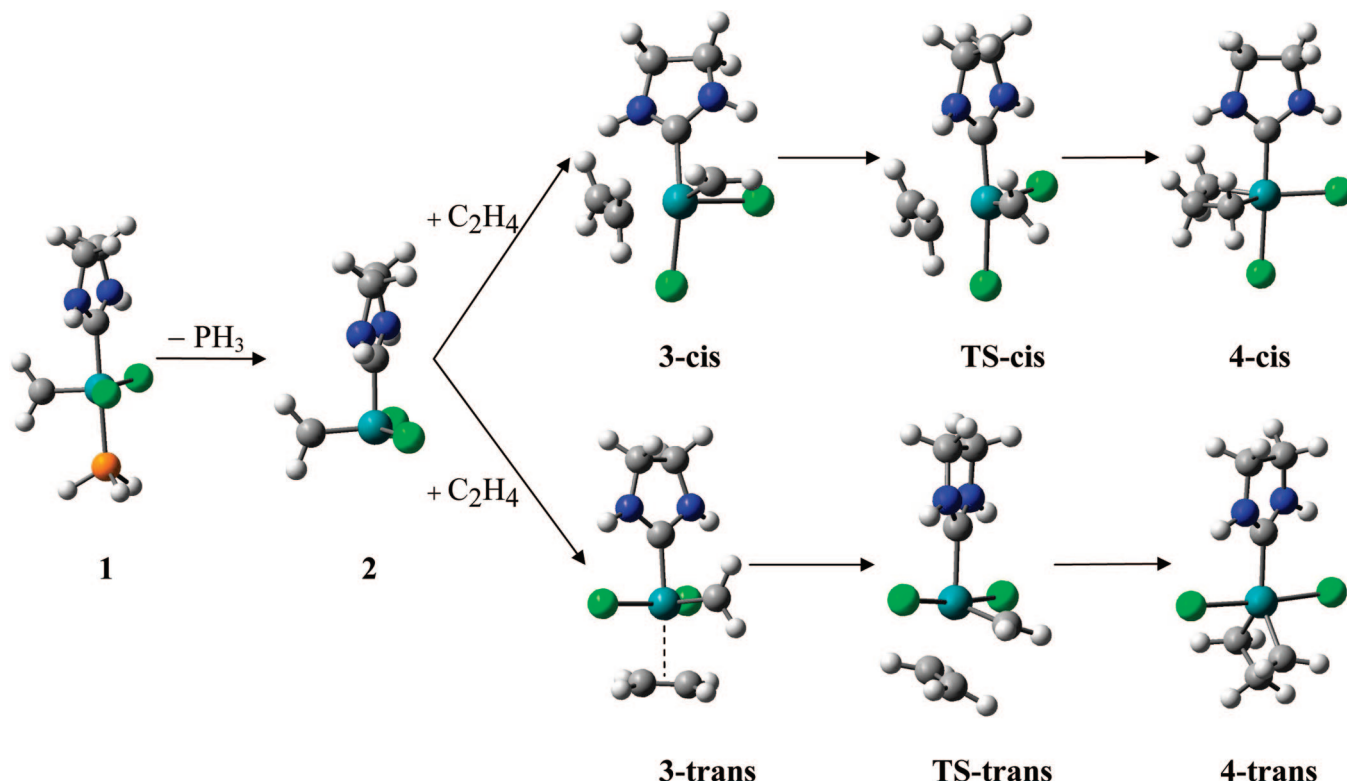
**2.2. Geometries and Energies.** The geometries of all stationary points in the catalytic cycle (Figure 1) of the model system [(PH<sub>3</sub>)(C<sub>3</sub>H<sub>6</sub>N<sub>2</sub>)Cl<sub>2</sub>Ru=CH<sub>2</sub>] (**1**) were optimized at the M06-L/TZQS level. For the purpose of comparison, we also carried out geometry optimizations at the M06-L/DZQ, M06/TZQS, BP86/TZQS, and B3LYP/TZQS levels.

All energies in the present paper are Born–Oppenheimer electronic energies including nuclear repulsion but not including zero-point vibrational energies or thermal vibrational–rotational energy.

**2.3. Benchmark Calculations.** Even for the model system [(PH<sub>3</sub>)(C<sub>3</sub>H<sub>6</sub>N<sub>2</sub>)Cl<sub>2</sub>Ru=CH<sub>2</sub>] (**1**), the CCSD(T)/AQZQ level of theory is too computationally demanding. We estimated the CCSD(T)/AQZQ+d relative energies for all stationary points by using a composite approach:

$$E(\text{est. CCSD(T)/AQZQ+d}) = E(\text{MP2/AQZQ+d}) + (E(\text{CCSD(T)/ATZQ}) - E(\text{MP2/ATZQ})) \quad (1)$$

**2.4. Optimization of a Spin-Component-Scaled Second-Order Møller–Plesset Perturbation Theory for Olefin Metathesis.** Second-order Møller–Plesset (MP2)<sup>64,65</sup> perturbation theory is the simplest and least expensive first-



**Figure 1.** Structures of the stationary points in the catalytic cycle of the model Grubbs II catalyst  $[(\text{PH}_3)(\text{NHC})\text{Cl}_2\text{Ru}=\text{CH}_2]$  (1).

principles WFT method that incorporates dynamical electron correlation in a systematic way. Recently Grimme<sup>66</sup> developed a spin-component-scaled MP2 (SCS-MP2) method by separate scaling of parallel- and antiparallel-spin pair correlation energies, in order to improve the accuracy over standard MP2 theory. (This represents a generalization of the SAC-MP2 method<sup>67</sup> in which both components of the MP2 correlation energy are scaled with the same factor.) The total SCS-MP2 energy can be written as

$$E_{\text{SCS-MP2}} = E_{\text{HF}} + p_S E_S + p_T E_T \quad (2)$$

where  $E_{\text{HF}}$  is the Hartree–Fock energy,  $E_S$  is the singlet (antiparallel spin) pair correlation energy, and  $E_T$  is the triplet (parallel spin) correlation energy. The scaling factors used by Grimme are  $p_S = 6/5$  and  $p_T = 1/3$ . Subsequently Hill and Platts<sup>68</sup> reoptimized these two parameters for weak and stacking interaction energies, and they obtained a method called SCSN-MP2, with  $p_S = 0$  and  $p_T = 1.76$ . More recently Distasio and Head-Gordon<sup>69</sup> optimized  $p_S$  and  $p_T$  for intermolecular interaction energies with different basis sets, and they named the new methods SCS(MI)-MP2, in which MI stands for molecular interaction. In the present study, we reoptimized  $p_S$  and  $p_T$  by least-squares fitting to high-level CCSD(T) energetic data in the model Grubbs II metathesis reactions with the AQZQ basis set. We used both the relative energies and absolute total energies in the optimization because we found that the optimization would produce unphysical parameters if the total energies were not included in the training set. Thus the final training set has 10 absolute energies and 7 relative energies. The resulting new parameters are  $p_S = 1.363$  and  $p_T = 0.568$ , and we call the new method SCSC-MP2, which stands for SCS-MP2

optimized for catalysis. The  $p_S$  and  $p_T$  parameters in SCS(MI)-MP2 or SCSN-MP2 have been optimized against the main-group noncovalent interaction energies, so these parameters are weighted toward the prediction of noncovalent relative energies. The training data for SCSC-MP2 include energetic data for covalent and noncovalent interactions and for transition states involving transition metals. Thus the SCSC-MP2 model chemistry is optimized against a more diverse set of data than SCS(MI)-MP2 or SCSN-MP2, and the optimized values of  $p_S$  and  $p_T$  look very reasonable in magnitude.

**2.5. Tested Methods.** We test 17 WFT methods, including HF, MP2, SCS-MP2, SCSC-MP2, SCS(MI)-MP2, SCSN-MP2, CCSD, and CCSD(T) with three basis sets.

We tested 39 density functionals with the M06-L/TZQS geometries. The tested functionals can be classified according to various rungs of “Jacob’s ladder”<sup>70</sup> The lowest rung is the local spin density approximation (LSDA), in which the density functional depends only on spin densities, and the second rung is the generalized gradient approximation (GGA, in which the density functional depends on spin densities and their reduced gradient). The third rung is meta-GGA, in which the functional also depends on the spin kinetic energy densities. The fourth rung is hyper GGA,<sup>70</sup> which employs full or partial exact Hartree–Fock (HF) exchange. There are two types of hyper GGAs on the fourth rung, namely the hybrid GGAs (HF + GGA) and hybrid meta-GGAs (HF + meta-GGA). In this work, the tested functionals include 8 GGAs (BLYP,<sup>71,72</sup> BP86,<sup>71,73</sup> G96LYP,<sup>72,74</sup> HCTH,<sup>75</sup> mPWLYP,<sup>72,76</sup> mPWPW,<sup>76</sup> OLYP,<sup>72,77</sup> and PBE<sup>78</sup>), 6 meta-GGAs (BB95,<sup>79</sup> M06-L,<sup>80</sup> mPWB95,<sup>76,79</sup> TPSS,<sup>81,82</sup>

**Table 2.** WFT Relative Energetics (kcal/mol)<sup>a</sup>

method	cost <sup>b</sup>	1	3- <i>cis</i>	3- <i>trans</i>	TS- <i>cis</i>	TS- <i>trans</i>	4- <i>cis</i>	4- <i>trans</i>	MSE	MUE
best estimate <sup>c</sup>		-25.15	-7.69	-17.64	-7.22	-8.84	-20.39	-19.21		
CCSD(T)/ATZQ	250	-24.91	-9.13	-18.73	-8.78	-10.07	-21.73	-20.27	-1.07	1.14
SCSC-MP2/AQZQ <sup>d</sup>	92	-25.37	-6.80	-16.39	-6.52	-4.87	-23.28	-22.08	0.12	1.83
SCSC-MP2/AQZQ+d <sup>d</sup>	93	-25.55	-6.74	-16.36	-6.45	-4.79	-23.21	-22.00	0.15	1.87
SCSC-MP2/ATZQ <sup>d</sup>	1.5	-25.45	-8.44	-17.65	-8.24	-6.15	-24.78	-23.21	-1.11	1.88
CCSD/ATZQ	113	-22.56	-3.58	-15.50	-4.55	-8.26	-18.73	-18.92	2.01	2.01
SCS-MP2/ATZQ <sup>e</sup>	1.5	-21.79	-1.04	-13.17	-1.99	-3.72	-17.36	-18.24	4.12	4.12
SCS-MP2/AQZQ <sup>e</sup>	92	-21.72	0.41	-12.06	-0.48	-2.56	-16.04	-17.22	5.21	5.21
SCS-MP2/AQZQ+d <sup>e</sup>	93	-21.88	0.47	-12.03	-0.41	-2.48	-15.98	-17.14	5.24	5.24
MP2/AQZQ+d	93	-29.75	-14.51	-21.03	-14.54	-10.09	-31.27	-27.11	-6.02	6.02
MP2/AQZQ	92	-29.54	-14.58	-21.05	-14.62	-10.16	-31.34	-27.18	-6.05	6.05
MP2/ATZQ	1.5	-29.50	-15.95	-22.12	-16.10	-11.31	-32.62	-28.17	-7.09	7.09
SCS(MI)-MP2/AQZQ <sup>f</sup>	92	-32.78	-20.26	-24.44	-21.80	-16.30	-37.45	-30.80	-11.10	11.10
SCSN-MP2/AQZQ <sup>g</sup>	92	-35.49	-25.26	-27.44	-27.20	-20.06	-42.67	-34.07	-15.15	15.15
SCSN-MP2/ATZQ <sup>g</sup>	1.5	-35.18	-25.84	-27.91	-27.94	-20.77	-43.24	-34.60	-15.62	15.62
SCS(MI)-MP2/ATZQ <sup>h</sup>	1.5	-35.74	-27.15	-28.73	-28.65	-20.48	-44.48	-35.51	-16.37	16.37
HF/ATZQ	1.0	-12.67	18.39	-1.31	12.21	-1.25	1.67	-5.06	16.88	16.88
HF/AQZQ	85	-12.80	18.69	-1.06	12.54	-0.83	1.92	-4.71	17.13	17.13

<sup>a</sup> All energies are relative to the 14-electron active catalyst **2**. M06-L/TZQS geometries are used for the calculations involved in this table.

<sup>b</sup> The cost for each method is measured by the computer time for a single point energy calculation of **1** divided by the computer time for an HF/AVTZ energy calculation with the *NWChem* program and 512 processors on the MPP2 computer of EMSL. <sup>c</sup> The best estimates for the relative energies are obtained with eq 1. <sup>d</sup> The optimized scaling factors for SCSC-MP2 are  $p_S = 1.363$  and  $p_T = 0.568$ . <sup>e</sup> The scaling factors for SCS-MP2 are  $p_S = 1.2$  and  $p_T = 1/3$ .<sup>66</sup> <sup>f</sup> The scaling factors used for SCS(MI)-MP2/AQZQ are  $p_S = 0.31$  and  $p_T = 1.46$ . Note that these parameters were optimized for the cc-pVQZ basis set.<sup>69</sup> <sup>g</sup> The scaling factors for SCSN-MP2 are  $p_S = 0$  and  $p_T = 1.76$ .<sup>68</sup> <sup>h</sup> The scaling factors used for SCS(MI)-MP2/ATZQ are  $p_S = 0.17$  and  $p_T = 1.75$ . Note that these parameters were optimized for the cc-pVTZ basis set.<sup>69</sup>

VSXC,<sup>83</sup> and  $\tau$ -HCTH<sup>84</sup>), 13 hybrid GGAs (B3LYP, B97-1,<sup>75</sup> B97-2,<sup>85</sup> B97-3,<sup>86</sup> B98,<sup>87</sup> BHandH,<sup>88</sup> BHandH-LYP,<sup>88</sup> MPW1K,<sup>89</sup> mPW1PW,<sup>76</sup> MPW3LYP,<sup>72,76,90</sup> O3-LYP,<sup>77,91</sup> PBEh,<sup>78</sup> X3LYP<sup>92</sup>), and 12 hybrid meta-GGAs (B1B95,<sup>79</sup> BB1K,<sup>93</sup> BMK, M05,<sup>94</sup> M05-2X,<sup>95</sup> M06,<sup>96</sup> M06-2X,<sup>96</sup> M06-HF,<sup>97</sup> MPW1B95,<sup>90</sup> MPWB1K,<sup>90</sup> TPSSH,<sup>81,82</sup> and  $\tau$ -HCTH<sup>84</sup>).

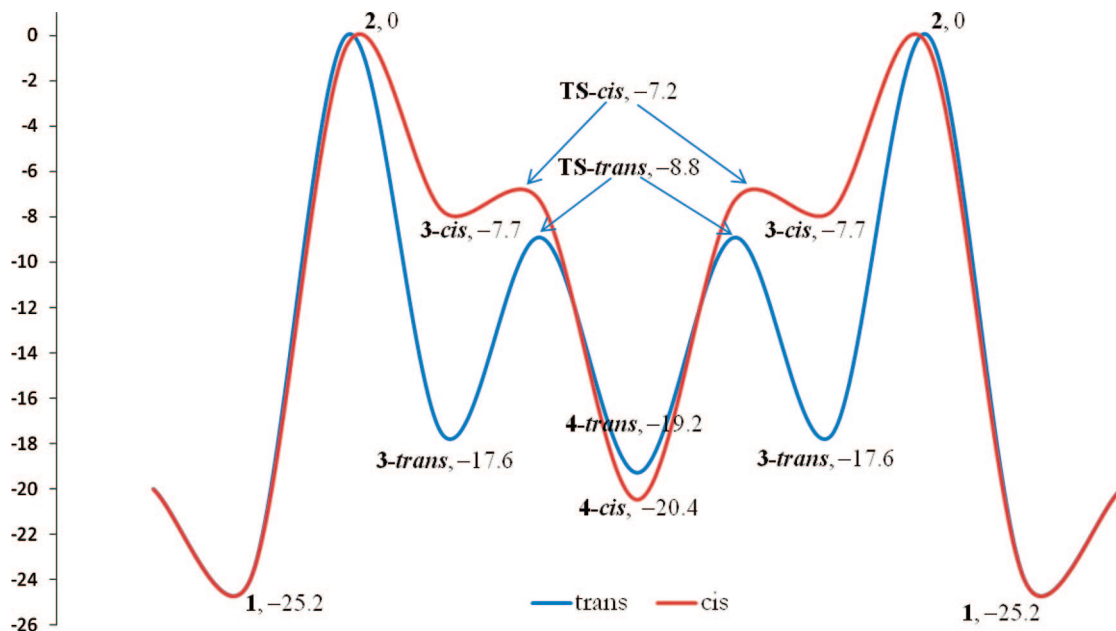
**2.6. Software.** All DFT calculations were carried out using a locally modified *Gaussian03*<sup>98,99</sup> program, and the MP2, SCS, and CCSD(T) calculations were performed with the *NWChem*<sup>100</sup> program.

**2.7. Timings.** Although computer timings are only approximate measures of cost because their exact value depends on the computer program, the computer, the computer's load, and other uncontrolled variables, relative timings calculated with the same program on the same number of processors of the same computer for the same system can be useful for approximately gauging the additional effort required for going to a higher level or different level of theory. Therefore, for each method (a method is a model chemistry,<sup>101</sup> that is, a combination of a theory level and a basis set), for structure **1**, we computed the ratio of the computer time for the single-point energy calculation at hand and a single-point energy calculation at the HF/AVTZ level with the same program on the same number of processors of the same computer. The relative timings are included in the tables where they are labeled as "cost". One technical issue that has a significant effect on the timings is that the computer programs we used (see Section 2.6) employ the resolution of the identity<sup>102,103</sup> for nonhybrid DFT but not for MP2. If one employed a program such as TURBOMOLE for the MP2 and SCS-MP2 calculations, their timings would be smaller. In addition to all the above caveats, the user should be aware that the ratios of timings also depend on the size of the molecule (with the slower methods usually scaling less

efficiently with system size than the faster methods). In light of these considerations, a factor of 1.5 or 2 between two such costs is not very significant, but a factor of more can 2 be a very significant consideration in choosing a method.

### 3. Results and Discussion

**3.1. WFT Calculations.** Table 2 presents the results for WFT methods. As discussed in Section 2.3, the best estimates in Table 2 are calculated with a composite approach (eq 1). The best estimates in Table 2 employ the AQZQ+d basis set (fifth basis of Table 1) for the MP2 component. We also carried out a full set of calculations (not presented in detail) in which we used the AQZQ basis set (fourth basis set of Table 1) for the MP2 step. Comparing these two composite calculations, the largest difference in the seven numbers in the first row of Table 2 is 0.20 kcal/mol, with a mean unsigned deviation of 0.09 kcal/mol. The benchmark relative energetics are illustrated in Figure 2. Figure 2 shows that the *trans*-bounded  $\pi$  complex (**3-trans**) is energetically more favorable than the *cis*-bounded  $\pi$  complex (**3-cis**), and the energy of **3-trans** is lower than **3-cis** by 10 kcal/mol. Furthermore the transition state for the *cis* pathway (**TS-cis**) is energetically less favorable than the *trans*-pathway by 1.6 kcal. The results for the  $\pi$  complexes and transition states seem to favor the *trans*-pathway. Actually the experimental NMR study of Romero and Piers<sup>14</sup> supports the *trans*-bound Ru-cyclobutane model compound. However, as shown in Figure 2, the energy of the *cis*-Ru-cyclobutane (**4-cis**) is energetically more favorable than the *trans* one (**4-trans**) by about 1.2 kcal/mol. Thus one cannot rule out the *cis*-pathway from the high-level WFT calculations on the small gas-phase model system for Grubbs II olefin metathesis catalysts. Indeed, an experimental paper by Ung et al.<sup>13</sup> supported the *cis*-pathway. From the results in Figure 2, we can say that



**Figure 2.** Estimated CCSD(T)/AQZQ potential energy surface (kcal/mol) for the metathesis reaction of the model Grubbs II catalyst  $[(\text{PH}_3)(\text{NHC})\text{Cl}_2\text{Ru}=\text{CH}_2]$  (**1**).

the *trans*-path way is kinetically more favorable ( $E(3\text{-cis}) > E(3\text{-trans})$ ;  $E(\text{TS-cis}) > E(\text{TS-trans})$ ) but thermochemically less favorable ( $E(4\text{-cis}) < E(4\text{-trans})$ ).

We now turn to the performance of the lower-cost WFT methods. The best performing WFT method is CCSD(T)/ATZQ, which gives a small mean unsigned error of 1.05 kcal/mol. However, the computational cost of CCSD(T)/ATZQ is still prohibitively high to be applied to real Grubbs II catalysts. The SCSC-MP2 method optimized in this work gives the second best performance. This is not surprising since we optimized two scaling parameters in SCSC-MP2 against the reference data. A somewhat surprising (but very encouraging) result is that although the two parameters in SCSC-MP2 were optimized with the AQZQ basis set, they work equally well for the ATZQ basis set.

Table 2 also shows that CCSD/ATZQ is slightly worse than SCSC-MP2, but it performs much better than SCS-MP2. Comparing the results of canonical MP2 to those of SCS-MP2, we found that standard MP2 overestimates the stability of all intermediate stationary points on the potential energy surface (PES), whereas SCS-MP2 overcorrects the MP2 method and it underestimates them. The MUE of SCS-MP2 is smaller than that for the canonical MP2 method. Perhaps because the SCS(MI)-MP2 and SCSN-MP2 methods have been optimized only for noncovalent interactions, both methods overestimate the stabilities of all stationary points much more severely than the standard MP2.

**3.2. DFT Calculations.** Table 3 presents the results for DFT methods. Table 3 also includes an *X* column for the percentage of Hartree–Fock (HF) exchange in each functional and a rung column that assigns each functional to a rung of Jacob’s ladder. The best performer in Table 3 is M06, which gives an MUE of 1.18 kcal/mol, surprisingly comparable to the performance of CCSD(T)/ATZQ (MUE = 1.14 kcal/mol). Note that M06 is a much more affordable method than CCSD(T) or MP2.

The second best method is MPW1B95, which is available in all versions of the *Gaussian 03* program. The best performing hybrid GGA is PBEh, which performs very well for this data set, although previous tests<sup>81</sup> have shown PBEh performs worse than B3LYP for main-group thermochemistry. The M06-L local functional, which does not have HF exchange, performs almost as well as PBEh. This is encouraging, since local functionals are well suited for calculations on large molecules where more efficient algorithms<sup>102,104–111</sup> can be employed in the absence of Hartree–Fock exchange. Indeed, with the resolution-of-identity algorithm, the M06-L/AQZQ calculation on the precatalysts  $[(\text{PH}_3)(\text{C}_3\text{H}_6\text{N}_2)\text{Cl}_2\text{Ru}=\text{CH}_2]$  (**1**) is 48 times faster than M06/AQZQ on the same molecule with 4 cores on an IBM BladeCenter Linux cluster.

Our previous tests<sup>55,56,95,96</sup> have shown that high-Hartree–Fock functionals sometimes perform poorly for transition-metal compounds, because the small unsaturated transition-metal species used as test cases often have large near-degeneracy correlation effects. However, for the systems in the present study, the high-HF functionals, such as MPWB1K, BB1K, M06-HF, M05-2X, MPW1K, and M06-2X perform quite well for modeling the model Grubbs II catalyst; they have an MUE less than 4 kcal/mol. This is probably due to the fact that near-degeneracy correlation effects are not dominant in any of the species in the present study. Nevertheless, we originally recommended<sup>96</sup> using M06 or M06-L rather than M06-HF or M06-2X for systems containing transition metals, and the results in Table 3 confirm that that was a good recommendation.

As can be seen from Table 3, the best performing GGA is PBE, and the MUE of PBE is 2.5 kcal/mol smaller than that of BP86.

Another surprising result in Table 3 is that the LYP correlation functional is problematic for describing the relative energetics in the model Grubbs II olefin metathesis

**Table 3.** DFT Relative Energetics (kcal/mol)<sup>a</sup>

method	X	rung <sup>b</sup>	cost <sup>c</sup>	1	3-cis	3-trans	TS-cis	TS-trans	4-cis	4-trans	MSE	MUE
best estimate <sup>d</sup>				-25.15	-7.69	-17.64	-7.22	-8.84	-20.39	-19.21		
M06	27	4 (HM)	0.27	-22.58	-5.76	-16.62	-6.17	-8.65	-19.82	-20.17	0.91	1.18
MPW1B95	31	4 (HM)	0.27	-20.93	-3.12	-14.16	-4.81	-9.68	-20.64	-21.64	1.60	2.60
PBEh	25	4 (HG)	0.24	-21.08	-2.42	-12.93	-3.73	-7.83	-19.39	-20.70	2.58	3.01
M06-L	0	3	0.12	-19.68	-4.84	-14.66	-4.36	-5.18	-18.87	-17.49	3.01	3.01
MPWB1K	44	4 (HM)	0.27	-21.77	-2.71	-14.82	-5.63	-11.90	-22.57	-24.09	0.38	3.27
BB1K	42	4 (HM)	0.27	-20.56	-1.35	-13.37	-4.10	-10.20	-20.79	-22.32	1.92	3.31
M06-HF	100	4 (HM)	0.27	-24.68	-2.61	-18.60	-7.59	-19.61	-19.81	-25.60	-1.77	3.52
M05-2X	56	4 (HM)	0.27	-21.48	-1.30	-15.83	-4.15	-12.74	-17.79	-22.50	1.48	3.53
B1B95	28	4 (HM)	0.27	-19.42	-1.50	-12.39	-2.92	-7.51	-18.36	-19.32	3.53	3.56
MPW1K	42	4 (HG)	0.24	-21.13	0.06	-12.26	-2.92	-9.18	-19.31	-21.71	2.81	3.62
TPSSh	10	4 (HM)	0.27	-20.77	-3.11	-12.86	-3.34	-6.77	-15.86	-16.71	3.82	3.82
PBE	0	2	0.10	-20.12	-4.30	-12.75	-3.28	-4.52	-16.51	-16.40	4.04	4.04
mPWb95	0	3	0.12	-19.24	-4.76	-13.20	-3.58	-4.83	-16.21	-15.51	4.12	4.12
M06-2X	54	4 (HM)	0.27	-18.34	0.90	-13.78	-1.88	-8.89	-16.92	-20.63	3.80	4.22
TPSS	0	3	0.12	-20.53	-3.98	-12.92	-3.30	-5.62	-14.71	-14.97	4.30	4.30
mPW1PW91	25	4 (HG)	0.24	-20.05	-0.62	-11.51	-1.90	-6.27	-16.80	-18.45	4.36	4.36
BMK	42	4 (HM)	0.27	-20.02	0.64	-13.12	-6.83	-14.25	-24.87	-29.15	-0.21	5.46
VSXC	0	3	0.12	-25.06	-15.09	-25.79	-14.16	-15.03	-26.30	-23.53	-5.55	5.57
$\tau$ -HCTHh	15	4 (HM)	0.27	-19.47	-0.13	-11.11	-0.84	-5.10	-13.59	-15.67	5.75	5.75
B97-1	21	4 (HG)	0.24	-19.01	0.85	-10.66	-0.36	-5.16	-13.32	-15.75	6.10	6.10
mPWPW	0	2	0.10	-18.80	-2.09	-10.94	-1.04	-2.54	-13.41	-13.66	6.24	6.24
BB95	0	3	0.12	-17.44	-2.38	-10.96	-1.22	-2.52	-13.69	-13.12	6.40	6.40
M05	28	4 (HM)	0.27	-15.88	1.90	-8.91	0.87	-2.98	-16.52	-18.37	6.61	6.61
BP86	0	2	0.10	-18.49	-1.79	-10.77	-0.77	-2.39	-12.65	-12.98	6.61	6.61
B98	21.98	4 (HG)	0.24	-18.40	2.01	-9.82	0.75	-4.31	-11.63	-14.28	7.21	7.21
BHandH	50	4 (HG)	0.24	-29.24	-10.92	-22.57	-14.48	-20.64	-33.12	-34.42	-8.46	8.46
BHandHLYP	50	4 (HG)	0.24	-18.22	6.59	-8.70	3.43	-5.31	-9.02	-13.83	8.73	8.73
B97-2	21	4 (HG)	0.24	-16.68	4.00	-7.44	3.14	-1.31	-10.64	-12.87	9.19	9.19
MPW3LYP	20	4 (HG)	0.24	-17.63	3.52	-8.92	3.17	-2.11	-7.58	-10.80	9.40	9.40
X3LYP	21.8	4 (HG)	0.24	-17.26	4.15	-8.44	3.61	-1.86	-7.30	-10.61	9.77	9.77
B97-3	26.93	5 (HG)	0.24	-15.86	5.35	-6.82	3.91	-1.04	-9.70	-12.44	9.93	9.93
B3LYP	20	4 (HG)	0.24	-16.33	5.23	-7.31	4.85	-0.46	-5.78	-9.09	11.04	11.04
mPWLYP	0	2	0.10	-15.75	3.40	-7.56	4.98	1.82	-3.10	-5.41	12.08	12.08
$\tau$ -HCTH	0	3	0.12	-14.72	4.43	-5.76	5.53	2.97	-5.71	-7.18	12.24	12.24
BLYP	0	2	0.10	-13.97	5.74	-5.36	7.28	4.09	-0.63	-3.07	14.32	14.32
O3LYP	11.61	4 (HG)	0.24	-11.11	9.42	-1.10	9.43	6.14	-4.39	-6.34	15.46	15.46
HCTH	0	2	0.10	-11.53	8.53	-1.59	9.95	7.77	-3.24	-4.95	15.87	15.87
G96LYP	0	2	0.10	-11.91	8.13	-2.71	9.42	6.52	0.78	-1.37	16.43	16.43
OLYP	0	2	0.10	-8.95	10.56	1.13	11.60	9.75	-1.51	-2.73	18.00	18.00

<sup>a</sup> All energies are relative to the 14-electron active catalyst **2**. The structures of all stationary points are shown in Figure 2. M06-L/TZQS geometries are used for the calculations involved in this table, and the basis set used for the single-point energies is TZQS. <sup>b</sup> GGAs are rung 2, meta functionals that contains spin kinetic energy density are rung 3, hybrid GGAs and hybrid meta functionals are rung 4. On rung 4, hybrid meta GGAs are denoted 4 (HM), and hybrid GGA are denoted 4 (HG). <sup>c</sup> The cost for each method is measured by the computer time for a single point energy calculation of **1** divided by the computer time for an HF/AVTZ energy calculation with the *Gaussian03* program and 4 cores on a IBM BladeCenter Linux cluster. <sup>d</sup> Taken from Table 2.

catalysts; all functionals containing LYP perform poorly. B3LYP gives an MUE of 11.04 kcal/mol, and it underestimates the stability of the stationary points for the *cis*-pathway by a large margin. A similar failure of the LYP correlation functional for Ag clusters has also been reported by Zhao et al.<sup>112</sup>

It is interesting to note that whereas the M06 functional is a many-parameters functional,<sup>96</sup> the MPW1B95 and PBEh functionals were each obtained<sup>78,90</sup> by starting with the exchange and correlation component of previously existing functionals<sup>76,78,79</sup> and combining them with only one new parameter. Although the PBEh functional performs well for the model system, it performs poorly for the real Grubbs catalyst considered in Section 3.4. Therefore it will not be highly recommended for further use in studies of olefin metathesis catalysis.

It is especially striking to compare the cost columns of Tables 2 and 3. Excluding the cost of HF/ATZQ, the costs in Table 3 are a factor of 6–15 times lower than the smallest

cost value in Table 2 and 930–2500 times smaller than the largest cost value in Table 2. This makes the good performance of M06, MPW1B95, and M06-L particularly striking.

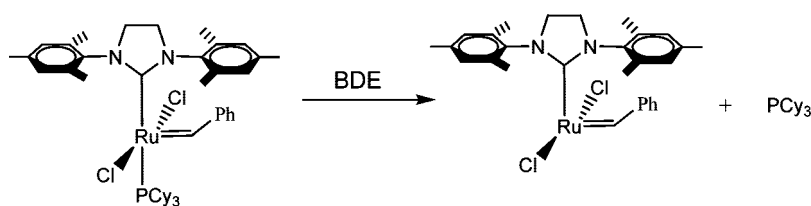
**3.3. Effect of Geometries and Basis Sets.** In previous sections, we based our discussions on single-point energies calculated with the M06-L/TZQS geometries and the TZQS basis set. In this section we compare results with the geometries optimized at the M06-L/DZQ, M06/TZQS, BP86/TZQS, and B3LYP/TZQS levels of theories. We also examine the sensitivity to the size of basis set. The results are shown in Table 4.

As shown in Table 4, M06/TZQS/opt gives a MUE of 1.06 kcal/mol, which is comparable to the CCSD(T)/ATZQ//M06-L/TZQS level. M06/TZQS//M06-L/DZQ gives a slightly smaller MUE than M06/TZQS//M06-L/TZQS, and this is also encouraging because M06-L/DZQ is a much faster method for optimization than the M06-L/TZQS method. Another encouraging result is that M06-L/DZQ gives an MUE of 1.75 kcal/mol, which is smaller than the MUE of

**Table 4.** Effects of Geometries and Basis Sets

method	cost <sup>b</sup>	1	3-cis	3-trans	TS-cis	TS-trans	4-cis	4-trans	MSE	MUE
best estimate <sup>a</sup>		-25.15	-7.69	-17.64	-7.22	-8.84	-20.39	-19.21		
M06/TZQS//opt <sup>c</sup>	0.27	-22.59	-6.25	-16.69	-6.20	-8.85	-19.91	-20.15	0.78	1.06
M06/TZQS//M06-L/DZQ	0.27	-22.46	-5.85	-16.63	-6.26	-9.01	-19.99	-20.33	0.80	1.17
M06/TZQS//M06-L/TZQS	0.27	-22.58	-5.76	-16.62	-6.17	-8.65	-19.82	-20.17	0.91	1.18
M06-L/DZQ//opt <sup>c</sup>	0.03	-20.61	-5.98	-16.44	-5.16	-7.99	-18.60	-19.15	1.75	1.75
M06-L/AQZQ//M06-L/TZQS <sup>d</sup>	0.30	-19.11	-4.45	-13.44	-4.44	-5.97	-19.43	-18.88	2.92	2.92
M06-L/TZQS//M06-L/TZQS	0.12	-19.68	-4.84	-14.66	-4.36	-5.18	-18.87	-17.49	3.01	3.01
BP86/TZQS//opt <sup>c</sup>	0.10	-18.42	-1.73	-12.25	-0.74	-2.30	-12.84	-13.20	6.38	6.38
BP86/AQZQ//BP86/TZQS <sup>d</sup>	0.18	-17.55	-1.33	-10.72	-0.65	-2.65	-13.43	-14.21	6.51	6.51
BP86/TZQS//M06-L/TZQS	0.10	-18.49	-1.79	-10.77	-0.77	-2.39	-12.65	-12.98	6.61	6.61
B3LYP/TZQS//opt <sup>c</sup>	0.24	-16.20	4.71	-7.50	4.51	-1.07	-6.20	-9.19	10.74	10.74
B3LYP/TZQS//M06-L/TZQS	0.24	-16.33	5.23	-7.31	4.85	-0.46	-5.78	-9.09	11.04	11.04

<sup>a</sup> Taken from Table 2. <sup>b</sup> The cost for each method is measured by the computer time for a single point energy calculation of **1** divided by the computer time for an HF/AVTZ energy calculation with the *Gaussian03* program and 4 cores on a IBM BladeCenter Linux cluster. <sup>c</sup> //opt denotes a consistently optimized geometry. <sup>d</sup> The calculations with the AQZQ basis set employ the resolution of identity (or density fitting) algorithm.

**Table 5.** Phosphine Dissociation Energies (kcal/mol) in the Real Grubbs II Catalyst

Expt: BDE = 40.2 kcal/mol<sup>a</sup>

	noCp <sup>b</sup>	Cp <sup>b</sup>
M06	42.84	40.29
M06-L	41.80	39.40
MPW1B95	32.13	30.38
PBEh	25.11	23.20
PBE	20.49	18.64
TPSS	17.80	15.93
BP86	14.49	12.69
B3LYP	12.59	10.84

<sup>a</sup> Obtained from the experimental<sup>16</sup> collision-induced dissociation energy (36.9 kcal/mol) and the scaled vibrational zero-point-energy correction and thermal vibrational-rotational energy at the M06-L/MIDI! level (scale factor = 0.982, which is determined by an approach described in ref 96). <sup>b</sup> noCP denotes the results that are calculated without counterpoise (Cp) correction for basis set superposition error, whereas Cp denotes counterpoise corrected results. All DFT calculations in this table employ the TZQS basis set and M06-L/DZQ geometries.

the WFT-based SCSC-MP2 method. Note that this good performance might due in part to the cancelation of errors, because M06-L/AQZQ//M06-L/TZQS gives an MUE of 2.94 kcal/mol, just slightly better than the M06-L/TZQS method.

For B3LYP and BP86, the difference in MUEs is small (<0.5 kcal/mol) between the calculations with consistently optimized geometries and the calculations with the M06-L/TZQS geometries.

**3.4. Extension to Real Catalysts.** Although the model catalyst studied here provides a useful model for coordinate covalent bonding in Grubbs II catalysts, it does not include the bulky and polarizable substituents of real Grubbs catalysts. Therefore we also carried out some single point calculations for the analog of the **1** → **2** step in a real Grubbs II catalyst. The best estimate was made by removing zero-point vibrational energy and thermal vibrational rotational energy from the experimental value of Torker et al.<sup>16</sup> As shown in Table 5, only M06-L and M06 give good performance for the phosphine dissociation energy in the real

Grubbs II catalyst. The counterpoise corrected M06 gives an error of 0.1 kcal/mol, and M06-L gives an error of -0.8 kcal/mol. All other functionals give much larger errors (>8 kcal/mol). Strikingly, the popular BP86 and B3LYP functionals give errors larger than 25 kcal/mol; therefore, they are not reliable for the studies of mechanisms in the Grubbs catalysts.

The fact that functionals like PBEh, PBE, TPSS, BP86, and B3LYP become worse for large molecules is not completely surprising since several previous studies<sup>36-38,42-44,47-49</sup> have shown a deterioration in performance of several density functionals for large molecules; the M06 family of functionals has been shown though not to suffer from this problem of deteriorating for larger-size molecules.<sup>30,53,113-115</sup>

## 4. Concluding Remarks

In the present study, we developed a benchmark data set for relative energetics in the catalytic cycle of a model system



for Grubbs second-generation olefin metathesis catalysts. The benchmark data were determined by a composite approach, and they are of CCSD(T)/QZ quality. The benchmark data were used as a training set to develop a new SCSC-MP2 method designed for catalysis. We employed the benchmark data and experimental data on a real Grubbs catalyst to test 17 WFT methods and 39 density functionals. We found the following:

1) The SCSC-MP2 method has improved performance for modeling Grubbs II olefin metathesis model catalysts as compared to canonical MP2 or SCS-MP2.

2) Among the tested density functionals, M06 is the best performing functional. M06/TZQS//M06-L/DZQ gives an MUE of only 1.15 kcal/mol for the benchmark data on model compounds and 0.1–0.6 kcal/mol for the phosphine dissociation energy in the real Grubbs catalyst, and it is a much more affordable method than the SCSC-MP2 method or any other correlated WFT method.

3) The best performing meta-GGA is M06-L; M06-L/DZQ gives an MUE of only 1.77 kcal/mol for the benchmark data on model compounds and 0.8–1.6 kcal/mol for the real Grubbs catalyst.

4) PBEh is the best performing hybrid GGA, with an MUE of 3.01 kcal/mol for the benchmark data on model compounds, but this error increases to 16–18 kcal/mol for the real Grubbs catalyst.

5) B3LYP and BP86 are not accurate for modeling Grubbs II olefin metathesis catalysts; BP86 perform relatively better than B3LYP. B3LYP underestimates the stability of the *cis*-pathway by a large margin and has an error of 28–29 kcal/mol for the real Grubbs catalyst.

From the assessments, we recommend the M06 and M06-L functionals for modeling Grubbs II olefin metathesis catalysts. The local M06-L method is especially efficient for calculations on large systems. BP86, B3LYP, and eight other functionals containing the LYP correlation functional should be avoided.

We have shown that M06-class functionals also give a good performance for describing interactions in zeolite model complexes<sup>115</sup> and adsorptions of CO on the Mg(001) surface<sup>116</sup> as well as performing well here. Thus the M06 functional suite has been shown to be broadly used for modeling catalysis.

**Acknowledgment.** This work was supported in part by the Air Force Office of Scientific Research (orbital-dependent density functionals), by the National Science Foundation under grant no. CHE07-04974 (complex systems), by the Office of Naval Research under award number N00014-05-0538 (software tools), and by a Molecular Science Computing Facility Computational Grand Challenge grant at Environmental Molecular Science Laboratory of Pacific Northwestern National Laboratory.

**Supporting Information Available:** Cartesian coordinates for all molecular systems. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Deshmukh, P. H.; Blechert, S. *Dalton Trans.* **2007**, 2479.
- (2) Hoveyda, A. H.; Zhugralin, A. R. *Nature* **2007**, *450*, 243.
- (3) Schrodi, Y.; Pederson, R. L. *Aldrich. Acta* **2007**, *40*, 45.
- (4) Chauvin, Y. *Angew. Chem., Int. Ed.* **2006**, *45*, 3740.
- (5) Grubbs, R. H. *Angew. Chem., Int. Ed.* **2006**, *45*, 3760.
- (6) Schrock, R. R. *Angew. Chem., Int. Ed.* **2006**, *45*, 3748.
- (7) [http://nobelprize.org/nobel\\_prizes/chemistry/laureates/2005/press.html](http://nobelprize.org/nobel_prizes/chemistry/laureates/2005/press.html) (accessed September 15, 2008).
- (8) Cortez, G. A.; Baxter, C. A.; Schrock, R. R.; Hoveyda, A. H. *Org. Lett.* **2007**, *9*, 2871.
- (9) Scholl, M.; Ding, S.; Lee, C. W.; Grubbs, R. H. *Org. Lett.* **1999**, *1*, 953.
- (10) Bielawski, C. W.; Grubbs, R. H. *Angew. Chem., Int. Ed.* **2000**, *39*, 2903.
- (11) Huang, J.; Stevens, E. D.; Nolan, S. P.; Peterson, J. L. *J. Am. Chem. Soc.* **1999**, *121*, 2674.
- (12) Weskamp, T.; Kohn, F. J.; Hieringer, W.; Gleich, D.; Hermann, W. A. *Angew. Chem., Int. Ed.* **1999**, *38*, 2416.
- (13) Ung, T.; Hejl, A.; Grubbs, R. H.; Schrodi, Y. *Organometallics* **2004**, *23*, 5399.
- (14) Romero, P. E.; Piers, W. E. *J. Am. Chem. Soc.* **2005**, *127*, 5030.
- (15) Wenzel, A. G.; Grubbs, R. H. *J. Am. Chem. Soc.* **2006**, *128*, 16048.
- (16) Torke, S.; Merki, D.; Chen, P. *J. Am. Chem. Soc.* **2008**, *130*, 4808.
- (17) Volland, M. A. O.; Hofmann, P. *Helv. Chim. Acta* **2001**, *84*, 3456.
- (18) Cavallo, L. *J. Am. Chem. Soc.* **2002**, *124*, 8965.
- (19) Vyboishchikov, S. F.; Bühl, M.; Thiel, W. *Chem. Eur. J.* **2002**, *8*, 3942.
- (20) Adlhart, C.; Chen, P. *Angew. Chem., Int. Ed.* **2002**, *41*, 4484.
- (21) Costabile, C.; Cavallo, L. *J. Am. Chem. Soc.* **2004**, *126*, 9592.
- (22) Suresh, C. H.; Koga, N. *Organometallics* **2004**, *23*, 76.
- (23) Tsepis, A. C.; Orpen, A. G.; Harvey, J. N. *Dalton Trans.* **2005**, 2849.
- (24) Benitez, D.; Goddard, W. A. *J. Am. Chem. Soc.* **2005**, *127*, 12218.
- (25) Cavallo, L.; Correa, A.; Costabile, C.; Jacobsen, H. *J. Organomet. Chem.* **2005**, *690*, 5407.
- (26) Schoeller, W. W.; Schroeder, D.; Rozhenko, A. B. *J. Organomet. Chem.* **2005**, *690*, 6079.
- (27) Jordaan, M.; Helden, P. v.; Sittert, C.G. C. E. v.; Vosloo, H. C. M. *J. Mol. Catal. A: Chem.* **2006**, *254*, 145.
- (28) Corsea, A.; Cavallo, L. *J. Am. Chem. Soc.* **2006**, *128*, 13352.
- (29) Lord, R. L.; Wang, H.; Vieweger, M.; Baik, M.-H. *J. Organomet. Chem.* **2006**, *691*, 5505.
- (30) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2007**, *9*, 1967.
- (31) Piacenza, M.; Hyla-Kryspin, I.; Grimme, S. *J. Comput. Chem.* **2007**, *28*, 2275.
- (32) Getty, K.; Delgado-Jaime, M. U.; Kennepohl, P. *J. Am. Chem. Soc.* **2007**, *129*, 15774.

- (33) Jonas, V.; Thiel, W. *J. Chem. Phys.* **1995**, *102*, 8474.
- (34) Li, J.; Schreckenbach, G.; Ziegler, T. *J. Am. Chem. Soc.* **1995**, *117*, 486.
- (35) González-Blanco, O.; Branchadell, C. *J. Chem. Phys.* **1999**, *110*, 778.
- (36) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.
- (37) Woodcock, H. L.; Schaefer, H. F.; Schreiner, P. R. *J. Phys. Chem. A* **2002**, *106*, 11923.
- (38) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.
- (39) Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624.
- (40) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.
- (41) Carlier, P. R.; Deora, N.; Crawford, T. D. *J. Org. Chem.* **2006**, *71*, 1592.
- (42) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.
- (43) Schreiner, P. R.; Fokin, A. A., Jr.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.
- (44) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.
- (45) Izgorodina, E. I.; Coote, M. L. *J. Phys. Chem. A* **2006**, *110*, 2486.
- (46) Izgorodina, E. I.; Coote, M. L. *Chem. Phys.* **2006**, *324*, 96.
- (47) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42.
- (48) Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, *72*, 2118.
- (49) Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217.
- (50) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 10754.
- (51) Paier, J.; Marsman, M.; Kresse, G. *J. Chem. Phys.* **2007**, *127*, 24103.
- (52) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.
- (53) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- (54) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
- (55) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388.
- (56) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.
- (57) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
- (58) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612.
- (59) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (60) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.
- (61) Woon, D. E.; Dunning, J., T. H. *J. Chem. Phys.* **1993**, *98*, 1358.
- (62) Peterson, K. A.; Figgen, D.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2007**, *126*, 124101.
- (63) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.
- (64) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (65) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem. Symp.* **1976**, *10*, 1.
- (66) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.
- (67) Gordon, M. S.; Truhlar, D. G. *J. Am. Chem. Soc.* **1986**, *5412*.
- (68) Hill, J. G.; Platts, J. A. *J. Chem. Theory Comput.* **2006**, *3*, 80.
- (69) Distasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073.
- (70) Perdew, J. P.; Schmidt, K. *In Density Functional Theory and Its Applications to Materials*; Van-Doren, V., Alsenoy, C. V., Geerlings, P., Eds.; American Institute of Physics: New York, 2001; p 1.
- (71) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (72) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (73) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (74) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.
- (75) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.
- (76) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (77) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.
- (78) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (79) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
- (80) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (81) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (82) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (83) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- (84) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.
- (85) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.
- (86) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.
- (87) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.
- (88) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Wong, M. W.; Foresman, J. B.; Robb, M. A.; Head-Gordon, M.; Replogle, E. S.; Gomperts, R.; Andres, J. L.; Raghavachari, K.; Binkley, J. S.; Gonzalez, C.; Martin, R. L.; Fox, D. J.; Defrees, D. J.; Baker, J.; Stewart, J. J. P.; Pople, J. A. *Gaussian 92/DFT; Revision F. 2*; Gaussian, Inc.: Pittsburgh, PA, 1993.
- (89) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.
- (90) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.
- (91) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.
- (92) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.

- (93) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.
- (94) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
- (95) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (96) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (97) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
- (98) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03; Revision D.01*; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (99) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module - Version 40beta*; University of Minnesota: Minneapolis, MN, 2006.
- (100) Bylaska, E. J.; Jong, W. A. d.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Aprà, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anshell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsels, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendenning, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; Lenthe, J. v.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers; Version 5.1*; Pacific Northwest National Laboratory: Richland, WA, U.S.A., 2007.
- (101) Pople, J. A. *Rev. Mod. Phys.* **1999**, *71*, 1267.
- (102) Dunlap, B. I. *J. Chem. Phys.* **1983**, *78*, 3140.
- (103) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (104) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.
- (105) Zheng, Y. C.; Almlöf, J. E. *THEOCHEM* **1996**, *388*, 277.
- (106) Skylaris, C.-K.; Gagliardi, L.; Handy, N. C.; Ioannou, A. G.; Spencer, S.; Willetts, A. *THEOCHEM* **2000**, *501–502*, 229.
- (107) Glaesemann, K. R.; Gordon, M. S. *J. Chem. Phys.* **2000**, *112*, 10738.
- (108) Te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- (109) Sierka, M.; Hoge Kamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.
- (110) Sodt, A.; Subotnik, J. E.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 194109.
- (111) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 44103.
- (112) Zhao, S.; Li, Z.-H.; Wang, W.-N.; Liu, Z.-P.; Fan, K.-N.; Xie, Y.; Schaefer, H. F. *J. Chem. Phys.* **2006**, *124*, 184102.
- (113) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2006**, *8*, 5753.
- (114) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 1095.
- (115) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 6860.
- (116) Valero, R.; Gomes, J. R. B.; Truhlar, D. G.; Illas, F. J. *Chem. Phys.* **2008**, *129*, 124710.

CT800386D

## Electrostatic Potentials from Self-Consistent Hirshfeld Atomic Charges

Sofie Van Damme, Patrick Bultinck,\* and Stijn Fias

*Department of Inorganic and Physical Chemistry, Ghent University,  
Krijgslaan 281 (S3), 9000 Gent, Belgium*

Received September 22, 2008

**Abstract:** It is shown that molecular electrostatic potentials obtained from iterative or self-consistent Hirshfeld atomic point charges agree remarkably well with the ab initio computed electrostatic potentials. The iterative Hirshfeld scheme performs nearly as well as electrostatic potential derived atomic charges, having the advantage of allowing the definition of the atom in the molecule, rather than just yielding charges. The quality of the iterative Hirshfeld charges for computing electrostatic potentials is examined for a large set of molecules and compared to other commonly used techniques for population analysis.

### Introduction

The electrostatic potential (ESP)<sup>1</sup> plays a very important role in the study of chemical reactivity.<sup>2–7</sup> Indeed, when a molecule nears another molecule, the first thing it “notices” from the other molecule is its ESP. Hence, the ESP plays a fundamental role in theories that aim at explaining chemical reactivity, for instance, in so-called conceptual or chemical density functional theory (DFT).<sup>8</sup> The ESP also attracts attention because it has been shown that atomic<sup>9</sup> and molecular energies<sup>10,11</sup> can be expressed in terms of electrostatic potentials at the nuclei and the nuclear charges, such that

$$E = f(\{V_{0,A}, Z_A\}) \quad (1)$$

When used in chemical reactivity, studying where the minima in a molecular ESP occur allows, for example, the prediction of where an electrophile is most likely to attack.<sup>2,3,5,6</sup> Naturally, this is only a first approximation, and one needs to include many more effects when the molecules approach each other more. The interaction energy is then best described in terms of a Taylor expansion of the energy with as variables the number of electrons and the external potential.<sup>12,13</sup> Even if for the latter part the ESP is a good first approximation, it should be taken into account that once two molecules approach each other sufficiently close, the electrostatic potential of the molecules is altered in the electronic polarization process, and as such, one should compute it for

every new mutual arrangement of the molecules. In other words, the electrostatic potential of an isolated molecule should only be used as a reactivity index for the very first stages when two distant molecules start approaching each other. Also, if charge transfer happens between the molecules, other reactivity indices need to be studied as well in order to interpret or predict the reactivity. Nevertheless, the ESP remains a very valuable field and is used to understand, e.g., interactions between biomolecules.<sup>4,14–20</sup> A frequently used option in studying the intermolecular interaction is to examine the electrostatic potential at and beyond some minimum distance from the molecule, e.g., on the van der Waals radius.

The ESP of a molecule  $V(\mathbf{r})$  can be computed relatively easily given the positions  $\mathbf{R}_A$  and charges  $Z_A$  of the nuclei in the molecule and the electronic density function  $\rho(\mathbf{r})$ :

$$V(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} - \int \frac{\rho(\mathbf{r}') \, d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \quad (2)$$

As the ESP is of such importance, it is a ubiquitously used field in 3D-QSAR.<sup>19,21</sup> In classical QSAR, often molecular descriptors are used that are related to the ESP.<sup>14,15,22</sup> A good example is atomic charges.

$V(\mathbf{r})$  can also be computed from a sum of contributions from all the multipoles of the charge distribution. One can either take the entire molecular charge distribution or work in terms of certain parts of the molecule, e.g., regions of space associated to certain atoms in the molecule. It follows from Coulomb’s law that any distribution of electrical charge creates a potential  $V(\mathbf{r})$  at each point in the surrounding space.

\* Corresponding author. E-mail: Patrick.Bultinck@UGent.be.

In the elementary example of the discrete charge distribution of a collection of point atomic charges, an approximate electrostatic potential can be computed as

$$V(\mathbf{r}) \approx V'(\mathbf{r}) = \sum_A \frac{q_A}{|\mathbf{r} - \mathbf{R}_A|} \quad (3)$$

Such an approximation naturally comes at the cost of a loss of quality of the ESP. Nevertheless, when high throughput of molecules is needed or for very large molecules, such approximate ESP are often used, e.g., in comparative molecular field analysis (COMFA).<sup>21</sup> When compared to the ab initio ESP, quite large errors can occur, depending on the technique used for computing the atomic charge. An exception is ESP-derived atomic charges. These naturally reproduce relatively well the ESP as they are obtained via a fit of  $V'(\mathbf{r})$  versus  $V(\mathbf{r})$  with the atomic charges as variables under some constraints such as that the sum of the atomic charges needs to be equal to the molecular charge and optionally the requirement that the atomic charges reproduce some molecular multipole moments. Different algorithms are commonly used, known under names such as CHELP(G)<sup>23,24</sup> or the Merz–Kollman–Singh<sup>25,26</sup> method. It is important to note that these atomic charges are obtained as a statistical fit to the ab initio potential and as such the charges are not the product of a true “atom in the molecule” (AIM). That is, there is no AIM density function from which the charge is derived as

$$q_A = Z_A - \int \rho_A^{\text{AIM}}(\mathbf{r}) \, d\mathbf{r} \quad (4)$$

Moreover, there are quite important statistical problems with ESP-derived charges due to rank problems.<sup>27–29</sup> The wording atom in the molecule (abbreviated AIM) is used in a more general meaning than Bader’s technique based on zero flux surface analysis.<sup>30,31</sup> As is clear from the above, the AIM plays an important role in the evaluation of  $V'(\mathbf{r})$ . The importance of the AIM, however, is far from limited to this use. The AIM forms a cornerstone of chemistry and is quantum chemical object with a density function attached to it. This is also the main distinction between a true AIM method and a population analysis method. The latter methods only yield atomic charges whereas these are only one quantity derivable from the density function. Despite the role of the AIM, its precise nature is still subject of debate<sup>32–34</sup> and many different AIM methods have been described. Recently, one of the authors has described the iterative Hirshfeld method<sup>35,36</sup> (Hirshfeld-I). At that time, it was found that the atomic charges resulting from the Hirshfeld-I AIM density function, correlate well with electrostatic potential derived charges. This raises the question whether this is merely a chance correlation between the two charge sets or whether there is also a good correlation between the Hirshfeld-I based  $V'(\mathbf{r})$  and ab initio  $V(\mathbf{r})$ .

The purpose of the present study is thus to examine how good  $V'(\mathbf{r})$  derived from (3) is compared to the ab initio potential. The reason why such a study is highly relevant is that, if the Hirshfeld-I AIM definition derived from solely the molecular density function, gives fairly good  $V'(\mathbf{r})$ , this suggests that it is a promising source for AIM condensed

reactivity indices in general. In other words, even a simple approximate ESP from (3) using Hirshfeld-I charges is able to provide insight in molecular reactivity. Moreover, if the quality of the approximate ESP is roughly similar to the approximate ESP obtained with ESP-derived charges, the Hirshfeld-I method has the important advantage of being an AIM method, rather than merely a technique for population analysis. Moreover, it is a technique that allows defining the atom in the molecule and which does not suffer from numerical instability, as ESP-derived charges do.

## Theoretical Background

As stated earlier, the ESP is especially useful for the first stages of an electrostatic interaction, with  $V(\mathbf{r})$  being computed on an outer surface of the molecule. A common approach is to use a set of intersecting spheres centered on each atom with some atom specific radius. This may be a number of times the van der Waals radius, or some other appropriate radius. Such a surface can be generated readily for any molecule. Alternatively, one can also use an isodensity surface, as was suggested by Bader et al.<sup>30</sup> The advantage of using such a surface is the fact that it is more molecule specific than using a surface based on intersecting spheres where each atom of a specific element shares the same radius. In the present study, a series of molecular surfaces is generated based on intersecting fixed radius spheres around the different nuclei. This choice is based on the fact that this is also the method used for computing ESP-derived charges and that the Hirshfeld-I based approximate ESP is to be compared to that based on ESP-derived charges. The latter are known to depend to sometimes significant extent on the selection of the surface, hence a fair comparison should be made based on the same appropriate point selection scheme.

In this study it will be examined how well different AIM methods perform for predicting  $V'(\mathbf{r})$  compared to ab initio  $V(\mathbf{r})$ . Most of these methods have already been described in detail in the literature. The exception is the iterative Hirshfeld-I method. In the Hirshfeld-I method, the AIM density function is obtained as follows. First, for a real molecule a promolecule is constructed in exactly the way proposed originally by Hirshfeld.<sup>37</sup> The promolecular density function  $\rho_{\text{Mol}}^0(\mathbf{r})$  is the union of the density functions  $\rho_A^0(\mathbf{r})$  of the isolated atoms A put on exactly the same place in space as in the molecule. At each point in space, the weight  $w_A(\mathbf{r})$  of an AIM A is computed as

$$w_A(\mathbf{r}) = \frac{\rho_A^0(\mathbf{r})}{\rho_{\text{Mol}}^0(\mathbf{r})} = \frac{\rho_A^0(\mathbf{r})}{\sum_A \rho_A^0(\mathbf{r})} \quad (5)$$

This positive definite weight summed over all atoms equals unity. The density function for an AIM A (denoted)  $\rho_A^{\text{AIM}}(\mathbf{r})$  is then obtained from the molecular density function  $\rho_{\text{Mol}}(\mathbf{r})$  via

$$\rho_A^{\text{AIM}}(\mathbf{r}) = w_A(\mathbf{r})\rho_{\text{Mol}}(\mathbf{r}) \quad (6)$$

Different authors<sup>32,35,38–40</sup> have shown that there is a fundamental problem with the original Hirshfeld method, as

**Table 1.** Atomic Charges (in au) in **54** Computed from Different AIM Methods and Methods for Population Analysis

	MKS-VBF	Hirshfeld	Hirshfeld-I	Mulliken	NPA
H	0.1042	0.0273	0.1078	0.1510	0.1797
C	-0.3003	-0.0721	-0.4018	-0.3881	-0.5315
H	0.0699	0.0223	0.0947	0.1242	0.1690
C	0.2782	0.0532	0.2534	-0.1136	0.0673
H	0.0657	0.0277	0.1036	0.1198	0.1765
O	-0.5256	-0.2095	-0.4316	-0.1717	-0.6712
H	0.0320	0.0302	0.0400	0.1473	0.1517
H	-0.0062	0.0193	0.0140	0.1020	0.1257
C	0.3972	0.0516	0.2679	-0.0925	0.0637
C	-0.4205	-0.0681	-0.3897	-0.5204	-0.5134
H	-0.0122	0.0182	0.0116	0.1168	0.1258
H	-0.0072	0.0191	0.0178	0.1104	0.1277
H	0.0872	0.0255	0.0977	0.1218	0.1711
H	0.1231	0.0278	0.1077	0.1454	0.1794
H	0.1181	0.0274	0.1069	0.1475	0.1786

**Table 2.** Root-Mean-Square Error (rmse, in Units  $10^{-3}$  au), Slope ( $a$ ) and Intercept ( $b$ , in Units  $10^{-3}$  au) of the Regression Line, and  $R^2$  for  $V'(\mathbf{r})$  versus  $V(\mathbf{r})$  for Molecule **54** and Different AIM Methods and Methods for Population Analysis<sup>a</sup>

	MKS-VBF	Hirshfeld	Hirshfeld-I	Mulliken	NPA
rmse	2.0	5.8	2.8	7.0	10.6
$a$	0.97	0.49	0.90	0.76	1.81
$b$	0.0	-0.4	-0.5	-0.6	-0.7
$R^2$	0.97	0.93	0.94	0.62	0.93

<sup>a</sup> Hirshfeld denotes the original Hirshfeld method and Hirshfeld-I the iterative version.

the AIM density function depends quite strongly on the isolated atom densities used. For instance, in a molecule like  $N_2$ , Davidson et al.<sup>32</sup> have shown that different atomic charges are found for the nitrogen atoms depending on whether the promolecule was constructed of two neutral nitrogen atoms or a positive and a negative one. Bader et al.<sup>40</sup> criticized the method because it, indeed, also gives very small atomic charges on highly ionic compounds. These problems were addressed by Bultinck et al.<sup>35,36</sup> by iteratively changing the isolated atoms used for computing the AIM density function. Starting from some chosen set of isolated atom fragments (neutral or charged),  $w_A(\mathbf{r})$  is computed according to (5). Using this weight function, AIM densities for all A are computed via (6). From this density function, the AIM electronic population is computed. For this precise electronic population, a new  $\rho_A^0(\mathbf{r})$  is computed for all A. Then, a new  $w_A(\mathbf{r})$  can be computed, yielding new AIM density functions. The procedure is repeated until convergence, meaning that the atomic electronic populations used for computing  $w_A(\mathbf{r})$  no longer differ from those obtained by integrating  $\rho_A^{\text{AIM}}(\mathbf{r})$ . For the exact details of the method, called Hirshfeld-I, including the method of how to deal with noninteger electronic populations, the reader is referred to Bultinck et al.<sup>35,36</sup> Hirshfeld-I AIM density functions no longer depend on the starting set of atomic densities and always produce the same unique solution and are fairly independent of the basis set used.<sup>36</sup>

## Computational Methods

In order to assess the quality of  $V'(\mathbf{r})$  from (3) by comparison to the exact  $V(\mathbf{r})$ , the molecular ESP is computed ab initio at the RHF/6-311++G\*\* level. As explained above, the ESP is a reactivity descriptor for the first stages of approach between two molecules. Hence, the ESP is computed on a grid of points on an outer surface of the molecule. A self-written program allows computing points on a single layer or on different layers around the molecule in a similar way as in the Merz–Singh–Kollman scheme introduced by U. C. Singh and P. A. Kollman.<sup>25</sup> In the present study, the grid points are located on four layers around the molecule. The first layer is constructed from atom-centered spherical surfaces with all Merz–Kollman radii of the spheres scaled by a factor of 1.4. The grid density on these surfaces is chosen at 5 points per unit surface. For the more distant layers, the radii of the atom centered spheres are increased as prescribed in ref 26. On the resulting total set of points, the ab initio ESP is computed and stored as reference.

From the molecular density function, the AIM are computed in each molecule using different techniques: the Mulliken<sup>33</sup> method, the original Hirshfeld method, and Hirshfeld-I and natural population analysis<sup>41</sup> (NPA).  $V'(\mathbf{r})$  are computed from (3) with atomic charges obtained from (4).  $V'(\mathbf{r})$  are also computed from atomic charges derived from electrostatic potential fitting. For the latter, a least-squares regression fit is performed between  $V'(\mathbf{r})$  and  $V(\mathbf{r})$  with the atomic charges as variables. Considering the grid used in the present study, the resulting atomic charges should be quite close to results obtained with the Merz–Kollman–Singh algorithm, as the grid points are sampled according to the Merz–Kollman–Singh scheme. Our calculations revealed this to be the case. Results of this fit with our own grid will be denoted as MKS-VBF.

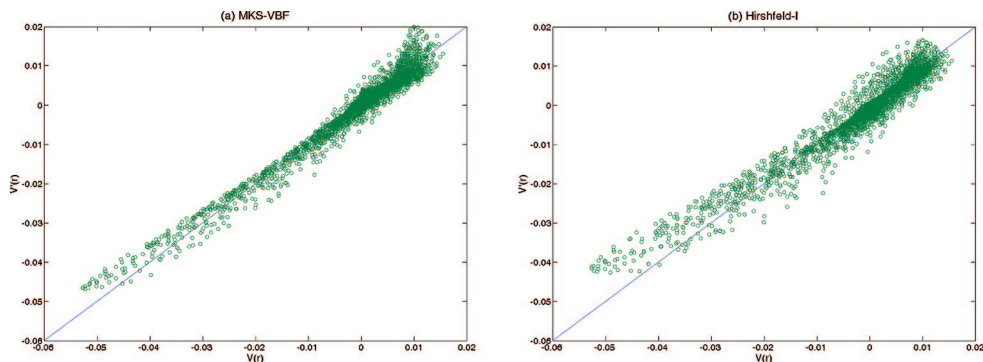
In order to be able to properly assess the quality of  $V'(\mathbf{r})$ , a set of 158 molecules is used. These molecular structures for the test set are available in the Supporting Information. For each molecule individually, the root-mean-square error between  $V'(\mathbf{r})$  and RHF/6-311++G\*\*  $V(\mathbf{r})$  is computed as well as the squared correlation coefficient  $R^2$  between both. The average  $R^2$  considering all molecules is also computed, together with its standard deviation. Finally, also the  $R^2$  over all grid points in all molecules is computed.

All molecular density functions in the present work were computed at the RHF/6-311++G\*\* level of theory using Gaussian-03.<sup>42</sup> AIM density functions and atomic charges were computed with self-developed programs, except for the NPA atomic charges that were taken from Gaussian-03 directly.

## Results and Discussion

In order to make the presentation of the results as clear as possible, each step is illustrated by just one molecule, namely ethoxyethane (molecule **54**), after which the results over the entire molecular set are presented. Table 1 gives the atomic charges on all atoms of molecule **54** for the different types of AIM methods or population analysis.

As was reported previously by Bultinck et al., Hirshfeld-I charges correlate relatively well with the ESP-derived



**Figure 1.** Correlation between  $V'(r)$  and  $V(r)$  (in au) for molecule **54** for (a) MKS-VBF and (b) Hirshfeld-I.

charges. The  $R^2$  between MKS-VBF and Hirshfeld-I is 0.81. Having computed the atomic charges and the exact ab initio molecular ESP  $V(\mathbf{r})$ ,  $V'(\mathbf{r})$  is computed from (3). Table 2 gives for each method the root-mean-square error (rmse) between both potentials for molecule **54**, as well as the slope and intercept of the regression line  $V(\mathbf{r}) = aV'(\mathbf{r}) + b$  and  $R^2$  between the ab initio and approximate ESP.

As expected, Table 2 reveals that the rmse is quite low for the MKS-VBF method and that the  $R^2$  is quite high. Interestingly, the rmse is quite low for the Hirshfeld-I technique as well. Also, the  $R^2$  is quite good. The correlation between  $V(\mathbf{r})$  and  $V'(\mathbf{r})$  is shown in Figure 1 for both MKS-VBF and Hirshfeld-I.

Figure 1 shows that the regression line is even relatively close to the bisector. The other sets of atomic charges perform less good. At least for molecule **54**, Hirshfeld-I is nearly as good as the MKS-VBF method. There is also a clear improvement for Hirshfeld-I compared to Hirshfeld, despite the fact that the squared correlation coefficient and intercept are very similar. The slope is completely different, which may have important consequences when making comparisons among different molecules (see below). This shows that it is important to include the iterative procedure in the Hirshfeld method.

Naturally, a single molecule does not suffice to draw general conclusions and so the entire molecular set must be considered. The table giving the statistical characteristics for each molecule in the molecular set can be found in the Supporting Information. When considering the entire set, it is found that for some molecules, e.g., NPA charges give very good  $R^2$  but relatively high rmse. This means that the regression line differs quite strongly from the bisector. Even with individual high  $R^2$  for every molecule separately, one can have poor performance for the comparison of different molecules. In order to establish to what extent the different methods perform over the entire test set, Table 3 gives the average rmse with the associated standard deviation, the average  $R^2$  with the standard deviation and  $R^2$ . The latter value is the squared correlation coefficient over the union of all grid points in all molecules. This value indicates whether there is a common correlation beneath a set of individually good correlations per molecule. Also included are the average slope and intercept for the regression equation  $V(\mathbf{r}) = aV'(\mathbf{r}) + b$  and the standard deviations on these values.

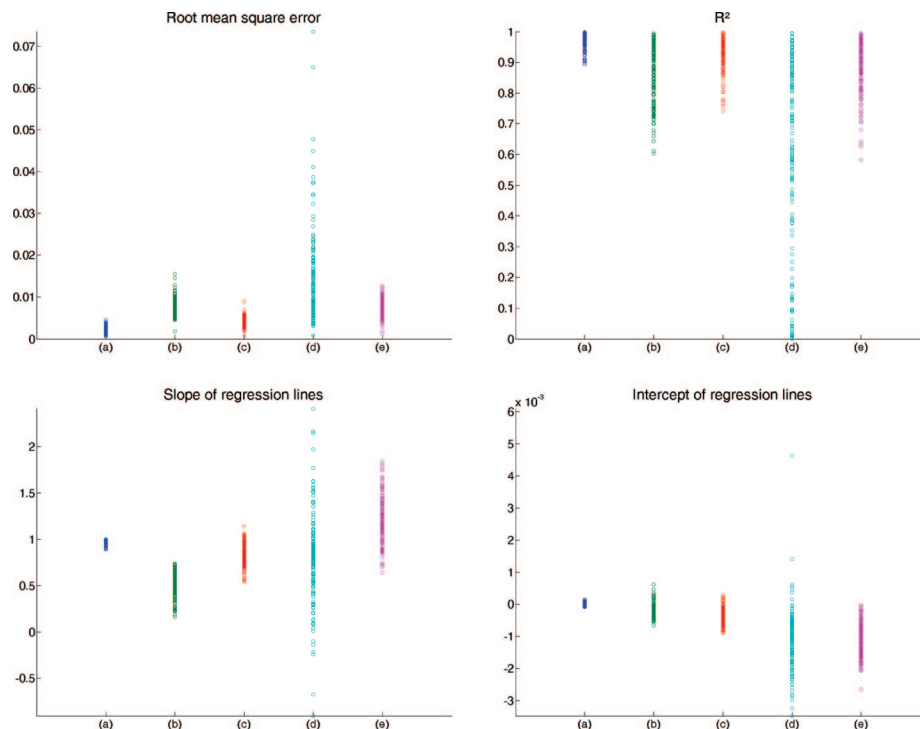
**Table 3.** Average Root Mean Square Error (rmse, in Units  $10^{-3}$  au), Slope ( $a$ ) and Intercept ( $b$ , in Units  $10^{-3}$  au),  $R^2$  over All Molecules with Standard Deviation, and  $R^2$  over all Grid Points in All Molecules<sup>a</sup>

	MKS-VBF	Hirshfeld	Hirshfeld-I	Mulliken	NPA
rmse	2.2 0.7	8.11 0.9	4.4 1.0	14.4 10.9	7.6 2.0
$a$	0.97 0.02	0.52 0.14	<i>0.83</i> <i>0.12</i>	0.81 0.48	1.22 0.23
$b$	0.0 0.0	-0.2 0.2	-0.4 0.2	-1.1 1.1	-1.1 0.5
$R^2$	0.97 0.02	0.87 0.09	<i>0.93</i> <i>0.05</i>	0.61 0.32	0.89 0.08
$R^2_{\text{all}}$	0.98	0.86	<i>0.93</i>	0.36	0.89

<sup>a</sup> The first number denotes the average, the second the standard deviation. For each parameter, the best value is italicized (excluding MKS-VBF).

Table 3 clearly shows that the average rmse is quite low for MKS-VBF, as expected. Also, the standard deviation is relatively low, meaning that for all molecules the rmse is roughly equally good. As MKS-VBF is a least-squares fit between the ab initio and approximate ESP, all statistical parameters describing their mutual relationship are the best over all methods. The original Hirshfeld method performs acceptably for both rmse and its standard deviation. The Mulliken method performs poorly, with NPA in between. Hirshfeld-I performs very well, as will be discussed below in more detail. Roughly the same picture is found based on the  $R^2$ . The average  $R^2$  is best for MKS-VBF and Hirshfeld-I. Occasionally, Mulliken or NPA can also give high  $R^2$  for individual molecules, but not in a consistent fashion. This is clearly revealed by the large standard deviation in  $R^2$  for these methods. As a graphical representation of the performance of the different methods, Figure 2 shows the values for the rmse,  $R^2$ , slope, and intercept for every molecule individually for all methods used.

As Figure 2 clearly shows, the ranges in all statistical parameters are relatively small for MKS-VBF and Hirshfeld-I. This is clearly not the case for Mulliken and NPA. A further interesting comparison lies in the slope and intercept of the regression lines between  $V(\mathbf{r})$  and  $V'(\mathbf{r})$ . As Table 3 shows, the average slope is relatively close to 1 for MKS-VBF, Hirshfeld-I, NPA, and surprisingly also Mulliken. For the Mulliken method, however, this high slope comes with a very large standard deviation, which is not the case for MKS-VBF and Hirshfeld-I. The average slope for the Hirshfeld method deviates quite a lot from 1.0. The average



**Figure 2.** Rmse (au),  $R^2$ , slope, and intercept (au) for each molecule for MKS-VBF (a), Hirshfeld (b), Hirshfeld-I (c), Mulliken (d), and NPA (e).

intercept for the different methods points out that both Hirshfeld and Hirshfeld-I perform very well.  $R^2$  confirms the conclusions from the other statistical parameters, namely that MKS-VBF and Hirshfeld-I perform consistently well and Mulliken performs very badly.

The Hirshfeld-I method, although not fitted to the ESP, performs nearly as well as MKS-VBF. This is a very interesting observation as the Hirshfeld-I method is an AIM method and not merely a method for atomic charges. This is very important as, given an AIM density function, one can compute all the expectation values that can be computed from a density function in the usual quantum mechanical way. Atomic dipole moments can, e.g., be computed. This is not the case with ESP-derived charges as one has no density function.

When comparing to the original Hirshfeld method it is seen that the inclusion of the iterative procedure in Hirshfeld-I improves the quality of the fit between  $V(\mathbf{r})$  and  $V'(\mathbf{r})$ . All parameters describing the quality of the fit are significantly better for Hirshfeld-I than for the original method except for a slightly better average intercept. The difference, however, is very small. Bultinck et al.<sup>35</sup> previously pointed out that there are significant differences in atomic charges between Hirshfeld-I and the original Hirshfeld method. These results are opposite to those of Nalewajski et al.,<sup>43</sup> who report only a weak dependence of resulting AIM charges on the promolecule chosen. According to them, using an ionic promolecule leads to only a slightly larger charge separation for NaCl. Our calculations reveal that this effect is quite large. For NaCl the difference in charge separation amounts roughly 0.5 between the Hirshfeld and Hirshfeld-I schemes. The present results show that the inclusion of the self-

consistently obtained promolecule in the Hirshfeld-I procedure also has a marked effect on the agreement between  $V(\mathbf{r})$  and  $V(\mathbf{r})$ .

Finally, the question arises to what extent the good performance of the Hirshfeld-I scheme allows one to conclude it as the best or final AIM method. We believe that such a far-reaching conclusion cannot be drawn from this study. The present study only shows that, among the methods tested, it is the best performing AIM method when it comes to yielding approximate ESP's that are most similar to the ab initio ones. Good performance in an approximate scheme cannot be used as an argument for more general conclusions. We do, however, stress that the present study is a clear indication that for atom condensed reactivity indices, which are almost always computed in some approximate way,<sup>44</sup> the very good performance of Hirshfeld-I shows that they may be the best choice for computing such indices. Other AIM techniques that already do not perform well for a simple field like electrostatic potential can hardly be expected to perform well for more subtle fields. The main advantage of the Hirshfeld-I approach to ESP as opposed to the electrostatic potential derived charges based ESP is that no statistical rank problems exist with the Hirshfeld-I method and that it gives a true AIM density function instead of merely a set of atomic charges.

## Conclusions

It has been shown that the quality of electrostatic potentials derived from a simple monopole approximation compared to the true ab initio electrostatic potentials on a surface surrounding the molecule depends strongly on the atoms in molecules method or population analysis technique. ESP-



derived atomic charges perform well, as expected, but have the drawback that they do not define the atom in the molecule. The Hirshfeld-I method does give an atom in the molecule with its own density function and the atomic charges derived from it give a statistical fit to the true electrostatic potential that is nearly as good as that for the ESP-derived charges. This suggests that the Hirshfeld-I method could be a very good candidate for use in other atom condensed reactivity indices.

**Acknowledgment.** The authors thank Ghent University and the Fund for Scientific Research-Flanders (Belgium) for their grants to the Quantum Chemistry group at Ghent University. S.V.D. thanks the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for a Ph.D. grant.

**Supporting Information Available:** Structures of the molecules used in the study and the statistical characteristics for each molecule individually are listed. This information is available free of charge via the Internet at <http://pubs.acs.org>.

### References

- (1) Politzer, P.; Murray, J. S. Molecular Electrostatic Potentials. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker, Inc: New York, 2004; pp 213–234.
- (2) Politzer, P.; Daiker, K. C. Models for chemical reactivity In *The Force Concept in Chemistry*; Deb, B. M., Ed.; Van Nostrand Reinhold: New York, 1981; pp 294–387.
- (3) Scrocco, E.; Tomasi, J. Electronic Molecular Structure, Reactivity and Intermolecular Forces: An Euristic Interpretation by Means of Electrostatic Molecular Potentials. *Adv. Quantum Chem.* **1979**, *11*, 115.
- (4) Politzer, P.; Truhlar, D. G. *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Plenum: New York, 1981.
- (5) Scrocco, E.; Tomasi, J. The electrostatic molecular potential as a tool for the interpretation of molecular properties. *Top. Curr. Chem.* **1973**, *42*, 95.
- (6) Politzer, P.; Murray, J. S. Molecular Electrostatic Potentials and Chemical Reactivity. *Rev. Comput. Chem.* **1991**, *2*, 273.
- (7) Narayszabo, G.; Ferenczy, G. G. Molecular Electrostatics. *Chem. Rev.* **1995**, *95*, 829.
- (8) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual density functional theory. *Chem. Rev.* **2003**, *103*, 1793.
- (9) Foldy, L. L. A note on atomic binding energies. *Phys. Rev.* **1951**, *83* (2), 397.
- (10) Politzer, P.; Parr, R. G. Some New Energy Formulas for Atoms and Molecules. *J. Chem. Phys.* **1974**, *61*, 4258.
- (11) Wilson, E. B. 4-Dimensional Electron Density Function. *J. Chem. Phys.* **1962**, *36*, 2232.
- (12) Ayers, P. W.; Anderson, J. S. M.; Bartolotti, L. J. Perturbative perspectives on the chemical reaction prediction problem. *Int. J. Quantum Chem.* **2005**, *101*, 520.
- (13) Chattaraj, P. K.; Nath, S.; Maiti, B. Reactivity Descriptors. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker, Inc.: New York, 2004; pp 295–322.
- (14) Brinck, T.; Jin, P.; Ma, Y. G.; Murray, J. S.; Politzer, P. Segmental analysis of molecular surface electrostatic potentials: application to enzyme inhibition. *J. Mol. Model.* **2003**, *9*, 77.
- (15) Murray, J. S.; bu-Awwad, F.; Politzer, P.; Wilson, L. C.; Troupin, A. S.; Wall, R. E. Molecular surface electrostatic potentials of anticonvulsant drugs. *Int. J. Quantum Chem.* **1998**, *70*, 1137.
- (16) Carrupt, P. A.; Eltayar, N.; Karlen, A.; Testa, B. Molecular Electrostatic Potentials for Characterizing Drug Biosystem Interactions. *Methods Enzymol.* **1991**, *203*, 638.
- (17) Politzer, P. Computational Approaches to the Identification of Suspect Toxic Molecules. *Toxicol. Lett.* **1988**, *43*, 257.
- (18) Murray, J. S.; Politzer, P. Electrostatic Potentials of Some Dibenzo-P-Dioxins in Relation to Their Biological-Activities. *Theor. Chim. Acta* **1987**, *72*, 507.
- (19) Murray, J. S.; Zilles, B. A.; Jayasuriya, K.; Politzer, P. Comparative-Analysis of the Electrostatic Potentials of Dibenzofuran and Some Dibenzo-Para-Dioxins. *J. Am. Chem. Soc.* **1986**, *108*, 915.
- (20) Politzer, P.; Laurence, P. R.; Jayasuriya, K. Molecular Electrostatic Potentials - An Effective Tool for the Elucidation of Biochemical Phenomena. *Environ. Health Perspect.* **1985**, *61*, 191.
- (21) Kubinyi, H. *3D QSAR in Drug Design: Theory Methods and Applications*; ESCOM Science Publishers: Leiden, The Netherlands, 1993; Vol 1.
- (22) Karelson, M. Quantum-Chemical descriptors in QSAR In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker Inc: New York, 2004; pp 641–668.
- (23) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials—the Need for High Sampling Density in Formamide Conformational-Analysis. *J. Comput. Chem.* **1990**, *11*, 361.
- (24) Chirlian, L. E.; Francl, M. M. Atomic Charges Derived from Electrostatic Potentials - A Detailed Study. *J. Comput. Chem.* **1987**, *8*, 894.
- (25) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11* (4), 431–439.
- (26) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129.
- (27) Sigfridsson, E.; Ryde, U. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *J. Comput. Chem.* **1998**, *19*, 377.
- (28) Francl, M. M.; Chirlian, L. E. The pluses and minuses of mapping atomic charges to electrostatic potentials. *Rev. Comput. Chem.* **2000**, *14*, 1.
- (29) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. Charges fit to electrostatic potentials 0.2. Can atomic charges be unambiguously fit to electrostatic potentials. *J. Comput. Chem.* **1996**, *17*, 367.
- (30) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. Properties of Atoms in Molecules—Atomic Volumes. *J. Am. Chem. Soc.* **1987**, *109*, 7968.

- (31) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon: Oxford, Great Britain, 1990.
- (32) Davidson, E. R.; Chakravorty, S. A Test of the Hirshfeld Definition of Atomic Charges and Moments. *Theor. Chim. Acta* **1992**, *83*, 319.
- (33) Mulliken, R. S. Electronic Population Analysis on Lcao-Mo Molecular Wave Functions 0.1. *J. Chem. Phys.* **1955**, *23*, 1833.
- (34) Parr, R. G.; Ayers, P. W.; Nalewajski, R. F. What is an atom in a molecule. *J. Phys. Chem. A* **2005**, *109*, 3957.
- (35) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phys.* **2007**, *126*, 14111.
- (36) Bultinck, P.; Ayers, P. W.; Fias, S.; Tiels, K.; Van Alsenoy, C. Uniqueness and basis set dependence of iterative Hirshfeld charges. *Chem. Phys. Lett.* **2007**, *444*, 205.
- (37) Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge-Densities. *Theor. Chim. Acta* **1977**, *44*, 129.
- (38) Francisco, E.; Pendas, A. M.; Blanco, M. A.; Costales, A. Comparison of direct and flow integration based charge density population analyses. *J. Phys. Chem. A* **2007**, *111*, 12146.
- (39) Matta, C. F.; Bader, R. F. W. An experimentalist's reply to "What is an atom in a molecule". *J. Phys. Chem. A* **2006**, *110*, 6365.
- (40) Bader, R. F. W.; Matta, C. F. Atomic charges are measurable quantum expectation values: A rebuttal of criticisms of QTAIM charges. *J. Phys. Chem. A* **2004**, *108*, 8385.
- (41) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular Interactions from A Natural Bond Orbital, Donor-Acceptor Viewpoint. *Chem. Rev.* **1988**, *88*, 899.
- (42) *Gaussian 03, Revision B.06*; Gaussian Inc.: Pittsburgh, PA, 2007.
- (43) Nalewajski, R. F.; Loska, R. Bonded atoms in sodium chloride—the information-theoretic approach. *Theor. Chem. Acc.* **2001**, *105*, 374.
- (44) Bultinck, P.; Fias, S.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. Critical thoughts on computing atom condensed Fukui functions. *J. Chem. Phys.* **2007**, *127*, 34102.

CT800394Q

# JCTC

Journal of Chemical Theory and Computation

## Solvatochromic Shifts on Absorption and Fluorescence Bands of *N,N*-Dimethylaniline

Ignacio Fdez. Galván,<sup>\*,†</sup> M. Elena Martín,<sup>†</sup> Aurora Muñoz-Losa,<sup>‡</sup> and Manuel A. Aguilar<sup>†</sup>

*Química Física, Edif. José María Viguera Lobo, Universidad de Extremadura, Avda. de Elvas s/n, 06071 Badajoz, Spain, and Dipartimento di Chimica e Chimica Industriale, Università degli Studi di Pisa, Via Risorgimento 35, 56126 Pisa, Italy*

Received October 15, 2008

**Abstract:** A theoretical study of the absorption and fluorescence UV/vis spectra of *N,N*-dimethylaniline in different solvents has been performed, using a method combining quantum mechanics, molecular mechanics, and the mean field approximation. The transitions between the three lowest-lying states have been calculated in vacuum as well as in cyclohexane, tetrahydrofuran, and water. The apparent anomalies experimentally found in water (a blue shift in the absorption bands with respect to the trend in other solvents, and an abnormally high red shift for the fluorescence band) are well reproduced and explained in view of the electronic structure of the solute and the solvent distribution around it. Additional calculations were done with a mixture of cyclohexane and tetrahydrofuran as solvent, which displays a nonlinear solvatochromic shift. Results, although not conclusive, are consistent with experiment and provide a possible explanation for the nonlinear behavior in the solvent mixture.

### 1. Introduction

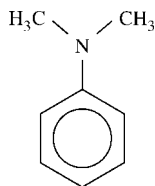
The nature and relative energies of the electronic states of a molecule determine its photophysical and photochemical properties. The environment in which a molecule is immersed can alter these states, which in turn modifies the properties, giving rise, for instance, to solvatochromic shifts in absorption and emission UV/vis spectra.<sup>1</sup> The experimental study of solvent effects on UV/vis spectra provides an important insight on the electronic properties of molecules, while their theoretical study represents an important challenge, since it requires both an accurate description of the internal structure of the solute and an appropriate modeling of the solvent structure and the solute–solvent interaction. The development of high-quality quantum methods capable of describing excited states (CASPT2, TD-DFT, etc.), together with convenient solvent models (PCM, RISM, MD, etc.), has allowed theoretical calculations of solvent effects to reach a high accuracy.

In our group, we have developed a method, called ASEP/MD (Averaged Solvent Electrostatic Potential from Molecular Dynamics) for including the solvent influence on quantum calculations.<sup>2–4</sup> This method has been successfully applied to the study of diverse properties and processes,<sup>5–10</sup> including UV/vis spectra.<sup>11–14</sup> In this paper, we carry out a study of solvent effects on the absorption and emission spectra of *N,N*-dimethylaniline (DMA), Figure 1. The solvatochromic shifts of the absorption and emission maxima of DMA in different solvents are in general proportional to the polarity function of the solvent ( $f(\epsilon) = 2(\epsilon - 1)/(2\epsilon + 1)$ ), but in water and other protic solvents this trend is broken. Additionally, in cyclohexane/tetrahydrofuran solvent mixtures, the solvatochromic shifts do not vary linearly with the molar fractions, as would be expected from the dielectric properties of the solvent.<sup>1</sup> We expect the ASEP/MD method to be able to correctly reproduce and explain these apparently anomalous behaviors, since it takes into account the explicit structure of the solvent and allows the use of accurate quantum methods. To attain these goals, it has been necessary to extend the method to work with solvent mixtures of arbitrary composition, which required only minimal changes in the previous software.

\* E-mail: jellby@unex.es.

<sup>†</sup> Universidad de Extremadura.

<sup>‡</sup> Università degli Studi di Pisa.



**Figure 1.** *N,N*-Dimethylaniline (DMA).

In section 2 we present a description of the methods and models used in this work, along with computational details. Section 3 contains the obtained results and discussion, divided into subsections gas phase, pure solvents, and solvent mixture. Finally, our conclusions are presented in section 4.

## 2. Methods and Details

Solvent effects on the DMA UV/vis spectra were calculated with ASEP/MD method. This is a sequential quantum mechanics/molecular mechanics (QM/MM) method implementing the mean field approximation. It combines, alternately, a high-level quantum mechanics (QM) description of the solute with a molecular mechanics (MM) description of the solvent. One of its main features is the fact that the solvent effect is introduced into the solute's wave function as an average perturbation. Details of the method have been described in previous papers,<sup>2–4</sup> so here we will only present a brief outline.

As mentioned above, ASEP/MD is a method combining QM and MM techniques, with the particularity that full QM and MD (molecular dynamics) calculations are alternated and not simultaneous. During the MD simulations, the intramolecular geometry and charge distribution of all molecules is considered as fixed. From the resulting data, the average electrostatic potential generated by the solvent on the solute (ASEP) is obtained. This potential is introduced as a perturbation into the solute's quantum mechanical Hamiltonian, and by solving the associated Schrödinger equation, one gets a new charge distribution for the solute, which is used in the next MD simulation. This iterative process is repeated until the electron distribution of the solute and the solvent structure around it are mutually equilibrated.

The ASEP/MD framework can also be used to optimize the geometry of the solute molecule.<sup>5</sup> At each step of the ASEP/MD procedure, the gradient and Hessian on the system's free-energy surface (including the Van der Waals contribution) can be obtained, and so they can be used to search for stationary points on this surface by some optimization method. After each MD simulation, the solute geometry is optimized within the fixed "average" solvent structure by using the free-energy derivatives. In the next MD simulation, the new solute geometry and charge distribution are used. This approach allows the optimization of the solute geometry simultaneously to the solvent structure.

For calculating transition energies, the iterative process is performed on the initial state of the transition (the ground-state for absorption, the excited-state for emission), i.e., the atomic charges for the MD and the energy derivatives for the geometry optimization of the solute are calculated with the initial state wave function. Then, with a frozen solvent

model, the transition energies between the different states are obtained. It is also possible to calculate transition energies with a polarizable solvent model; in this case, once the solute and solvent structure have been optimized for the initial state of the solute, each state energy and wave function is calculated with the same solvent structure, but where the solvent molecules' charges are replaced by gas-phase charges plus a molecular polarizability.<sup>11,13</sup> In this work we used a nonpolarizable solvent model in all cases, as test calculations with polarizable solvent did not show an important enough influence to compensate the increased computational effort required.

With the transition energies calculated in solution and in gas phase, the solvent shift  $\delta$  can be obtained as the difference:

$$\begin{aligned} \delta &= \Delta E - \Delta E^0 \\ &= (\langle \Psi_f | \hat{H}_{QM} + \hat{V} | \Psi_f \rangle - \langle \Psi_i | \hat{H}_{QM} + V | \Psi_i \rangle) - \\ &\quad (\langle \Psi_f^0 | \hat{H}_{QM}^0 | \Psi_f^0 \rangle - \langle \Psi_i^0 | \hat{H}_{QM}^0 | \Psi_i^0 \rangle) \\ &= (\langle \Psi_f | \hat{H}_{QM} + \hat{V} | \Psi_f \rangle - \langle \Psi_f^0 | \hat{H}_{QM}^0 | \Psi_f^0 \rangle) - \\ &\quad (\langle \Psi_i | \hat{H}_{QM} + \hat{V} | \Psi_i \rangle - \langle \Psi_i^0 | \hat{H}_{QM}^0 | \Psi_i^0 \rangle) \end{aligned} \quad (1)$$

where the subindices *i* and *f* denote the initial and final state,  $\hat{H}_{QM}$  is the QM Hamiltonian of the solute at the in-solution geometry, without the solute–solvent interaction,  $\hat{V}$ , and  $\hat{H}_{QM}^0$  is the QM Hamiltonian at the gas-phase geometry;  $\Psi$  and  $\Psi^0$  are, respectively, the wave functions optimized in solution and in gas phase. This solvent shift can be partitioned in different contributions, namely a geometry contribution  $\delta_{\text{geo}}$ , an electronic distortion contribution  $\delta_{\text{dist}}$ , and an electrostatic solute–solvent contribution  $\delta_{\text{elec}}$ . If we introduce  $\Psi'$  as the wave function optimized for the  $\hat{H}_{QM}$  Hamiltonian:

$$\begin{aligned} \delta &= \delta_{\text{geo}} + \delta_{\text{dist}} + \delta_{\text{elec}} \\ \delta_{\text{geo}} &= (\langle \Psi'_f | \hat{H}_{QM} | \Psi'_f \rangle - \langle \Psi_f^0 | \hat{H}_{QM}^0 | \Psi_f^0 \rangle) - \\ &\quad (\langle \Psi'_i | \hat{H}_{QM} | \Psi'_i \rangle - \langle \Psi_i^0 | \hat{H}_{QM}^0 | \Psi_i^0 \rangle) \\ \delta_{\text{dist}} &= (\langle \Psi_f | \hat{H}_{QM} | \Psi_f \rangle - \langle \Psi'_f | \hat{H}_{QM} | \Psi'_f \rangle) - \\ &\quad (\langle \Psi_i | \hat{H}_{QM} | \Psi_i \rangle - \langle \Psi'_i | \hat{H}_{QM} | \Psi'_i \rangle) \\ \delta_{\text{elec}} &= \langle \Psi_f | \hat{V} | \Psi_f \rangle - \langle \Psi_i | \hat{V} | \Psi_i \rangle \end{aligned} \quad (2)$$

Thus,  $\delta_{\text{geo}}$  is the solvent shift due to the change in geometry between gas phase and solution,  $\delta_{\text{elec}}$  corresponds to the difference in solute–solvent interaction energy between the initial and final states, and  $\delta_{\text{dist}}$  corresponds to the difference in the wave function distortion energy. For convenience, fluorescence energies are reported as positive values, although they would be negative when eq 1 is applied. Similarly the  $\delta$  values for fluorescence are given as positive numbers for blue shifts and negative for red shifts. Note that the Van der Waals component of the interaction energy is not included in the above expressions, since we adopt the approximation of considering it constant for all electronic states of the solute, and therefore it vanishes when vertical transition energies are considered.

The quantum calculations of the solute molecule were done with the complete active space self-consistent field (CAS-SCF) method,<sup>15</sup> using the 6–311G\*\* basis set. Gas-phase calculations were also done with 6–31G\*\*, cc-pVDZ, and

6-311++G\*\* basis sets. The active orbitals were the six  $\pi$  and  $\pi^*$  orbitals of the phenyl ring and the nonbonded orbital of the nitrogen, and eight electrons were included in these orbitals, for an (8,7) total active space. Geometry optimizations in gas phase and in solution were performed on pure roots (the ground state,  $S_0$ , or the first excited singlet state,  $S_1$ ), but transition energies were always calculated with a state-average (SA) calculation of the first three singlet states,  $S_0$ ,  $S_1$ , and  $S_2$ . To obtain accurate transition energies, it is known that the inclusion of dynamic correlation in the quantum calculations is necessary, which we did with the complete active space second-order perturbation (CASPT2) method,<sup>16,17</sup> using the SA-CASSCF(8,7) wave functions as reference. A new IP-EA shifted zeroth-order Hamiltonian has been recently proposed for CASPT2 calculations,<sup>18</sup> which is supposed to reduce systematic overestabilization errors in open-shell systems (as in the excited states). We did all CASPT2 with the proposed IP-EA shift of  $0.25 E_h$  (CASPT2(0.25)) as well as with no IP-EA shift (CASPT2(0.00)). To minimize the appearance of intruder states, an additional imaginary shift of  $0.1i E_h$  was used. No symmetry was assumed in any case.

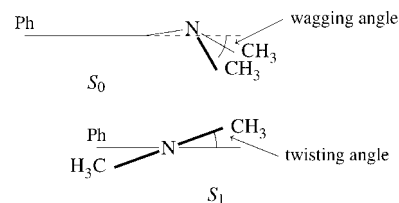
The MD simulations were carried out with rigid molecules; cyclohexane, tetrahydrofuran, and water were used as solvents. Lennard-Jones parameters and solvent atomic charges were taken from the OPLS-AA force field,<sup>19</sup> and solute atomic charges were calculated from the quantum calculations with the CHELPG method.<sup>20</sup> The geometry of cyclohexane and tetrahydrofuran were optimized with B3LYP/6-311G\*\*; for water, the TIP3P model was employed. An amount of 216 solvent molecules and the solute were included in a cubic simulation box (800 water molecules for aqueous solution) at the experimental density of the solvent.<sup>21</sup> Periodic boundary conditions were applied, and spherical cut-offs were used to truncate the interatomic interactions at 12 Å; long-range interactions were calculated using the Ewald sum technique. The temperature was fixed at 298 K by using the Nosé-Hoover thermostat. A time step of 0.5 fs was used during the simulations, and each one was run for 100 ps after 25 ps equilibration.

At each step of the ASEP/MD procedure, 500 configurations evenly distributed from the MD run were used to calculate the ASEP and a radius of  $15 a_0$  ( $12 a_0$  for water) was used for including explicit solvent charges. Each ASEP/MD run was continued until the energies and solute geometry and charges are stabilized for at least five iterations; results are reported as the average of these last five iterations.

For in-solution calculations, the ASEP/MD software<sup>3</sup> was used, with the needed modifications to allow the use of more than one solvent species. During the ASEP/MD runs, quantum calculations (CASSCF optimizations) were performed with the Gaussian 98 package.<sup>22</sup> The final SA-CASSCF and CASPT2 calculations were done with Molcas 6.4.<sup>23</sup> All MD simulations were performed using Moldy.<sup>24</sup>

### 3. Results and Discussion

**3.1. Gas Phase.** The geometry of DMA was optimized in gas phase, at CASSCF/6-311G\*\* level, for both the



**Figure 2.** Scheme showing the wagging and twisting angles in the ground and excited states of DMA.

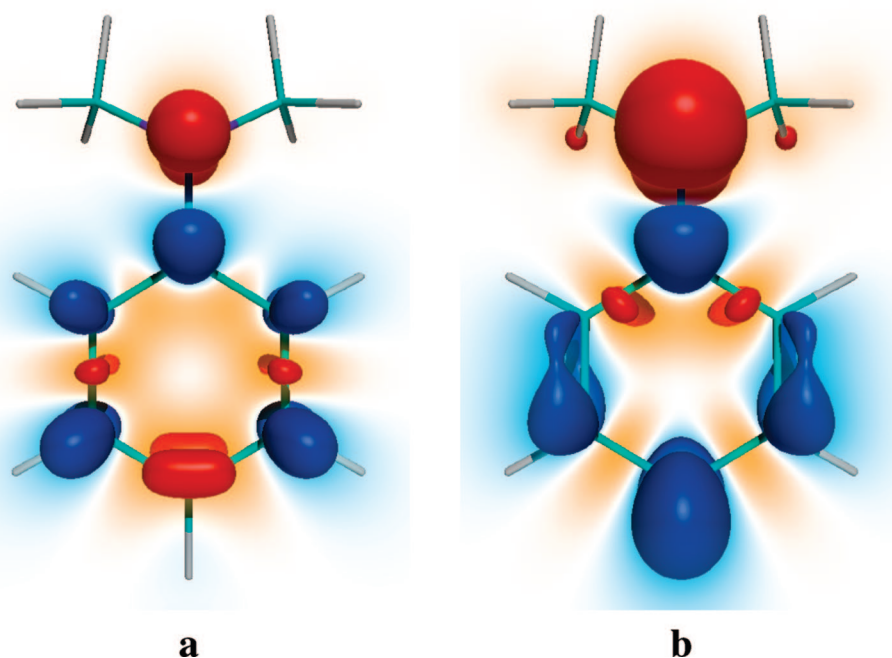
**Table 1.** Absorption Energies Calculated in Gas Phase, in eV (oscillator strength in parentheses)

	SA-CASSCF		CASPT2(0.25)		CASPT2(0.00)	
	$S_0 \rightarrow S_1$	$S_0 \rightarrow S_2$	$S_0 \rightarrow S_1$	$S_0 \rightarrow S_2$	$S_0 \rightarrow S_1$	$S_0 \rightarrow S_2$
6-31G**	4.82	7.11	4.77 (0.010)	5.73 (0.267)	4.41 (0.010)	5.27 (0.246)
cc-pVDZ	4.79	6.99	4.71 (0.011)	5.57 (0.268)	4.33 (0.010)	5.09 (0.245)
6-311G**	4.80	6.99	4.69 (0.010)	5.54 (0.269)	4.30 (0.009)	5.06 (0.245)
6-311++G**	4.78	6.85	4.65 (0.008)	5.32 (0.277)	4.30 (0.007)	4.87 (0.253)
experimental <sup>26</sup>	4.30 (0.044)	5.16 (0.256)				

ground state ( $S_0$ ) and the first excited state ( $S_1$ ). In agreement with experimental results,<sup>25</sup> the obtained  $S_0$  geometry is pyramidal in the N, with a  $\text{CH}_3\text{-N-CH}_3$  angle of  $114.7^\circ$  (experimental:  $114^\circ$ ) and a wagging angle (the angle between the phenyl ring plane and the  $\text{CH}_3\text{-N-CH}_3$  plane) of  $28.4^\circ$  (experimental:  $27.0^\circ$ ), the N atom being slightly ( $0.059 \text{ \AA}$ ) out of the phenyl ring plane (see Figure 2). These geometrical parameters are maintained (to within  $0.4^\circ$ ,  $0.6^\circ$ , and  $0.004 \text{ \AA}$ , respectively) when the optimization is carried out with the 6-31G\*\*, cc-pVDZ, and 6-311++G\*\* basis sets.

The transition energies to the  $S_1$  and  $S_2$  states, at the CASSCF optimized  $S_0$  geometry were calculated with a state-average CASSCF method (including the first three roots), and with perturbation theory using both CASPT2(0.25) and CASPT2(0.00). The results are displayed in Table 1; it is clear that both absorption energies are overestimated at SA-CASSCF level, but the CASPT2 method yields results in good agreement with the experiment. As expected, the transition energies with CASPT2(0.25) are larger than with CASPT2(0.00), the latter results being closer to the experimental values. However, given that CASPT2(0.25) results approach the experimental reference when the basis set quality is improved, the good performance of CASPT2(0.00) in this case is probably due to error cancelation, especially for the  $S_0 \rightarrow S_2$  transition.

The oscillator strengths for the two transitions  $S_0 \rightarrow S_1$  and  $S_0 \rightarrow S_2$  are also in very good agreement with the experimental estimations and are much less dependent on the basis set and method. They indicate that the transition to  $S_1$  has a weak intensity while that to  $S_2$  is much more favored. According to the assignment of Kimura et al.,<sup>26</sup> the main contribution to the  $S_1$  state would correspond to a local excitation in the phenyl ring, while  $S_2$  stems from an intramolecular charge transfer between the  $\text{N}(\text{CH}_3)_2$  electron donor and the phenyl acceptor. This assignment is confirmed by the calculated dipole moments of the three states, being at CASPT2(0.25)/6-311G\*\* level, 1.33 D for  $S_0$ , 1.66 D for  $S_1$ , and 5.98 D



**Figure 3.** Electron density change in the  $S_0 \rightarrow S_1$  transition (a) and in the  $S_0 \rightarrow S_2$  transition (b). Isosurfaces for a change of  $\pm 0.0032$ , red for a decrease in density, blue for an increase. Densities calculated at SA-CASSCF/6-311G\*\* level.

for  $S_2$ , in all cases directed from the phenyl ring to the nitrogen and toward the side of the ring plane where the methyl groups lie. Electron density differences between the ground state and  $S_1$  and  $S_2$  are displayed in Figure 3; they clearly show the important charge-transfer nature of the  $S_2$  state. There is also some transfer component in  $S_1$ , but it is not so drastic. The Mulliken populations confirm a flux of 0.28 electrons from  $N(\text{CH}_3)_2$  to the phenyl for the  $S_0 \rightarrow S_2$  transition and only 0.05 electrons for  $S_0 \rightarrow S_1$ .

The CASSCF/6-311G\*\* optimization of the  $S_1$  state yields a planar structure of the N atom, but the  $\text{CH}_3\text{-N-CH}_3$  plane is now twisted  $19.5^\circ$  with respect to the phenyl ring (Figure 2) and the  $\text{CH}_3\text{-N-CH}_3$  angle is  $115.9^\circ$ . Again the other basis sets give similar results. This planar and twisted structure in the excited state agrees with the interpretation of the experimental spectrum given by Saigusa et al.,<sup>27</sup> who conclude a torsion angle of  $26^\circ$ . These authors suggest a pyramidal N atom (with a wagging angle of  $13^\circ$ ) but with an inversion barrier so low that it would lie below zero-point energy, and thus the  $S_1$  state of DMA could be considered planar in the N atom.

Table 2 collects the calculated band origins (0-0 transition) and fluorescence energies ( $S_1 \rightarrow S_0$ ) obtained with the different methods and basis sets, with the optimized  $S_0$  and  $S_1$  geometries. Similarly to the absorption energies, SA-CASSCF overestimates the transition energies and the difference between CASPT2(0.25) and CASPT2(0.00) is quite constant, around 0.3–0.4 eV. Again, with increasing basis set quality CASPT2(0.25), results seem to improve.

It was also possible to optimize an untwisted pyramidal geometry for  $S_1$ , similar to the  $S_0$  structure, with a wagging angle of  $19.5^\circ$ . At CASPT2(0.00)//CASSCF/6-311G\*\* level, this wagged minimum is 0.03 eV higher in energy than the planar twisted one, its  $S_1 \rightarrow S_0$  transition energy is 0.12 eV larger, and its dipole moment is  $\sim 0.2$  D lower. The

**Table 2.** Band Origins and Fluorescence Energies Calculated in Gas Phase, in eV (oscillator strength in parentheses)

	SA-CASSCF		CASPT2(0.25)		CASPT2(0.00)	
	0-0	$S_1 \rightarrow S_0$	0-0	$S_1 \rightarrow S_0$	0-0	$S_1 \rightarrow S_0$
6-31G**	4.61	4.34	4.52	4.26 (0.015)	4.15	3.90 (0.014)
cc-pVDZ	4.59	4.32	4.40	4.20 (0.018)	4.01	3.82 (0.016)
6-311G**	4.60	4.32	4.41	4.17 (0.018)	4.01	3.79 (0.016)
6-311++G**	4.57	4.31	4.37	4.14 (0.016)	4.00	3.79 (0.015)
experimental	4.08 <sup>a</sup>	3.69 <sup>b</sup> $\sim 3.87^c$				

<sup>a</sup> Reference 27. <sup>b</sup> Reference 28 in *n*-hexane. <sup>c</sup> Reference 29 in *n*-hexane (estimated from graph).

lower energy of the twisted minimum and its fluorescence energy more in agreement with the experimental results available make this structure the most likely for the excited-state of DMA, in line with the conclusions of Saigusa et al.<sup>27</sup> Moreover, the higher dipole moment would additionally favor the twisted minimum in solution, as it would be better stabilized by the solvent. The wagged minimum may be an artifact of the CASSCF optimization and it might not appear if the optimization were performed at CASPT2 level. In the rest of this paper we always consider the planar twisted structure for the optimized  $S_1$  state.

**3.2. Pure Solvents.** The DMA geometry was also optimized in solution, using cyclohexane (CH), tetrahydrofuran (THF), and water as solvents. As in the gas-phase study, the  $S_0 \rightarrow S_1$  and  $S_0 \rightarrow S_2$  absorption energies were calculated with the optimized  $S_0$  structure, while the  $S_1 \rightarrow S_0$  fluorescence energy was calculated only with the planar twisted  $S_1$  structure.

**Table 3.** Characteristic Angles (in degrees) and Dipole Moments (at CASPT2(0.00)/6-311G\*\* level, in D) for Optimized Geometries of DMA<sup>a</sup>

	<i>S</i> <sub>0</sub> geometry				<i>S</i> <sub>1</sub> geometry		
	wag	$\mu(S_0)$	$\mu(S_1)$	$\mu(S_2)$	twist.	$\mu(S_0)$	$\mu(S_1)$
gas	28.4	1.34	1.68	5.98	19.5	1.62	2.19
cyclohexane	28.7	1.34	1.67	5.95	19.0	1.64	2.21
CH/THF (0.5)	28.5	1.41	1.77	6.05	18.5	1.91	2.53
tetrahydrofuran	28.5	1.56	1.93	6.25	18.1	2.10	2.78
water	34.0	1.56	1.77	5.83	15.8	3.23	4.56

<sup>a</sup> For the *S*<sub>0</sub> geometry, the wagging angle is given; for the *S*<sub>1</sub> geometry, the twisting angle is given.

The optimized wagging and twisting angles, as well as the dipole moments of the different states in the solvents considered are given in Table 3. As with the gas-phase calculations, the geometry was optimized with the CASSCF method, energies and dipoles were then calculated at SA-CASSCF and CASPT2 level, and only the 6-311G\*\* basis set was used. The table shows a trend in the gas phase, cyclohexane, and tetrahydrofuran results: CH values are very similar to gas phase, while THF, with stronger polarity, originates an increase in the dipole moments, more important in the *S*<sub>1</sub> optimization. The changes in the wagging and twisting angles are negligible. In water, however, the behavior is different. In the *S*<sub>0</sub> geometry the pyramidalization of the N is enhanced and the dipole moments do not increase from the THF values; on the contrary, they decrease for the excited states. In the *S*<sub>1</sub> geometry, the changes in the twisting angle and in the dipole moments go in the same direction as with the other solvents, but they are much more important. These results already indicate a certain anomaly for DMA when dissolved in water, as will be seen in the transition energies.

Different estimations for the dipole moment difference between the ground and excited states, based on experimental solvatochromic and thermochromic shifts, have proposed values of 3.5 D,<sup>30</sup> 3.27 D,<sup>31</sup> or 1.89 D–1.99 D.<sup>32</sup> Our results cast doubt on the validity of these estimations, as we obtain a dipole moment difference between 0.9 and 1.2 D (*S*<sub>0</sub> and *S*<sub>1</sub> at their respective minima), and much lower if we consider the dipole moment increase upon excitation (*S*<sub>0</sub> and *S*<sub>1</sub> at the ground-state minimum). Only in water is the dipole moment difference 3 D, but the experimental data refer only to less polar solvents. In our opinion, the disagreement between our values and the experimental estimations shows the errors associated to the assumptions of the above-mentioned works, which basically rely on the Onsager solvation model.

The different transition energies calculated in solution are detailed in Table 4. As expected, the values obtained in cyclohexane are almost identical to the gas-phase results, with just a very slight (0.01 eV) blue shift in the absorption bands. This contrasts with the somewhat more sizable red shift (~0.1 eV) found experimentally in all three transitions studied.<sup>31,33</sup> There are several possible sources for this error.

(1) The calculations did not consider the solvent electronic polarization in response to the electron transition in the solute. We did some test calculations with the polarizable version of ASEP/MD, in cyclohexane, and we obtained only a very

**Table 4.** Transition Energies, in eV, Calculated in Solution at CASPT2(0.00)/6-311G\*\* level (experimental values in parentheses)

	<i>S</i> <sub>0</sub> → <i>S</i> <sub>1</sub>	<i>S</i> <sub>0</sub> → <i>S</i> <sub>2</sub>	<i>S</i> <sub>1</sub> → <i>S</i> <sub>0</sub>
gas	4.30 (4.30) <sup>a</sup>	5.06 (5.16) <sup>a</sup>	3.79
cyclohexane	4.31 (4.22) <sup>b</sup>	5.07 (5.02) <sup>b</sup>	3.79 (~3.72) <sup>c</sup>
CH/THF (0.5)	4.30	5.05	3.75
tetrahydrofuran	4.29	5.03	3.73
water	4.39 (4.28) <sup>d</sup>	5.15 (~5.10) <sup>e</sup>	3.47 (3.40) <sup>d</sup>

<sup>a</sup> Reference 26. <sup>b</sup> Reference 31. <sup>c</sup> Reference 33 (estimated from graph). <sup>d</sup> Reference 34. <sup>e</sup> Reference 35 (estimated from graph).

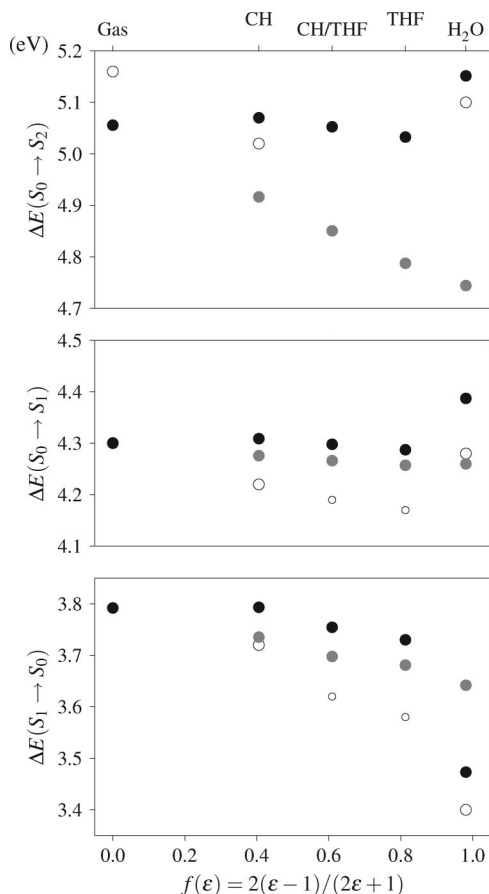
small red shift (~0.02 eV) with respect to the nonpolarizable calculations. This is therefore not enough to explain the discrepancy between the experimental and the calculated transition energies in solution.

(2) The neglect of the dispersion component of the transition energies. This component is known<sup>36</sup> to decrease transition energies in solution, since, in general, excited states are better stabilized by dispersion. There is, however, no accurate way to include the dispersion component in the calculations other than including a number of solvent molecules in the quantum system, which makes it difficult to estimate the contribution of this component. Nevertheless, the dispersion component depends mainly, in what regards the solvent, on the refractive index, and since this is quite constant in the studied solvents (1.33–1.43), we can expect the dispersion contribution to be similar in all cases. This would result in solvent differences and trends being well reproduced.

The transition energies obtained in tetrahydrofuran show a small red shift with respect to the cyclohexane values. The shift is larger for the *S*<sub>1</sub>→*S*<sub>0</sub> transition (0.06 eV) and smaller for the *S*<sub>0</sub>→*S*<sub>1</sub> transition (0.02 eV). This red shift is expected, considering the higher dipole moment of the excited states of DMA and the increased polarity of THF. The experimental data available<sup>1</sup> confirm the increased red shift both in absorption and fluorescence bands.

This trend, higher solvent polarity gives a larger red shift, is broken when the solvent is water (see Figure 4). In this case there are “anomalies” both in the absorption and emission energies, as happened with the geometry and dipole moments, commented above. In the absorption bands there is a blue shift of 0.08 eV when the cyclohexane and water solvents are compared, which would not be expected on the basis of the solvent polarity alone. In the fluorescence band, the red shift observed in water is much larger (0.32 eV) than what could be expected from polarity, too. These two anomalies are also found experimentally. The blue shift in absorption is also observed with other protic solvents such as alcohols, while the extraordinarily high red shift in fluorescence is only found in water.<sup>1</sup>

It is interesting to note that the error in the calculated values of the transition energies is very similar in cyclohexane and in water, despite being such disparate solvents. This fact points to the dispersion component as mainly responsible for the error in the computed transition energies in solution,



**Figure 4.** Transition energies for DMA in gas phase and different solvents, from Tables 4 and 5. Grey circles are PCM values, black circles are ASEP/MD or gas-phase values, and white circles are experimental values. The small white circles are obtained from Figure 12 in ref 1, considering the difference with respect to cyclohexane.

**Table 5.** Transition Energies, in eV, Calculated in Solution, with PCM, at CASPT2(0.00)/6-311G\*\* Level

	$S_0 \rightarrow S_1$	$S_0 \rightarrow S_2$	$S_1 \rightarrow S_0$
cyclohexane	4.28	4.92	3.74
CH/THF (0.5)	4.27	4.85	3.70
tetrahydrofuran	4.26	4.79	3.68
water	4.26	4.74	3.64

since, as noted above, the magnitude of this component is expected to be quite similar in the different solvents. Thus, the trends in solvation are very well reproduced, as can be seen in Figure 4 if the differences with respect to cyclohexane are considered. Also, the error is similar for the absorption and emission energies, which translates in the calculated Stokes shifts being in excellent agreement with experimental values: 0.52 eV (exp. 0.50 eV) in cyclohexane, 0.92 eV (exp. 0.88 eV) in water. It is also worth mentioning that CASPT2(0.25) values for the transition energies (not given in Table 4) were in all cases 0.39 eV higher for the  $S_0 \rightarrow S_1$  and  $S_1 \rightarrow S_0$  transitions, and 0.48 eV higher for the  $S_0 \rightarrow S_2$  absorption.

The observed anomalies are not explained by continuum models, such as the Polarizable Continuum Model (PCM).<sup>37,38</sup> For comparison, we carried out PCM calculations of the three studied transitions; the results are shown in Table

5. As before, the geometries were optimized at CASSCF(8,7)/6-311G\*\* level, and the final energies were calculated with SA-CASSCF and CASPT2. To compare with the nonpolarizable ASEP/MD calculations, the fast polarization component in PCM was neglected, i.e., all solute states were calculated with the solvent charges in equilibrium with the initial state ( $S_0$  for absorption,  $S_1$  for fluorescence). The  $S_0 \rightarrow S_1$  and  $S_1 \rightarrow S_0$  transition energies are slightly smaller than with ASEP/MD, but the differences between cyclohexane and tetrahydrofuran are very similar. In the  $S_0 \rightarrow S_2$  transition, the difference is larger and the calculated values are further from experiments. As expected, in all cases, the results with water follow the general trend and do not show the anomalies described above (see Figure 4). We also calculated the transition energies in vacuo with the PCM-optimized solute geometries, and we did not find significant differences, in any of the solvents, compared to the gas-phase transition energies. The increase in wagging angle in the ground-state in water is significantly smaller with PCM (2.5°) than with ASEP/MD (5.6°). For the excited state, the change in the twisting angle is stronger with PCM, but this is compensated for with a less out-of-plane position of the hydrogens in the ortho positions.

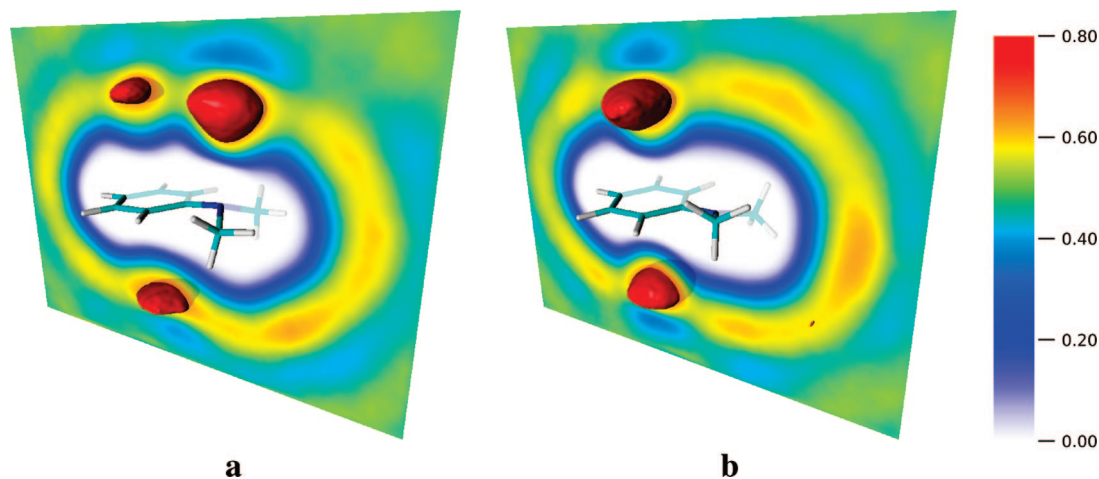
The behavior of the electron transitions in water must be therefore associated to specific interactions between the water molecules and the solute and not only to the bulk properties (polarity) of the solvent. The described anomalies are compatible with (a) a specific strong stabilization of the ground-state through O–H···N hydrogen bonds, which is lost when the excitation to  $S_1$  or  $S_2$  occurs, and (b) an increased stabilization of the  $S_1$  state before fluorescence, probably through solvation of the phenyl ring, which is also lost when the relaxation to  $S_0$  takes place.

In order to gain a deeper insight on the reasons for the behavior in water, we first performed gas-phase calculations with the geometries optimized in solution, which allowed us to obtain the solvent shifts components calculated according to eq 2, given in Table 6. The results for the two absorption energies are 4.35 and 5.11 eV. These values are halfway between the gas phase and the aqueous solution (4.30–4.39 eV and 5.06–5.15 eV) and already show a blue shift of  $\sim 0.05$  eV ( $\delta_{\text{geo}}$ ). Thus, an important part of the effect of water on the absorption spectrum of DMA can be ascribed to the influence on the molecular geometry: an increased wagging angle originates larger transition energies (a similar dependence was already described for the *p*-cyano derivative<sup>37</sup>). The other  $\sim 0.05$  eV of blue shift is then due to the difference in stabilization of the electron density in the ground and excited states ( $\delta_{\text{dist}} + \delta_{\text{elec}}$ ).

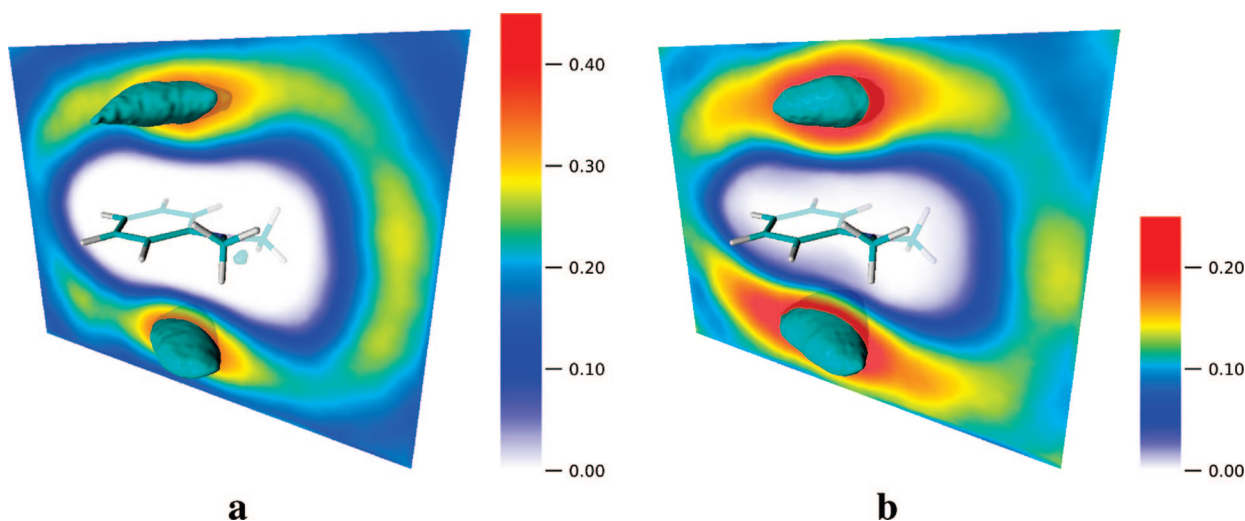
In the  $S_1$  structure, the geometry change in the solute is slightly smaller, but it also has an important effect on the transition energy. With the optimized geometry in solution, we obtain a gas-phase fluorescence energy of 3.72 eV, with a shift of  $-0.08$  eV ( $\delta_{\text{geo}}$ ). In this case, however, the effect of the solvent on the electron density stabilization is much higher, accounting for an additional shift of  $-0.24$  eV ( $\delta_{\text{dist}} + \delta_{\text{elec}}$ ).

By examining the distribution of water molecules around the solute, the effect of solvation on the transition energies





**Figure 5.** Occupancy maps of water oxygen atoms (considered as Van der Waals spheres, as calculated by VMD<sup>38</sup>) around DMA for (a) the optimized  $S_0$  structure, and (b) the optimized  $S_1$  structure. Solid isosurfaces shown for values of 0.64.



**Figure 6.** Occupancy maps of THF  $C_\beta$  atoms (considered as Van der Waals spheres, as calculated by VMD<sup>38</sup>) around DMA in the optimized  $S_1$  state for (a) pure THF (isosurface value 0.35), and (b) THF/CH mixture (isosurface value 0.22). Note the different color scales in a and b.

**Table 6.** Solvent Shifts and Their Components, in eV, in Water, Calculated at CASPT2(0.00)/6-311G\*\* Level

	$\delta$	$\delta_{\text{geo}}$	$\delta_{\text{dist}}$	$\delta_{\text{elec}}$
$S_0 \rightarrow S_1$	0.087	0.045	0.005	0.036
$S_0 \rightarrow S_2$	0.096	0.058	-0.020	0.057
$S_1 \rightarrow S_0$	-0.319	-0.075	-0.156	-0.087

can be further understood. Figure 5 shows in red the regions of space where oxygen atoms are more frequently found. There is a clear high concentration of water molecules near the N atom in the  $S_0$  structure, indicating the existence of a hydrogen bond. This hydrogen bond stabilizes in particular the ground state, while the excited states, characterized by an electron density loss in the N, are less stabilized. Thus, the electrostatic contribution leads to a larger energy difference between the states, giving rise to a blue shift in the absorption bands, which is indicated by the positive sign of  $\delta_{\text{elec}}$ . There are also regions of high oxygen concentration at both sides of the phenyl ring, solvating its partial negative charge (through the hydrogens, not shown). These solvent

molecules contribute to stabilize in preference the excited states and somewhat counter the effect of the N atom solvation.

In the optimized  $S_1$  structure, only the high oxygen concentration regions at both sides of the phenyl ring are found, and they are closer to the solute and stronger than in the  $S_0$  structure. As before, these solvent molecules contribute to stabilizing the excited state more than the ground state. Moreover, the absence of water molecules solvating the N atom means that there is no counter stabilization of the ground state, and thus  $\delta_{\text{elec}}$  is negative and larger in absolute value than for the absorptions.

**3.3. Solvent Mixture.** We also studied the behavior of DMA in a mixture of cyclohexane and tetrahydrofuran, with a molar fraction of 0.5. It is found experimentally that the solvatochromic shift, especially of the fluorescence band, is clearly nonlinear with the molar fraction, although the solvent mixture itself shows an almost ideal dielectric behavior,<sup>1</sup> where the polarity function  $f(\epsilon) = 2(\epsilon - 1)/(2\epsilon + 1)$  varies linearly with the molar fraction of the components.

The obtained results are included in Tables 3 and 4, all values are intermediate between those of cyclohexane and tetrahydrofuran, as expected. Regarding the transition energies, although the studied variations are rather small (0.02–0.06 eV), some nonlinearity can be observed in the fluorescence energies, where the maximum in the solvent mixture is closer to the value in THF than to that in CH. Both the direction and the amount of the nonlinearity are in agreement with experiment.<sup>1</sup>

This effect would be compatible with a preferential solvation of DMA by THF, meaning that the local concentration of this solvent around the solute should be higher than its bulk concentration. However, we find the opposite effect: the average number of tetrahydrofuran molecules within 3 Å of the solute is 5.1, while the number of cyclohexane molecules is 6.7 (a local THF molar fraction of 0.43). But, as it was shown for water (Figure 5b), solvation of the  $S_1$  state occurs mainly at both sides of the phenyl ring. If we place one point at 3.5 Å at either side of the ring and consider only the solvent molecules within 1 Å of these points, we get in turn that the local THF molar fraction in these regions is 0.54. Thus, the preferential solvation by THF is observed in the regions most important for the stabilization of the excited-state of the solute, while in other regions THF is depleted. This is shown in Figure 6, taking into account that the partial density of THF in the solvent mixture is one-half of the pure solvent. The volumes inside the isosurfaces are similar, but the occupancy value for the mixture is 63%, more than one-half, of the value for pure THF. Likewise, the change in the color scale allows comparison of the occupancies in relation to the partial THF density.

Again, we compare with the results obtained with PCM, in Table 5. Somewhat surprisingly, the same nonlinearity in the  $S_1 \rightarrow S_0$  transition is found in this case. The nonlinear behavior cannot be attributed here to the solvent response, since it is modeled as a linear-response continuum, so it must be due to the solute. In fact, the vacuum emission energy obtained with the PCM-optimized geometry in the solvent mixture is 0.02 eV lower than with the geometry in THF and 0.01 eV lower than in CH, and this can explain the nonlinearity in the final values. In any case, the energy variations are probably too small to draw definitive conclusions: a difference of only  $\sim 0.01$  eV separates linear and nonlinear behavior.

#### 4. Conclusions

A theoretical study of the lowest-energy electron transitions in *N,N*-dimethylaniline has been performed. The first absorption transition has a very low intensity and implies mainly a local excitation on the phenyl ring, similarly to the fluorescence transition; the transition to the second excited state has a significant charge transfer component and consequently an enhanced intensity. Results in gas phase agree with experiments and support a pyramidal ground state and a twisted planar excited state for the DMA molecule.

In solution, a red shift of the absorption and fluorescence bands is found in polar nonprotic solvents, which is more important in the  $S_0 \rightarrow S_2$  transition. The anomalous behavior experimentally found in water is well reproduced: a blue shift

in the absorption bands seems to be due to the strong stabilization of the ground state through hydrogen bonds between water and the amine nitrogen, with an important contribution from the geometrical distortion of the solute; the strong red shift in the fluorescence band corresponds to an increased solvation of the phenyl ring in the excited state.

For the first time, calculations with a solvent mixture (cyclohexane and tetrahydrofuran) were performed with the ASEP/MD method. These calculations reproduce the nonlinearity found in the solvent shift with the mixture composition, and, although the magnitude of the effect does not allow to draw definitive conclusions, the results point to a local increase of the concentration of THF only in the regions perpendicular to the phenyl ring, where solvation of the excited state occurs, as a possible cause for the nonlinearity.

In summary, these results are in good agreement with experimental findings and show the ability of the ASEP/MD method to correctly describe the solute–solvent interactions involved in solvent shifts of absorption and emission bands. Moreover, the detailed representation of the system allows a more complete analysis of those interactions than with other models.

**Acknowledgment.** I.F.G. acknowledges the Junta de Extremadura and the European Social Fund for financial support. This work was supported by the CTQ2008-06224/BQU Project from the Ministerio de Ciencia e Innovación of Spain.

#### References

- (1) Suppan, P. *J. Photochem. Photobiol. A* **1990**, *50*, 293–330.
- (2) Sánchez, M. L.; Aguilar, M. A.; Olivares del Valle, F. J. *J. Comput. Chem.* **1997**, *18*, 313–322.
- (3) Fdez. Galván, I.; Sánchez, M. L.; Martín, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *Comput. Phys. Commun.* **2003**, *155*, 244–259.
- (4) Aguilar, M. A.; Sánchez, M. L.; Martín, M. E.; Fdez. Galván, I. An Effective Hamiltonian Method from Simulations: ASEP/MD. In *Continuum Solvation Models in Chemical Physics*, 1st ed.; Mennucci, B., Cammi, R., Eds., Wiley: New York, 2007; Chapter 4.5, pp 580–592.
- (5) Fdez. Galván, I.; Sánchez, M. L.; Martín, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2003**, *118*, 255–263.
- (6) Muñoz Losa, A.; Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *J. Phys. Chem. B* **2003**, *107*, 5043–5047.
- (7) Fdez. Galván, I.; Olivares del Valle, F. J.; Martín, M. E.; Aguilar, M. A. *Theor. Chem. Acc.* **2004**, *111*, 196–203.
- (8) Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *J. Comput. Chem.* **2004**, *25*, 1227–1233.
- (9) Fdez. Galván, I.; Aguilar, M. A.; Ruiz-López, M. F. *J. Phys. Chem. B* **2005**, *109*, 23024–23030.
- (10) Martín, M. E.; Muñoz Losa, A.; Fdez. Galván, I.; Aguilar, M. A. *J. Mol. Chem. Struct. (THEOCHEM)* **2006**, *775*, 81–86.
- (11) Martín, M. E.; Muñoz Losa, A.; Fdez Galván, I.; Aguilar, M. A. *J. Chem. Phys.* **2004**, *121*, 3710–3716.
- (12) Muñoz Losa, A.; Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *J. Phys. Chem. B* **2006**, *110*, 18064–18071.

- (13) Muñoz Losa, A.; Fdez. Galván, I.; Aguilar, M. A.; Martín, M. E. *J. Phys. Chem. B* **2007**, *111*, 9864–9870.
- (14) Muñoz Losa, A.; Fdez. Galván, I.; Martín, M. E.; Aguilar, M. A. *J. Phys. Chem. B* **2008**, *112*, 8815–8823.
- (15) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157–173.
- (16) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.
- (17) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (18) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (19) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (20) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (21) Lepori, L.; Matteoli, E. *Fluid Phase Equilib.* **1998**, *145*, 69–87.
- (22) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, Ö.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C. S.; Adamo, C.; Clifford, S.; Ochterski, J. W.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T. A.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andrés, J. L.; González, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98 (Revision A.11.3)*; Gaussian, Inc.: Pittsburgh, PA, 2001.
- (23) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.
- (24) Refson, K. *Comput. Phys. Commun.* **2000**, *126*, 310–329.
- (25) Cervellati, R.; Borgo, A. D.; Lister, D. G. *J. Mol. Struct.* **1982**, *78*, 161–167.
- (26) Kimura, K.; Tsubomura, H.; Nagakura, S. *Bull. Chem. Soc. Jpn.* **1964**, *37*, 1336–1346.
- (27) Saigusa, H.; Miyakoshi, N.; Mukai, C.; Fukagawa, T.; Kohtani, S.; Nakagaki, R.; Gordon, R. *J. Chem. Phys.* **2003**, *119*, 5414–5422.
- (28) Shanmugapriya, T.; Selvaraju, C.; Ramamurthy, P. *Spectrochim. Acta A* **2007**, *66*, 761–767.
- (29) Kawski, A.; Kukliński, B.; Bojarski, P. *Z. Naturforsch.* **2003**, *58a*, 411–418.
- (30) Ghoneim, N.; Suppan, P. *J. Chem. Soc., Faraday Trans.* **1990**, *86*, 2079–2081.
- (31) Prabhuramirashi, L. S.; Kutty, D. K. N.; Bhide, A. S. *Spectrochim. Acta A* **1983**, *39*, 663–668.
- (32) Kawski, A.; Kukliński, B.; Bojarski, P. *Chem. Phys.* **2006**, *320*, 188–192.
- (33) Tobita, S.; Kamiyama, R.; Takehira, K.; Yoshihara, T.; Yotoriyama, S.; Shizuk, H. *Anal. Sci.* **2001**, *17*, s50–s52.
- (34) Oshima, J.; Shiobara, S.; Naoumi, H.; Kaneko, S.; Yoshihara, T.; Mishra, A. K.; Tobita, S. *J. Phys. Chem. A* **2006**, *110*, 4629–4637.
- (35) Weidemaier, K.; Tavernier, H. L.; Fayer, M. D. *J. Phys. Chem. B* **1997**, *101*, 9352–9361.
- (36) Linder, B. Reaction-Field Techniques and Their Applications to Intermolecular Forces. In *Intermolecular Forces*; Hirschfelder, J. O., Ed.; Advances in Chemical Physics 12; Interscience Publishers: New York, 1967; Chapter 5, pp 225–281.
- (37) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.
- (38) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.
- (39) Serrano-Andrés, L.; Merchán, M.; Roos, B. O.; Lindh, R. *J. Am. Chem. Soc.* **1995**, *117*, 3189–3204.
- (40) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

# JCTC

Journal of Chemical Theory and Computation

## Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations

David L. Mobley,<sup>\*,†</sup> Christopher I. Bayly,<sup>§</sup> Matthew D. Cooper,<sup>§</sup> Michael R. Shirts,<sup>||</sup>  
and Ken A. Dill<sup>‡</sup>

*Department of Chemistry, University of New Orleans, New Orleans, Louisiana 70148  
Department of Pharmaceutical Chemistry, University of California at San Francisco,  
San Francisco, California 94158 Merck-Frosst Canada Ltd., 16711 TransCanada  
Highway, Kirkland, Quebec, Canada H9H 3L1, and Department of Chemical Engineering,  
University of Virginia, P.O. Box 400741, Charlottesville, Virginia 22904-4741*

Received September 30, 2008

**Abstract:** Using molecular dynamics free energy simulations with TIP3P explicit solvent, we compute the hydration free energies of 504 neutral small organic molecules and compare them to experiments. We find, first, good general agreement between the simulations and the experiments, with an rms error of 1.24 kcal/mol over the whole set (i.e., about 2 kT) and a correlation coefficient of 0.89. Second, we use an automated procedure to identify systematic errors for some classes of compounds and suggest some improvements to the force field. We find that alkyne hydration free energies are particularly poorly predicted due to problems with a Lennard-Jones well depth and find that an alternate choice for this well depth largely rectifies the situation. Third, we study the nonpolar component of hydration free energies—that is, the part that is not due to electrostatics. While we find that repulsive and attractive components of the nonpolar part both scale roughly with surface area (or volume) of the solute, the total nonpolar free energy does not scale with the solute surface area or volume, because it is a small difference between large components and is dominated by the deviations from the trend. While the methods used here are not new, this is a more extensive test than previous explicit solvent studies, and the size of the test set allows identification of systematic problems with force field parameters for particular classes of compounds. We believe that the computed free energies and components will be valuable to others in the future development of force fields and solvation models.

### I. Introduction

Aqueous solvation (hydration) of molecules is important for much of chemistry and biochemistry. Many experimental hydration free energies are available, providing a wonderful opportunity for testing force fields and computational treatments of solvation.

There have been a number of extensive tests of hydration free energies computed using continuum representations of water and static solute conformations.<sup>1–4</sup> One recent study extended this by sampling ensembles of solute conformations using classical molecular dynamics and using these to compute hydration free energies.<sup>5</sup> Continuum representations of solvent, however, have known limitations,<sup>6,7</sup> and explicit treatment of solvent provides a “gold standard” for molecular simulations. Early explicit solvent hydration free energy studies were limited by computational cost to a few compounds and, more recently, by the availability of

\* Corresponding author e-mail: dmobley@gmail.com.

† University of New Orleans.

§ Merck-Frosst Canada Ltd.

|| University of Virginia.

‡ University of California at San Francisco.

parameters for small molecules. Thus a variety of studies have looked at hydration free energies of amino acid side chain analogs in explicit solvent (for example, refs 8–11), but few have studied a more diverse set.

With recent computational and methodological developments, both of these hurdles—computational cost and parameters—are now at least partially surmountable. Hydration free energy calculations can now be conducted more efficiently,<sup>8,12</sup> and computers are faster. Recent developments also make possible semiautomatic parametrization of small molecules, using general small molecule force fields like the general Amber force field (GAFF)<sup>13</sup> and parameter assignment tools like Antechamber.<sup>14</sup> Thus, two recent studies have examined hydration free energies of a total of roughly 60 small molecules in explicit solvent.<sup>4,12</sup>

Here, we perform a much more extensive test of explicit solvent modeling, on a test set of 504 molecules previously used for implicit solvent hydration free energy calculations<sup>5</sup>—more than 10 times larger than the largest previous explicit solvent tests.<sup>12</sup> Because this test is so extensive, we believe it provides a good benchmark for the best results that can currently be expected from molecular dynamics models of hydration. We also hope that others will find this compilation of computational and experimental results useful for analysis and force field parametrization efforts.

## II. Simulation Methods

**A. General Simulation Parameters.** In this work, we use alchemical free energy calculations to compute hydration free energies in explicit solvent for 504 small molecules, using the compound set from a previous implicit solvent study.<sup>5</sup> Simulation protocols were similar to those used in previous explicit solvent studies.<sup>4,12</sup> Hydration free energies were computed using the Bennett acceptance ratio (BAR).<sup>15</sup> A brief summary of the methods follows, and we note the deviations from the previous studies.<sup>4,12</sup>

Here, starting molecular conformations were the same as those for the previous implicit solvent study,<sup>5</sup> except that here a single starting conformation was used for each molecule (rather than 5) due to computational limitations relating to the size of the set. Simulations were performed in GROMACS 3.3.1<sup>16,17</sup> using the GAFF small molecule parameters<sup>13</sup> as assigned by Antechamber<sup>14</sup> (as in the implicit solvent study).<sup>5</sup> Here, AM1-BCC<sup>18,19</sup> partial charges were assigned using the Merck-Frosst implementation of AM1-BCC.

This data set contains several nitro-containing compounds which did not have improper torsions for the nitro-ring system in the GAFF parameter set, specifically improper torsions for GAFF types ca-o-no-o and c3-o-no-o. We added these using generic GAFF values (that is, the values used for the majority of the improper torsions in GAFF)—a barrier height of 2.2 kcal/mol, a phase shift of 180°, and a periodicity of 2.

After setup in Antechamber and Leap, small molecule parameters were converted to GROMACS topology and coordinate files using a Perl conversion script developed

previously.<sup>20</sup> Small molecules were then solvated using GROMACS utilities in a dodecahedral water box with at least 1.2 nm from the solute to the nearest box edge using the TIP3P model of water.<sup>21</sup> The number of water molecules varied depending on the solute size. Simulations were performed separately at a variety of different alchemical intermediate  $\lambda$  values, with the number of  $\lambda$  values and the amount of equilibration as described previously.<sup>12</sup> Production simulations were 5 ns in length at each  $\lambda$  value, and free energies and uncertainties were computed as described previously.<sup>4,12</sup> Uncertainties were computed using the block bootstrap procedure described previously. Cutoffs and simulation parameters were as described previously except that the real-space electrostatic cutoff was 10 Å rather than 9 Å.

We computed the electrostatic and nonpolar components of solvation. The electrostatic component was computed as the free energy of turning on the solute partial charges in water, less the free energy of the same transformation in vacuum. The nonpolar component was the free energy of turning on the Lennard-Jones interactions between the uncharged solute and water, as in previous studies.<sup>4,12</sup> Alternative definitions of the nonpolar component are possible.<sup>44</sup>

**B. Analysis of the Nonpolar Component.** In implicit solvent models, the nonpolar component of solvation is often assumed to correlate with the surface area and/or the volume based on theoretical arguments relating to cavity creation cost.<sup>22–26</sup> To explore this we computed the solvent accessible surface area and volume for all of the solutes considered here using GROMACS tool `g_sas` with a probe radius of 1.4 nm.

We also further dissected the nonpolar part (due the Lennard-Jones interactions) into repulsive and attractive components using the Weeks-Chandler-Andersen (WCA) separation.<sup>27</sup> To do this, we implemented the WCA separation in a modified version of GROMACS 3.3.1.<sup>45</sup>

In our main study, we simply computed the total nonpolar component and retained the trajectories. The attractive component for each solute was then obtained by applying the WCA separation to stored trajectories of the fully interacting solute and reprocessing these simulations with the attractive interactions turned off to re-evaluate the energies. We computed the free energy for turning off the attractive interactions using exponential averaging (the Zwanzig relation)<sup>28</sup> and standard error analysis. This assumes that phase-space overlap is good between the ensemble where the solute has attractive interactions with water and that where it does not. Error analysis should tell us if this is not the case. We further tested this by recomputing the attractive contribution using simulations at a series of separate  $\lambda$  values (where  $\lambda$  modifies only the attractive interactions) for selected solutes (phenol, *p*-xylene, pyridine, and toluene) and found that computed free energies were within uncertainty of the values computed using exponential averaging, indicating overlap was sufficient.

With these attractive components, we then obtained repulsive components by subtracting the attractive component from the total nonpolar component. This probably results in slightly larger uncertainties in computed repulsive compo-

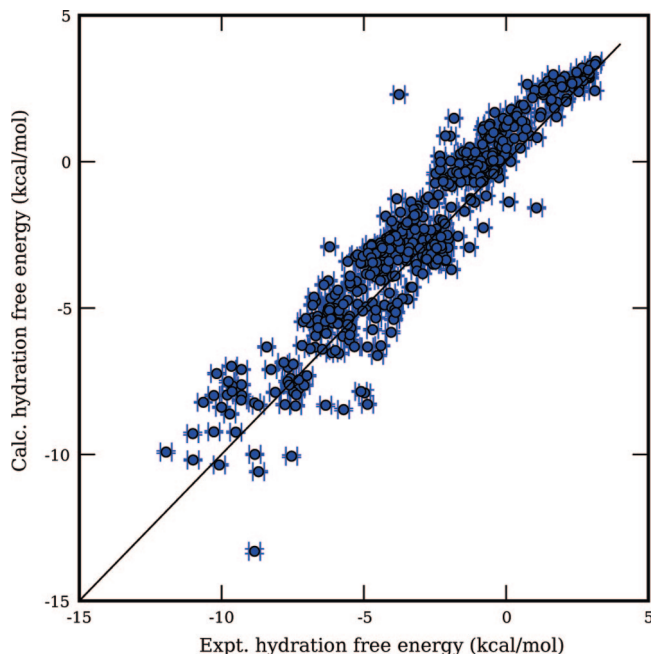
nents than would have resulted from computing the repulsive component separately, but it also saves a large amount of computer time since we had already computed the total nonpolar component, and the repulsive portion of the calculation is the most difficult to converge.

**C. Identification of Systematic Errors.** Some functional groups may lead to systematic errors, resulting in errors which are larger for some types of molecules than for other types. Alternatively, there might be no systematic errors. We seek an approach to easily identify systematic errors and prioritize functional groups which have the largest errors.

We make a list of compounds and sort it by rms error, from largest error to smallest error. Following a method that is often used to determine *enrichment factors* for drug discovery, we look at the cumulative distribution function (CDF) for each functional group—the probability of compounds with that functional group having a ranked rms error up to rank  $x$ . Those functional groups that are systematically wrong will tend to cluster at high rms error and will result in a rapid rise in the CDF versus  $x$ . This can be assessed easily by computing the area under the CDF, biased by a weighting function to give the most weight to high rms errors. Here, we do this using the recently developed BEDROC metric,<sup>29</sup> which evaluates the integral of the CDF multiplied by an exponentially decaying weighting factor and then rescales this to run from 0 to 1. Chemical groups which occur most often in compounds with high rms errors will have larger BEDROC values, while chemical groups which have more random errors will have smaller BEDROC values (the expected BEDROC value for a uniform distribution can be computed analytically).<sup>29</sup> Chemical groups that only occur in compounds with low rms errors have the smallest BEDROC values. In Section III we report BEDROC values for a variety of chemical groups and atom types. Uncertainties were computed using the standard deviation of the mean for 40 iterations of a bootstrap procedure where BEDROC values for each chemical group are recomputed using a new list of compounds made up of a random selection of compounds from the original list.

Here, BEDROC values were computed using a weighting factor of  $\alpha = 1.0$ . This value was obtained empirically by experimenting with different  $\alpha$  values to see what gave the best ability to recognize functional groups which differ substantially from random. If  $\alpha$  is too large, the weighting is too strong, and only compounds at the very highest rms errors matter. If  $\alpha$  is too small, making BEDROC equivalent to the ROC metric, the weighting of the early part of the curve is too weak, also apparently reducing the ability to recognize systematic errors.  $\alpha = 1.0$  was a good compromise.

To avoid having to assign functional groups to all of the compounds in the test set by hand, we used the program Checkmol,<sup>30</sup> which automatically assigns chemical groups to molecules. We used MDL molfiles generated by OpenEye's OEChem toolkit as input. This resulted in an extremely large set of chemical groups, so we retained only those chemical groups which occurred in at least 5 molecules. We also combined some small groups. For example, we made a single group of amines, containing all types of amines. We also did the same for amides, ethers, esters, thiols, acids, and



**Figure 1.** Calculated hydration free energies versus experiment. Shown are the calculated hydration free energies versus experiment for the full test set. The diagonal line is  $x = y$ . Vertical error bars denote computed uncertainties, and horizontal error bars are a conservative estimate.

alcohols. We also manually created a “hypervalent S” group and included the appropriate compounds in this group. The resulting list of molecules assigned to chemical groups was used to generate BEDROC values for these chemical groups.

We also tried using Student's  $t$ -test to look for systematic errors to supplement the BEDROC approach. We used our own implementations of the  $t$ -test and SciPy's implementation of the incomplete beta function for computing the significance. Results from this are discussed below.

### III. Results and Discussion

**A. The Mean Error Relative to Experiment Is Less than 1 kcal/mol.** Here we evaluate the agreement between computed hydration free energies and the experimental values for the full test set. A previous study on the same 504 small-molecule test set compared the accuracy of several different implicit solvent models<sup>5</sup> using molecular dynamics free energy calculations. rms errors ranged from  $2.014 \pm 0.008$  kcal/mol to  $2.433 \pm 0.002$  kcal/mol depending on the implicit solvent model, with correlation coefficients ( $r^2$ ) from  $0.685 \pm 0.001$  to  $0.774 \pm 0.001$ . In all four solvent models tested, the computed hydration free energies were systematically too negative relative to experiments (the solutes preferred the water phase too much in the simulations), so the mean error was negative ( $-0.65 \pm 0.09$  to  $-1.1 \pm 0.1$ ).

Here, using explicit TIP3P water, we find an rms error of  $1.26 \pm 0.01$  kcal/mol, with a correlation coefficient of  $0.889 \pm 0.006$  and a mean error of  $0.676 \pm 0.002$  (Figure 1). Hence, on average, explicit solvent simulations give significantly better agreement with experiments than our earlier implicit solvent study, consistent with an earlier comparison.<sup>4</sup> Interestingly, the systematic errors of explicit and implicit

solvent studies are in opposite directions. In explicit solvent, the hydration free energies here are systematically too positive. These differences are likely due to the solvent models rather than the force field parameters, since the solute parameters are very similar in the two cases. Systematic errors in other explicit solvent models tended to be in the same direction as the explicit solvent deviation here,<sup>8</sup> so perhaps limitations of the water model are playing a role. Another potential source of such differences is the neglect, in implicit solvent models, of asymmetries in the response of water to solutes of different polarities.<sup>7</sup> Another origin of differences could be the nonpolar term in the implicit models. That is, the term  $\gamma \times A$  (where  $A$  is the surface area) in implicit solvent models involves an adjustable parameter which can change the errors. A table of the full results from this study is available in the Supporting Information.

**B. Improving the Alkyne Lennard-Jones Parameters and Identifying Other Systematic Errors.** Are there systematic errors in the force field parameters for molecules in our test set? We found that the computed hydration free energies for alkynes were systematically much too positive (Figure 1 and the Supporting Information). There were six alkynes in the set, and the mean error was  $1.92 \pm 0.21$  kcal/mol. All of the computed hydration free energies were actually around 2 kcal/mol, while experimental values are close to zero. For all of the alkynes, the electrostatic component of hydration is quite small ( $-0.8$  to  $-0.9$  kcal/mol), since these molecules are largely nonpolar. We reasoned that errors in alkyne parameters are thus not likely to be in the electrostatic terms, nor are the errors expected to come from the bonded parameters (bond stretching, angle bending, etc.), which should affect hydration free energies only weakly. Hence, we focused on the alkyne Lennard-Jones parameters. In GAFF, the alkyne carbon Lennard-Jones parameters are identical to those for all carbons except selected sp<sup>2</sup> carbons (the 'c2' atom type) and are taken directly from comparable carbons in older AMBER force fields.<sup>13</sup> We were particularly concerned about the parameters for the GAFF "c1" atom type, for the triple bonded carbons in alkynes. These apparently originated with the work of Howard et al., where they "were obtained by analogy to the Weiner et al. and Cornell et al. force fields".<sup>31</sup> In that work, those Lennard-Jones parameters were taken to be the same as for the other carbons.

Many AMBER Lennard-Jones parameters were originally taken from the OPLS force field, so we examined the OPLS choices for triple bonded carbons. It turned out that OPLS uses several different atom types for alkyne carbons, originating from simulations of linear and substituted alkynes,<sup>32-34</sup> and some of these have much stronger dispersion interactions than those for the GAFF c1 type, which is intuitively reasonable. It seemed likely that missing dispersion interactions could account for at least part of the error we were seeing for alkynes, thus we examined modifying the Lennard-Jones well-depth for alkynes in GAFF.

We sought to avoid adding additional atom types to GAFF, but OPLS has several different carbon well-depths for alkynes, depending on whether the carbon is terminal ( $\epsilon = 0.086$  kcal/mol), nonterminal with an attached atom having

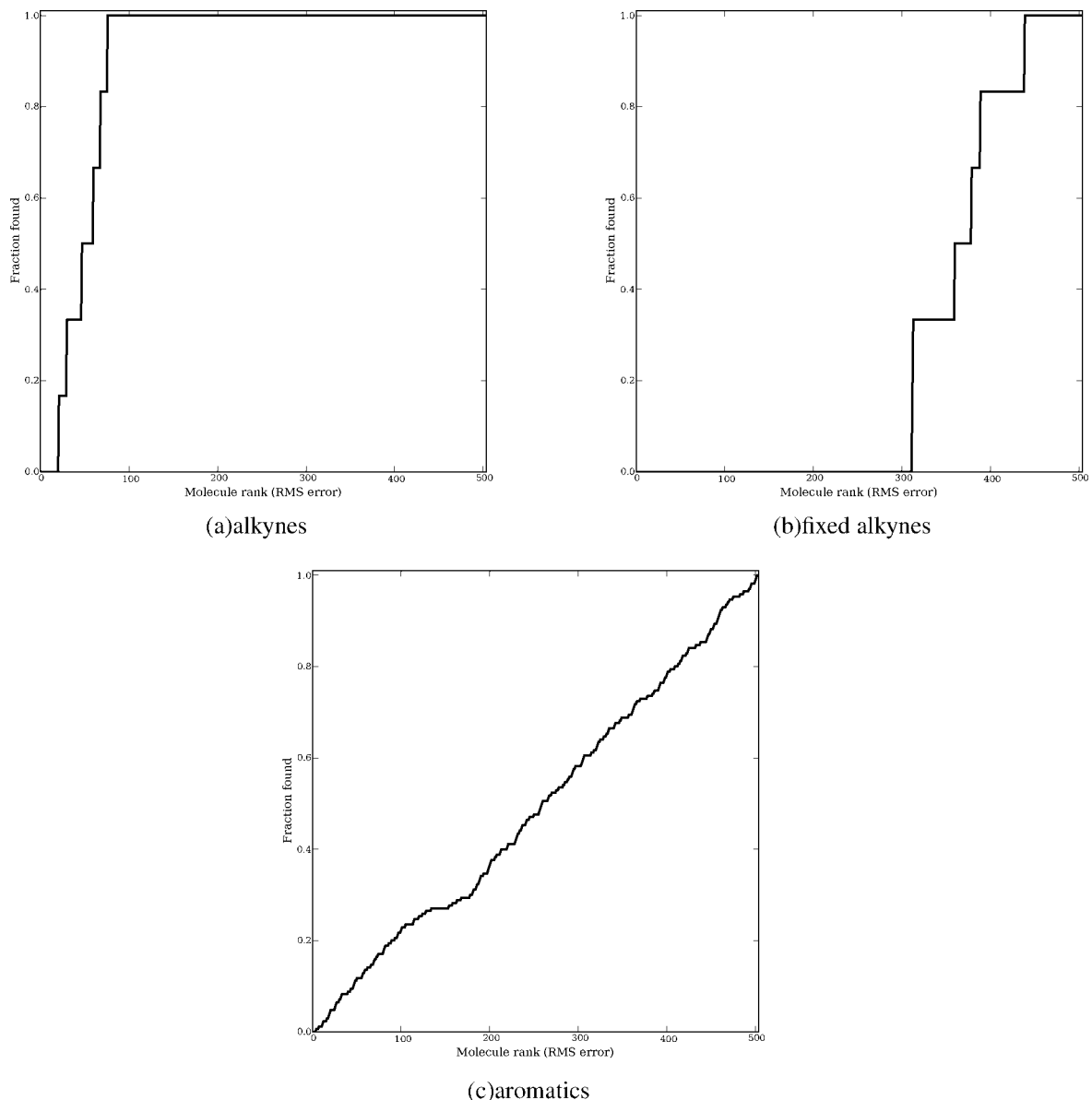
two or three hydrogens ( $\epsilon = 0.210$  kcal/mol), nonterminal with an attached atom having one hydrogen ( $\epsilon = 0.135$  kcal/mol), or nonterminal with an attached phenyl or other atom having no hydrogens ( $\epsilon = 0.100$  kcal/mol).<sup>32-34</sup> To avoid adding additional atom types to GAFF, we needed to pick just one of these, so we chose the one which gave the most accurate hydration free energies when used for all alkyne triple bonded carbons. This was  $\epsilon = 0.210$  kcal/mol. The original GAFF well depth was  $\epsilon = 0.086$  kcal/mol.

Using this new  $\epsilon$  value for triple-bonded carbons, the computed hydration free energies for alkynes are much closer to zero (although still slightly positive); now the mean error is  $0.49 \pm 0.07$  kcal/mol, down from  $1.92 \pm 0.21$  kcal/mol initially. Increasing the well depth further could reduce this somewhat more, but this might cause other inconsistencies within the force field. Nevertheless, the systematic error here on alkynes is compelling, and we recommend that future GAFF studies use a well depth of  $\epsilon = 0.210$  kcal/mol for triple bonded carbons (GAFF types c1, cg, and ch).<sup>46</sup>

The alkynes also provide an example of how the BEDROC metric works for identifying systematic errors. Before the adjustment of the well depth for alkynes, the BEDROC value (with  $\alpha = 1$ ) for alkynes was  $0.90 \pm 0.02$  (compared to 0.49 for a random distribution with this  $\alpha$ ),<sup>47</sup> indicating that alkynes were systematically wrong. After the fix, the BEDROC value was  $0.26 \pm 0.05$ , indicating that alkynes now actually are considerably better than other typical compounds (Figure 2). Although our correction of  $\epsilon$  was done without regard for the carbonitriles, the change results in a decrease in BEDROC for the carbonitriles from  $0.86 \pm 0.05$  to  $0.73 \pm 0.06$  (compared to 0.49 for uniform). So carbonitriles are now improved too but still have substantial systematic errors. With this change, the overall rms error decreases slightly to  $1.24 \pm 0.01$  kcal/mol, and the correlation coefficient remains essentially the same ( $0.891 \pm 0.006$ ). In all that follows we report values computed with the new well depth.

We believe that the approach utilized here (looking for compounds that are over-represented at the highest rms errors) is a general and useful strategy for identifying systematic flaws in the energy functions used for molecular modeling simulations and prioritize reparameterization efforts. Functional groups which tend to cause significant errors should occur frequently at the high-rms error end of the set, while functional groups which are not necessarily linked to the errors should be roughly randomly distributed over the test set. For example, one would intuitively expect that whether a compound is aromatic or not will have little to do with whether it is systematically mispredicted. Indeed, aromatic compounds have a BEDROC value of  $0.48 \pm 0.03$ , roughly randomly distributed (Figure 2). BEDROC values by functional group for the set are shown in Table 1. These BEDROC values show that cyclic hydrocarbons, alkynes (with the fix), alkanes, aldehydes, and ketones are now particularly well predicted. On the other hand, there appear to be systematic errors for alcohols, alkyl bromides, and carbonitriles.

We also tried another approach for identifying systematic errors involving Student's  $t$ -test, which compares the means



**Figure 2.** CDFs for selected functional groups versus error. Shown are cumulative distribution functions for finding compounds with particular functional groups at a given ranked error. Compounds found far to the left have very large errors; compounds far to the right have very small errors. An ideal random distribution of errors would give rise to a linear rise in the CDF. CDFs are shown for (a) alkynes before fixing the Lennard-Jones well-depth, (b) alkynes after fixing the Lennard-Jones well-depth, and (c) aromatics.

of two distributions and provides a measure of the significance of any difference in the means. We applied this approach in two different ways:

(1) We compared the mean experimental value for each functional group with the mean calculated value for each functional group (Supporting Information, Table 5). This proved not to be particularly useful, as these means are significantly different for almost every functional group. This is not surprising given the fact that the mean error across the entire test set is  $0.676 \pm 0.002$ , so most computed values (in all functional groups) are too positive. This does show that results could be improved across the entire set by addressing this systematic offset, but it does not provide any insight into which functional groups are particularly problematic.

(2) We compared the error for the compounds in each functional group with the error for the entire set (Table 2). This shows which functional groups have a significantly *different* performance than the overall set, though this performance could be better or worse. We also show the mean error for each functional group in Table 2; functional groups with mean errors around 0.676 kcal/mol are typical, while those with larger mean errors are worse than average, and those with smaller mean errors are better than average. The *t*-test tells us which of these differences are significant, and many are. This appears to be a useful analysis that complements the BEDROC analysis. The advantage of the BEDROC analysis is that it tells us which functional groups have the worst errors, while this analysis can tell us which functional groups have the most significant errors.



**Table 1.** BEDROC Values by Functional Group for the Different Functional Groups Represented in the Test Set, Compared to What Would Be Expected for the Same Number of Compounds Distributed Randomly Across the Test Set<sup>a</sup>

functional group	number	BEDROC
acid	73	0.48 ± 0.03
alcohol	38	0.76 ± 0.03
aldehyde	20	0.22 ± 0.04
alkanes	28	0.16 ± 0.03
alkene	35	0.55 ± 0.04
alkyl bromide	17	0.72 ± 0.08
alkyl chloride	31	0.61 ± 0.05
alkyl iodide	9	0.44 ± 0.06
alkyne	6	0.26 ± 0.04
amine	44	0.47 ± 0.04
aromatic compound	170	0.48 ± 0.03
aryl chloride	20	0.54 ± 0.05
carbonitrile	12	0.73 ± 0.07
cyclic hydrocarbon	8	0.14 ± 0.03
ester	8	0.46 ± 0.11
ether	42	0.60 ± 0.04
halogen derivative	22	0.58 ± 0.07
heterocyclic compound	48	0.60 ± 0.04
hypervalent S	5	0.62 ± 0.20
ketone	25	0.26 ± 0.06
nitro compound	17	0.63 ± 0.08
other	29	0.62 ± 0.06
phenol or hydroxyhetarene	33	0.60 ± 0.05
thiol	5	0.46 ± 0.04

<sup>a</sup> Functional groups with high BEDROC values (relative to the value for random, roughly 0.5 here) are overrepresented in compounds with high RMS errors.

The study done here uses one particular charge model. Charge model may affect which compounds are particularly poorly predicted, though in two recent tests, the compounds which were poorly predicted tended to be poorly predicted by most charge models.<sup>4,35</sup> Still, our analysis here does not in general point to a specific source of error. Errors may be due to the charge model, Lennard-Jones parameters, or bonded parameters, some combination, or even due to the water model. In the case of the alkynes, we can be fairly confident that the source of error is the Lennard-Jones parameters for the reasons noted above. But for the other cases noted here, further work will be required to determine the source of error.

**C. The Total Nonpolar Component Does Not Correlate with Surface Area or Volume.** We examined the nonpolar components (the nonelectrostatic component of the hydration free energy) for our data set. The total nonpolar contribution to the solvation free energy is typically assumed to correlate with surface area or volume in implicit solvent models. Yet we find that there is essentially no correlation. Plots of nonpolar components versus surface area and volume are shown in Figure 3. The correlation of the nonpolar component with surface area is  $r^2 = 0.019 \pm 0.001$ , and that with volume is  $r^2 = 0.011 \pm 0.001$ . The molecules in this test set are small enough that surface area and volume are highly correlated ( $r^2 = 0.991 \pm 0.001$ ).

We further dissect the nonpolar component using the WCA separation of the Lennard-Jones potential energy (and thus the nonpolar component) into attractive and repulsive parts. The potential is split based on the sign of the force, as discussed in the Methods section. We find that both the

**Table 2.** Statistics from Applying Student's *t*-Test to the Difference between the Calculated and Experimental Means by Functional Group<sup>a</sup>

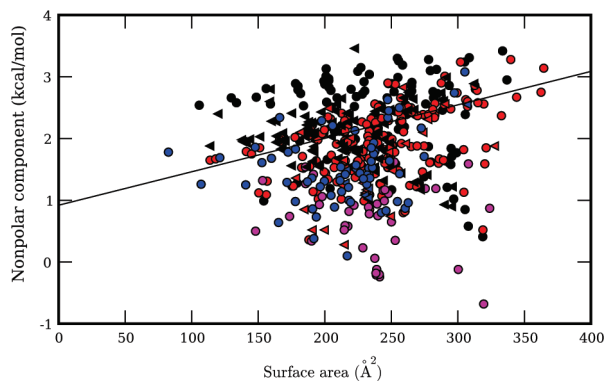
functional group	number	<i>t</i> -value	significance	mean error
acid	73	-7.43	4e-13	-0.34
alcohol	38	3.62	0.0003	1.29
aldehyde	20	-3.04	0.003	-0.07
alkanes	28	-1.69	0.09	0.31
alkene	35	2.34	0.02	1.07
alkyl bromide	17	3.31	0.001	1.50
alkyl chloride	31	2.31	0.02	1.09
alkyl iodide	9	0.59	0.6	0.86
alkyne	6	-0.38	0.7	0.49
amine	44	-0.65	0.5	0.55
aromatic compound	170	-1.05	0.3	0.55
aryl chloride	20	1.65	0.1	1.04
carbonitrile	12	3.22	0.001	1.63
cyclic hydrocarbon	8	-1.18	0.2	0.21
ester	8	-1.69	0.09	0.02
ether	42	2.18	0.03	1.01
halogen derivative	22	0.32	0.8	0.73
heterocyclic compound	48	2.38	0.02	1.02
hypervalent S	5	-4.55	7e-06	-1.50
ketone	25	-2.77	0.006	0.05
nitro compound	17	1.86	0.06	1.13
other	29	-0.48	0.6	0.55
phenol or hydroxyhetarene	33	2.72	0.007	1.16
thiol	5	0.51	0.6	0.89

<sup>a</sup> Shown are the number of compounds in each functional group, the calculated *t* value, the computed significance (probability that *t* could be this large or larger by chance), and the mean error for this group (in kcal/mol). The overall mean error is  $0.676 \pm 0.002$  kcal/mol, so groups with mean errors smaller than this may be significantly better than average (until the mean error becomes negative), while those with mean errors larger than this may be significantly worse.

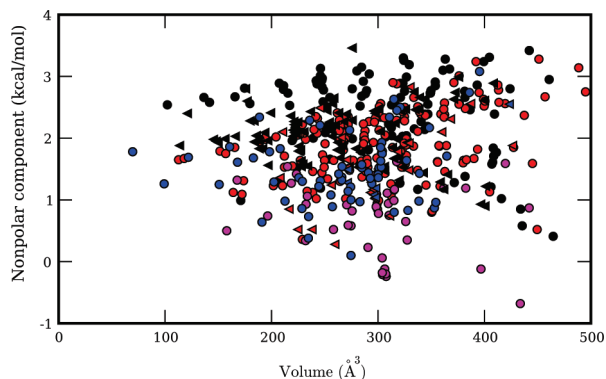
attractive and repulsive components individually correlate strongly with surface area and volume (repulsive:  $r^2 = 0.964 \pm 0.002$  with surface area,  $r^2 = 0.952 \pm 0.002$  with volume; attractive:  $r^2 = 0.945 \pm 0.002$  with surface area,  $r^2 = 0.946 \pm 0.002$  with volume; Figure 4), and it is only the total (the small difference of the two large individual components) that does not correlate well with surface area or volume. This is in accord with previous work on a smaller set of compounds.<sup>36</sup> Essentially, the total nonpolar component is the sum of two anticorrelated quantities, and so the total ends up being dominated by the scatter in these quantities. It is interesting to note that the minimum in the Lennard-Jones potential is precisely where these two forces, the attractive and repulsive components, are very well balanced, so it is perhaps not surprising that the attractive and repulsive components correlate so well.

The observed poor correlation, and the importance of attractive interactions, is consistent with several previous studies which have found that the nonpolar component of solvation does not correlate well with surface area.<sup>36-39</sup>

Why is the correlation with surface area so poor? In Figure 3, it is apparent that compounds containing only carbon and hydrogen have a nonpolar component that is less favorable to solvation than molecules of an equal size which additionally contain nitrogen and/or oxygen. The likely reason for this is that nitrogen and oxygen atoms tend to have stronger attractive dispersion interactions with their environment than



(a) Nonpolar component versus surface area



(b) Nonpolar component versus volume

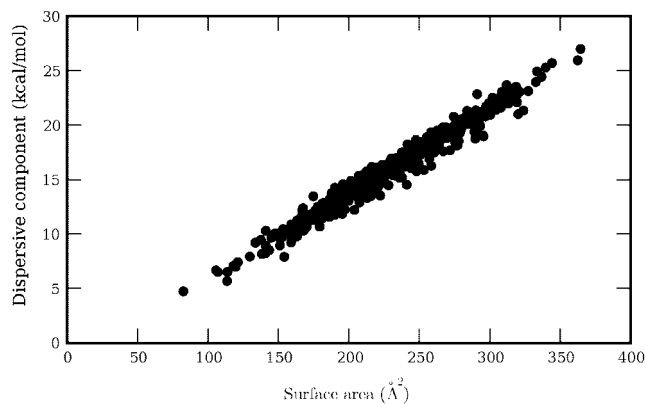
**Figure 3.** Nonpolar components versus solvent accessible surface area and volume. Shown are the calculated nonpolar component of the hydration free energies versus solvent accessible surface area and volume for the compounds in the set. Carbon and hydrogen containing compounds are black, those with oxygen additionally are red, those with nitrogen additionally are blue, and those with nitrogen and oxygen both are magenta. Compounds with diamond symbols contain other elements in addition to C, H, N, and O. In the surface area plot, the line is a typical implicit solvent nonpolar component estimate of  $G_{np} = (0.00542 \cdot SA + 0.92)$  kcal/mol<sup>1</sup>.

do carbon and hydrogen. Several other studies have noted that dispersion interactions play an important role in nonpolar solvation.<sup>36–39</sup> Even interior solute atoms contribute to these attractive interactions in an important way.<sup>40</sup> Other factors may also contribute to the poor correlation with surface area. For example, geometric effects may play an important role as well.

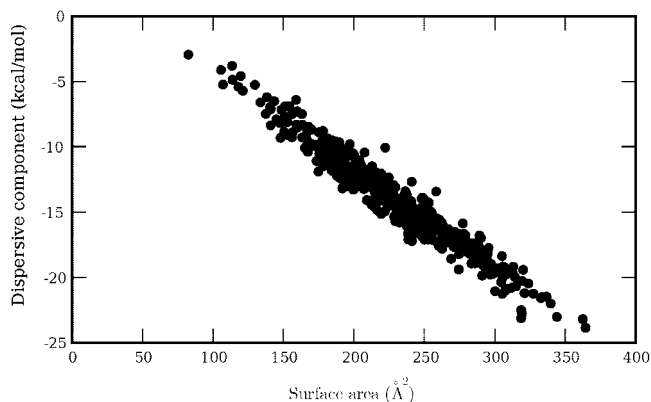
Overall, our results strongly support the growing consensus that implicit solvent models should move beyond the simple surface area model for treatment of the nonpolar component, perhaps at least to include a treatment of dispersion interactions. A number of alternate models have already been proposed.<sup>36,38,39,41,42</sup>

#### IV. Conclusions

We used molecular dynamics simulations in explicit TIP3P water to compute the hydration free energies for a set of 504 neutral compounds. We compared the results with experimental data in the most extensive such test in explicit solvent to date. We find a good correlation ( $r^2$  of 0.891 ±



(a) Repulsive part of nonpolar component versus surface area



(b) Attractive part of nonpolar component versus surface area

**Figure 4.** Repulsive and attractive parts of the nonpolar component versus surface area. Shown are the repulsive (a) and attractive (b) parts of the nonpolar component, as calculated using the WCA separation, plotted versus the solvent accessible surface area for solutes in the test set. Similar plots comparing the repulsive and attractive components to volume are given in the Supporting Information.

0.06) and an rms error of  $1.24 \pm 0.01$  kcal/mol or roughly 2 kT. We believe this is representative of the accuracy that can be expected from the best current physical models for hydration free energies. It may be possible to develop new models which can do somewhat better, though we expect that it may be very hard to increase accuracies past 1 kT. A key finding is that these explicit solvent free energies are considerably more accurate than the corresponding implicit solvent values for the same data set.

At the same time, many of the molecules in this test set are relatively small and simple compared to typical druglike molecules, which may be highly polyfunctional. Recent work suggests that overall performance of the approach applied here may be significantly worse in tests where the compounds involved are more polar and polyfunctional.<sup>4,35</sup> This may suggest we need much more hydration free energy data on more polyfunctional, druglike molecules in order to refine our force fields.

Here, we also propose a way to identify systematic errors in force field parameters for particular functional groups. We do this using the BEDROC method.<sup>29</sup> Using this approach,

we were able to fix a systematic problem with alkyne Lennard-Jones parameters. We also identified several other classes of compounds which appear to have systematic errors, and for which further force field development should be done. Having a method to systematically identify problematic compound classes provides good opportunities for force field improvements.

In addition, we studied the nonpolar component of the hydration free energy for the compounds in the test set. We find that while the large repulsion and attraction terms both correlate well with the size (area or volume) of the solute, the total nonpolar component, which is a small difference between these two quantities, does not. This strongly suggests that implicit solvent models need to move away from treating the nonpolar component as simply dependent on the surface area. The data additionally suggest that new models must address the nonlinear behavior arising from the delicate balance of repulsive and attractive nonpolar terms. Furthermore, implicit solvent models that have been parametrized to match experimental hydration free energies using a simple surface area-based nonpolar term may need to be reparameterized.

Here, the real value of this study is not the methods presented—the methods were used in previous work. Rather, it is the extensive nature of the test, which provides the opportunity to actually identify systematic errors in the force field descriptions of particular functional groups. It also provides guidance into what compounds are likely to be particularly challenging to study computationally with current force fields.

Because we believe the real value of this study is these results, we have deposited the full set of computed free energies, components, starting molecular structures, and parameters for this work in the Supporting Information. We hope that others find this experimental data set and the computational results to be useful in future studies of solvation and for force field development.

**Acknowledgment.** We thank John D. Chodera (Stanford University) for helpful discussions. We appreciate the support of NIH grant GM 63592 to K.A.D.

**Supporting Information Available:** Coordinate files (mol2) with AM1-BCC partial charges for the small molecules in the test set used here; computed hydration free energies, electrostatic and nonpolar components, and the experimental values; AMBER parameter and coordinate files for all of the molecules in the test set; plots of attractive and repulsive components versus solute volume; a table mapping the names used for the files to IUPAC names; a table of computed solvent accessible surface area and volume for each solute; and results from Student's *t*-test comparing the mean experimental and calculated values for each functional group. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.
- (2) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2002**, *106*, 11009–11015.
- (3) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532–6542.
- (4) 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*, 769–778.
- (5) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2008**, *111*, 938–946.
- (6) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.
- (7) Mobley, D. L.; Barber, A. E., II; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.
- (8) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
- (9) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.
- (10) Deng, Y.; Roux, B. *J. Chem. Phys.* **2004**, *108*, 16567–16576.
- (11) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.
- (12) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- (13) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (14) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *26*, 247260.
- (15) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (16) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (17) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
- (18) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.
- (19) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (20) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902.
- (21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (22) Stillinger, F. H. *J. Solution Chem.* **1973**, *2*, 141–158.
- (23) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.
- (24) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.
- (25) Chothia, C. *Nature* **1974**, *248*, 338.
- (26) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2925.
- (27) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237–5247.
- (28) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (29) Truchon, J.-F.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (30) Haider, N. Checkmol. <http://merian.pch.univie.ac.at/nhaider/cheminf/cmmm.html> (accessed July 20, 2007).
- (31) Howard, A. E.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 243–261.
- (32) Jorgensen, W. L., personal communication, 2007.
- (33) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

- (34) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.
- (35) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *J. Phys. Chem. B* Accepted for publication.
- (36) Tan, C.; Tan, Y.-H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.
- (37) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243–9244.
- (38) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.
- (39) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.
- (40) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.
- (41) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.
- (42) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.
- (43) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- (44) Here, the alchemical pathway used turns off all solute partial charges, meaning that the nonpolar component is calculated without solute intramolecular electrostatic interactions. An alternative pathway<sup>8,43</sup> involves turning off only intermolecular electrostatic interactions between the solute and its environment, while maintaining intramolecular electrostatic interactions. While the two pathways must give equivalent hydration free energies, the breakdown into electrostatic and nonpolar components will be slightly different, since the conformational ensemble sampled during the nonpolar component of the calculation will be altered by the presence (or lack thereof) of intramolecular electrostatic interactions.
- (45) Other separation schemes and probe radii are possible, as are other surface definitions, but our main conclusions here should not depend significantly on these factors, as suggested by the work of Tan et al.<sup>36</sup>
- (46) We do not believe it is necessary to perform additional tests to examine other alkyne properties (such as pure liquid properties) before making this recommendation for two reasons. First, the AMBER force field does not typically use these properties to inform the parameterization process, so including them would be a deviation from AMBER parameterization strategies. Second, the AMBER force field and GAFF claim (in the force field files) to use the OPLS Lennard-Jones parameters for alkynes. The suggested modification simply makes this claim true and brings AMBER/GAFF back into conformity with OPLS.
- (47) BEDROC values for a random distribution actually depend on the number of compounds being considered relative to the total. But here the BEDROC value for the random distribution is 0.49 for all of the sizes of our chemical groups except for aromatics, where it is 0.50. To simplify our tables, then, we simply compare all BEDROC values to 0.49.

CT800409D

# JCTC

Journal of Chemical Theory and Computation

## Incorporating Phase-Dependent Polarizability in Nonadditive Electrostatic Models for Molecular Dynamics Simulations of the Aqueous Liquid–Vapor Interface

Brad A. Bauer, G. Lee Warren, and Sandeep Patel\*

*Department of Chemistry and Biochemistry, University of Delaware, Newark, Delaware 19716*

Received August 5, 2008

**Abstract:** We discuss a new classical water force field that explicitly accounts for differences in polarizability between liquid and vapor phases. The TIP4P-QDP (4-point transferable intermolecular potential with charge-dependent polarizability) force field is a modification of the original TIP4P-FQ fluctuating charge water force field of Rick et al. [*J. Chem. Phys.* 1994, 101, 6141] that self-consistently adjusts its atomic hardness parameters via a scaling function dependent on the *M*-site charge. The electronegativity ( $\chi$ ) parameters are also scaled in order to reproduce condensed-phase dipole moments of comparable magnitude to TIP4P-FQ. TIP4P-QDP is parametrized to reproduce experimental gas-phase and select condensed-phase properties. The TIP4P-QDP water model possesses a gas phase polarizability of 1.40 Å<sup>3</sup> and gas-phase dipole moment of 1.85 Debye, in excellent agreement with experiment and high-level ab initio predictions. The liquid density of TIP4P-QDP is 0.9954 ( $\pm$  0.0002) g/cm<sup>3</sup> at 298 K and 1 atm, and the enthalpy of vaporization is 10.55 ( $\pm$  0.12) kcal/mol. Other condensed-phase properties such as the isobaric heat capacity, isothermal compressibility, and diffusion constant are also calculated within reasonable accuracy of experiment and consistent with predictions of other current state-of-the-art water force fields. The average molecular dipole moment of TIP4P-QDP in the condensed phase is 2.641 ( $\pm$  0.001) Debye, approximately 0.02 Debye higher than TIP4P-FQ and within the range of values currently surmised for the bulk liquid. The dielectric constant,  $\epsilon = 85.8 \pm 1.0$ , is 10% higher than experiment. This is reasoned to be due to the increase in the condensed phase dipole moment over TIP4P-FQ, which estimates  $\epsilon$  remarkably well. Radial distribution functions for TIP4P-QDP and TIP4P-FQ show similar features, with TIP4P-QDP showing slightly reduced peak heights and subtle shifts toward larger distance interactions. Since the greatest effects of the phase-dependent polarizability are anticipated in regions with both liquid and vapor character, interfacial simulations of TIP4P-QDP were performed and compared to TIP4P-FQ, a static polarizability analog. Despite similar features in density profiles such as the position of the GDS and interfacial width, enhanced dipole moments are observed for the TIP4P-QDP interface and onset of the vapor phase. Water orientational profiles show an increased preference (over TIP4P-FQ) in the orientation of the permanent dipole vector of the molecule within the interface; an enhanced *z*-induced dipole moment directly results from this preference. Hydrogen bond formation is lower, on average, in the bulk for TIP4P-QDP than TIP4P-FQ. However, the average number of hydrogen bonds formed by TIP4P-QDP in the interface exceeds that of TIP4P-FQ and observed hydrogen bond networks extend further into the gaseous region. The TIP4P-QDP interfacial potential, calculated to be  $-11.98 (\pm 0.08)$  kcal/mol, is less favorable than that for TIP4P-FQ by approximately 2% as a result of a diminished quadrupole contribution. Surface tension is calculated within a 1.3% reduction from the experimental value. Results reported demonstrate TIP4P-QDP as a model comparable to the popular TIP4P-FQ while accounting for a physical effect neglected by many other classical water models. Further refinements to this model, as well as future applications are discussed.

### I. Introduction

The study of liquid–vapor interfacial systems has enjoyed a rich history of experimental and theoretical investigation.<sup>2–11</sup> Recent advances in experimental methodologies and protocols<sup>12–14</sup> including sum frequency generation (SFG) and second

harmonic generation (SHG) spectroscopies as well as improvements in computational modeling<sup>15,16</sup> continue to elucidate atomically resolved structural, dynamical, and thermodynamic aspects of such systems. Aqueous solution–vapor interfaces, in particular, have generated intense interest due to the importance of such systems in atmospheric, environmental, and biological chemistry.<sup>15–17</sup>

\* Corresponding author. E-mail: sapatel@udel.edu.

Computational approaches to the atomistic modeling of liquid–vapor interfaces, such as molecular dynamics and Monte Carlo techniques, have become viable in recent decades due to advances in computational hardware and improvements in simulation algorithms. Such techniques employ simplified empirical interaction models, or force fields, which are classical models parametrized to properties derived from experiment or first principles calculations on carefully selected training systems.<sup>18</sup> In these models, the electrostatic contribution to the intermolecular interaction potential is described by a Coulombic interaction between molecular charge distributions which are constructed from fixed atomic partial charges or multipole moments placed throughout each molecule.<sup>19–21</sup> Unfortunately, fixed-moment representations of these classical interaction models also entail a number of shortcomings<sup>22–25</sup> which limit overall simulation accuracy. In particular, nonadditive polarization and induction effects are ignored and there is no explicit provision for describing charge transfer effects. Consequently, community interest in polarizable force fields is growing and the development of polarizable models for inorganic ions,<sup>26–31</sup> small molecules,<sup>1,24,26,32–39</sup> and larger biologically relevant macromolecules<sup>40–50</sup> is rapidly increasing in pace even though such models have not yet realized the popularity enjoyed by fixed-charge models. This heightened interest has fostered several different approaches for modeling atomic and molecular polarization including point-dipole (and higher-order multipole) polarizable models,<sup>35,51,52</sup> Drude oscillator models,<sup>30,36,37,53–55</sup> and charge equilibration/fluctuating charge models.<sup>1,24,38,39,41,42,48,56–68</sup>

Polarizable interaction models that incorporate dipole induction effects have already proven to be an indispensable tool for obtaining an accurate theoretical estimation of solution structure and thermodynamics in interfacial systems such as aqueous solutions of inorganic salts.<sup>27,32,34,69–73</sup> The success of such models stems from a dipole induction response that is sensitive to the local electrostatic environment. However, one particular aspect of such models that has not received specific attention is the variation of *molecular polarizability* with phase which recent theoretical investigations have demonstrated is decreased in the condensed phase environment relative to the gas phase. Instead, many current polarizable force fields parametrized specifically for condensed phase environments employ a fixed molecular polarizability which is reduced in magnitude relative to the gas phase value in order to achieve stable dynamics and acceptable condensed-phase properties.<sup>34,36,37,53,64,66,74</sup> While such an approach is perhaps adequate in an isotropic bulk environment, a description based on a fixed, scaled molecular polarizability can be questioned in the presence of anisotropic environments such as the aqueous liquid–vapor interface where the bulk environment transitions to the vacuum over molecular length scales. Indeed, some dipole polarizable models such as the AMOEBA model<sup>73</sup> which employ gas phase polarizabilities also incorporate Thole-type damping at short-range to prevent unstable overpolarization in the condensed phase.

Currently, there are few models that are able to explicitly account for this effect within the context of a molecular dynamics or Monte Carlo simulation and, to our knowledge, none that explicitly consider a dynamically responsive

molecular polarizability. Consequently, our objective in the present work is to present a water force field that effectively allows for the variation of molecular polarizability with phase; more specifically, as will be discussed, molecular polarizability is coupled to variable atomic partial charges which, for the specific aim of modeling the neat water liquid–vapor interface, offers a simple and continuous phase-dependent parameter to which polarizabilities may be coupled.

Section II presents the development of trends necessary for establishing the charge dependent polarizable (TIP4P-QDP) water model (II.A), implications of applied scaling based on these trends within the charge equilibration formalism (II.B), and the details of the condensed phase and liquid–vapor interfacial simulations (II.C). Section III presents the parametrization of this model (III.A), results of the condensed phase (III.B), and liquid–vapor interfacial (III.C) simulations and offers a comparison of the TIP4P-QDP (4-point transferable intermolecular potential with charge-dependent polarizability) model to the original TIP4P-FQ model. We conclude our study with a general discussion and perspectives on future work in Section IV.

## II. Theoretical Methods and Force Fields

**A. Phase-Dependent Polarizabilities.** A variety of recent theoretical investigations involving *ab initio* calculations with polarizable continuum solvent, the partitioning of cluster polarizabilities, and the temperature/density dependence of dielectric constants of fluids reasonably establish that the surrounding condensed phase environment can significantly affect the polarizability of a solvated molecule. Krishtal et al. have previously reported that the average intrinsic polarizability of water molecules decreases as the size of a cluster increases and also as the number and types of hydrogen bonds on a molecule increases.<sup>75</sup> The notion of decreasing polarizability in condensed regions is further supported by the *ab initio* calculations of Morita involving water clusters<sup>74</sup> which suggest that the condensed-phase polarizability of water should be 7–9% lower than that of the gas phase value. The spatial constraints imposed by condensed-phase environments limit the number of accessible excited states and diffuse character of the electron density distribution as dictated by Pauli's exclusion principle.<sup>30,74</sup> A recent study by Schropp and Tavan<sup>76</sup> further suggests that the average effect of the inhomogeneous electric fields *within* the molecular volume of a single water molecule are consistent with classical parametrizations of polarizable water force fields in which the molecular polarizability is assigned a value around 68% of the gas-phase value.

While these results indicate a reduction of polarizability within the condensed phase, the implications for the rate and nature of the decrease remain unclear. Similarly, a self-consistent analytic formalism capable of correlating changes in molecular polarizability to atomic or molecular properties remains undetermined. While it has been observed that metrics such as aggregation number, hydrogen bonding, or local density are associated with a phase-dependent decrease in molecular polarizability, such metrics are impractical from

the perspective of a molecular dynamics simulation. Consequently, a relationship between the polarizability and an atomic property that smoothly and monotonically transitions from one phase to another is desirable in attempting to establish a simple functional form for polarizability change between phases.

One potentially useful parameter for modeling phase-dependent changes in polarizability is the dipole moment of the molecule. Both experiment and theoretical calculations such as ab initio molecular dynamics simulations demonstrate a difference in molecular dipole moments between the condensed phase and gas phase environments.<sup>16,77,78</sup> Although there is no consensus on an exact value of the average condensed-phase dipole moment of water, it is accepted that the average dipole moment increases upon condensation. Since dipole moments within classical molecular dynamics simulations are readily obtained from the atomic positions and partial charges, no additional information based on neighboring molecules is explicitly required. If a rigid water geometry is chosen, the calculation is even further simplified in that the dipole moment may be determined solely from the magnitude of the associated atomic partial charges.

Within the charge equilibration/fluctuating charge formalism, atomic hardnesses determine molecular polarizability. Thus, a plausible approach to modeling a phase-dependent polarizability in water lies in coupling the atomic charges to the atomic hardness parameters. A similar approach has been previously implemented by Rappé and Goddard for the hydrogen atom in which a linear charge dependence is introduced into the corresponding atomic hardness value.<sup>58</sup> Most generally, each atomic hardness function will depend simultaneously on all partial charges within the molecule; however, this introduces an unnecessary level of complexity into the model. A more simplistic approach entails modulating or scaling all of the atomic hardness values within a molecule based on a single parameter based on the polarization state of the molecule. This parameter may be chosen to be an instantaneous function of all atomic charges within the molecule (such as the dipole moment). However, for water, the average molecular dipole moment appears to be correlated with an increased negative partial charge on the oxygen atom. Consequently, the model may be significantly simplified by coupling the atomic hardnesses directly to the oxygen partial charge.

**B. Charge-Dependent Polarizable Model.** The charge equilibration formalism, based on Sanderson's idea of electronegativity equalization,<sup>79</sup> offers one convenient route to incorporating a local chemical environmental dependence of the molecular polarizability. Polarization of the electronic density (modeled classically as a distribution of atomic partial charges) is affected by the redistribution of charge density within the molecule in an effort to equalize the instantaneous electrostatic chemical potential in the presence of external electric fields arising from nearby molecules. The directionality and ease of charge redistribution is determined by parametrized physical properties of individual atoms. Further details regarding the specifics of charge equilibration methods are available in the literature.<sup>1,24,39,48,58–60,68,79–82</sup>

The charge equilibration electrostatic energy of an  $N$ -atom molecule in the absence of an external electric field, each atom carrying partial charge  $Q_i$ , is

$$E(Q) = \sum_{i=1}^N \left( \chi_i Q_i + \frac{1}{2} \eta_i Q_i^2 \right) + \sum_{i<j}^N Q_i Q_j J_{ij} + \lambda \left( \sum_{i=1}^N Q_i - Q_{\text{total}} \right) \quad (1)$$

where the  $\chi_i$ 's are atom electronegativities and the  $\eta_i$ 's are atomic hardnesses. The  $J_{ij}$  terms represent the interatomic hardness terms for each pair of atoms  $i$  and  $j$  within a molecule. A standard Coulomb interaction is employed between each pair of atoms located on different molecules. The last term in eq 1 describes a molecular charge constraint applied to the entire molecule and enforced via the Lagrange multiplier  $\lambda$ . In the following, we will specifically focus on a TIP4P-FQ water molecular geometry which consists of three charge carrying sites: two hydrogen sites and one off-atom  $M$ -site located along the angle bisector.

In order to establish an appropriate correspondence between charge and polarizability, we consider the polarizability expression for a TIP4P-FQ molecule within the charge equilibration formalism.<sup>83,84</sup>

$$\alpha = \mathbf{R}^T \mathbf{J}^{-1} \mathbf{R} \quad (2)$$

where  $\alpha$  is the  $3 \times 3$  polarizability tensor, and  $\mathbf{R}$  is the  $3 \times 4$  position matrix.  $\mathbf{J}$  is the  $4 \times 4$  hardness matrix comprised of the diagonal  $\eta$  terms and the off-diagonal  $J$ -terms and augmented by a molecular charge neutrality constraint as

$$\mathbf{J} = \begin{pmatrix} \eta_M & J_{MH} & J_{MH} & 1 \\ J_{MH} & \eta_H & J_{HH} & 1 \\ J_{MH} & J_{HH} & \eta_H & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \quad (3)$$

Similarly, the position matrix is also augmented to ensure proper dimensions for matrix operations. Equations 2 and 3 illustrate that the polarizability may be directly related to the molecular geometry and inverse atomic hardnesses. For a rigid molecule such as TIP4P-FQ where the position matrix  $\mathbf{R}$  is fixed, a charge-dependent polarizability may be obtained by introducing an explicit charge dependence into the corresponding hardness matrix elements. Most generally, we modify both the hardness and electronegativity parameters as a function of charge to incorporate the desired phase-dependent polarizable effect:

$$E(Q) = \sum_{i=1}^N \chi_i(Q_M) Q_i + \frac{1}{2} \sum_{i=1}^N \eta_i(Q_M) Q_i^2 + \sum_{i<j}^N J_{ij}(Q_M) Q_i Q_j + \lambda \left( \sum_{i=1}^N Q_i - Q_{\text{total}} \right) \quad (4)$$

These modifications are also accounted for in the corresponding energy derivatives (Appendix).

Since the molecular polarizability response is determined by the atomic hardnesses, this suggests that one may readily modulate the effect by applying an appropriate charge-dependent scaling function,  $g(Q_M)$ , to the atomic hardnesses to yield the following definition for charge-dependent hardness:

$$\mathbf{J}(Q_M) = g(Q_M)\mathbf{J} \quad (5)$$

In order to maintain proper gas- and condensed-phase charge distributions, it is also necessary to scale the electronegativity values in addition to the hardnesses. While reasonable results may be obtained by employing the same scaling factor for both the electronegativities and the hardnesses, finer control of the condensed phase dipole moment distribution is afforded by a tunable factor for the electronegativity scaling. Such flexibility also permits better control of the bulk dielectric constant which may be obtained from the fluctuations of the condensed phase dipole moment.<sup>85</sup> Empirically, we have chosen to introduce the  $\chi$ -scaling

$$\chi(Q_M) = \frac{g(Q_M)}{h(Q_M)}\chi = [(1-p) + pg(Q_M)]\chi \quad (6)$$

where  $p$  is an empirical parameter that controls the extent to which  $\chi$  is scaled relative to the hardness scaling function. For  $p = 1$ , the scaling on electronegativity values is equivalent to that in the hardnesses; similarly, a value of  $p = 0$  would correspond to no scaling (i.e., a constant electronegativity with no charge-dependent scaling).

Unfortunately, the explicit introduction of a charge dependence into the molecular hardness matrix has introduced the additional complication that the polarizability expression in eq 2 is no longer exact. Consequently, the corresponding system of equations for the equilibrium charges (and polarizabilities) is now a nonlinear system and must be solved by an iterative approach. In situations where the equilibrium charge distribution is not strongly perturbed by the implicit charge dependence, a slightly modified version of eq 2

$$\alpha_{\beta\gamma}(Q_M) \approx \frac{\alpha_{\beta\gamma}}{\xi(Q_M)} - \left( \frac{\nabla_M g(Q_M) \langle R_\beta | \mathbf{J}^{-1}(r) | \hat{\mathbf{M}} \rangle}{|g(Q_M)|^2 h(Q_M)} \right) \times \left[ p\chi_M \langle R_\beta | \mathbf{J}^{-1}(r) | \hat{\mathbf{M}} \rangle + \frac{1}{2}\mu_\gamma \right] \quad (7)$$

(derived in the Appendix) is convenient for obtaining a leading-order approximation of the polarizability in the absence of a fully nonlinear treatment. In the above expression,  $\alpha_{\beta\gamma}$  is the  $\beta\gamma$ -element of the gas-phase molecular polarizability tensor,  $\nabla_M g(Q_M)$  is the derivative of the scaling function with respect to  $Q_M$ ,  $R_\beta$  is the  $\beta$ -position vector,  $\hat{\mathbf{M}}$  is a matrix that selects elements associated with the  $M$ -site (since we have chosen our hardness elements to only depend on the  $M$ -site charge), and  $\mu_\gamma$  is the  $\gamma$ -component of the dipole moment. We see the charge-dependent polarizability differs from the unscaled (gas-phase) value by a multiplicative factor  $\xi(Q_M) = g(Q_M)h(Q_M)$  and additive terms, which are related to the  $M$ -site hardness and dipole moment, respectively. These additive terms of equal magnitude and opposite sign are small compared to the first term and do not greatly influence  $\alpha(Q_M)$  (see Figure 8). If we neglect the additive terms and consider the limit in which  $p = 1$ , eq 7 reduces to

$$\alpha_{\beta\gamma}(Q_M) \approx \frac{\alpha_{\beta\gamma}}{g(Q_M)} \quad (8)$$

from which it is clearly seen that  $\alpha(Q_M)$  modulates the gas-phase polarizability via an inverse relationship with the scaling function,  $g(Q_M)$ . While eq 8 is effective for illustrative purposes, we have employed eq 7 for calculations of the condensed-phase polarizability within this work.

Having now introduced an explicit charge-dependent polarizability via a simple scaling function  $g(Q_M)$ , it is relevant to discuss the nature and form of this scaling function. While there is no formal theory connecting charge and polarizability, general trends provide guiding insight. In prior work, Rappé and Goddard have employed atomic hardnesses which depend linearly on charge.<sup>58</sup> In the context of molecular dynamics simulations, such an approach would necessitate some degree of charge bounding to prevent unfavorable overpolarization or underpolarization and to establish consistent polarizabilities in the gaseous and condensed phases. In light of this we have chosen to employ an error function form which applies constant scaling in the purely condensed-phase and gaseous regions and approximately linear scaling in the intermediate region. The use of the error function is also preferred as it allows smooth transitions between each region, which is necessary to avoid discontinuities in the forces. Thus, we choose an error function of the form

$$g(Q_M) = a - b \operatorname{erf}(c(d - Q_M)) \quad (9)$$

as the scaling function since it incorporates additional empirical parameters which can be utilized to model the desired relationship between polarizability and charge. Parameters  $a$  and  $b$  collectively define the polarizability at the gaseous and condensed-phase limits. The rate of polarizability change with charge for nonisolated molecules is controlled by  $c$ . Collectively,  $c$  and  $d$  describe the onset of scaling and the range of charges over which polarizability changes. The selection of these parameters and comments regarding the parametrization of this model are discussed further in section III.A.

**C. Molecular Dynamics Simulations.** Condensed phase simulations of the TIP4P-QDP model are conducted at constant pressure and temperature ( $T = 298$  K) using CHARMM.<sup>86,87</sup> For comparative purposes, analogous simulations of TIP4P-FQ under matching conditions are also performed. 216 molecules of each model are included in their respective simulations. Simulations 25 ns in length are performed for the TIP4P-QDP and TIP4P-FQ models. Conditionally convergent long-range interactions are accounted for using particle mesh Ewald<sup>88</sup> with  $\kappa = 0.37$  and 20 grid points in each dimension (FFT grid spacing). Fictitious charge degrees of freedom are assigned masses of 0.000 069 kcal/(mol $\cdot$ ps<sup>2</sup>). The Nose–Hoover<sup>89</sup> method is implemented to couple the charge degrees of freedom to a thermostat at 1 K; this thermostat has a mass of 0.005 kcal/(mol $\cdot$ ps<sup>2</sup>). A 0.5-fs time step is used for propagating the classical equations of motion using a Verlet leapfrog integrator.

Liquid–vapor interface simulations are performed for 1024 molecules at constant volume and constant temperature ( $T = 298$  K). The dimensions used for the box are 24.0  $\times$  24.0  $\times$  130.0 Å. Particle mesh Ewald parameters are modified



**Table 1.** Summary of Simulation Parameters

parameter	condensed-phase	liquid–vapor interface
$T$ (K)	298.	298.
$N$ (molecules)	216.	1024.
simulation length (ns)	25.	30.
time step (fs)	0.5	0.5
QMAS kcal/(mol·ps <sup>2</sup> )	0.000069	0.000069
TMAS kcal/(mol·ps <sup>2</sup> )	0.005	0.005
$\kappa$	0.37	0.33
grid points (1 Å spacing)	20 × 20 × 20	30 × 30 × 120

**Table 2.** Parameters Used for the Scaling Functions of the TIP4P-QDP model<sup>a</sup>

parameter	TIP4P-QDP	QDP-P1
$a$	1.18022	1.18022
$b$	0.17985	0.17985
$c$	−2.33071	−2.33071
$d$	−1.49180	−1.49180
$p$	0.80000	1.00000

<sup>a</sup> Parameters  $a$ – $d$  are coefficients used in the error function (eq 9), and  $p$  is the additional scaling factor applied to the scaling of  $\chi$  (eq 6).

from the condensed phase simulation, with 30 grid points in the transverse directions, 120 grid points in the longitudinal direction, and  $\kappa = 0.33$ . Simulation lengths of 30 and 75 ns are used for the two models. All other simulation parameters are equivalent to those listed for the condensed phase simulations in Table 1.

### III. Results

**A. Parameterization of TIP4P-QDP Model.** The TIP4P-QDP model is based on the application of the charge-dependent scaling function to the original TIP4P-FQ model; therefore, we retain the TIP4P-FQ geometry in the TIP4P-QDP model. To construct the TIP4P-QDP model, we first modify the TIP4P-FQ hardness values which were originally chosen to mimic a reasonable condensed-phase polarizability for stable bulk simulations; thus, the TIP4P-FQ model has a static polarizability that does not correspond to any theoretical or experimental gas phase value. Since it is desired to establish a polarizability gradient between the gaseous and condensed phases and since the gas-phase polarizability is well-known experimentally, we reparameterize the hardness values to reproduce a reasonable gas-phase polarizability of 1.40 Å<sup>3</sup>. We note that the resulting TIP4P-QDP gas-phase hardnesses maintain approximately the same relative magnitudes as the original TIP4P-FQ hardnesses. The error function form (eq 9) is then parametrized with the caveat that the gas phase polarizability remains unchanged. In the condensed phase, the hardnesses are scaled by a value of  $g(Q_M) > 1$  for charges greater than the equilibrium gas phase charges. The parameters of the scaling function which influence the height, slope, and inflection are determined empirically such that the resulting polarizability distribution is centered about an anticipated condensed phase polarizability 7–9% less than the gas phase value (Table 2). A broad polarizability distribution allows for a description of a greater range of local chemical environments.

**Table 3.** Comparison of Potential Parameters for the TIP4P-QDP and TIP4P-FQ Models

parameter	TIP4P-FQ <sup>a</sup>	TIP4P-QDP	QDP-P1
$\epsilon$ (kcal/mol)	0.28620	0.29012	0.350120
$F_{\min}$ (Å)	3.54586	3.55	3.5646
$\theta$ (deg)	104.52	104.52	104.52
$r_{\text{OH}}$ (Å)	0.9572	0.9572	0.9572
$r_{\text{OM}}$ (Å)	0.15	0.15	0.15
$\chi_{\text{M}} - \chi_{\text{H}}$ (kcal/(mol·e))	68.49	60.63	59.91
$J_{\text{MM}}$ (kcal/(mol·e <sup>2</sup> ))	371.6	309.92	309.92
$J_{\text{HH}}$ (kcal/(mol·e <sup>2</sup> ))	353.0	295.36	295.36
$J_{\text{MH}}(r_{\text{MH}})$ (kcal/(mol·e <sup>2</sup> ))	286.4	239.47	239.47
$J_{\text{HH}}(r_{\text{HH}})$ (kcal/(mol·e <sup>2</sup> ))	203.6	181.91	181.91

<sup>a</sup> Reference 1.

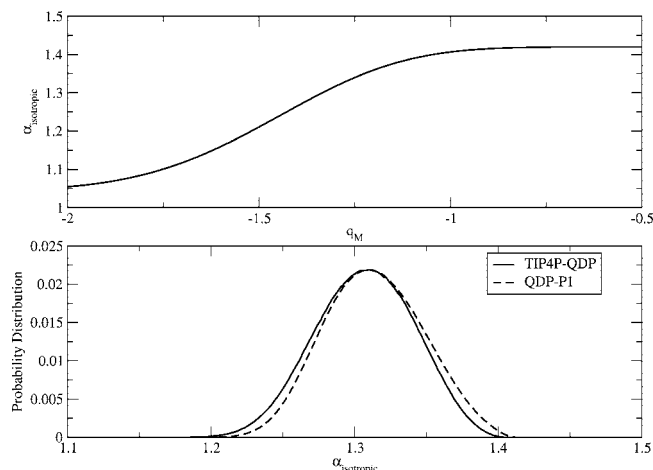
**Table 4.** Gas-Phase Properties from TIP4P-FQ, TIP4P-QDP, and Experiment

property	TIP4P-FQ <sup>a</sup>	TIP4P-QDP	QDP-P1	experiment
$\mu$ (Debye)	1.85	1.85	1.85	1.85 <sup>b</sup>
$\bar{\alpha}$ (Å <sup>3</sup> )	1.12	1.40	1.40	1.47 <sup>c</sup>
$E_{\text{dimer}}$ (kcal/mol)	−4.50	−4.67	−4.44	−5.4 ± 0.7 <sup>d</sup>
dimer O–O length (Å)	2.92	2.91	2.98	2.98 <sup>d</sup>

<sup>a</sup> Reference 1. <sup>b</sup> Reference 113. <sup>c</sup> Reference 114. <sup>d</sup> Reference 115.

Regarding electronegativity scaling, a final value of the  $p$ -parameter is determined to be  $p = 0.80$ ; this value generates a condensed phase dipole moment distribution with an average of 2.641 (± 0.001) Debye, similar to that exhibited by the TIP4P-FQ model. Introduction of the scaling function and modification of the hardnesses further necessitated slight reparameterization of the remaining electrostatic and nonbonded parameters. The electronegativities were then reparameterized such that a single water molecule in vacuum minimizes to the experimental dipole moment of 1.85 D. Since the polarizability of TIP4P-QDP in the condensed phase is higher than that of TIP4P-FQ, the Lennard-Jones parameters required minor modification to prevent overpolarization while still reproducing reasonable densities and energetics. The Lennard-Jones parameters were parametrized based on fitting to gas-phase water dimer binding energies and geometries (bond distances). A comparison of the electrostatic and nonbonded parameters for TIP4P-QDP and TIP4P-FQ is presented in Table 3, while the gas phase properties for these two models are compared in Table 4.

For further comparisons, a model consisting of full scaling  $p = 1.0$  was also developed and parametrized. The results of this model (hereafter referred to as QDP-P1) are also included in this work as a reference in order to more fully clarify differences between the TIP4P-FQ and TIP4P-QDP models. We point out that the best parametrization of QDP-P1 featured comparable density and polarizability to TIP4P-QDP, but had notably higher dipole moments in the condensed phase ( $\langle \mu \rangle \approx 2.75$ ), an anticipated consequence of scaling electronegativity and hardness equivalently. As will be discussed, the enhanced dipole moments are responsible for increased intermolecular cohesion which ultimately reduced the quality of the QDP-P1 parametrization and necessitated additional scaling of the atomic electronegativities.



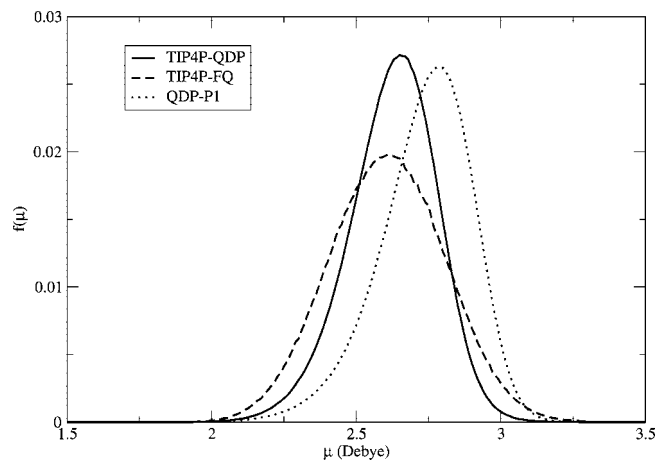
**Figure 1.** (a) Molecular polarizability ( $\text{\AA}^3$ ) as a function of  $Q_M$ . (b) Distribution of molecular polarizabilities within the condensed phase. Polarizabilities were calculated using eq 7 and the charges from simulation.

**B. Condensed-Phase Properties. 1. Density.** The density of the condensed phase was determined via the expression

$$\rho = \frac{NW}{N_A \langle V \rangle} \quad (10)$$

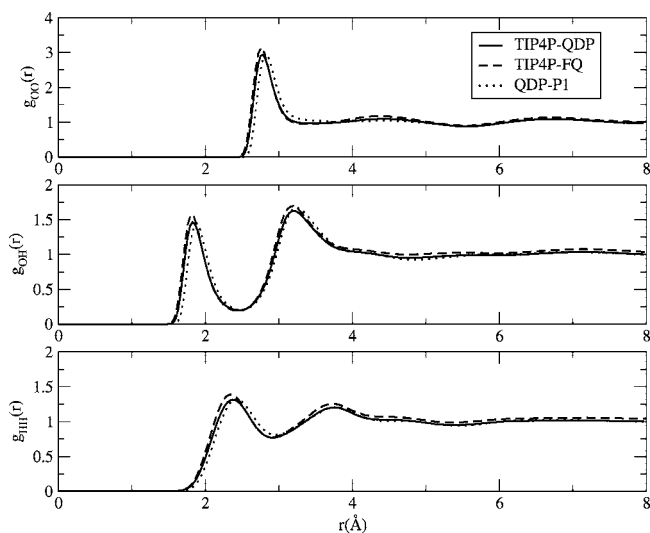
where  $N$  is the number of molecules in the simulation,  $\langle V \rangle$  is the average volume of the simulation cell,  $W$  is the molecular mass of water, and  $N_A$  is Avogadro's number. The average density at 298 K was calculated to be  $0.9954 (\pm 0.0002) \text{ g/cm}^3$  for the TIP4P-QDP model. This value is in agreement with the experimental value  $0.997 \text{ g/cm}^3$  and reflects the general quality of widely used water models.<sup>1,54,90–94</sup> As discussed below, the parametrization of the Lennard-Jones parameters to give this density for TIP4P-QDP was balanced with the need to accurately reproduce  $\Delta H_{\text{vap}}$ .

**2. Condensed-Phase Polarizability.** Since the TIP4P-QDP model adjusts the molecular polarizability dynamically in response to the (electro)chemical environment, we consider the distribution of molecular polarizabilities in the condensed phase. As previously discussed within the context of formulating this model, an isolated TIP4P-QDP molecule was parametrized to a polarizability of  $1.40 \text{ \AA}^3$ , a value comparable to the experimental gas phase polarizability. The scaling function allows for TIP4P-QDP molecules to have polarizabilities as low as  $1.05 \text{ \AA}^3$ . The polarizability distribution observed in the condensed phase is presented in (Figure 1).  $M$ -site charges for every molecule from each snapshot of a trajectory were used in conjunction with eq 7 to generate the polarizability distribution. The average polarizability in the condensed phase is calculated to be  $1.309 (\pm 0.001) \text{ \AA}^3$ , approximately 17% higher than the static condensed phase polarizability of TIP4P-FQ and about 11% less than the experimental gas phase value. The average condensed-phase value approximately reflects a 6.5% reduction in the molecular polarizability relative to the TIP4P-QDP gas phase value, which agrees well with the estimated range of 7–9% reduction calculated by Morita from first principles.<sup>74</sup> The distribution also exhibits a width of observed molecular polarizabilities of about  $0.20 \text{ \AA}^3$ , which is consistent with the width of the distribution for QDP-P1.



**Figure 2.** Dipole moment distributions for TIP4P-QDP, TIP4P-FQ, and QDP-P1.

**3. Dipole Moment Distribution.** The dipole moment distributions for both the TIP4P-QDP and TIP4P-FQ models are presented in Figure 2. The distribution for TIP4P-QDP has an average value of  $2.641 (\pm 0.001)$  Debye, almost indistinguishable from that of TIP4P-FQ. While refinement of the  $p$ -value could improve the agreement with TIP4P-FQ, such a refinement would be arbitrary due to the uncertainty in the *true* value of the average condensed-phase dipole moment for water. Experimental estimates<sup>78</sup> of the dipole moment of condensed-phase water ( $2.96 \pm 0.6 \text{ D}$ ) span a broad range of values; this range is also observed in *ab initio* molecular dynamics simulations.<sup>16,77</sup> The TIP4P-QDP distribution also features skewed symmetry, with the population of molecules higher dipole moment more sharply declining than those with lower dipole moments. TIP4P-QDP also features a somewhat more narrow distribution than TIP4P-FQ. Although it might seem reasonable that adjustment of the  $p$ -value could result in a wider distribution for TIP4P-QDP, various  $p$ -values in the range considered during parametrization show essentially the same shape and width, as evident by comparison of the TIP4P-QDP and QDP-P1 distributions. *Ab initio* studies of bulk water dipole moments demonstrate the importance of local environment, particularly the hydrogen bond coordination, on a molecule's dipole moment.<sup>95,96</sup> Furthermore, as a molecule becomes more coordinated with hydrogen bonds, the range of accessible dipole moments increases due to an increase in possible structural variations. It is therefore argued that systems with a larger fraction of highly coordinated water molecules will feature wider dipole moment distributions. As is noted in section III.C.3, the average number of hydrogen bonds in the condensed phase is lower for TIP4P-QDP than TIP4P-FQ. In particular, the ratio of water molecules with four hydrogen bonds to molecules with three hydrogen bonds is lower in TIP4P-QDP than TIP4P-FQ (1.4 to 1.8, respectively). Such a change in hydrogen bonding can be reasoned to impact condensed phase properties (such as the higher diffusion constant observed for TIP4P-QDP) and the narrowed dipole moment distribution. We also note that the ratio of hydrogen bonds is consistent between QDP models, further supporting the link between hydrogen bond coordination and dipole moment distribution width. The slight skewed



**Figure 3.** Radial distribution functions for TIP4P-QDP, TIP4P-FQ, and QDP-P1.

character of the TIP4P-QDP dipole moment distribution can be attributed, at least in part, to the selective attenuation of the largest oxygen partial charges due to the significant scaling of the hardnesses in the condensed phase.

**4. Radial Distribution Function.** Radial distribution functions for TIP4P-QDP and TIP4P-FQ are presented in Figure 3. Although the distributions share similar features, there are subtle differences between the two models. There is a slight reduction in peak height and structure for the TIP4P-QDP model compared to TIP4P-FQ. This feature which is common to the three distributions presented corresponds to a decrease in four-coordinate hydrogen bond formation. This notion of decreased hydrogen bond formation is further discussed below. A subtle shift of the TIP4P-QDP distributions toward larger separation distances is also observed in each distribution. Although this is a rather minor difference from the TIP4P-FQ model, the shift is in better agreement with neutron diffraction data<sup>97,98</sup> which also features peaks centered at greater distances than predicted by TIP4P-FQ.

**5. Enthalpy of Vaporization.** Enthalpy of vaporization is defined as

$$\Delta H_{\text{vap}} = \Delta E_{\text{vap}} + \Delta(PV)_{\text{vap}} \quad (11)$$

Noting that the state change in vaporization is from the liquid phase to the vapor phase allows for the expansion to liquid and vapor terms. Making the assumption that the change in volume of the liquid is negligible compared to that of the gas and assuming ideality in the gas phase results in the final expression:

$$\Delta H_{\text{vap}} = E_{\text{gas}} - E_{\text{liq}} + RT \quad (12)$$

Here, the  $E_{\text{gas}}$  is the energy of a single minimized molecule in vacuum and  $E_{\text{liq}}$  is the average system energy from condensed-phase simulations. The original TIP4P-FQ model was parametrized to exhibit excellent agreement with the experimental enthalpy of vaporization. The current parametrization of TIP4P-QDP predicts a vaporization enthalpy of 10.55 ( $\pm$  0.12) kcal/mol. Although higher than the experiment (10.51 kcal/mol) and TIP4P-FQ (10.49 kcal/mol)

values, the TIP4P-QDP result still agrees well with these two values, exhibiting a 0.4% increase over the experimental value. Additionally, further improvement of this value via modification of the Lennard-Jones parameters appears to have adverse effects on the density of the condensed phase. Thus, in order to ensure reasonable accuracy in both properties, a slight compromise in the model's enthalpy of vaporization was deemed acceptable.

**6. Dielectric Constant.** The dielectric constant for each system was calculated using the relation:

$$\epsilon = \epsilon_{\infty} + \frac{4\pi}{3k_{\text{B}}T\langle V \rangle} (\langle M^2 \rangle - \langle \mathbf{M} \rangle \cdot \langle \mathbf{M} \rangle) \quad (13)$$

where  $\mathbf{M}$  is the dipole moment of the simulation cell. The term  $\epsilon_{\infty}$  is the infinite frequency (optical) dielectric constant, estimated using the approach outlined in ref 54. To summarize, charge dynamics simulations were performed on static configurations generated from an *NVT* simulation at the average density of QDP predicted from constant pressure simulations. The charge degrees are propagated at 1 K, and the Kirkwood fluctuation formula is applied to determine the optical dielectric:

$$\epsilon_{\infty} = 1 + \frac{4\pi}{3k_{\text{B}}T\langle V \rangle} (\langle M^2 \rangle - \langle \mathbf{M} \rangle \cdot \langle \mathbf{M} \rangle) \quad (14)$$

We obtain a dielectric constant of 85.8 ( $\pm$  1.0) where the contribution from the  $\epsilon_{\infty}$  is approximately 2.1; the total dielectric constant is 10% higher than experiment, though still of generally acceptable quality in comparison with several fixed-charge models.<sup>90–94</sup> We note that the optical dielectric is slightly higher than the value for TIP4P-FQ reported by Rick et al.,<sup>1</sup> as well as experimental estimates. However, the value estimated for TIP4P-FQ using the present approach,  $\epsilon_{\infty} = 1.775$ , overestimates the previously reported result of  $\epsilon_{\infty} = 1.592$ . This suggests the tendency for the Kirkwood approach to overestimate the optical dielectric constant. The increased value of the optical dielectric constant for TIP4P-QDP over that of TIP4P-FQ is attributed to the higher molecular polarizability, which influences  $\epsilon_{\infty}$  as dictated by the Clausius–Mossotti relation.

On the basis of the results of QDP-P1 which yielded  $\epsilon \approx 98$ , it is suggested that the average condensed phase dipole moment plays an important role in obtaining the correct dielectric constant. In this regard, we find that  $\langle \mu_{\text{liq}} \rangle \approx 2.6$  is sufficient for the QDP model to approach the experimental dielectric constant—an observation also made by Sprik.<sup>85</sup> Therefore, the higher dielectric constant of TIP4P-QDP can be attributed to the slightly larger average dipole moment.

**7. Diffusion Constant.** The self-diffusion constant was calculated using the Einstein relationship applied to constant volume and temperature simulations of pure liquid:

$$D_s = \lim_{t \rightarrow \infty} \frac{1}{6t} \langle (r(t) - r(0))^2 \rangle \quad (15)$$

The volume of the simulation cell was set to reproduce the average density of TIP4P-QDP from the constant *NPT* simulations. The diffusion constant for TIP4P-QDP was calculated to be 2.20 ( $\pm$  0.04)  $\times 10^{-9}$  m<sup>2</sup>/s which is quite close to the experimental value of 2.30  $\times 10^{-9}$  m<sup>2</sup>/s. This

**Table 5.** Condensed Phase and Interfacial Properties

property	TIP4P-FQ <sup>a</sup>	TIP4P-QDP	QDP-P1	experiment
$\rho_{\text{liq}}$ (g/cm <sup>3</sup> )	1.0001 (0.0003)	0.9954 (0.0002)	0.9951 (0.0002)	0.997 <sup>b</sup>
$\langle \mu_{\text{liq}} \rangle$ (Debye)	2.623 (0.001)	2.641 (0.001)	2.752 (0.001)	2.96 (0.60) <sup>c</sup>
$\Delta H_{\text{vap}}$ (kcal/mol)	10.49 <sup>d</sup>	10.55 (0.12)	10.96 (0.12)	10.51 <sup>b</sup>
$\langle \alpha_{\text{iso,liq}} \rangle$ Å <sup>3</sup>	1.12 <sup>d</sup>	1.309 (0.001)	1.323 (0.001)	1.34 <sup>e</sup>
$D_s$ (10 <sup>-9</sup> m <sup>2</sup> /s) <sup>j</sup>	1.93 (0.05), 2.15	2.20 (0.04), 2.46	1.83(0.05), 2.04	2.30 <sup>f</sup>
$\kappa_T$ (10 <sup>-10</sup> Pa <sup>-1</sup> )	3.877 (0.098)	4.013 (0.062)	3.409 (0.051)	4.524 <sup>b</sup>
$C_p$ (cal/mol K)	21.0 (5.5)	16.4 (3.5)	18.5 (2.2)	18.0 <sup>b</sup>
$\epsilon_\infty$	1.775, 1.592 <sup>d</sup>	2.128	2.057	1.79 <sup>g</sup>
$\epsilon$	79. (8) <sup>d</sup>	85.8 (1.0)	97.6 (0.2)	78. <sup>h</sup>
$\Delta\Phi$ (kcal/mol)	-12.21 (0.05)	-11.98 (0.08)	-12.87 (0.05)	
$\gamma$ (dyne/cm)	72.7 (1.5)	71.0 (2.7)	81.2 (3.1)	71.9 <sup>i</sup>

<sup>a</sup> Values presented are based on calculations from this work, unless otherwise noted. <sup>b</sup> Reference 116. <sup>c</sup> Reference 78. <sup>d</sup> Reference 1. <sup>e</sup> Estimated condensed-phase isotropic polarizability based on the gas-phase value of 1.47 Å<sup>3</sup> from ref 114 and assuming a 9% reduction in polarizability as deduced by Morita in ref 74. <sup>f</sup> Reference 117. <sup>g</sup> Reference 118. <sup>h</sup> Reference 119. <sup>i</sup> Reference 120. <sup>j</sup> Values as calculated for an  $N = 216$  system (left) and corrected for extrapolation to infinite system size (right).

reflects an improvement over TIP4P-FQ calculated here to be  $1.93 (\pm 0.05) \times 10^{-9}$  m<sup>2</sup>/s and elsewhere<sup>1</sup> to be  $1.9 (\pm 0.1) \times 10^{-9}$  m<sup>2</sup>/s. As previously mentioned, the enhanced diffusion constant relative to TIP4P-FQ is likely to be the result of reduced structure in the condensed phase as suggested by decreased hydrogen bonding. We mention the diffusion constant calculated for QDP-P1 was lower than both TIP4P-QDP and TIP4P-FQ. It is reasoned that the enhanced dipole moment of QDP-P1 is influenced by higher cohesive intermolecular forces which limit the dynamics in the bulk. However, little change in the average hydrogen bonding relative to the TIP4P-QDP model suggests that the strong cohesive interactions did not result in enhanced organization of the fluid structure. From the QDP-P1 RDF, it is evident that QDP-P1 lacks key structural features that are common to both TIP4P-QDP and TIP4P-FQ, in particular the depletion of the first minimum in the O–O distribution.

Finally, it has been suggested that diffusion constants computed from molecular dynamics simulations have an inherent sensitivity to system size.<sup>99</sup> Hence, in order to make a valid comparison to experimental data, the diffusion constant should be extrapolated for an infinitely large system. Simulations of a larger system ( $N = 988$ ) of TIP4P-QDP were performed to assess the extent to which system size influenced the calculation. For this larger system, a diffusion constant of  $2.40 (\pm 0.02) \times 10^{-9}$  m<sup>2</sup>/s was calculated. Employing the linear extrapolation method used by Miller and Manolopoulos,<sup>99</sup> we estimate the diffusion constant for an infinitely large system to be approximately  $2.46 \times 10^{-9}$  m<sup>2</sup>/s, suggesting a 1.1 scaling factor from the  $N = 216$  to the  $N = \infty$  system. Applying this factor to the TIP4P-FQ and TIP4P-QDP values yields better agreement with the experimental value. Although the corrected diffusion constant of TIP4P-QDP now overestimates experiment, the deviation from experiment is relatively consistent with the uncorrected value. The diffusion constants as calculated for the  $N = 216$  system and extrapolated for  $N = \infty$  are included in Table 5.

**8. Isobaric Heat Capacity.** The isobaric heat capacity was calculated via numerical differentiation method utilized by Horn et al.:<sup>91</sup>

$$C_p = \left( \frac{\partial H}{\partial T} \right)_p \approx \frac{\langle H_2 \rangle - \langle H_1 \rangle}{T_2 - T_1} \quad (16)$$

where  $\langle H \rangle$  is the average enthalpy calculated from  $NPT$

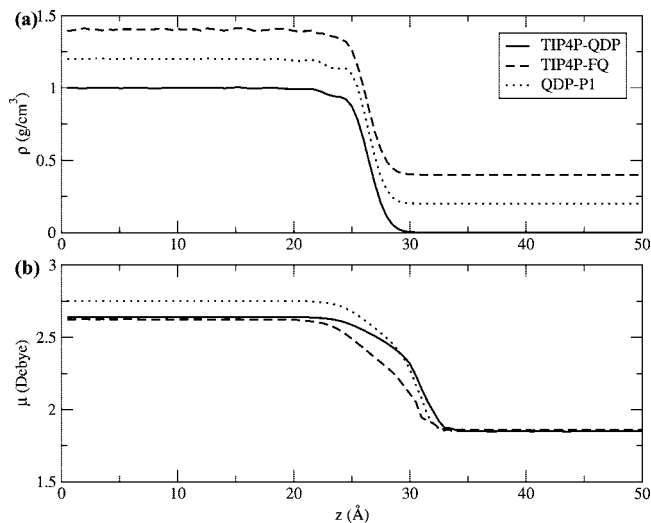
simulations. Additional simulations at  $T = 297$  and  $299$  K were utilized to compute  $C_p$  at 298 K. TIP4P-QDP has a heat capacity of  $16.4 (\pm 3.5)$  which underestimates experiment by about 9%. This agreement with experiment is slightly better than TIP4P-FQ, which overestimates experiment by about 17%. It is further noted that eq 16 provides only an approximate  $C_p$  value that is not corrected for quantum effects. Regardless, such effects are not expected to greatly influence the results reported here since they are anticipated to be less than the magnitude of uncertainty for each value.

**9. Isothermal Compressibility.** The isothermal compressibility was calculated using the following equation:

$$\kappa_T = \frac{\sigma_v^2}{\langle V \rangle k_B T} \quad (17)$$

where  $\langle V \rangle$  and  $\sigma_v$  denote the average and the standard deviation of the total system volume over the course of the simulation,  $T$  is the temperature, and  $k_B$  represents Boltzmann's constant. For TIP4P-QDP, a value of  $4.013 (\pm 0.062) \times 10^{-10}$  Pa<sup>-1</sup> was calculated. Although this is  $0.5 \times 10^{-10}$  Pa<sup>-1</sup> lower than the experimental value, it showed a 3% improvement over the value calculated for TIP4P-FQ. As is anticipated due to the dependence of this property on  $\sigma_v$ , QDP-P1 exhibited a value of  $\kappa_T \approx 3.3$ , which is notably less than both the TIP4P-QDP and TIP4P-FQ. The reduced value of  $\kappa_T$  results from reduced fluctuations in volume, a characteristic expected from increased cohesive forces resulting from the enhanced dipole moments as previously suggested.

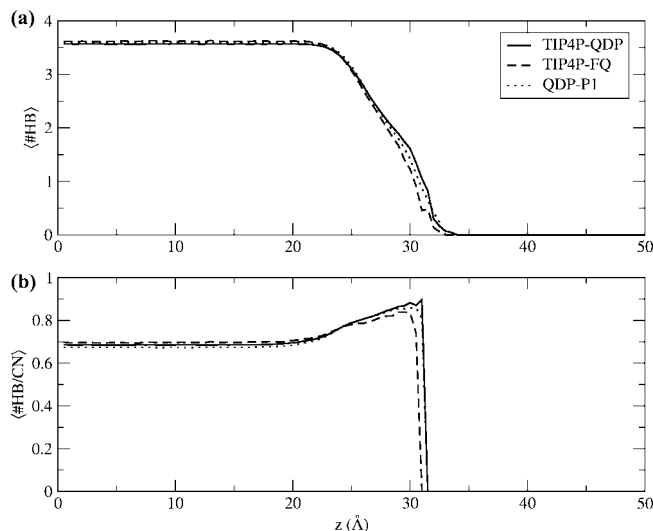
**C. Liquid–Vapor Interfacial Properties. 1. Density Profile.** The water density profiles as a function of the  $z$ -position relative to the center of mass of the water slab are presented in Figure 4. From this data, the Gibbs dividing surface (GDS) is calculated to occur at 26.01 Å<sup>3</sup> from the system's center of mass for both TIP4P-QDP and TIP4P-FQ. The GDS is calculated as the point in which the surface excess is zero. Using the “10–90” criteria for interfacial thickness,<sup>100</sup> the interfacial region is considered to be the region in which the density transitions from 10% to 90% of the bulk condensed-phase density. We estimate the interfacial thickness to be approximately 3.3 Å for TIP4P-QDP which is commensurate with that for TIP4P-FQ in this work. The TIP4P-FQ interfacial thickness has been previously reported



**Figure 4.** Interfacial profiles as a function of  $z$ -position relative to the center of mass for (a) the density and (b) the dipole moment of TIP4P-QDP, TIP4P-FQ, and QDP-P1. A  $0.2 \text{ g/cm}^3$  offset was applied to the density profiles to distinguish unique features.

as  $3.5 \text{ \AA}$ ,<sup>101</sup> which agrees reasonably well with the value calculated here. One notable difference between the density profiles of the two models is a distinguishable “notch” present in the TIP4P-QDP profile at the onset of the interfacial region. This notch represents a region prior to the onset of the interface, but features a reduced density of approximately  $0.93 \text{ g/cm}^3$ . This notch is a feature that was also observed for QDP-P1. While the TIP4P-FQ profile shows a region of reduced density between the pure bulk phase and interface, this region features a gradual slope, different than the sharp drop-off and leveling observed in the QDP models. This suggests that the introduction of dynamical polarizability allows for the formation a stable transitional region between the condensed phase and interface (as defined by the 10–90 criteria). The nature of hydrogen bonding (see also section III.C.3) between the two models at this depth is reasoned to cause this feature. While a general trend of decreasing hydrogen bonds beginning at the onset of this region is observed for both model types, the ratio of hydrogen bonds to the coordination number becomes notably higher for the QDP models than TIP4P-FQ at this depth. Additional spatial requirements for hydrogen bond networking, as well as the increased stability of such a network help explain this region of reduced density exhibited by the QDP models.

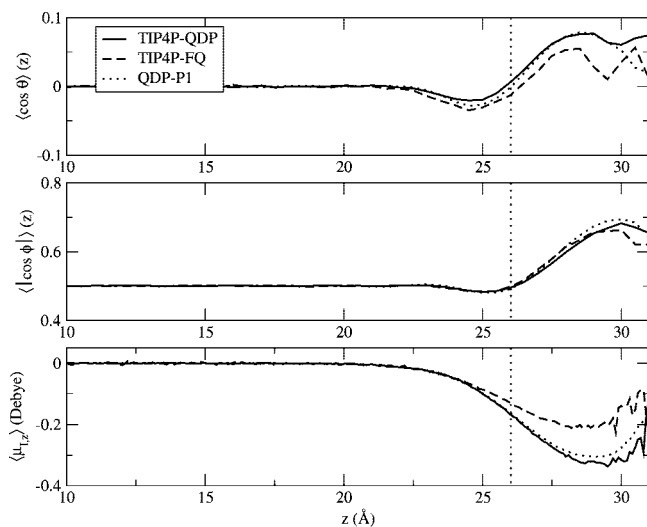
**2. Dipole Moment Profile.** The dipole moment profile is presented in Figure 4 for TIP4P-QDP and TIP4P-FQ. Consistent with the condensed-phase simulations, the average condensed-phase dipole moment is approximately  $2.641 (\pm 0.001) \text{ D}$  for TIP4P-QDP, which is slightly higher than the TIP4P-FQ value of  $2.623 (\pm 0.001) \text{ D}$ . The enhanced TIP4P-QDP dipole moments in the region extending from the interface into the gas-phase are an anticipated result from the increasing molecular polarizability in this region relative to the bulk phase. It is noted, however, that the dipole moment of both models ultimately reduce to the appropriate value of  $1.85 \text{ Debye}$  within the gas-phase. Furthermore, this dipole moment enhancement suggests hydrogen bond net-



**Figure 5.** Hydrogen bond profile. (a) Average number of hydrogen bonds as a function of  $z$ -position for TIP4P-QDP, TIP4P-FQ, and QDP-P1. (b) Probability of hydrogen bond formation as a function of  $z$ -position as calculated from the ratio of hydrogen bonds to coordination number. Definitions of O–O distance less than  $3.5 \text{ \AA}$  and H–O–O angle less than  $30^\circ$  were used as the hydrogen bond criteria.

works extending into the gas-phase as supported in the following section. Similar results were seen in QDP-P1, suggesting this is a common effect due to the modulation of polarizability in this region.

**3. Hydrogen Bond Profile.** For the hydrogen bond profile (Figure 5), the average number of hydrogen bonds formed by a water molecule as a function of its  $z$ -position relative to the center of mass was calculated. The definition for a hydrogen bond was based on the geometric criteria used by Liu et al.<sup>101</sup> We select the geometric definition over an energetic one<sup>102</sup> such that a comparison could be drawn to the results of Liu et al. and because this definition has shown more reliable simulation results.<sup>103</sup> Using an O–O distance of  $3.5 \text{ \AA}$  as a distance criteria, prospective hydrogen bonding pairs were tagged. Among these tagged pairs, an angular requirement for the HO–O bond to be less than  $30^\circ$  was implemented to define a hydrogen bond pair. Within the bulk region, water molecules formed an average of  $3.57$  hydrogen bonds for TIP4P-QDP while the average for TIP4P-FQ was  $3.62$  hydrogen bonds. Also presented in Figure 5 is the probability of hydrogen bond formation as a function of  $z$ -position relative to the center of mass. This probability was calculated as the ratio of the number of hydrogen bonds formed by a molecule divided by its coordination number. The coordination number was defined as the number of water molecules having an O–O distance less than  $3.5$  from the water molecule of consideration. The probability of hydrogen bond formation increases significantly in the interfacial region compared to the bulk for both models. This is consistent with the observations of Liu et al. for TIP4P-FQ.<sup>101</sup> It is noted, however, that the hydrogen bond probability for TIP4P-QDP is reduced in the condensed phase and enhanced in the interfacial region, relative to TIP4P-FQ. This reduced hydrogen bonding within the condensed region was anticipated based on the reduced structure indicated by the RDFs

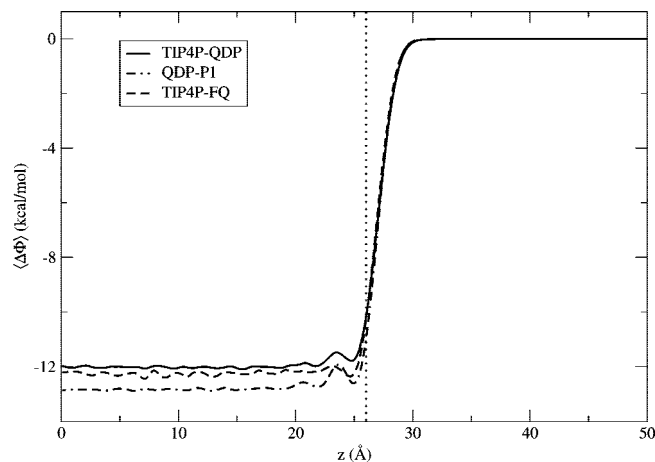


**Figure 6.** (a) Profile of  $\langle \cos \theta \rangle$ , where  $\theta$  is angle formed between the permanent dipole vector of the water molecule and the fixed  $z$ -axis. (b) Profile of  $\langle \cos \phi \rangle$ , where  $\phi$  is the angle formed between the molecular plane of water and the fixed  $z$ -axis. (c) Profile of the  $z$ -induced dipole moment. All  $z$ -values are relative to the center of mass of the system. The GDS is represented on each plot as a dashed vertical line.

of TIP4P-QDP. It is likely the combination of enhanced polarizability relative to the bulk and a more favorable dimer energy than TIP4P-FQ prolong a water molecule's ability to remain hydrogen bonded as it leaves the interface for the gas-phase.

**4. Molecular Orientation.** The orientational structure of TIP4P-QDP and TIP4P-FQ are analyzed via the distribution of two angular coordinates,  $\theta$  and  $\phi$ , as a function of  $z$ -position with respect to the GDS. For convenience, we present the orientational distributions as a function of the average value of the cosine of these angles. The angle  $\theta$  is defined here as the angle between the permanent molecular dipole vector (as determined from the gas-phase dipole moment) and the fixed  $z$ -axis of the simulation cell (the vector perpendicular to the surface). For molecules below the GDS in which  $\theta = 0^\circ$ , both hydrogen sites are directly pointing toward the gas-phase. Molecules having  $\theta = 90^\circ$  are those in which the molecular symmetry axis is parallel to the surface. The second angular component under consideration,  $\phi$  is the angle made between the molecular plane and the  $z$ -axis of the simulation cell. A value  $\cos \phi = 0$  represents a molecule lying parallel to the GDS, while  $\cos \phi = \pm 1$  represents configurations perpendicular to the GDS. Due to the indistinguishability of the two perpendicular configurations, the absolute value of this quantity is considered.

The orientational profiles with respect to the depth relative to the GDS are presented in Figure 6. The  $\langle \cos \theta \rangle$  profile suggests TIP4P-QDP has a stronger orientational preference for the molecules' permanent dipole vector in the outermost portion of the interface than TIP4P-FQ. There is, however, a diminished orientational preference for TIP4P-QDP within the condensed phase. This profile suggests hydrogen atoms point into the gas-phase for values of  $z$  above the GDS, while there is a lesser preference for hydrogen atoms to point into solution below the GDS. Furthermore,



**Figure 7.** Total interfacial potential as a function of  $z$  relative to the center of mass of the water slab.

there is essentially no orientational preference in the condensed-phase region ( $\langle \cos \theta \rangle$ ), which is expected from the bulk isotropic environment. The  $\langle \cos \phi \rangle$  profile demonstrates similar features. Within the condensed-phase region, there is essentially no net structural effect (a value of 0.5 represents the average of the two extreme values for this property). Again, a notable well forms in the region several angstroms below the GDS in the  $\langle \cos \phi \rangle$  profile, suggesting a strong preference for the water to lie parallel to the surface but slightly tipped such that the hydrogens point toward the bulk. A peak in this profile above the GDS indicates a preference for the water molecule to orient perpendicular to the surface; this preference is essentially the same for the different models.

Within the context of orientation profiles, the  $z$ -induced dipole moment can also be discussed. The  $z$ -induced dipole moment was calculated by subtracting the fixed  $z$ -component of the dipole moment for the  $z$ -projection of the total dipole moment for each molecule. The profile of this quantity as a function of  $z$ -position relative to the center of mass is also featured in Figure 6. There is significant (over 50%) increase in the maximum induced dipole moment for TIP4P-QDP over TIP4P-FQ. This increase is expected based on the increased  $\theta$  preference of the TIP4P-QDP model since  $\theta$  is connected to the permanent dipole moment vector.

**5. Interfacial Potential.** The interfacial potential is the potential drop associated with moving a volumeless positive test charge from the vapor phase into solution. This is a measure of the combined electrostatic effects of the water orientation and induced dipole moment distributions at the interface. The liquid–vapor interfacial potential is calculated by the integration of  $z$ -component of charge density<sup>11,33,66,69,104</sup>  $\rho(z)$ , as

$$\Delta \Phi(z) = \Phi(z) - \Phi(z_0) = - \int_{z_0}^z dz' \int_{z_0}^{z'} dz'' \rho(z) \quad (18)$$

where  $z$  is the direction perpendicular to the interface and  $z_0$  is center of mass position of the bulk phase. The integration is performed numerically by evaluation of the charge density for 1 Å-wide segments in the  $z$ -direction. The interfacial potential for TIP4P-QDP converges to  $-11.98 (\pm 0.08)$  kcal/

mol, while a more favorable potential of  $-12.21 (\pm 0.05)$  kcal/mol is calculated for TIP4P-FQ. Currently, there is no experimental consensus on the value, or even the sign of interfacial potential since values have been presented in the range  $\pm 1500$  mV ( $\pm 34.6$  kcal/mol).<sup>105,106</sup> Furthermore, we remark that studies of the aqueous liquid–vapor interface utilizing molecular dynamics have reported more consistent values in the range of  $-400$  to  $-600$  mV (approximately  $-9$  to  $-14$  kcal/mol).<sup>54,69,107–110</sup> We note that the values calculated for all water models in this study fall in the typical range for MD simulations, with the TIP4P-FQ value demonstrating excellent agreement with the previously reported value of  $12.20 (\pm 0.05)$  kcal/mol.<sup>110</sup> Integration of charge density allows for the partitioning of the interfacial potential into dipole and quadrupole moment contributions.<sup>10</sup> The similar dipole contributions for the two models, approximately  $13.74$  kcal/mol, indicate the quadrupole contributions offer the greatest differences between the models. The interfacial potential profile for TIP4P-QDP and TIP4P-FQ is featured in Figure 7. A distinguishing feature of the TIP4P-QDP profile is the exaggerated peak in the region just below the interface, which based on the aforementioned comments is likely a consequence of a locally enhanced quadrupole moment contribution.

**6. Surface Tension.** As a final comparison between the TIP4P-QDP and TIP4P-FQ models in the liquid–vapor interface, the surface tension was calculated from the difference of the normal and tangential elements of the internal pressure tensor<sup>111</sup>

$$\gamma(z) = \frac{L_z}{2} \left( P_{zz} - \frac{P_{xx} + P_{yy}}{2} \right) \quad (19)$$

where  $P_{xx}$ ,  $P_{yy}$ , and  $P_{zz}$  are the diagonal elements of the internal pressure tensor and  $L_z$  is the length of the simulation cell in the  $z$ -direction (normal to the surface). Surface tension computed for TIP4P-QDP model is  $71.0 (\pm 2.7)$  dyne/cm, 2.3% less than the TIP4P-FQ estimate of  $72.7$  dyne/cm and 1.3% less than the experimental value of  $71.9$  dyne/cm. This slight reduction is consistent with the more favorable dimer energy than TIP4P-FQ which allows a reduced energetic penalty for water molecules leaving the condensed phase for the vapor phase.

#### IV. Conclusions

In this work, we have presented a new water model, TIP4P-QDP, which explicitly accounts for the polarizability gradient between thermodynamic phases. The model is built upon the charge equilibration formalism and the TIP4P-FQ model of Rick et al.<sup>1</sup> Although there are numerous possible paths to introduce phase-dependence in the context of molecular dynamics simulations, we chose the development of a multiplicative scaling function that is based on the  $M$ -site partial charge. This notion is based on the considerable gradient in dipole moment from gas to liquid phases, which could be coupled to a change in molecular polarizability between these two phases. An error function was chosen as the functional form due to its ability to apply constant scaling in regions considered purely gaseous or condensed, while providing

nearly linear modulation between these two phases. Moreover, atomic electronegativities were also scaled in order to maintain self-consistency with the hardness scaling. Here additional scaling parameter,  $p$ , was used to control the amount by which  $\chi$  was scaled. Although a value of  $p = 1$  is expected to reduce the charge-dependent expressions to their charge-independent analogs for an isolated molecule, condensed phase effects result in undesirable increases in cohesive forces. Hence, a reduced  $p$ -value allowed for an appropriate average condensed-phase dipole moment and an accurate depiction of intermolecular forces in the condensed phase as suggested by the agreement of these properties with experiment.

A reparameterization of the hardness, electronegativity, and Lennard-Jones parameters was necessary to correct the gas phase polarizability, dipole moment, and dimer energies. Selection of the specific parameter set was based on the ability to simultaneously match the density and enthalpy of vaporization to experiment. Ultimately, the density at ambient conditions,  $0.9954 (\pm 0.0002)$  g/cm<sup>3</sup>, and enthalpy of vaporization,  $10.55 (\pm 0.12)$  kcal/mol, demonstrate excellent agreement with experiment. Isothermal compressibility, diffusion constants, and isobaric heat capacity are also commensurate with the experimental and TIP4P-FQ results. The dielectric constant of  $\epsilon = 85.8$  overestimates experiment by 10%. However, the comparison of TIP4P-QDP as presented in this work to its analog with full scaling on  $\chi$  ( $p = 1$ ) suggests that the average dipole moment in the condensed phase is an important consideration for replicating the appropriate dielectric constant. Interfacial properties are qualitatively similar to the TIP4P-FQ results. Consistent observations made among *charge*-dependent models with different  $p$ -values suggest features that are a direct consequence of accounting for the polarizability gradient between phases. A hydrogen bond network that extends further into the gaseous region than TIP4P-FQ, as well as a corresponding extension of enhanced dipole moments, suggest enhanced cohesion within the interfacial region. This is anticipated on the basis of increased polarizability and more favorable dimer energy of the TIP4P-QDP model. A stronger orientational preference of the TIP4P-QDP model's permanent dipole vector is also a common feature of the QDP-scheme that results in an enhanced  $z$ -induced dipole moment.

Acknowledging that the scaling function used for TIP4P-QDP is not rigorously bound to experimental or quantum mechanical results, we stress that the primary focus of this work is a preliminary analysis of how accounting for the difference in polarizability between phases affects the physics of the liquid–vapor interface. Furthermore, relationships dictating the nature of change of polarizability via a convenient simulation parameter such as  $Q_M$  are not definite, limiting the level to which this model can replicate such phenomena. The coupling of the scaling function to the  $M$ -site charge is a computationally efficient and convenient method of incorporating a phase-dependent molecular polarizability in molecular dynamics simulations. Future generations of phase-dependent models may rely on hydrogen

bond coordination, which has been shown to correspond strongly to both the molecular dipole moment<sup>95,96</sup> and molecular polarizability.<sup>75</sup>

The first step toward exploring the implications of such physics within the context of classical molecular dynamics simulations, we have demonstrated an approach to incorporate phase dependence of molecular polarizability. The resulting TIP4P-QDP model also demonstrates intriguing physical properties, most notably, the enhanced structure of the liquid–vapor interface. One promising approach toward improving this model will be to introduce an additional scaling of the Lennard-Jones parameters.<sup>112</sup> Such modifications may capture more precise structural features that were ignored in the current treatment. This may in turn further improve upon the condensed-phase properties calculated here. Future incorporation of additional atomic sites will allow for treatment of out of plane polarizability, a feature neglected in the original TIP4P-FQ model and the TIP4P-QDP model proposed in this work. Future studies involving TIP4P-QDP (or further refined versions of this model) will focus on how phase-dependent polarizability affects interfacial simulations involving nonpolarizable and polarizable ions. Furthermore, the extent to which this model can replicate the liquid–vapor coexistence curve will also be studied. The dynamical nature of polarizability as exhibited in TIP4P-QDP is a crucial element that has been lacking in studies involving the latter, and it is therefore suggested that the TIP4P-QDP model (and future versions) may have success in such applications.

**Acknowledgment.** The authors gratefully acknowledge support from the National Institute of Health sponsored COBRE (Center of Biomedical Research) Grant P20-RR015588 (Department of Chemical Engineering) at the University of Delaware. One of the authors, S.P., also acknowledges the University of Delaware for startup funds.

## Appendix: Derivation of QDP Expressions

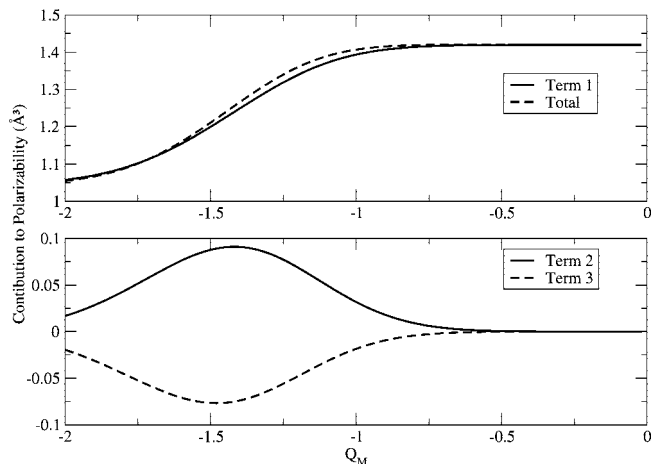
The charge-dependent polarizable water model introduces a multiplicative scaling factor to each term in the Hamiltonian. In addition to the  $Q$ -dependent scaling it is desired to control the magnitude of scaling on the chi parameter, which is done by the introduction of an empirical  $p$ -parameter. The generalized energy expression with  $Q$ -dependent hardness and  $Q, p$ -dependent electronegativities is given by

$$E(\mathbf{Q}) = \langle \chi(\mathbf{Q}, p) | \mathbf{Q} \rangle + \frac{1}{2} \langle \mathbf{Q} | \mathbf{J}(\mathbf{Q}, \mathbf{r}) | \mathbf{Q} \rangle \quad (\text{A1})$$

By taking a derivative with respect to the  $k$ th charge and setting the resulting expression

$$\nabla_k E(\mathbf{Q}) = \chi_k(\mathbf{Q}, p) + \langle \mathbf{J}_k(\mathbf{Q}, \mathbf{r}) | \mathbf{Q} \rangle + \left\langle \nabla_k \chi(\mathbf{Q}, p) + \frac{1}{2} \mathbf{Q} [\nabla_k \mathbf{J}(\mathbf{Q}, \mathbf{r})] | \mathbf{Q} \right\rangle \quad (\text{A2})$$

equal to zero, we may obtain a set of simultaneous equations, one for each charge. In the above equation,  $\chi_k$  is the  $k$ th site electronegativity and  $\mathbf{J}_k$  is the  $k$ th row of the hardness matrix.



**Figure 8.** Contribution of each term in eq A18 to the total scaling function,  $g(Q_M)$ . The top panel features the first term of eq A18 (solid line) as it compares to the total scaling function (dashed line) as a function of  $Q_M$ . The lower panel features the second (solid line) and third (dashed line) terms of eq A18. Terms 2 and 3 are approximately equal in magnitude while opposite in sign, which results in the first term's dominance in  $g(Q_M)$ .

The first two terms in the above expression are the usual terms one obtains when  $\chi$  and  $\mathbf{J}$  are independent of  $\mathbf{Q}$ . The last term accounts for the charge dependence of  $\chi$  and  $\mathbf{J}$ . We may now make the specific assumptions

$$\chi(\mathbf{Q}, p) = \chi[(1-p) + pg(Q_M)] \quad (\text{A3})$$

and

$$\mathbf{J}(\mathbf{Q}, \mathbf{r}) = \mathbf{J}(\mathbf{r})g(Q_M) \quad (\text{A4})$$

which tie the scaling of the hardnesses and electronegativities to a factor  $g(Q_M)$  based on the  $M$ -site partial charge. Introduction of the  $p$  dependence in eq A3 acts to control the extent of scaling due to the  $g(Q_M)$  term; a value of  $p = 0$  introduces no  $Q_M$ -scaling, while a value of  $p = 1$  introduces full scaling equivalent to that applied to the hardness matrix.

If we now consider a single water molecule and the cases in which  $k$  corresponds to a hydrogen site, only the first two terms of eq A2 survive (since  $\chi$  and  $\mathbf{J}$  only depend on  $Q_M$ ) and we may write

$$-\chi_H[(1-p) + pg(Q_M)] = g(Q_M) \langle \mathbf{J}_H(\mathbf{r}) | \mathbf{Q} \rangle \quad (\text{A5})$$

Letting  $h(Q_M) = g(Q_M)/[(1-p) + pg(Q_M)]$  and rearranging, we obtain the simplified expression

$$-\chi_H = h(Q_M) \langle \mathbf{J}_H(\mathbf{r}) | \mathbf{Q} \rangle \quad (\text{A6})$$

which, in the limit  $p = 1$ , reduces to the usual expression

$$-\chi_H = \langle \mathbf{J}_H(\mathbf{r}) | \mathbf{Q} \rangle \quad (\text{A7})$$

For the  $M$ -site ( $k = M$ ) and using the fact that  $\nabla_M[(1-p) + pg(Q_M)] = p \nabla_M g(Q_M)$ , we obtain

$$-\chi_M = h(Q_M) \langle \mathbf{J}_M(\mathbf{r}) | \mathbf{Q} \rangle + \left[ h(Q_M) \left( \frac{\nabla_M g(Q_M)}{g(Q_M)} \right) \times \left( p \chi_M \langle \hat{\mathbf{M}} | + \frac{1}{2} \langle \mathbf{Q} | \mathbf{J}(\mathbf{r}) | \right) | \mathbf{Q} \right] \quad (\text{A8})$$



where  $\langle \hat{\mathbf{M}} |$  is a unit vector that picks out the  $M$ -site charge. The full system of equations is constructed from the equations for both hydrogen sites and the  $M$ -site as

$$-|\chi\rangle = h(Q_M) \left[ \mathbf{J}(\mathbf{r}) + \left( \frac{\nabla_M g(Q_M)}{g(Q_M)} \right) \times \left( p\chi_M \mathbf{M} + \frac{1}{2} |\hat{\mathbf{M}}\rangle \langle \mathbf{Q} | \mathbf{J}(\mathbf{r}) | \right) \right] |\mathbf{Q}\rangle = h(Q_M) \mathbf{H}(\mathbf{Q}, \mathbf{r}) |\mathbf{Q}\rangle \quad (\text{A9})$$

where the matrix  $\mathbf{M}$  denotes the outer product  $|\hat{\mathbf{M}}\rangle \langle \hat{\mathbf{M}}|$ . This form allows us to recover eqs A6 and A8 upon dotting a particular unit vector such as  $\langle \hat{\mathbf{M}} |$ ,  $\langle \hat{\mathbf{H}}_1 |$ , or  $\langle \hat{\mathbf{H}}_2 |$  into this expression from the left. To compute the polarizability, we make the assumption

$$\mathbf{H}(\mathbf{Q}, \mathbf{r}) \approx \mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r}) \quad (\text{A10})$$

which constructs a static (charge-independent) version of the hardness matrix from the true equilibrium charges. In the limit of no external field and  $p = 1$  ( $h(Q_M) = 1$ ), this assumption is exact since the same equilibrium charges are recovered and self-consistent:

$$h(Q_M) \mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r}) |\mathbf{Q}\rangle = -|\chi\rangle \quad (\text{A11})$$

$$\langle \mathbf{Q} \rangle = -[\mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r})]^{-1} |\chi\rangle$$

In the presence of an external field, this assumption essentially requires that the charge-induced changes to the hardness matrix be small which should be a valid approximation in the weak field limit. This makes it unnecessary to iteratively solve the nonlinear system of equations. Therefore, the equilibrium charges in the presence of some external field  $\epsilon_\gamma$  are given by

$$h(Q_M) \mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r}) |\mathbf{Q}\rangle_\epsilon \approx -|\chi\rangle + \frac{1}{g(Q_M)} \epsilon_\gamma |\mathbf{R}_\gamma\rangle$$

$$\langle \mathbf{Q} \rangle_\epsilon \approx -[h(Q_M) \mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r})]^{-1} |\chi\rangle + \left( \frac{[\mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r})]^{-1}}{g(Q_M) h(Q_M)} \right) |\epsilon_\gamma \mathbf{R}_\gamma\rangle \quad (\text{A12})$$

The induced dipole in the  $\beta$  direction is given by the difference in charges due to the presence of the field

$$\mu_\beta^{\text{ind}} = \langle \mathbf{R}_\beta | \mathbf{Q} \rangle_\epsilon - \langle \mathbf{R}_\beta | \mathbf{Q} \rangle \approx \langle \mathbf{R}_\beta | \frac{[\mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r})]^{-1}}{g(Q_M) h(Q_M)} |\epsilon_\gamma \mathbf{R}_\gamma\rangle \quad (\text{A13})$$

and the derivative of the induced dipole moment with respect to the field then yields the  $\beta\gamma$  component of the polarizability:

$$\alpha_{\beta\gamma}(Q_M) = \frac{\partial \mu_\beta^{\text{ind}}}{\partial \epsilon_\gamma} \approx \frac{1}{g(Q_M) h(Q_M)} \langle \mathbf{R}_\beta | [\mathbf{H}(\langle \mathbf{Q} \rangle, \mathbf{r})]^{-1} |\mathbf{R}_\gamma\rangle \quad (\text{A14})$$

By manipulating the definition of  $\mathbf{H}(\mathbf{Q}, \mathbf{r})$  in eq A9, we may obtain an expression for the inverse as

$$[\mathbf{H}(\mathbf{Q}, \mathbf{r})]^{-1} = \mathbf{J}^{-1}(\mathbf{r}) \left[ 1 + \left( \frac{\nabla_M g(Q_M)}{g(Q_M)} \right) \times \left( p\chi_M \mathbf{M} \mathbf{J}^{-1}(\mathbf{r}) + \frac{1}{2} |\hat{\mathbf{M}}\rangle \langle \mathbf{Q} | \right) \right]^{-1} \quad (\text{A15})$$

which relates the charge-dependent hardness matrix to the charge-independent hardness matrix at leading order. Assuming the correction is small, we can perform a series expansion of the bracketed term to yield

$$[\mathbf{H}(\mathbf{Q}, \mathbf{r})]^{-1} = \mathbf{J}^{-1}(\mathbf{r}) \left[ 1 - \left( \frac{\nabla_M g(Q_M)}{g(Q_M)} \right) \times \left( p\chi_M \mathbf{M} \mathbf{J}^{-1}(\mathbf{r}) + \frac{1}{2} |\hat{\mathbf{M}}\rangle \langle \mathbf{Q} | \right) \right] \quad (\text{A16})$$

Insertion of this definition into the polarizability expression yields

$$\alpha_{\beta\gamma}(Q_M) \approx \frac{1}{g(Q_M) h(Q_M)} \left\langle \mathbf{R}_\beta | \mathbf{J}^{-1}(\mathbf{r}) \left[ 1 - \left( \frac{\nabla_M g(Q_M)}{g(Q_M)} \right) \times \left( p\chi_M \mathbf{M} \mathbf{J}^{-1}(\mathbf{r}) + \frac{1}{2} |\hat{\mathbf{M}}\rangle \langle \mathbf{Q} | \right) \right] | \mathbf{R}_\gamma \right\rangle \quad (\text{A17})$$

which expresses the polarizability in terms of the original, unscaled matrix  $\mathbf{J}(\mathbf{r})$ . Thus, when the gradient of  $g(Q_M)$  is zero (constant  $g(Q_M)$ ) for  $p = 1$  ( $h(Q_M) = 1$ ), the new contribution disappears so that the usual expression is recovered to within a multiplicative constant depending on the constant scaling factor applied to the hardnesses. Since  $\mathbf{J}(\mathbf{r})$  is symmetric, this expression may be further simplified into

$$\alpha_{\beta\gamma}(Q_M) \approx \frac{\langle \mathbf{R}_\beta | \mathbf{J}^{-1}(\mathbf{r}) | \mathbf{R}_\gamma \rangle}{g(Q_M) h(Q_M)} - \left( \frac{\nabla_M g(Q_M)}{[g(Q_M)]^2 h(Q_M)} \right) \times \left[ p\chi_M \langle \mathbf{R}_\beta | \mathbf{J}^{-1}(\mathbf{r}) | \hat{\mathbf{M}} \rangle^2 + \frac{1}{2} \langle \mathbf{R}_\beta | \mathbf{J}^{-1}(\mathbf{r}) | \hat{\mathbf{M}} \rangle \langle \mathbf{Q} | \mathbf{R}_\gamma \rangle \right] \quad (\text{A18})$$

The first term is the usual polarizability expression, but scaled by a factor which becomes  $1/g(Q_M)$  in the limit  $p = 1$ . The second term always *increases* the polarizability if  $\chi_M$  is positive and the gradient is negative since the functions  $g(Q_M)$  and  $h(Q_M)$  are always positive. The last term works to decrease the polarizability with increasing magnitude of the dipole moment. While all terms should be included in order to accurately estimate the  $Q$ -dependent polarizability, the last two terms are typically an order of magnitude smaller than the first term and opposite in sign. The magnitude of each term as a function of  $Q_M$  is featured in Figure 8. Finally, we remark that eq 7 is a slightly modified form of eq A18, utilizing the substitutions  $\alpha_{\beta\gamma} = \langle \mathbf{R}_\beta | \mathbf{J}^{-1}(\mathbf{r}) | \mathbf{R}_\gamma \rangle$  and  $\mu_\gamma = \langle \mathbf{Q} | \mathbf{R}_\gamma \rangle$ .

## References

- (1) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141.
- (2) Matsumoto, M.; Kataoka, Y. *J. Chem. Phys.* **1989**, *90*, 2390.
- (3) Matsumoto, M.; Kataoka, Y. *J. Chem. Phys.* **1988**, *88*, 3233.
- (4) Petersen, P. B.; Saykally, R. J. *J. Chem. Phys. Lett.* **2004**, 397, 51.
- (5) Petersen, P. B.; Saykally, R. J. *Annu. Rev. Phys. Chem.* **2006**, *57*, 333.
- (6) Petersen, P. B.; Saykally, R. J.; Mucha, M.; Jungwirth, P. J. *Phys. Chem. B* **2005**, *109*, 10915.

- (7) Walker, D. S.; Richmond, G. L. *J. Am. Chem. Soc.* **2007**, *129*, 9446–9451.
- (8) Walker, D. S.; Richmond, G. L. *J. Phys. Chem. C* **2007**, *111*, 8321–8330.
- (9) Foster, K. L.; Plastringe, R. A.; Bottenheim, J. W.; Shepson, P. B.; Finlayson-Pitts, B. J.; Spicer, C. W. *Science* **2001**, *291*, 471.
- (10) Wilson, M. A.; Pohorille, A.; Pratt, L. R. *J. Chem. Phys.* **1989**, *90*, 5211.
- (11) Wilson, M. A.; Pohorille, A.; Pratt, L. R. *J. Chem. Phys.* **1988**, *88*, 3281.
- (12) Shen, Y. R. *Nature* **1989**, *337*, 519.
- (13) Salafsky, J. S.; Eienthal, K. B. *Chem. Phys. Lett.* **2000**, *319*, 435.
- (14) Salafsky, J. S.; Eienthal, K. B. *J. Phys. Chem. B* **2000**, *104*, 7752.
- (15) Vrbka, L.; Mucha, M.; Minofar, B.; Jungwirth, P.; Brown, E. C.; Tobias, D. J. *Curr. Opin. Colloid Interface Sci.* **2004**, *9*, 67.
- (16) Kuo, I. W.; Mundy, C. J.; Eggiman, B. L.; McGrath, M. J.; Siepmann, J. I.; Chen, B.; Viecelli, J.; Tobias, D. J. *J. Phys. Chem. B* **2006**, *110*, 3738.
- (17) Kuo, I. W.; Mundy, C. J. *Science* **2004**, *303*, 658.
- (18) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.
- (19) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.
- (20) Chirlian, L. E.; Fancl, M. M. *J. Comput. Chem.* **1987**, *8*, 894.
- (21) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (22) Bucher, D.; Raugei, S.; Guidoni, L.; DalPeraro, M.; Rothlisberger, U.; Carloni, P.; Klein, M. L. *Biophys. Chem.* **2006**, *124*, 292.
- (23) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Bio.* **2001**, *11*, 236.
- (24) Patel, S.; Brooks, C. L., III. *Mol. Sim.* **2006**, *32*, 231.
- (25) Koch, D. M.; Peslherbe, G. H. *J. Phys. Chem. B* **2008**, *112*, 636.
- (26) Smith, D. E.; Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 3757.
- (27) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 2481–2486.
- (28) Perera, L.; Berkowitz, M. L. *Z. Phys. D* **1993**, *26*, 166–168.
- (29) Perera, L.; Berkowitz, M. L. *J. Chem. Phys.* **1992**, *96*, 8288.
- (30) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2005**, *110*, 3308.
- (31) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671.
- (32) Dang, L. X. *J. Chem. Phys.* **1992**, *97*, 2659.
- (33) Dang, L. X.; Chang, T. M. *J. Chem. Phys.* **1997**, *106*, 8149.
- (34) Dang, L. X.; Chang, T. M. *J. Chem. Phys.* **2003**, *119*, 9851.
- (35) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.
- (36) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; Alexander, D.; MacKerell, J. *J. Chem. Theory Comput.* **2005**, *1*, 153.
- (37) Vorobyov, I. V.; Anisimov, V. M.; Alexander, D.; MacKerell, J. *J. Phys. Chem. B* **2005**, *109*, 18988.
- (38) Patel, S.; Brooks, C. L., III. *J. Comp. Chem.* **2004**, *25*, 1.
- (39) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672.
- (40) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *110*, 741.
- (41) Stern, H.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730.
- (42) Stern, H.; Rittner, F.; Berne, B. J.; Friesner, R. *J. Chem. Phys.* **2001**, *115*, 2237.
- (43) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.
- (44) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515.
- (45) Ledecq, M.; Lebon, F.; Durant, F.; Giessner-Prettre, C.; Marquez, A.; Gresh, N. *J. Phys. Chem. B* **2003**, *107*, 10640.
- (46) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856.
- (47) Gresh, N.; Garmer, D. R. *J. Comput. Chem.* **1996**, *17*, 1481.
- (48) Patel, S.; MacKerell, A. D.; Brooks, C. L., III. *J. Comp. Chem.* **2004**, *25*, 1504.
- (49) Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824.
- (50) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960.
- (51) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.
- (52) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.
- (53) Vorobyov, I. V.; Anisimov, V. M.; Greene, S.; Moser, R. M. V. A.; Pastor, R. W.; Alexander, D.; MacKerell, J. *J. Chem. Theory Comput.* **2007**, *3*, 1120.
- (54) Lamoureux, G., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.
- (55) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.
- (56) Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315.
- (57) Mortier, W. J.; Genechten, K. V.; Gasteiger, J. *J. Am. Chem. Soc.* **1985**, *107*, 829.
- (58) Rappé, A. K.; Goddard, W. A., III. *J. Phys. Chem.* **1991**, *95*, 3358.
- (59) Rick, S. W. *J. Chem. Phys.* **2001**, *114*, 2276.
- (60) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. J. *J. Mol. Liq.* **1995**, *65/66*, 31.
- (61) Olano, L. R.; Rick, S. W. *J. Comput. Chem.* **2005**, *26*, 699.
- (62) Patel, S.; Brooks, C. L., III. *J. Chem. Phys.* **2006**, *124*, 204706.
- (63) Nalewajski, R. F.; Korchowicz, J.; Zhou, Z. *Int. J. Quantum Chem.* **1988**, *22*, 349.
- (64) Patel, S.; Brooks, C. L., III. *J. Chem. Phys.* **2005**, *122*, 24508.
- (65) Ribeiro, M. C. C.; Almeida, L. C. J. *J. Chem. Phys.* **1999**, *110*, 11445.
- (66) Patel, S.; Brooks, C. L., III. *J. Chem. Phys.* **2005**, *123*, 164502.
- (67) Zhong, Y.; Warren, G. L.; Patel, S. *J. Comput. Chem.* **2008**, *29*, 1142.

- (68) Rick, S. W.; Stewart, S. J. Potentials and Algorithms for Incorporating Polarizability in Computer Simulations. In *Reviews of Computational Chemistry*; Lipkowitz, K. B. Boyd, D. B. Eds.; John Wiley & Sons: New York, 2002; p 89.
- (69) Dang, L. X.; Chang, T.-M. *J. Phys. Chem. B* **2002**, *106*, 235.
- (70) Yu, H.; Geerke, D. P.; Liu, H.; v. Gunsteren, W. F. *J. Comput. Chem.* **2006**, *27*, 1494.
- (71) Dang, L. X. *J. Phys. Chem. B* **2001**, *105*, 804.
- (72) Dang, L. X.; Chang, T. M.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **2002**, *117*, 3522.
- (73) Piquemal, J.-P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511.
- (74) Morita, A. *J. Comput. Chem.* **2002**, *23*, 1466.
- (75) Krishtal, A.; Senet, P.; Yang, M.; van Alsenoy, C. *J. Chem. Phys.* **2006**, *125*, 034312.
- (76) Schropp, B.; Tavan, P. *J. Phys. Chem. B* **2008**, *112*, 6233.
- (77) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308.
- (78) Badyal, Y. S.; Saboungi, M. L.; Price, D. L.; Shastri, S. D.; Haefner, D. R.; Soper, A. K. *J. Chem. Phys.* **2000**, *112*, 9206.
- (79) Sanderson, R. T. *Chemical Bonds and Bond Energy*; Academic Press: New York, 1976.
- (80) Sanderson, R. T. *Science* **1951**, *114*, 670.
- (81) Chelli, R.; Procacci, P. *J. Chem. Phys.* **2002**, *117*, 9175.
- (82) Itskowitz, P.; L.Berkowitz, M. *J. Chem. Phys.* **1997**, *101*, 5687.
- (83) Warren, G. L.; Davis, J. E.; Patel, S. *J. Chem. Phys.* **2008**, *128*, 144110.
- (84) Bauer, B. A.; Patel, S. *J. Mol. Liq.* **2008**, *142*, 32.
- (85) Sprik, M. *J. Chem. Phys.* **1991**, *95*, 6762.
- (86) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; Stages, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (87) Brooks, C. L., III.; Karplus, M.; Pettitt, B. M. *A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*; John Wiley & Sons: New York, 1988; Vol. LXXI.
- (88) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- (89) Nose, S. *Mol. Phys.* **1984**, *52*, 255.
- (90) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (91) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665.
- (92) Price, D. J.; Brooks, C. L., III. *J. Chem. Phys.* **2004**, *121*, 10096.
- (93) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910.
- (94) Alper, H. E.; Levy, R. M. *J. Chem. Phys.* **1989**, *91*, 1242.
- (95) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J. *Mol. Phys.* **2007**, *105*, 1411.
- (96) Kemp, D. D.; Gordon, M. S. *J. Phys. Chem. A* **2008**, *112*, 4885.
- (97) Soper, A. K.; Phillips, M. G. *Chem. Phys.* **1986**, *107*, 47.
- (98) Soper, A. K. *J. Phys.: Condens. Matter* **2007**, *19*, 335206.
- (99) Miller, T. F., III.; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *123*, 154504.
- (100) Beaglehole, D. *Fluid Interfacial Phenomena*; Wiley: New York, 1986.
- (101) Liu, P.; Harder, E.; Berne, B. J. *J. Phys. Chem. B* **2005**, *109*, 2949.
- (102) Benjamin, I. *Chem. Rev.* **1996**, *96*, 1449.
- (103) Starr, F.; Nielsen, J.; Stanley, H. *Phys. Rev.* **2000**, *62*, 579.
- (104) Pandit, S. A.; Bostick, D.; Berkowitz, M. L. *Biophys. J.* **2003**, *85*, 3120.
- (105) Parfenyuk, V. I. *Colloid J.* **2002**, *64*, 588.
- (106) Paluch, M. *Adv. Colloid Interface Sci.* **2000**, *84*, 27.
- (107) Warren, G. L.; Patel, S. *J. Chem. Phys.* **2007**, *064509*, 127.
- (108) Ishiyama, T.; Morita, A. *J. Phys. Chem. C* **2007**, *111*, 721.
- (109) Sokhan, V. P.; Tildesley, D. J. *Mol. Phys.* **1997**, *92*, 625.
- (110) Warren, G. L.; Patel, S. *J. Phys. Chem. B* **2008**, *112*, 11679.
- (111) Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1949**, *17*, 338.
- (112) Chen, B.; Xing, J.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 2191.
- (113) Clough, S. A.; Beers, Y.; Klein, G. P.; Rothman, L. S. *J. Chem. Phys.* **1973**, *59*, 2254.
- (114) Murphy, W. F. *J. Chem. Phys.* **1977**, *67*, 5877.
- (115) Odutola, J. A.; Dyke, T. R. *J. Chem. Phys.* **1980**, *72*, 5062.
- (116) *CRC Handbook of Chemistry and Physics*, 77th ed.; Lide, D. R., Ed.; CRC: Boca Raton, FL, 1997.
- (117) Krynicki, K.; Green, C. D.; Sawyer, D. W. *Discuss. Faraday Soc.* **1978**, *66*, 199.
- (118) Buckingham, A. D. *Proc. R. Soc. London Ser. A* **1956**, *238*, 235.
- (119) Watanabe, K.; Klein, M. L. *Chem. Phys.* **1989**, *131*, 157.
- (120) Cini, R.; Logio, G.; Ficalbi, A. *J. Colloid Interface Sci.* **1972**, *41*, 287.

## Theoretical Study of Hydrogen Storage in Ca-Coated Fullerenes

Qian Wang,<sup>†,‡</sup> Qiang Sun,<sup>\*,‡,§</sup> Puru Jena,<sup>‡</sup> and Yoshiyuki Kawazoe<sup>||</sup>

*School of Physical Science and Technology, Southwest University, Chongqing 400715, China, Department of Physics, Virginia Commonwealth University, Richmond, Virginia 23284, Department of Advanced Materials and Nanotechnology and Center for Applied Physics and Technology, Peking University, Beijing 100871, China, and Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan*

Received September 6, 2008

**Abstract:** First principles calculations based on gradient corrected density functional theory and molecular dynamics simulations of Ca decorated fullerene yield some novel results: (1) C<sub>60</sub> fullerene decorated with 32 Ca atoms on each of its 20 hexagonal and 12 pentagonal faces is extremely stable. Unlike transition metal atoms that tend to cluster on a fullerene surface, Ca atoms remain isolated even at high temperatures. (2) C<sub>60</sub>Ca<sub>32</sub> can absorb up to 62 H<sub>2</sub> molecules in two layers. The first 30 H<sub>2</sub> molecules dissociate and bind atomically on the 60 triangular faces of the fullerene with an average binding energy of 0.45 eV/H, while the remaining 32 H<sub>2</sub> molecules bind on the second layer quasi-molecularly with an average binding energy of 0.11 eV/H<sub>2</sub>. These binding energies are ideal for Ca coated C<sub>60</sub> to operate as a hydrogen storage material at near ambient temperatures with fast kinetics. (3) The gravimetric density of this hydrogen storage material can reach 6.2 wt %. Simple model calculations show that this density is the limiting value for higher fullerenes.

### Introduction

Hydrogen, the least complex and the most abundant element in the universe, is an energy carrier that is expected to play a critical role in a new, decentralized energy infrastructure with many important advantages over other fuels. Unlike fossil fuels such as oil, natural gas, and coal that contain carbon, produce CO<sub>2</sub>, contribute to global warming, and have limited supply, hydrogen is clean, abundant, nontoxic, renewable, and packs more energy per unit mass than any other fuel. However, the biggest challenge in a new hydrogen economy is finding materials that can store hydrogen with high gravimetric and volumetric density under favorable thermodynamic conditions and exhibit fast kinetics.<sup>1–7</sup> The current methods of storing hydrogen as compressed gas or in the liquid form does not meet the industry requirements

since the energy densities are much lower than that in gasoline. Moreover, there are issues of safety and cost involved in compressing hydrogen under high pressure or liquefying it at cryogenic temperatures.

Although storage of hydrogen in solid state materials offers an alternative, currently there are no materials that meet the industry requirement. This is because materials to store hydrogen with high gravimetric density (e.g., 9 wt %) have to be lighter than aluminum. Unfortunately, in these elements, hydrogen is bound either strongly as in complex light metal hydrides or weakly as in carbon based nanostructures, clathrates, zeolites, and metal organic frameworks.<sup>8–14</sup> The early promise of carbon nanotubes<sup>8</sup> as high density storage materials has not materialized.<sup>15–17</sup> Attention has, therefore, turned to the functionalized carbon fullerenes and nanotubes<sup>18–23</sup> where transition metal atoms uniformly distributed over the surface were shown to bind copious amounts of hydrogen in a quasi-molecular form through a novel mechanism where the adsorbed H<sub>2</sub> molecule donates electrons to the unfilled d-orbitals of the transition metals atoms which in turn back-

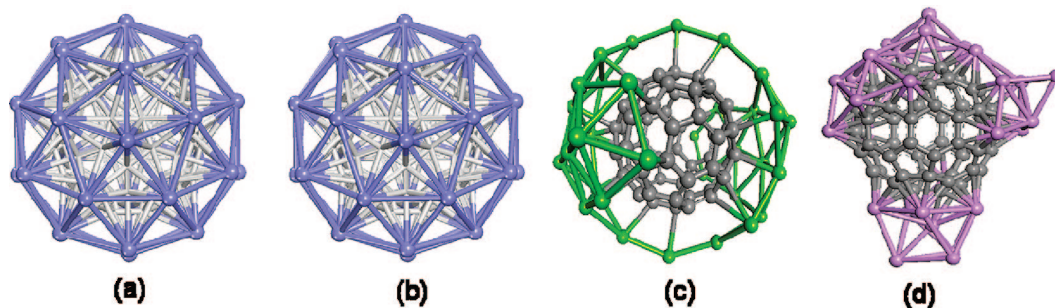
\* Corresponding author e-mail: sunq@coe.pku.edu.cn.

<sup>†</sup> Southwest University.

<sup>‡</sup> Virginia Commonwealth University.

<sup>§</sup> Peking University.

<sup>||</sup> Tohoku University.



**Figure 1.** (a) Optimized geometry of  $C_{60}Ca_{32}$ ; (b) geometry of  $C_{60}Ca_{32}$  after 5 ps MD simulation; (c) geometry of  $C_{60}Mg_{32}$  after 0.4 ps MD simulation; and (d) geometry of  $C_{60}Li_{32}$  after 0.5 ps MD simulation.

donate the electron to the antibonding orbital of the  $H_2$  molecule. Consequently, the  $H_2$  molecule does not dissociate but binds quasi-molecularly with a stretched H–H bond. The binding energy of about 0.5 eV/ $H_2$  molecule is in the ideal range for room temperature applications. Later studies<sup>21</sup> showed that these materials are not stable as the strong d-d interaction between transition metal atoms leads to clustering, which greatly affects their hydrogen storing capacity. Although Li atoms in  $C_{60}Li_{12}$  do not cluster due to strong Li–C bond and weak Li–Li bond,<sup>22</sup> the absorption energy of  $H_2$  is weak and hydrogen desorbs at low temperatures.

In this study we show that  $C_{60}Ca_{32}$  does not suffer from any of these shortcomings. First, it is a very stable cluster whose magicity has been established from gas-phase experiments.<sup>24</sup> Martin and co-workers found this cluster to have a conspicuous peak in the mass spectra which is characteristic of a magic cluster with high stability. Second, Ca atoms show no tendency for clustering. Third, this nanocluster can bind up to 124 hydrogen atoms with an average binding energy in the required range (0.1–1.0 eV) and with a weight percentage that can reach the Department of Energy’s 2010 target, namely 6 wt %.

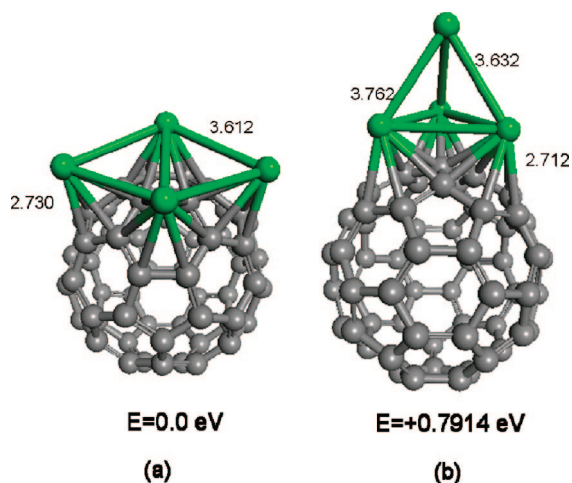
The above results are based on first principles calculations using density functional theory and generalized gradient approximation for exchange and correlation. We used a supercell approach where the cluster was surrounded by 15 Å of vacuum space along x, y, and z directions. The  $\Gamma$  point was used to represent the Brillouin zone due to the large supercell. The total energies and forces and optimizations of geometry were carried out using a plane-wave basis set with the projector augmented plane wave (PAW) method as implemented in the Vienna Ab initio Simulation Package (VASP).<sup>25</sup> The PW91 form was used for the generalized gradient approximation to exchange and correlation potential. The geometries of clusters were optimized without symmetry constraint using conjugate-gradient algorithm. The energy cutoff and the convergence in energy and force were set to 400 eV,  $10^{-4}$  eV, and  $1 \times 10^{-3}$  eV/Å, respectively. The accuracy of our numerical procedure for  $C_{60}$  and hydrogen has been demonstrated in our previous papers.<sup>21,22,26</sup> For the  $Ca_2$  molecule, we obtained the ground-state to be  $^1\Sigma_g^+$  with a bond length of 4.20 Å, in good agreement with the experimental value of 4.277 Å.<sup>27</sup>

Our fully optimized geometry of  $C_{60}Ca_{32}$  structure without symmetry constraint shown in Figure 1(a) agrees with the previous theoretical study.<sup>28</sup> Here the C–Ca and Ca–Ca

distances are respectively 2.752 Å and 3.681 Å. The bond lengths between C atoms at the pentagon-hexagon and hexagon-hexagon interface are respectively 1.465 and 1.446 Å. To confirm the stability of this structure, we have carried out molecular dynamics simulation by using Nose algorithm<sup>29</sup> at room temperature ( $T=300$  K) with 0.5 fs time steps. After 5 ps simulation, we found that the structure retains its identity. The resulting geometry is shown in Figure 1(b) which, except for some small fluctuations in bond length due to the thermal motion of atoms at finite temperature, is essentially the same as that in Figure 1(a). Based on these results, we can conclude that  $C_{60}Ca_{32}$  is indeed very stable. However, the HOMO–LUMO gap of this complex structure is almost zero, and the system is nearly metallic. Thus, the observed stability is not of electronic origin but purely due to the geometric effects,<sup>23,28</sup> which is different from what happened in  $Ca@C_{60}$ .<sup>30</sup> To further verify the stability of  $C_{60}Ca_{32}$ , we also performed simulated annealing from 300 K to 0 K, by starting from the geometry of Figure 1(b). After 5 ps simulations, it recovers back to the structure of Figure 1(a), indicating again that the later structure has high stability.

To further prove that Ca atoms do not cluster on the  $C_{60}$  surface as Ti atoms were found<sup>21</sup> to do, we carried out two separate calculations: In the first we placed four Ca atoms on the neighboring hexagonal and pentagonal sites of the  $C_{60}$  surface and second, the four atoms forming a tetrahedron. The optimized geometries with these as starting configurations are shown in Figure 2(a,b). Note that the configuration where the Ca atoms form a tetrahedron is 0.79 eV higher in energy than when they occupy the hollow sites on the  $C_{60}$  surface, providing ample evidence that Ca atoms do not cluster on  $C_{60}$ , and the core–shell-like geometry of  $C_{60}Ca_{32}$  is more stable. In the equilibrium geometry of  $C_{60}Ca_{32}$ , about 0.4 electrons per Ca atom are transferred to fullerene core. In Figure 3(a) we show these changes in charge distribution where yellow stands for missing charge and blue for charge gained. We also find that HOMO and LUMO originate mainly from the Ca coating shell, as shown in Figure 3(b,c). Thus, it is the Ca shell that would take part in any chemical activity.

Next we studied the interaction of a single hydrogen molecule with  $C_{60}Ca_{32}$  by considering three different configurations as shown in Figure 4. In the top configuration (Figure 4(a)),  $H_2$  is initially placed on top of a Ca atom with H–H and H–Ca distances set to 0.74 and 2.0 Å, respectively. Upon full optimization, the distance between  $H_2$  and



**Figure 2.** The optimized structures of  $\text{Ca}_4\text{C}_{60}$  with (a) the Ca atoms occupying the adjacent hollow sites and (b) the four forming a tetrahedron characteristic of a clustered configuration.

the Ca shell became 2.79 Å (Figure 4(b)). The hydrogen atoms remain molecularly bound, and the corresponding absorption energy is only 0.05 eV. In the bridge configuration (Figure 4(c)),  $\text{H}_2$  was placed on the Ca–Ca bridge with initial H–H and H–Ca distances of 0.74 and 2.03 Å, respectively. However after optimization, the hydrogen molecule is found to dissociate and atomically bound at the centers of a Ca–Ca–Ca triangle with a binding energy of 0.74 eV/H (Figure 4(e)). This is much less than the binding energy in conventional metal hydrides. When  $\text{H}_2$  is introduced on the hollow site of a Ca–Ca–Ca triangle (Figure 4(d)), the geometry also converged to that in Figure 4(e). Thus, in the preferred configuration, the hydrogen atoms bind dissociatively.

Now we consider an extreme situation, where all 60 triangles are occupied by H atoms. The optimized structure is given in Figure 5(a), where due to the insertion of H atoms, Ca–Ca and Ca–C distances have expanded to 3.75 and 2.86 Å, respectively, from the initial values of 3.68 and 2.75 Å. This increase in distance between fullerene-core and Ca-shell causes the C–C bond lengths to change from 1.465 and 1.446 Å to 1.447 and 1.425 Å, respectively. The binding energy of H atoms now reduces to 0.45 eV/H and is in the ideal thermodynamic range for the hydrogen storage materials' application in the mobile industry. With 60 H atoms bound to  $\text{C}_{60}\text{Ca}_{32}$ , the gravimetric density amounts to 3.0 wt %.

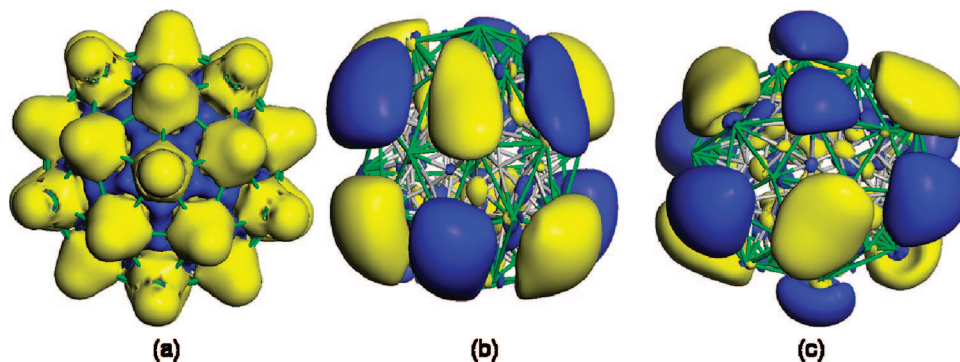
In Figure 5(b) we plot the charge distribution in the  $\text{C}_{60}\text{Ca}_{32}\text{H}_{60}$  complex. Here yellow represents missing charge, and blue for charge gained. Due to further charge transfer from Ca to H, Ca sites become more positively charged. In fact, each Ca site carries a charge of +1.30 e, and each H site carries a charge of –0.535 e.

The significant positive charge on the Ca atoms allows the possibility that further hydrogen atoms may be bound to the  $\text{C}_{60}\text{Ca}_{32}$  cluster. We note that Rao and Jena<sup>31–33</sup> had shown more than a decade ago that a positively charged atom can bind a large amount of hydrogen in quasi-molecular form through the charge polarization mechanism. To see if

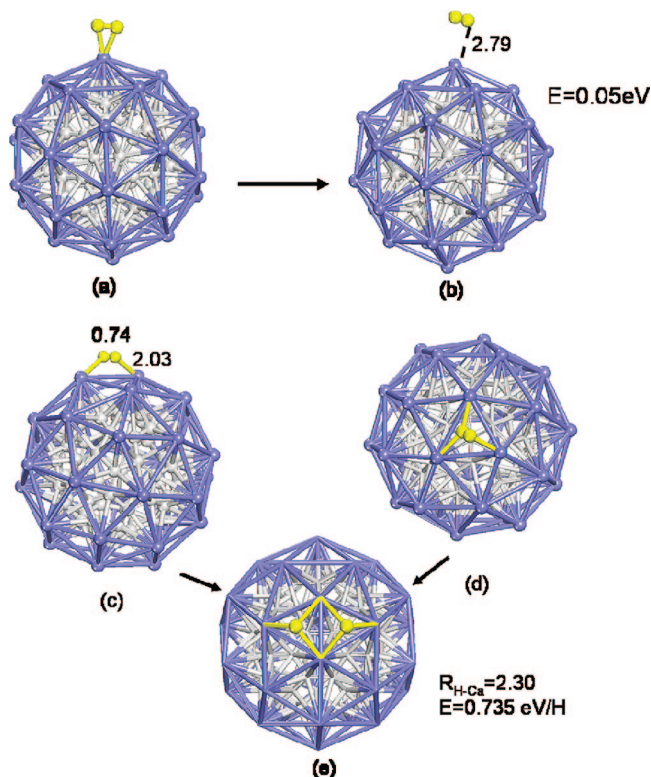
$\text{C}_{60}\text{Ca}_{32}\text{H}_{60}$  can bind more hydrogen atoms we first carried out a model calculation. We began the analysis with a small model of Ca– $\text{C}_2$ , as shown in Figure 6. When a Ca atom binds with  $\text{C}_2$ , 0.982 electrons are transferred to C atoms, and a  $\text{H}_2$  is bound molecularly on top site of Ca with a binding of only 0.051 eV. The distance between the Ca and H atom is 2.881 Å (Figure 6(a)). This is very similar to the absorption of  $\text{H}_2$  on the top site in  $\text{C}_{60}\text{Ca}_{32}$  as discussed above. Now if we add a H atom next to Ca in the  $\text{CaC}_2$  cluster the situation becomes very different. The charge transfer from Ca increases to +1.202e since the Ca atom now has to donate some charge to the H atom. The binding energy of  $\text{H}_2$  to the  $\text{HCaC}_2$  cluster increases to 0.171 eV, and correspondingly the distance between Ca and H decreases to 2.632 Å. The H–H bond also stretches from its molecular value of 0.74 Å to 0.757 Å. These results suggest that the presence of H atoms next to Ca may allow  $\text{C}_{60}\text{Ca}_{32}\text{H}_{60}$  to bind more hydrogen.

Following this clue, we added one  $\text{H}_2$  molecule on top of each of the 32 Ca sites in  $\text{C}_{60}\text{Ca}_{32}\text{H}_{60}$  and reoptimized the structure. The resulting geometry of the  $\text{C}_{60}\text{Ca}_{32}\text{H}_{60}\text{-}32\text{H}_2$  complex is shown in Figure 7. The distance between  $\text{H}_2$  and Ca shell is 2.621 Å, and the absorption energy of the second layer of hydrogen is 0.11 eV/ $\text{H}_2$ , similar to the values of the model system  $\text{C}_2\text{CaH-H}_2$  described above. The total weight percentage of the complex  $\text{C}_{60}\text{Ca}_{32}\text{H}_{60}\text{-}32\text{H}_2$  is now 6.2 wt %, providing hope that  $\text{C}_{60}\text{Ca}_{32}$  may be a suitable material for hydrogen storage.

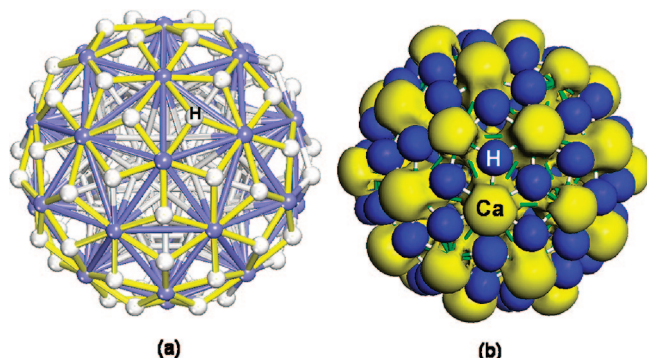
It is important to stress the advantages of a structure that derives stability from the geometric shell closure instead of electronic shell closure. For example, if the stability were of electronic origin, the cluster would have a large HOMO–LUMO gap, and as the size increases, this gap may close making the structure less stable. But for the geometric shell closure, there are no such limitations for the stability of a Ca coated higher fullerene. For example, in  $\text{C}_{60}$  there are 12 pentagons and 20 hexagons, as we have seen it can accommodate 32 Ca atoms in total. In  $\text{C}_{70}$ , there are 12 pentagons and 25 hexagons, so it should be able to accommodate 37 Ca atoms. Indeed,  $\text{Ca}_{37}\text{C}_{70}$  has also been found to exhibit a conspicuous peak in the mass spectra<sup>23</sup> and hence is a magic cluster. We can use this “counting the rings of fullerenes”<sup>23</sup> as a means to determine the magic number of Ca atoms that can decorate a higher fullerene, and hence the maximum number of hydrogen it can hold. For example, we consider the  $\text{C}_{720}$  fullerene which has 12 pentagons and 350 hexagons and thus can accommodate 362 Ca atoms. These 362 Ca atoms again result in 720 triangular faces, in analogy with Figure 5(a), it can first bind 720 H atoms in dissociated form. Then the 362 Ca atoms can further bind 362  $\text{H}_2$  molecules in quasi-molecular form, thus forming a cluster with the composition  $\text{C}_{720}\text{Ca}_{362}\text{H}_{720}\text{-}362\text{H}_2$  having a diameter of about 3.6 nm. The corresponding gravimetric density of hydrogen storage is 6.25%. It is only marginally larger than that in  $\text{C}_{60}$  fullerene. One can easily extend this rule to fullerenes of any size and determine the maximum hydrogen storing capacity of a Ca-coated fullerene system. Consider a fullerene with number of C atoms,  $N_{\text{C}}$ . The magic number,  $N_{\text{Ca}}$ , of Ca atoms it can accommodate is given by



**Figure 3.** (a) Charge difference and (b) HOMO and (c) LUMO of  $C_{60}Ca_{32}$ .

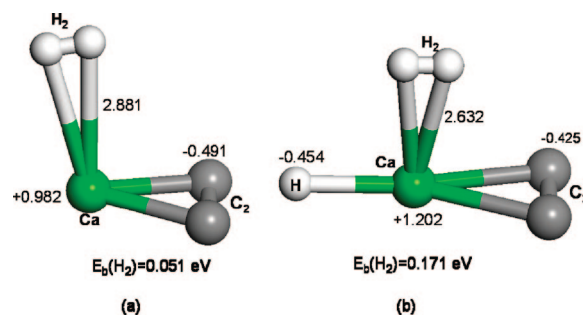


**Figure 4.** (a) Initial and (b) final geometry of  $H_2$  placed at the on-top configuration. Initial geometries of  $H_2$  placed on the (c) Ca–Ca bridge and (d) hollow site. The final optimized geometry is given in (e).

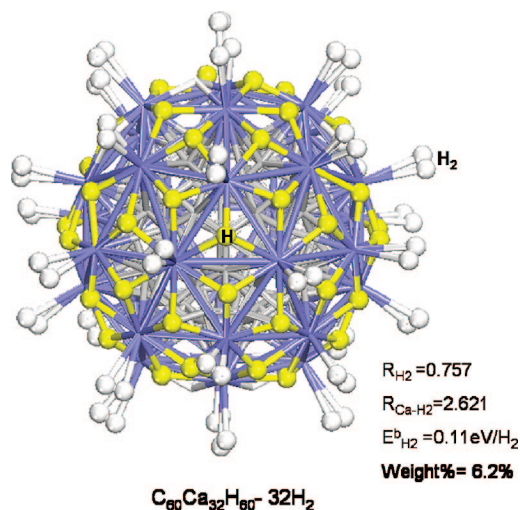


**Figure 5.** (a) Optimized geometry of and (b) charge distribution in  $C_{60}Ca_{32}H_{60}$ .

$N_{Ca} = N_C/2 + 2$ . The metal-loading atomic percentage  $N_{Ca}/N_C$  is  $(0.5 + 2/N_C) \times 100\%$ . The total number,  $N_H$ , of stored



**Figure 6.** Model analysis of the effect of H insertion on the absorption of  $H_2$ .



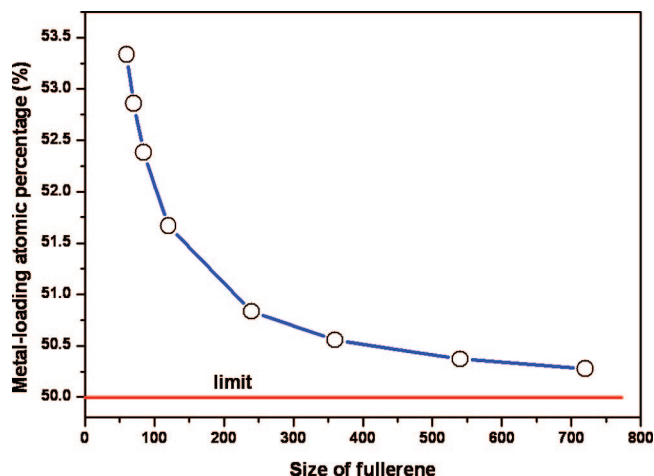
**Figure 7.** Geometry of 32  $H_2$  molecules absorbed on  $C_{60}Ca_{32}H_{60}$ .

hydrogen atoms in such a complex is then  $N_C + 2N_{Ca}$ . With this the hydrogen weight percentage  $W_H$  (%) can be calculated with following formula

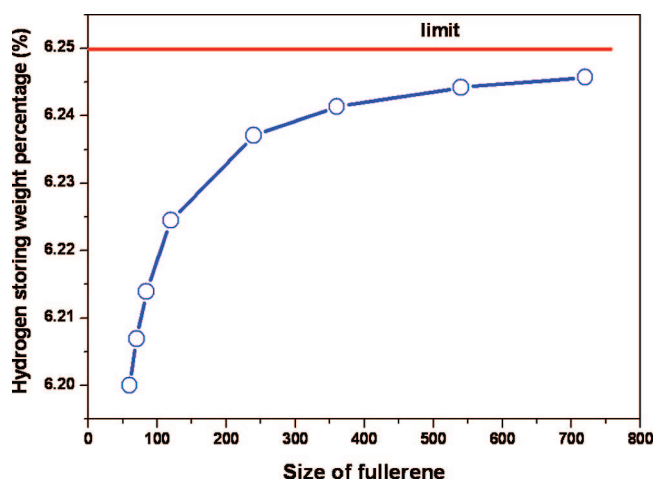
$$WH(\%) = [(2N_C + 4) \times 1.0] \times 100\% / [N_C \times 12.0 + (0.5 \times N_C + 2) \times 40.0]$$

In Table 1, we give the changes of  $N_{Ca}$ ,  $N_H$ ,  $A_{Ca}$ , and  $W_H$  with the size of fullerene. It is interesting to note that with the increase in fullerene size, the metal-loading atomic percentage decreases, saturating at 50% (Figure 8); while the weight percentage of stored hydrogen increases, terminating at 6.25% (Figure 9).

We also examined the possibility that  $C_{60}Mg_{32}$  can store hydrogen with larger gravimetric density than  $C_{60}Ca_{32}$  as Mg



**Figure 8.** Dependence of metal-loading atomic percentage with fullerene size.



**Figure 9.** Dependence of hydrogen gravimetric density with fullerene size.

**Table 1.** Fullerene Size ( $N_C$ ), Number of Ca Atom ( $N_{Ca}$ ), Total Number of Hydrogen ( $N_H$ ), Metal-Loading Atomic Percentage ( $A_{Ca}$ ), and Hydrogen Storage Weight Percentage ( $W_H$ )

$N_C$	$N_{Ca}$	$N_H$	$A_{Ca}$ (%)	$W_H$ (%)
60	32	124	53.3333	6.2000
72	37	144	52.8571	6.2069
84	44	172	52.3810	6.2139
120	62	244	51.6667	6.2245
240	122	484	50.8333	6.2371
360	182	724	50.5556	6.2414
540	272	1084	50.3704	6.2442
720	362	1444	50.2778	6.2457

belongs to the alkaline-earth series and is lighter than Ca. However only after 0.4 ps simulation, the structure is totally fractured as shown in Figure 1(c), indicating that due to the small size of Mg, 32 Mg atoms are not enough to cover the surface of  $C_{60}$ . We also studied the stability of  $C_{60}Li_{32}$ . Note that  $C_{60}Li_{12}$  has been known to be a magic cluster, and decorating the remaining 20 hexagonal sites on  $C_{60}$  will lead to  $C_{60}Li_{32}$ . After 0.5 ps simulations, the initial Li coating layer ruptured, and Li atoms formed some small clusters dotted on the surface of  $C_{60}$  as shown in Figure 1(d).

In summary, we show that  $C_{60}Ca_{32}$  is thermodynamically stable and can bind up to 6.2 wt % hydrogen with the first 3 wt % bound atomically with a reduced binding energy of 0.45 eV/H and the remaining 3.2 wt % quasi-molecularly with a binding energy of 0.11 eV/ $H_2$ . Our findings are different from the recent report,<sup>34</sup> where 92  $H_2$  molecules are stored corresponding to an uptake of 8.4 wt% with a binding energy of  $\sim 0.4$  eV/ $H_2$  within LDA and  $\sim 0.2$  eV/ $H_2$  within GGA. Furthermore, based on the fullerene counting rule,<sup>23</sup> we also show that as the fullerene size increases, the fully coated Ca fullerene cannot store more than 6.25 wt% hydrogen. Although Li and Mg are lighter in mass for a possible higher weight percentage of hydrogen storage, they cannot form stable and uniformly coated  $M_{32}C_{60}$  structures as confirmed by our MD simulations.

**Acknowledgment.** This work is partially supported by grants from the National Natural Science Foundation of China (NSFC-10744006, NSFC-10874007) and from the U.S. Department of Energy.

## References

- (1) Alper, J. *Science* **2003**, *299*, 1686.
- (2) Cortright, R. D.; Davada, R. R.; Dumesic, J. A. *Nature* **2002**, *418*, 964.
- (3) Chen, P.; Xiang, Z.; Luo, J. Z.; Tan, K. L. *Nature* **2002**, *420*, 302.
- (4) Rosi, N. L.; Eckert, J.; Eddaoudi, M.; Vodak, D. K.; Kim, J.; O'Keefe, M.; Yaghi, O. M. *Science* **2003**, *300*, 1127.
- (5) Schlappbach, L.; Zuttel, A. *Nature* **2001**, *414*, 353.
- (6) Chandrakumar, K. R. S.; Ghosh, S. K. *Nano Lett.* **2008**, *8*, 13.
- (7) Dillion, A. C.; Jones, K. M.; Bekkedahl, T. A.; Kiang, C. H.; Bethune, D. S.; Heben, M. J. *Nature* **1997**, *386*, 377.
- (8) Liu, C.; Fan, Y. Y.; Liu, M.; Cong, H. T.; Cheng, H. M.; Dresselhaus, M. S. *Science* **1999**, *286*, 1127.
- (9) Zarkevich, N. Z.; Johnson, D. D. *Phys. Rev. Lett.* **2008**, *100*, 040602.
- (10) Sun, Q.; Wang, Q.; Jena, P. *Nano Lett.* **2005**, *5*, 1273.
- (11) Lee, H.; Choi, W. I.; Ihm, J. *Phys. Rev. Lett.* **2006**, *97*, 056104.
- (12) Lochan, R. C.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1357.
- (13) Sun, Q.; Wang, Q.; Jena, P.; Reddy, B. V.; Marquez, M. *Chem. Mater.* **2007**, *19*, 3074.
- (14) Yoon, M.; Yang, S.; Wang, E. G.; Zhang, Z. Y. *Nano Lett.* **2007**, *7*, 2578.
- (15) Zuttel, A.; Sudan, P.; Mauron, P.; Kiyobayashi, T.; Emme-negger, C.; Schlappbach, L. *Int. J. Hydrogen Energy* **2002**, *27*, 203.
- (16) Nikitin, A.; Li, X.; Zhang, Z.; Ogasawara, H.; Dai, H.; Nilsson, A. *Nano Lett.* **2008**, *8*, 62.
- (17) Tibbetts, G. G.; Meisner, C. P.; Olk, C. H. *Carbon* **2001**, *39*, 2291.
- (18) Zhao, Y.; Kim, Y. H.; Dillon, A. C.; Heben, M. J.; Zhang, S. B. *Phys. Rev. Lett.* **2005**, *94*, 155504.



- (19) Yildirim, T.; Iniguez, J.; Ciraci, S. *Phys. Rev. B* **2005**, *72*, 153403.
- (20) Shin, W. H.; Yang, S. H.; Goddard, W. A.; Kang, J. K. *Appl. Phys. Lett.* **2006**, *88*, 053111.
- (21) Sun, Q.; Wang, Q.; Jena, P.; Kawazoe, Y. *J. Am. Chem. Soc.* **2005**, *127*, 14582.
- (22) Sun, Q.; Jena, P.; Wang, Q.; Marquez, M. *J. Am. Chem. Soc.* **2006**, *128*, 9742.
- (23) Sun, Q.; Wang, Q.; Jena, P. *Appl. Phys. Lett.* , In press.
- (24) Zimmermann, U.; Malinowski, N.; Näher, U.; Frank, S.; Martin, T. P. *Phys. Rev. Lett.* **1994**, *72*, 3542.
- (25) Kresse, G.; Heffner, J. *Phys. Rev. B* **1996**, *54*, 11169.
- (26) Sun, Q.; Wang, Q.; Jena, P.; Rao, B. K.; Kawazoe, Y. *Phys. Rev. Lett.* **2003**, *90*, 135503.
- (27) Vidal, C. R. *J. Chem. Phys.* **1980**, *72*, 1864.
- (28) Gong, X. G.; Kumar, V. *Chem. Phys. Lett.* **2001**, *334*, 238.
- (29) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511.
- (30) Wang, L. S.; Chai, A. Y.; Diener, M.; Zhang, J.; McClure, S. M.; Guo, T.; Scuseria, G. E.; Smalley, R. E. *Chem. Phys. Lett.* **1993**, *207*, 354.
- (31) Rao, B. K.; Jena, P. *Euro. Phys. Lett.* **1992**, *20*, 307.
- (32) Niu, J.; Rao, B. K.; Jena, P. *Phys. Rev. Lett.* **1992**, *68*, 2277.
- (33) Niu, J.; Rao, B. K.; Jena, P.; Manninen, M. *Phys. Rev. B* **1995**, *51*, 4475.
- (34) Yoon, M.; Yang, S.; Hicke, C.; Wang, E.; Geohegan, D.; Zhang, Z. *Phys. Rev. Lett.* **2008**, *100*, 206806.

CT800373G

## Charges for Large Scale Binding Free Energy Calculations with the Linear Interaction Energy Method

Göran Wallin,<sup>†</sup> Martin Nervall,<sup>†</sup> Jens Carlsson, and Johan Åqvist\*

Department of Cell and Molecular Biology, Uppsala University,  
Box 596, SE-751 24 Uppsala, Sweden

Received September 26, 2008

**Abstract:** The linear interaction energy method (LIE), which combines force field based molecular dynamics (MD) simulations and linear response theory, has previously been shown to give fast and reliable estimates of ligand binding free energies, suggesting that this type of technique could be used also in a high-throughput fashion. However, a limiting step in such applications is the assignment of atomic charges for compounds that have not been parametrized within the given force field, in this case OPLS-AA. In order to reach an automatable solution to this problem, we have examined the performance of nine different *ab initio* and semiempirical charge methods, together with estimates of solvent induced polarization. A test set of ten HIV-1 reverse transcriptase inhibitors was selected, and LIE estimates of their relative binding free energies were calculated using the resulting 23 different charge variants. Over 800 ns of MD simulation show that the LIE method provides excellent estimates with several different charge methods and that the semiempirically derived CM1A charges, in particular, emerge as a fast and reliable alternative for fully automated LIE based virtual screens with the OPLS-AA force field. Our conclusions regarding different charge models are also expected to be valid for other types of force field based binding free energy calculations, such as free energy perturbation and thermodynamic integration simulations.

### Introduction

Computational structure-based ligand design relies on accurate predictions of binding free energies of usually relatively small organic molecules upon binding to a macromolecular receptor. These predictions can then serve as guidelines for lead compound identification and optimization. The methods used in structure-based binding affinity prediction range between being theoretically stringent to more or less approximate, where there is always a tradeoff between accuracy and computational cost. For instance, free energy perturbation (FEP) is one of the rigorous but more time-consuming methods that often requires considerable initial preparation by the user followed by days of calculation. If infinite thermodynamic sampling could be attained, the method would in principle deliver the true binding free

energies given by the particular potential energy function. However, the limited sampling that can be achieved by computer simulations remains a serious problem, and this is the main reason why FEP applications are still rare in structure-based ligand design. In contrast, empirical scoring functions are much faster, typically requiring only fractions of a second per binding estimate, but then only because they describe ligand-protein interactions phenomenologically and usually do not rely on conformational sampling at all. In between these extremes there is a wide range of methods, reviewed in refs 1–3, and one of these is the linear interaction energy (LIE) method which is the focus of the present work.

The LIE method is a semiempirical approach which is faster than free energy perturbation, typically requiring a few hours per binding estimate, yet is more accurate than empirical scoring functions and has been employed for a number of biomolecular systems with good results.<sup>1,4–9</sup> The approximations behind the LIE method, namely electrostatic

\* Corresponding author phone: +46 18 471 4109; fax: +46 18 53 69 71; e-mail: aqvist@xray.bmc.uu.se.

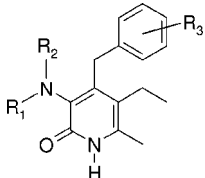
<sup>†</sup> These authors contributed equally to this work.

linear response together with a nonpolar binding contribution that depends linearly on ligand size (representing hydrophobic effect, translational/rotational entropy loss, etc.),<sup>10</sup> leads to a simple linear relation between the binding free energy and the difference in ligand-surrounding average potential energies between the bound and free states, i.e. between the compound immersed in water and enveloped in the binding pocket. These average energies are then calculated as arithmetic mean values from sufficiently long molecular dynamics (MD) or Monte Carlo runs. With continuously increasing computational power, the limiting step of free energy calculations with the LIE method has shifted from the actual simulations to system preparation and analysis. Hence, whereas LIE can conveniently handle hundreds of compounds today, the preparation process needs to be fully automated to further push this number to the tens or hundreds of thousands of compounds that will be computationally feasible in a not too distant future. At this point, our standard LIE scheme has recently been applied to screen about 1000 commercially available compounds for inhibitory activity against a potential drug target in tuberculosis, the 1-deoxy-D-xylulose-5-phosphate reductoisomerase, with promising results (Carlsson et al., unpublished). An efficient simplified LIE version based on energy minimization with a continuum solvent model, rather than explicit water simulations, has also been devised by Caffisch and co-workers and successfully applied for virtual screening on the order of 10<sup>5</sup> compounds.<sup>11,12</sup>

One of the major bottlenecks in the preparation process is the derivation and assignment of partial charges, and solvation free energies of organic compounds are clearly affected by this choice.<sup>13,14</sup> It is therefore of considerable interest to automate this process, and, herein, we investigate the precision and accuracy of several different charge schemes for use together with the OPLS-AA force field. These models include rigorous *ab initio* schemes (Mulliken,<sup>15</sup> Natural Population Analysis,<sup>16</sup> Atoms in Molecules topological analysis,<sup>17,18</sup> and ESP methods by Breneman<sup>19</sup> and Merz–Kollman–Singh<sup>20</sup>) as well as two fast parametrized methods, one based on semiempirical wave functions (CM1A<sup>21</sup> and its scaled version CM1A\*1.14) and one on the concept of electronegativity equalization (Vcharge<sup>22</sup>). The relative performance of the different schemes is evaluated with respect to ligand binding free energies as given by LIE and is compared to the standard charge method associated with the force field, in this case a simple rule based method of combining OPLS-AA fragments. In addition, the effect of solvent charge polarization on the *ab initio* wave functions is evaluated through the Conductor-like Polarizable Continuum Model (CPCM).<sup>23</sup>

Our main goal is thus to investigate whether there are readily automatized charge schemes that can be used in conjunction with the OPLS-AA force field for large scale binding free energy calculations, to eliminate the need for manual parametrization of new chemical structures or fragments. To this end, ten HIV-1 reverse transcriptase (RT) inhibitors were selected from a previous study by Carlsson et al.,<sup>24</sup> such that they span the relative ligand binding free energy space and provide reliable results from well con-

**Table 1.** HIV-1 Reverse Transcriptase Inhibitors Used in This Work<sup>a</sup>



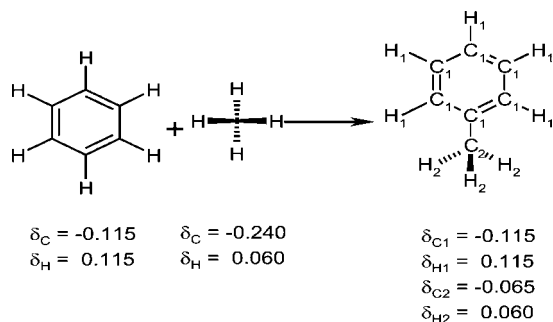
	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	IC <sub>50</sub>
<b>39</b>	CH <sub>3</sub>	C <sub>3</sub> H <sub>7</sub>	3,5-diCH <sub>3</sub>	0.016
<b>40</b>	CH <sub>3</sub>	CH(CH <sub>3</sub> )CH <sub>2</sub> OCH <sub>3</sub>	3,5-diCH <sub>3</sub>	0.006
<b>41</b>	CH <sub>3</sub>	(CH <sub>2</sub> ) <sub>3</sub> SCH <sub>3</sub>	3,5-diCH <sub>3</sub>	0.025
<b>46</b>	H	COC <sub>3</sub> H <sub>7</sub>	3,5-diCH <sub>3</sub>	100
<b>49</b>	H	C <sub>4</sub> H <sub>9</sub>	3,5-diCH <sub>3</sub>	0.126
<b>52</b>	H	CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	3,5-diCH <sub>3</sub>	0.251
<b>60</b>	CH <sub>3</sub>	(CH <sub>2</sub> ) <sub>3</sub> OH	3-CH <sub>3</sub>	0.003
<b>62</b>	CH <sub>3</sub>	(CH <sub>2</sub> ) <sub>2</sub> OCH <sub>3</sub>	3-CH <sub>3</sub>	0.001
<b>65</b>	CH <sub>3</sub>	(CH <sub>2</sub> ) <sub>2</sub> CN	3-CH <sub>3</sub>	0.016
<b>68</b>	H	NH-CS-NHC <sub>6</sub> H <sub>5</sub>	3-CH <sub>3</sub>	3.162

<sup>a</sup> Their experimental IC<sub>50</sub> values (μM) as well as their naming have been adopted from ref 25.

verged runs with LIE and OPLS-AA. These compounds consist of a pyridinone ring connected to a benzyl group, with two side chains differing along the ligand series (see Table 1).

## Methods

The ten RT ligands were selected from a previously published inhibitor series<sup>25</sup> and span 5 orders of magnitude in terms of experimentally observed IC<sub>50</sub>-values. The inhibitor-enzyme structures were adopted from earlier work<sup>24</sup> and originate from dockings with GOLD<sup>27</sup> that were carried out on a crystal structure<sup>26</sup> of HIV-1 RT in complex with compound **62** studied here (PDB code: 2BAN). MD simulations were conducted with the Q software package<sup>28</sup> in an 18 Å sphere centered on the inhibitor, using the OPLS-AA force field.<sup>29</sup> The three docked conformations with the highest ranking for each ligand were extracted and solvated with TIP3P waters.<sup>30</sup> The solvated systems were heated to 310 K in six consecutive steps, while at the same time releasing positional restraints applied to the heavy atoms of the enzyme. An equilibration phase of 50 ps was performed with no positional restraints before entering the collection phase, which was pursued for 1 ns with a time step of 1 fs. Since the ligand simulations in the free state converged much faster, a single simulation of 500 ps was considered to be sufficient for each ligand, thus yielding a total simulation time of 3.5 ns for each ligand. Given that there were ten ligands and that 23 distinct sets of partial charges were examined for each of them, this added up to a total simulation time of about 800 ns. During the collection phase ligand-surrounding energies were collected every 50 fs. The internal geometries of all solvent molecules were constrained with the SHAKE algorithm,<sup>31</sup> and the SCAAS model<sup>28,32</sup> was applied to solvent molecules close to the border to model the density and dipole angular distribution of bulk water. The nonbonded cutoff was set to 10 Å for all atoms inside the sphere, except for ligand atoms for which all nonbonded



**Figure 1.** Example of the OPLS-AA fragment based charge model, where a benzene and a methyl group are merged to form a toluene. The partial charges of all atom types in each molecular species are displayed under the corresponding molecule. The ambiguities that arise when the two groups are merged are dealt with by adjusting the charge of the aliphatic carbon ( $C_2$ ) in the toluene molecule.

interactions were explicitly calculated. Long-range electrostatic interactions were treated with the local reaction field multipole expansion approximation,<sup>33</sup> whereas atoms outside the simulation sphere, which only interacted through bonded terms, were subjected to strong positional restraints.

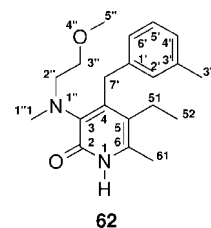
**Linear Interaction Energy.** In the LIE method, the binding free energy is estimated in analogy with solvation energies as the free energy of transfer between water and protein environments. Simulations are carried out for the ligand in water and in the solvated protein, and the Gibbs free energy of binding is calculated from the ligand-surrounding (*l-s*) electrostatic (*el*) and van der Waals (*vdW*) interaction energies through the LIE equation

$$\Delta G_{bind}^{LIE} = \alpha \Delta \langle U_{l-s}^{vdW} \rangle + \beta \Delta \langle U_{l-s}^{el} \rangle + \gamma \quad (1)$$

where the  $\Delta$ 's refer to differences in protein and water simulations. While the  $\beta \Delta \langle U_{l-s}^{el} \rangle$  term represents the polar contribution to the binding free energy and is based on a linear response approximation,  $\alpha \Delta \langle U_{l-s}^{vdW} \rangle + \gamma$  represents the nonpolar binding contributions. The latter term can be derived from the observation that both nonpolar solvation energies in different solvents and ligand-surrounding van der Waals interactions tend to scale linearly with solute size measures, such as molecular surface area or the number of heavy atoms in the ligand.<sup>7,10</sup> This leads to the following type of relationship between  $\Delta \langle U_{l-s}^{vdW} \rangle$  and the change in nonpolar solvation free energy,  $\Delta \Delta G_{sol}^{np}$ , between protein and water environments

$$\begin{aligned} \Delta \Delta G_{sol}^{np} &= a\sigma + b \\ \Delta \langle U_{l-s}^{vdW} \rangle &= c\sigma + d \\ \Rightarrow \Delta \Delta G_{sol}^{np} &= \frac{a}{c} (\Delta \langle U_{l-s}^{vdW} \rangle - d) + b = \alpha \Delta \langle U_{l-s}^{vdW} \rangle + \gamma \quad (2) \end{aligned}$$

where  $\sigma$  is a size measure, and  $a$ ,  $b$ ,  $c$ , and  $d$  are empirically derived parameters. From eq 2, the contributions from nonpolar solvation to  $\alpha$  and  $\gamma$  in eq 1 can in principle be identified as  $a/c$  and  $b-ad/c$ , respectively. Since our standard parametrization of LIE was performed using experimental binding free energies, the obtained value of  $\alpha = 0.18$  takes all size dependent contributions to binding into account, such as the hydrophobic effect and relative translational and



**Figure 2.** Chemical structure formula for compound 62.

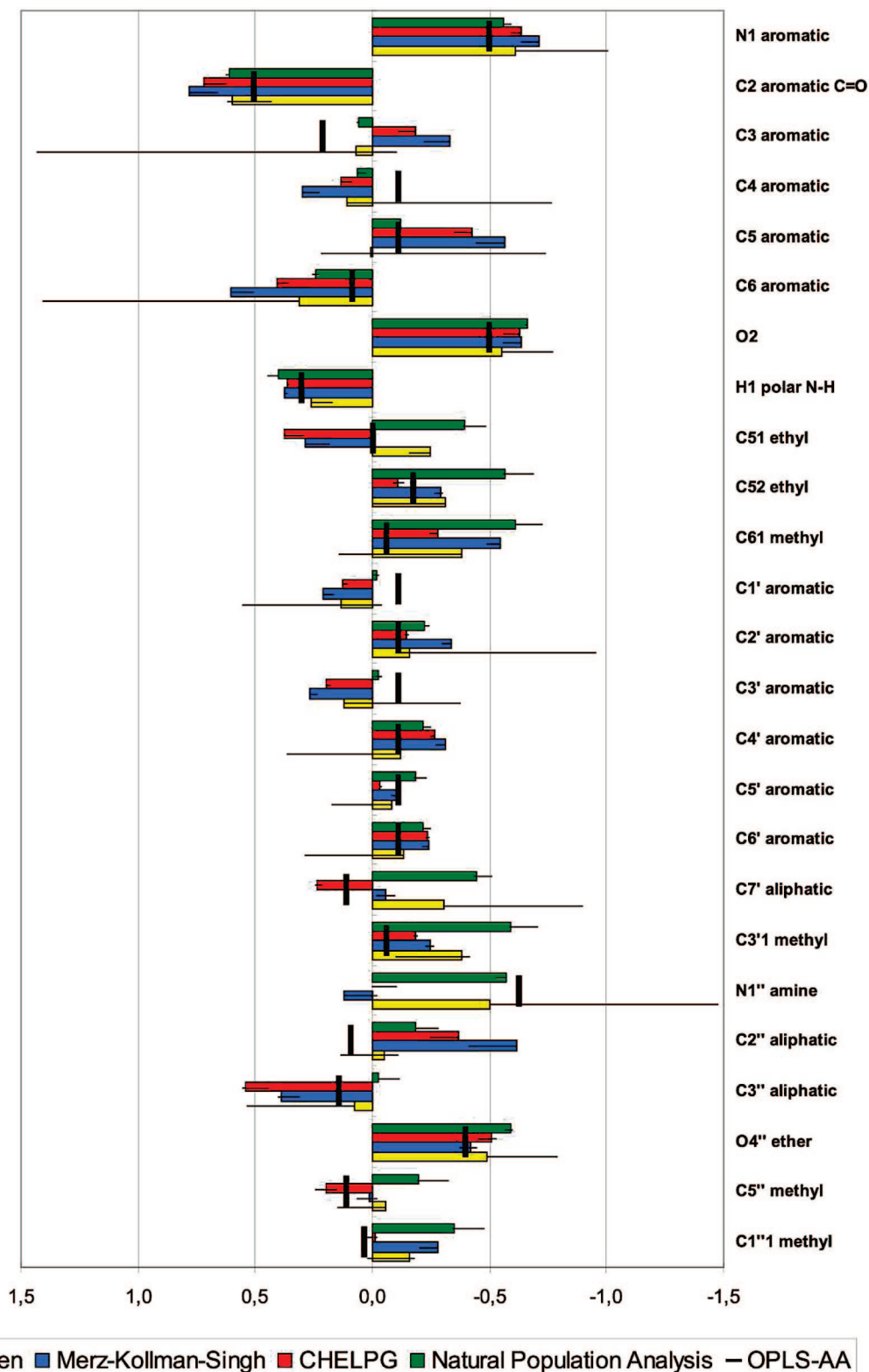
rotational entropies as well as van der Waals interactions. The constant offset  $\gamma$  has been shown to correlate with the hydrophobicity of the binding site pocket<sup>5</sup> and is thus generally protein specific but is freely optimized here for each charge set since only relative binding free energies can be extracted from the present experimental data.<sup>24</sup>

In our standard version of the LIE method, the ligand-surrounding electrostatic energies in both the protein and water simulations are scaled by the same factor, which assumes that the electrostatic response of the protein binding site is similar to that of water. The value of the  $\beta$  coefficient can be derived from linear response approximation, which predicts that  $\beta = 0.5$ .<sup>34</sup> However, based on rigorous FEP calculations in different solvents carried out by Åqvist and Hansson<sup>34</sup> (see also ref 5), the  $\beta$  value used for a ligand in the standard parametrization of the LIE method,  $\beta_{FEP}$ , is determined by its chemical groups. For the inhibitors studied here  $\beta_{FEP}$  is equal to 0.43 in all cases except for compound 60 (for which  $\beta_{FEP} = 0.37$ ).<sup>35</sup> Since the point of this study is to examine the impact of charge variation on the LIE method, and the  $\alpha$ - and  $\beta$ -values have been shown<sup>5</sup> to provide consistent results even with different force fields (Amber95, Gromos87, and OPLS-AA), this standard model with  $\alpha = 0.18$  was adopted, and  $\gamma$  was optimized for every charge set.

**Partial Charges.** Since partial atomic charges are not quantum mechanical observables and therefore cannot be measured by experiment, there is no unambiguous way of assigning them. In this study a few acknowledged methods have been chosen which will be briefly outlined below.

In the *ab initio* Mulliken population analysis,<sup>15</sup> the total molecular wave function is subdivided into net atomic and overlap populations, where the overlap populations are evenly distributed between the atoms. The Mulliken population of an atom  $A$  is thus given by the diagonal sum  $n_A = \sum_{\mu \in A} (PS)_{\mu\mu}$  where  $P$  and  $S$  denote the density and overlap matrix, respectively. The partial charge is then simply the difference between the atomic population and the nuclear charge. Although simple and straightforward, this scheme is overly basis set dependent, especially when compared to actual differences in the wave function.<sup>36</sup>

In contrast, the Natural Population Analysis (NPA) by Weinhold et al.<sup>16</sup> transforms the generally nonorthogonal basis set into an orthonormal basis of atomic orbitals by using *occupancy-weighted symmetry orthogonalization*. This is an eigen decomposition akin to the Löwdin transformation<sup>37</sup> but on atomic angular symmetry blocks in the density matrix and is weighted by orbital occupancy. To describe the atomic state in a molecule, the Rydberg states (i.e., the unoccupied orbitals in the free ground-state atom) are allowed to become



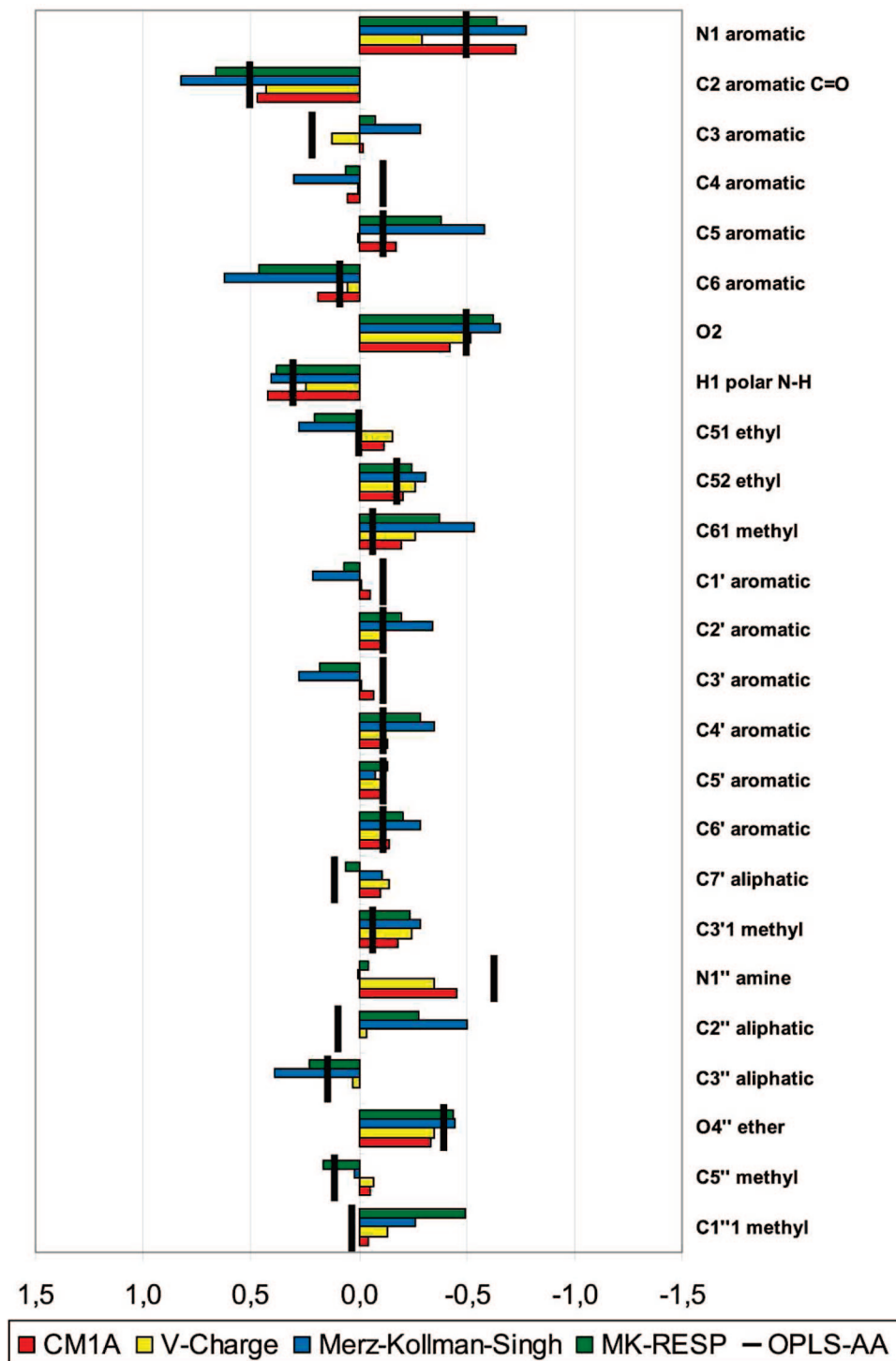
**Figure 3.** Calculated *ab initio* partial charges for compound **62** from Mulliken, Merz–Kollman-Singh, CHELPG, and Natural Population Analysis, compared to the OPLS-AA fragment charges. Charges derived from the most accurate description of the wave function, i.e. B3LYP/6–311+G(d,p) are shown in colored bars, except for the Mulliken charges where instead the B3LYP/6–31G(d,p) level is shown. The span between the maximum and minimum charges for the different basis sets, i.e. 6–31G(d,p), 6–31+G(d,p), 6–311G(d,p), and 6–311+G(d,p), are given by the error bars. The atomic numbering corresponds to that shown in Figure 2.

weakly populated. The populations are then given by the eigenvalues in this decomposition.

In Atoms in Molecules theory (AIM) by Bader<sup>17,18</sup> the nuclei of the molecule are attractors of the gradient density field, and an atom is consequently defined as being the basin

containing all the gradient trajectories terminating in its nucleus. Integrating the density over this basin then gives the atomic populations.

In lieu of populations, one may fit monopole charges to the overall molecular electrostatic potential (ESP). Originally



**Figure 4.** Calculated RESP, MK and semiempirical partial charges for compound **62**. The MK charges are identical to the ones in Figure 3 and were added for comparison with RESP. The atomic numbering corresponds to that shown in Figure 2.

introduced by Momany<sup>38</sup> and Cox and Williams,<sup>39</sup> these methods are based on the total density, which is a quantum mechanical observable, and will thus provide experimentally verifiable dipole and higher order multipole moments. In the Merz–Kollman–Singh (MK) variant, the potential is evaluated at points placed on the Connolly surface of the molecule<sup>40,41</sup> onto which the charges are fitted in a least-squares manner, with a total integer charge constraint by the method of Lagrange multipliers.<sup>20,42</sup> The Charges from Electrostatic Potentials scheme (CHELP) by Chirlian and

Franci<sup>43</sup> is similar but with the points placed in concentric, symmetric, and nearly spherical shells about the atoms. Breneman and Wiberg<sup>19</sup> suggested that the potential instead be evaluated in a grid of uniformly spaced points (CHELPG) to dampen its sensitivity toward conformational changes and is the variant that is used here.

Being evaluated at some distance (about 1.5–2.0 times the van der Waals-radii<sup>20</sup>), the charges of buried atoms (e.g.,  $sp^3$  carbons) have a tendency of being less important for the quality of the ESP fit than atoms closer to the molecular

**Table 2.** Static Average Polarizabilities and Dipole Moments in Gaseous and Aqueous Phase for the Compounds Studied Here, Using Two Levels of Theory

	B3LYP/6-31G(d,p)				B3LYP/6-311+G(d,p)			
	$\hat{\alpha}$ [ $a_0^3$ ]	$\mu(g)$ [D]	$\mu(aq)$ [D]	$\Delta\mu$	$\mu(g)$ [D]	$\mu(aq)$ [D]	$\Delta\mu$	
<b>39</b>	243	4.08	5.65	1.57	4.44	6.35	1.91	
<b>40</b>	259	3.84	5.26	1.42	4.03	5.65	1.62	
<b>41</b>	270	5.97	7.78	1.81	6.32	8.48	2.16	
<b>46</b>	244	4.02	5.86	1.84	4.36	6.64	2.28	
<b>49</b>	248	3.38	4.84	1.46	3.78	5.68	1.90	
<b>52</b>	273	4.23	5.69	1.46	4.68	6.54	1.87	
<b>60</b>	234	5.02	6.71	1.69	5.37	7.39	2.02	
<b>62</b>	233	4.81	6.71	1.89	5.29	7.65	2.36	
<b>65</b>	228	2.26	2.70	0.44	2.40	2.95	0.54	
<b>68</b>	294	5.56	8.40	2.84	5.76	9.07	3.31	

surface. For this reason, Bayly et al.<sup>44</sup> suggested that the charges on enveloped atoms be restrained with penalty functions to dampen arbitrary fluctuations and to enforce symmetry invariance, without deteriorating the overall description of the potential. Termed the Restrained Electrostatic Potential (RESP) method this procedure is used in the AMBER force field definition.<sup>45,46</sup>

To reduce the computational effort, Cramer, Truhlar, and co-workers introduced the Charge Model 1A (CM1A) which extracts Mulliken populations from NNDO Austin Model 1 (AM1)<sup>47</sup> semiempirical wave functions and subsequently maps them with a multilinear form to reproduce experimentally observed dipole moments.<sup>21</sup> These charges have been successfully used in solvation free energy calculations with the OPLS-AA force field.<sup>13,48,49</sup>

Charges can also be inferred from the concept of electronegativity equalization where the atomic electron densities are shifted to atoms with higher electronegativity upon bond formation, thus giving rise to partial charges. The subsequent increase in atomic radius corresponds to a lowering in electronegativity which continues until equilibrium has been reached. Since total equalization will result in e.g. all atoms in a molecule of the same sort having identical charges, Gasteiger and Marsili suggested a partial equalization based on orbital electronegativities<sup>50</sup> where the charge transfer is somewhat dampened.<sup>51</sup> In the Vcharge method, which is used here, Gilson et al. instead adjusted the initial electronegativities based on the properties of neighboring atoms, valence bond types, and a set of variables. The variables were parametrized on a set of compounds to reproduce the *ab initio* molecular ESP from the Hartree-Fock/6-31G(d) model chemistry.<sup>22</sup> These methods are computationally cheap and only require the chemical structure formula of the molecule.

Finally, charges have been derived in analogy with the OPLS-AA force field definition,<sup>29</sup> where e.g. charges of the toluene molecule is found by maintaining the symmetric benzene charges and then adjusting the methyl carbon charges to maintain the overall charge group neutrality, as shown in Figure 1. Such charges are henceforth referred to as the OPLS-AA fragment based charges. Although charge redistribution between fragments is not taken into account, these charges are most likely to remain balanced with the force field parameters of the water model and the protein.

It may be noted that the Maestro software from Schrödinger uses an automated fragment-based method with bond charge increments<sup>52</sup> to estimate junction atom charges. However, since this method requires optimized parameters for it to be applicable to the OPLS-AA force field and they are not publically available, this method will not be covered in the present work.

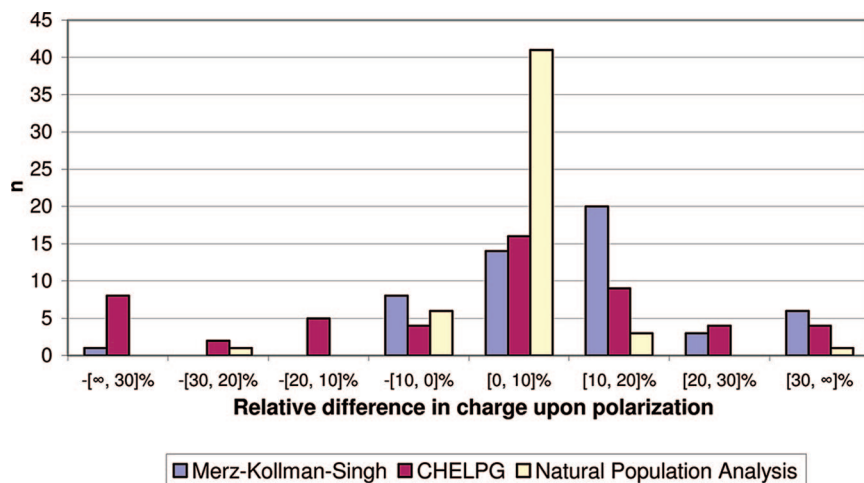
The *ab initio* Mulliken, MK, CHELPG, and NPA charges were derived from density functional theory (DFT) wave functions at two levels of theory on structures optimized in the gas phase—namely B3LYP/6-31G(d,p)//B3LYP/6-31(d,p) and B3LYP/6-31G(d,p)//B3LYP/6-311+G(d,p). That is, using the Becke three parameter hybrid functional<sup>53</sup> and the correlation functional of Lee, Yang, and Parr (B3LYP),<sup>54,55</sup> together with Pople's polarized split valence and triple split contracted Gaussian basis sets, augmented with diffuse functions in the latter case.<sup>56</sup> Furthermore, to get an estimate of the basis set dependence of the different charge schemes, wave functions were formed using the additional basis sets 6-31+G(d,p) and 6-311G(d,p) for compound **62**, from which populations were derived with the methods outlined above. Where nothing else is stated the Gaussian 03 package<sup>57</sup> was used for all *ab initio* calculations.

The AIM calculations were performed with Bader's original AIMPAC source code, using the PROMEGA algorithm bundled in PROAIMV.<sup>58</sup> However, since the original settings only support molecules of at most 50 atoms, and the ligands considered herein are somewhat larger, the MCENT parameter was increased throughout the source code to allow a maximum of 60 centers instead. The adopted numerical integration parameters in the PROMEGA algorithm were 64 phi planes, 48 theta planes, and 96 radial points per integration ray within the Beta sphere. These settings yielded an underestimation of the atomic basin populations in the order of 3 to 6 · 10<sup>-4</sup> e. To correct for these artificial positive charges, the overall deviation from neutrality was divided equally across all basins and were subsequently subtracted. Although the absolute atomic net charges were affected by this correction, their relative charge differences were preserved.

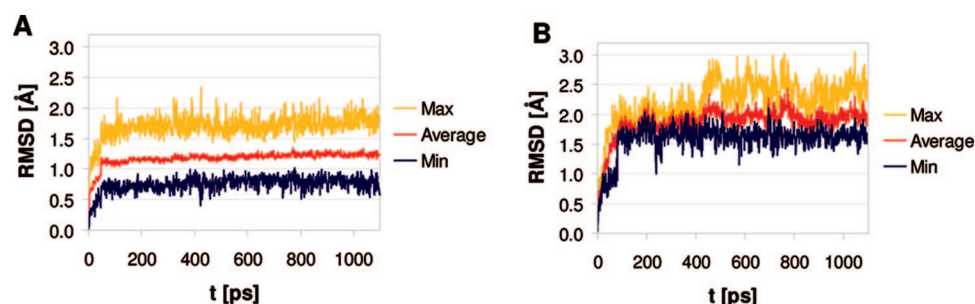
To comply with the recommendations regarding the RESP calculation scheme,<sup>44</sup> Hartree-Fock densities with the 6-31G(d) basis set were also formed from which the MK ESP was extracted. RESP fitting of these charges were performed in two steps with the antechamber utility from the AmberTools distribution,<sup>59,60</sup> using the ESP explicit output Gaussian internal option (IOP) 6/33=2. The ESP evaluation was set to 6 points per unit area using the IOP 6/42=6.

CM1A charges were calculated with the Amsol software,<sup>61</sup> and Gilson charges were obtained with Vcharge<sup>62</sup> (VC/2004 parameter set). As mentioned above, partial charges for the OPLS-AA charge set were assigned in analogy with the force field, except for the amine substituent of compound **68** where RESP charges were used.

**Charge Polarization.** The propensity of a compound to become polarized when exposed to surroundings can be gauged by calculating its static polarizability tensor  $\alpha$ . This



**Figure 5.** The impact of continuum solvent polarization on ESP charges (Merz–Kollman–Singh and CHELPG) as well as Natural Population Analysis for compound **62**. The histogram shows the number of atoms in this compound that have adjusted their charge within a certain range when exposed to the continuum solvent (e.g., for NPA, 41 out of 52 atoms have adjusted their partial charge by less than 10%).



**Figure 6.** Extremum and average RMSD values across protein equilibration and production runs with the OPLS-AA fragment based charges are shown in panel A, with the exception of the ligand **52** runs which are shown separately in panel B. The RMSD scales in these plots have been kept equal for comparison.

tensor as well as higher order hyperpolarizabilities can be obtained from *ab initio* calculations where they are identified with the analytic derivatives of the energy with respect to the electric field at vanishing field strength.<sup>63</sup> Furthermore, the average polarizability  $\bar{\alpha}$ , i.e. the mean of the tensor diagonal elements, is an invariant that can be inferred from experiments.<sup>64</sup>

Solvent induced polarization of the wave function can be modeled using Self Consistent Reaction Field theory (SCRF) where the solute molecule is placed in a cavity submersed in an infinite continuous polarizable medium. To this end, a conductor-like extension of Tomasi's apparent charge Polarizable Continuum Model<sup>65,66</sup> (PCM), similar to the COSMO method by Klamt and Schüürman<sup>67,68</sup> and termed Conductor-like PCM<sup>23,69</sup> (CPCM), was used here.

Polarizabilities were extracted from frequency calculations at the B3LYP/6–31G(d,p) level of theory. Where imaginary frequencies were detected, or the magnitude of the rotational low frequencies surpassed  $\sim 10 \text{ cm}^{-1}$ , the optimizations were resubmitted with increased accuracy in the integration grid (see e.g. ref 70). The pruned UltraFine grid, with its 99 radial shells and 590 angular points, proved to be sufficient in all such cases.

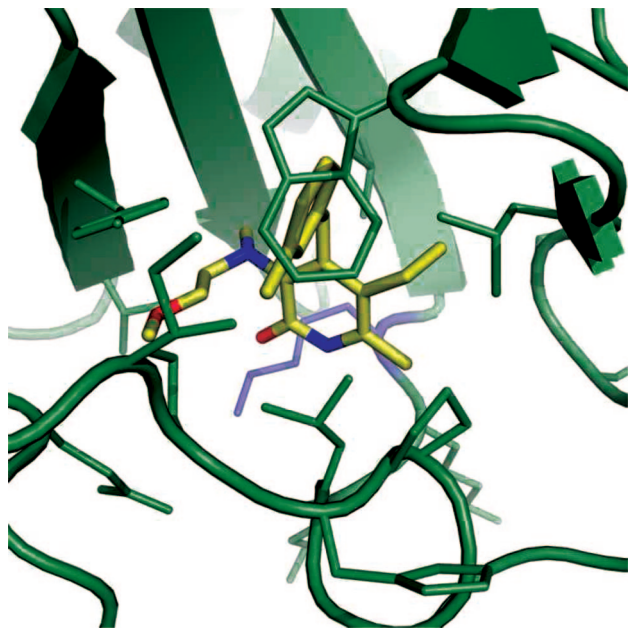
CPCM polarization was implemented with the Gaussian keyword, using the standard permittivity for water and the

default number of tesserae per sphere. Furthermore, the atomic radii of the United Atom Kohn–Sham topological model (UAKS) were used in all B3LYP calculations since these radii have been optimized with respect to DFT using the parameter free Perdew, Burke, and Ernzerhof functional (PBE0) at the 6–31G(d) level of theory.<sup>57</sup> However, in the RESP Hartree–Fock calculations the default UA0 radii were used.

**Statistical Measures.** The accuracy of the LIE binding free energy estimates with the different partial charge sets was measured with the coefficient of determination, denoted  $R^2$ . Since the coefficient of determination compares the variance of the predictions with the total variance in the data, it is commonly used to measure the extent to which the model explains the observed variance. For these purposes, however, it is enough to regard  $R^2$  as a simple indication of the goodness-of-fit of the experimental data to the LIE regression line. Appropriately, the LIE parameters that have been optimized here, e.g. the  $\gamma$  parameter in the LIE standard model, have been done so by least-squares regression, i.e. by minimizing the *SSE* with respect to experimental data.

Apart from gauging the extent of explained variance in the model, it is also of special interest to measure the degree of correspondence between the rankings, or the *rank cor-*





**Figure 7.** HIV-1 RT (green) in complex with compound **62** (yellow). The structure is a docking solution used as a starting structure in MD simulations. The figure shows the close proximity of Lys103 (blue) to the carbonyl oxygen of the pyridine.

relation,<sup>71</sup> between the observed and calculated binding free energies. A popular way of doing this in virtual high-throughput screening<sup>72</sup> is by computing Spearman's  $\rho$  value, given by

$$\rho = 1 - \frac{6S(d^2)}{n^3 - n} \quad (3)$$

where  $n$  is the number of observations, and  $S(d^2)$  denotes the sum of squared differences between the experimental and calculated ranking for each observation. Spearman's  $\rho$  will thus be +1 if there is perfect agreement between the rankings, -1 if the sets are anticorrelated, and 0 if there is no agreement at all.

The error of the calculated binding free energy was estimated with the standard error of the mean (SEM) taken with respect to the ensemble averages in the protein starting from the three docking poses of each ligand. Given that the sample based estimation of the standard deviation is  $s$  and the number of observations is  $n$ —in this case three—the error then reads

$$\text{Err}[\Delta G_{\text{calc}}] = \alpha \frac{s[\langle U_{\text{lig-surr}}^{\text{vdw}} \rangle_{\text{prot}}]}{\sqrt{n}} + \beta \frac{s[\langle U_{\text{lig-surr}}^{\text{el}} \rangle_{\text{prot}}]}{\sqrt{n}}$$

Here it is assumed that the  $\alpha$  and  $\beta$  parameters do not contribute significantly to the uncertainty of the estimation when the LIE standard model is used.

Furthermore, leave-one-out cross-validated  $R^2$ , commonly referred to as  $Q_{\text{LOO}}^2$ , has been used as a quantitative method to assess the predictive ability of the different LIE models and charge sets presented here. The definition of  $Q_{\text{LOO}}^2$  is similar to that of  $R^2$ , namely

$$Q_{\text{LOO}}^2 = 1 - \frac{\sum_i (\Delta G_i^{\text{obs}} - \Delta G_i^{\text{calc}})^2}{\sum_i (\Delta G_i^{\text{obs}} - \overline{\Delta G})^2} \quad (4)$$

The difference is that the binding energy of the  $i$ th ligand is estimated with a regression model on a data set where that particular compound was left out. If the model is dependent on the data points from which it was derived, this will have an impact on the  $Q_{\text{LOO}}^2$  value. Following common practice, charge sets that fall short of producing a  $Q_{\text{LOO}}^2 \geq 0.5$  will be ruled out. Although external validation is the most reliable method to assess predictability,<sup>73</sup> internal validation is deemed quite adequate for these purposes. However,  $Q^2$  can give a slight underestimation of the true predictive error when applied to small data sets.<sup>74</sup>

**Structural Stability and Convergence.** The structural stability of the ligand-protein complexes was estimated with the root-mean-square deviation (RMSD), given as

$$\text{RMSD}([\vec{u}_1 \dots \vec{u}_n], [\vec{v}_1 \dots \vec{v}_n]) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\vec{u}_i - \vec{v}_i)^2} \quad (5)$$

In this case the RMSD was taken with respect to the heavy atom position vectors of the ligand structures, which were kept unfitted since the ligands are more or less held in position by the protein. By this choice rotation and translation are expected to have a significant impact on the computed values.

Before considering the impact of varying charges, the structural stability of the OPLS-AA fragment based charge runs was measured with respect to the starting structures,  $[\vec{x}(0)_{1\dots n}]_i^j$ , of ligand  $i$  and pose  $j$ , hence  $\text{RMSD}(t) - ([\vec{x}(0)_{1\dots n}]_i^j, [\vec{x}(t)_{1\dots n}]_i^j)$ . Then the charge model (CM) stability was measured by forming production phase average structures,  $\langle \text{CM} \rangle$ , and computing their RMSD with respect to OPLS-AA. That is, for the  $n$ th charge model this reads  $\text{RMSD}(\langle \text{OPLSAA} \rangle_i^j, \langle \text{CM}_n \rangle_i^j)$ , for ligand  $i$  and pose  $j$ . The former comparison is meant to gauge the overall stability of the system *per se*, whereas the latter reflects the impact of charge variation on the structures.

Since the RMSD values were extracted from 690 MD runs, a few operations were defined to simplify the presentation of this data. For the initial OPLS-AA stability, extremum and average RMSD values across all ligands and poses as a function of time are presented. When comparing charge models with the OPLS-AA average structures, the RMSD values were summed over the poses, that is  $\sum_j \text{RMSD}(\langle \text{OPLSAA} \rangle_i^j, \langle \text{CM}_n \rangle_i^j)$ .

## Results and Discussion

To begin with, the actual differences in charge between the models will be investigated for a representative compound from the ligand set. Then the calculated polarizabilities and their impact on the calculated effective dipole moments, i.e. when subjected to implicit solvent, are presented. After this, the charge sets are culled, first with respect to the precision, as given by the convergence of the MD ensemble averages,

and then on the accuracy in their binding free energy estimates. Finally, the reliability of these results is examined through statistical internal validation.

**Comparing Charges.** To get an overview of how partial charges actually differ between the charge models, ligand **62** (see Figure 2) was chosen as a representative compound, and the derived *ab initio* charges of its first row atoms and the polar N–H hydrogen are displayed in Figure 3. To begin with, charge dependence on level of the model chemistry was gauged by performing DFT calculations with a range of basis sets, namely 6–31G(d,p) and 6–311G(d,p) with and without the addition of diffuse functions. Charges derived from the highest level of theory are displayed in Figure 3, with the exception of the Mulliken charges where instead the low level is shown, as well as the span between the charge assignments which is shown in the error bars. As expected, the sensitivity toward the choice of basis set is most apparent for the Mulliken charges, where for instance the C3 atom of the pyridinone ring (defined in Figure 2) ranges from being neutral to highly negative ( $\sim -1.5 e$ ) when adding diffuse functions. This is presumably due to the considerable overlap populations. In contrast, the NPA and the ESP charges are significantly more stable in this respect.

Apart from basis set impact, charge assignment from the different methods is seen to vary throughout, which is only natural considering that the atomic partial charges are not quantum mechanical observables and thus cannot be directly measured by experiment. Any charge assignment will therefore be more or less arbitrary. In particular, substituent atoms that link functional groups together seem to be difficult to assign by any method, e.g. C51, C7', the tertiary N1'', and its neighbors. However, there are a few cases where there is a remarkable agreement between all the methods, such as the ether O4'' and the pyridinone polar H1 atoms. Also, it is reassuring to see an overall similarity between the ESP charges from the MK and CHELPG schemes. Moreover, charge redistribution seems to be causing the most pronounced overall difference between the OPLS-AA fragment based charges and the *ab initio* sets. For instance, the aromatic C1'–C6' atoms are seen to be in fair agreement except where the aliphatic substituents are bound, i.e. C1' and C3'. This can also be seen in the pyridinone ring C3 and C4 atoms.

In order to compare the CM1A and Vcharge sets with OPLS-AA and *ab initio* charges, they are plotted along with the MK ESP charges in Figure 4. Also included in this plot are the RESP fitted charges, which were extracted from the Hartree–Fock ESP using the 6–31G(d) basis set, rather than the high level DFT wave functions. In spite of these differences, the RESP charges are seen to be very similar to the MK set.

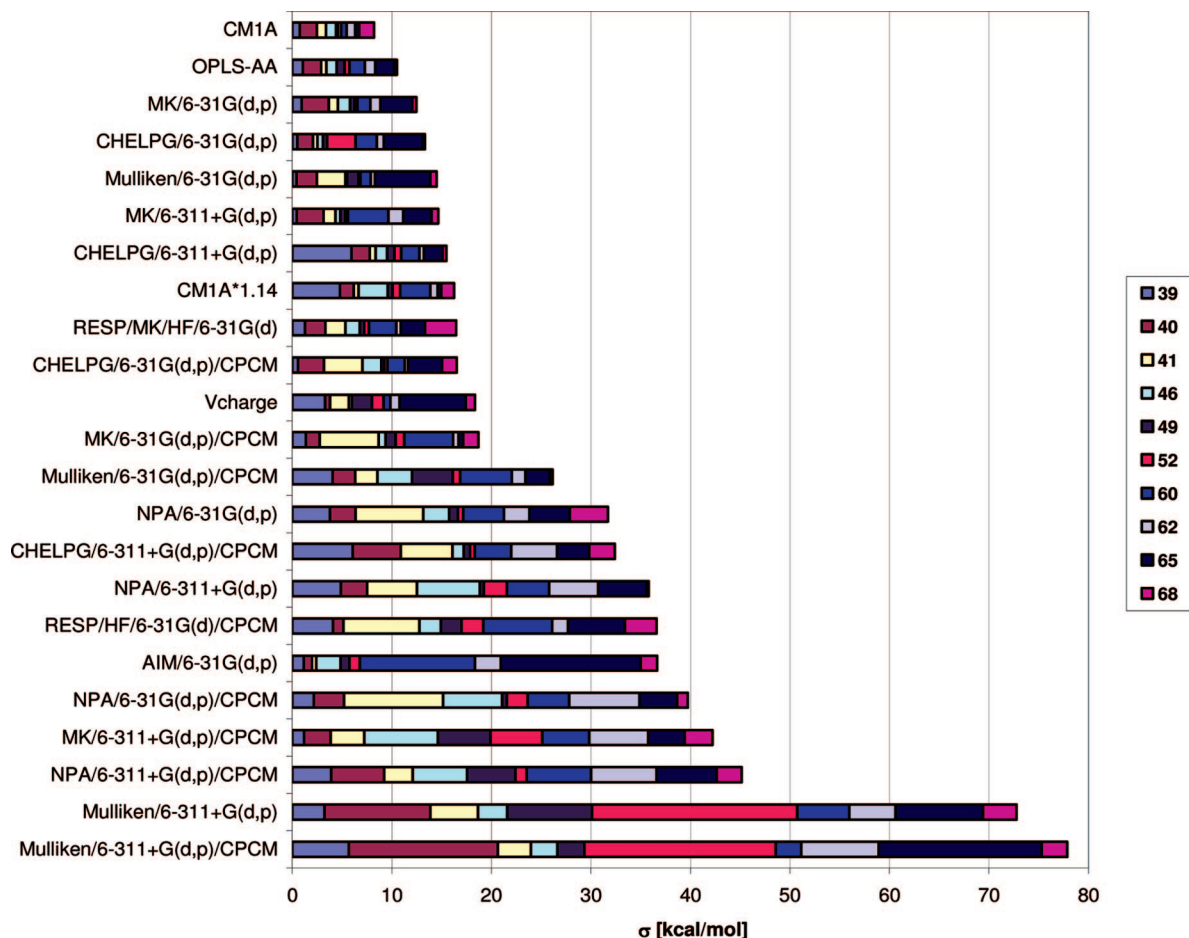
The overall picture that emerges in this figure is just about the same as in Figure 3. The charge redistribution that is disregarded in the fragment based method is indeed captured by the empirical models, where CM1A seems to lie closest to the OPLS-AA charges. The overall RMSD between the CM1A and Vcharge set is  $0.09 e$ , the largest difference being the pyridinone nitrogen where CM1A is substantially more polarized. There are a few cases where they deviate from

OPLS-AA, e.g. the amine N1'' and C7', but the impression is that they are more consistent with the force field than the MK and RESP charges.

**Polarizability and Solvent Effects.** The charges that were discussed in the previous section were all derived *in vacuo*. However, drug binding and hydration pertain to polar surroundings that can have a significant impact on the overall electrostatic properties of the compound due to polarization. For this reason, it seems useful to estimate the importance of charge polarization for these compounds, either by including an environment in the calculation or by studying their intrinsic static average polarizability in vacuum. In the latter case, a significant polarizability indicates that the charges may vary as the ligand is subjected to different surroundings. The polarizability of the wave function was thus calculated and compared with the change in dipole moment as the compounds are subjected to the implicit solvation model. As expected, the presence of aromatic groups results in significant polarizabilities ranging from about 200 to 300  $a_0^3$  (see Table 2), which may be compared with the polarizability of, for instance, water which is  $10.13 a_0^3$  ( $\Delta_{\text{calc-exp}} = -4.78 a_0^3$ ), methane  $16.52$  ( $\Delta_{\text{calc-exp}} = -3.84$ )  $a_0^3$ , and ammonia  $14.19$  ( $\Delta_{\text{calc-exp}} = -5.61$ )  $a_0^3$ , with the model chemistry used here (values from ref 75), where  $a_0$  is the Bohr radius. This is in turn reflected in the significant increase of the dipole moment, by up to 3 D, when exposed to the implicit solvent model (see Table 2).

It seems rather clear that the impact of the solvent on the total charge distribution, as represented by the dipole moments, should also appear as changes in the partial charges. Indeed, plotting the distributions of the relative changes in atomic charge when compound **62** is exposed to solvent shows that they change by 10% or more due to polarization when ESP methods are used (see Figure 5). In contrast, polarization does not seem to affect NPA charges as much, where only four atoms change by more than 10%. This difference can perhaps be understood by considering that the ESP is directly based on the overall charge distribution. The increase in the dipole moments and its impact on the partial charges substantiates the typical choice of using charges derived from the HF/6–31G(d) level of theory,<sup>44</sup> which usually overestimates the dipole moment of the molecule to an extent that roughly corresponds to its effective value in aqueous solution. However, it should be noted that this effect does not seem to be as pronounced for the B3LYP model chemistries that are used here.<sup>75</sup> Charge polarization is also the motivation behind linear scaling of charges present in the OPLS/CM1A force field,<sup>48</sup> i.e. where the CM1A charges are multiplied by a factor of 1.14 to enhance solvation free energy estimates. This linear scaling has also been included in this study.

**Stability and Convergence.** As outlined in the Methods section, the binding free energy estimates in the LIE method are calculated from ensemble averages of the ligand-surrounding interaction energies, which are extracted from molecular dynamics trajectories. Every ligand is simulated from three different docking poses for every charge set, and



**Figure 8.** Standard deviations of the triplicate  $\langle U_{el}^i \rangle$  protein–ligand electrostatic energies for each ligand, accumulated for each charge set.

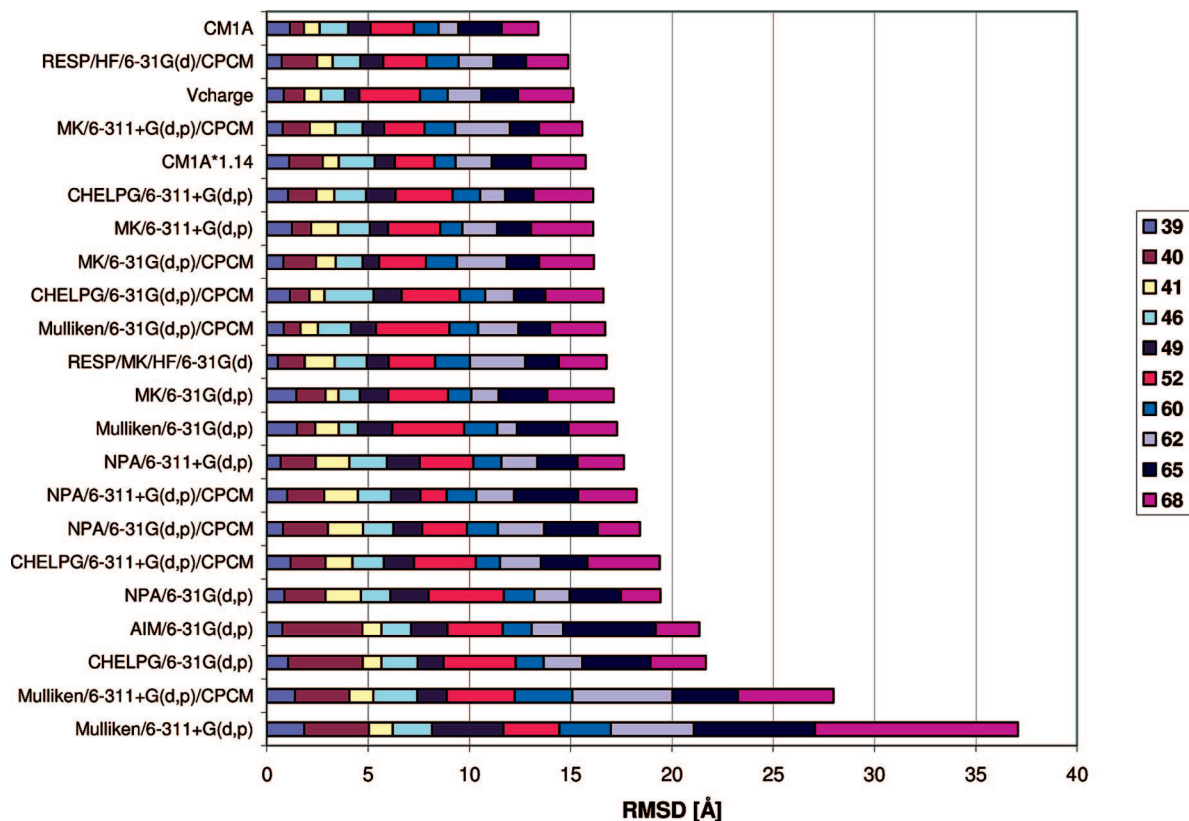
since these poses were selected from the same binding mode the resulting energies are given as the mean from the three simulations. It should be noted that in cases where poses cannot interchange during the course of the simulation it would be hard to justify that they are in thermal equilibrium and so the averages would have to be Boltzmann weighted.

It is important that the MD runs have converged structurally and energetically, both within one and the same run and between simulations with different docking poses, since this will determine the precision in the binding free energy estimates. Structurally, the simulations are stable with low RMSD of the ligands and only minor fluctuations in the protein. Specifically, in the reference OPLS-AA charge simulations the maximum unfitted RMSD among the ten inhibitors and their three poses is largest for compound **52**, namely 3.1 Å, whereas it does not exceed 2.3 Å for the other ligands. However, the largest contribution to this value occurs during equilibration, after which the ligands have settled into a position on average differing by 1.9 Å for compound **52** and 1.2 Å for the rest (Figure 6). After this initial displacement the fluctuations about their equilibrated positions are quite small. Such small variations are expected since the inhibitors are rigid and bind to a well-defined allosteric pocket. The main structural fluctuation, occurring only in a few cases, is found in the side chain of Lys103, which is located close to the carbonyl group of the pyridinone ring in the docked starting position (Figure 7) as well as in the

crystal structure.<sup>26</sup> In cases where this lysine leaves the carboxyl group and starts to interact with the solvent, the electrostatic ligand-surrounding energies are substantially decreased, which results in an increase in the predicted binding free energy.

Turning to the energies, convergence within runs is easily examined and is seen to be satisfactory in all cases. However, convergence between docking poses requires some additional figure of measure. To this end, the standard deviation of the protein–ligand electrostatic interaction energy between triplicates has been adopted, since it is the electrostatic energy that gives the largest contribution to the error and is mostly affected upon exchange of partial charges. Indeed, the van der Waals interaction energies were seen to be fairly constant throughout the simulations. As shown in Figure 8, the cumulative standard deviations of the ligand-protein electrostatic interaction energies span an order of magnitude, where CM1A and the standard OPLS-AA charge sets give the most converged estimations. Interestingly, CM1A charges are seen to give better convergence in this respect than the OPLS-AA fragment based method. On the other hand, charge polarization either by linear scaling or continuum solvent consistently increases the binding free energy error. As expected, Mulliken populations with diffuse basis sets stand out as a terrible choice by this measure.

At this point it is important to remember that precision by itself has little or no merit if the accuracy of the



**Figure 9.** RMSD values between average structures of the OPLS-AA and the different charge model simulations summed over poses, that is  $\sum_j RMSD(\langle OPLSAA \rangle_j, \langle CM_n \rangle_j)$ , accumulated for each charge model.

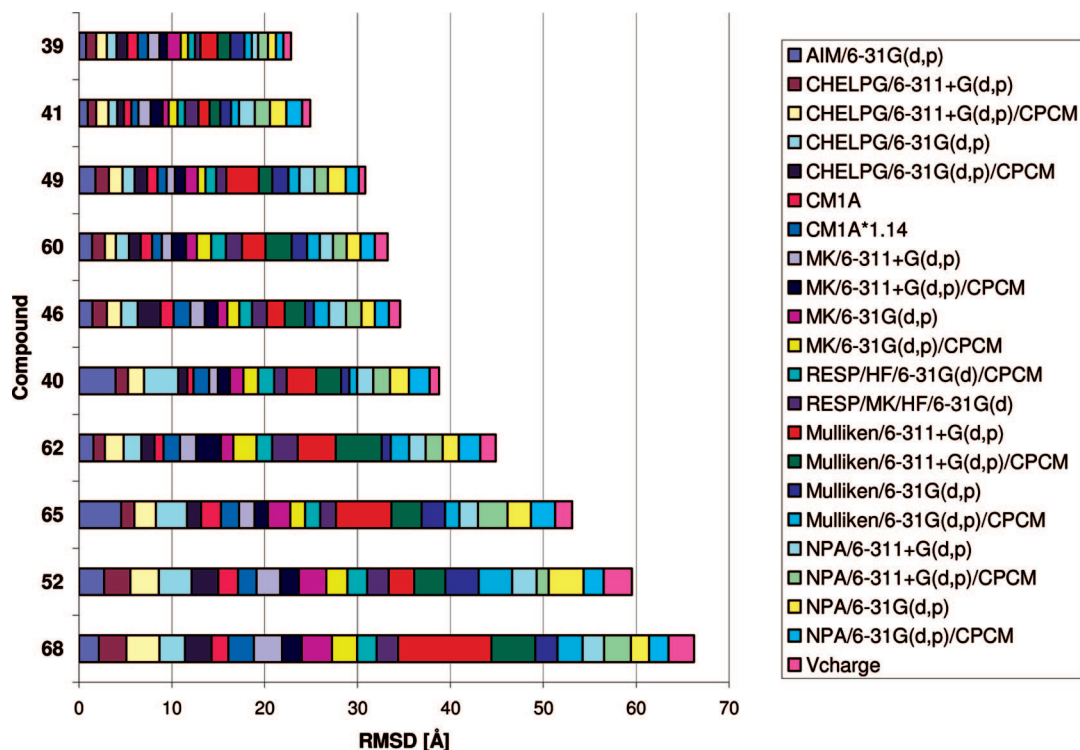
predictions is disregarded. For instance, with the present figure of measure one would obtain a very high precision by neutralizing all the charges but at a cost of severely reducing the accuracy of the predictions. Conversely, if there are models here that give highly accurate predictions but with a very low precision, they will be hard to distinguish from random noise—especially on such a limited number of observations. For this reason, it is argued that both precision and accuracy are of moment here, and charge sets with an accumulated standard deviation in ligand-surrounding electrostatic energies that exceeds 20 kcal/mol are therefore discarded henceforth.

Since there are considerable variations in the electrostatic ligand-surrounding energies, one may question whether this is reflected in the structural stability of the bound complex. As judged by the  $\sum_j RMSD(\langle OPLSAA \rangle_j, \langle CM_n \rangle_j)$  values shown in Figures 9 and 10 however, the structures seem not to be greatly affected by the choice of charges, except for the aforementioned rather extreme Mulliken populations. One may also note that the CHELPG/6-31G(d,p) model, with a rather large accumulated RMSD in Figure 9, still is among the most precise models with respect to ligand-surrounding electrostatic energies in Figure 8. Hence, the precision in the binding free energy estimation does not appear very sensitive to minor structural rearrangements of the ligand-protein complexes. Furthermore, from Figure 10 it seems that the structural variations are inherent to the ligand rather than the charge model. Indeed, plotting the most aberrant average structure in this respect, namely compound **68** with the Mulliken/6-311+G(d,p) charge set, together with the OPLS-AA average structure and its corresponding initial

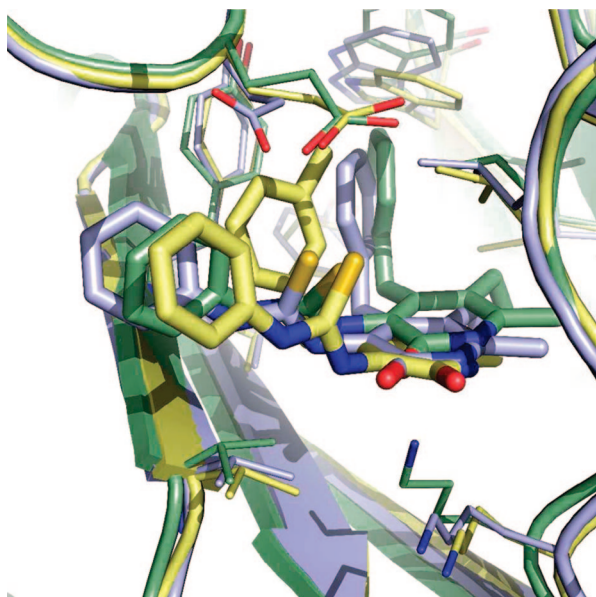
docking pose, reveals that the significant unfitted RMSD in this case is due to translation and rotation of the ligand and not a change in the actual binding mode (see Figure 11).

**Relative Binding Free Energies.** After optimization of the constant offset parameter  $\gamma$  in the LIE standard model, the accuracy of the calculated relative binding free energies from the different charge sets was estimated through the coefficient of determination ( $R^2$ ), Spearman's coefficient of rank correlation ( $\rho$ ), and the mean unsigned error ( $\langle |Error| \rangle$ ) with respect to experimental data. The results are given in descending order of accuracy with respect to  $R^2$  in Table 3 together with the rms difference in charge from the OPLS-AA fragment method. *Caveat lector*—as already mentioned the compounds were selected from our previous study<sup>24</sup> to give accurate binding free energies and to be well behaved during simulations with the OPLS-AA force field. However, since the  $R^2$  for all the compounds in the original study was 0.70 and the ten compounds selected here score better than average with an  $R^2$  of 0.90 (see Figure 12A), there is a slight statistical bias toward these charges. In the same way, the unsigned average error was 0.80 kcal/mol in the original data set, whereas it is 0.49 kcal/mol for this subset. This is not a major concern however, since our aim is rather to compare the remainder of the methods with respect to each other than with respect to the fragment based method. The performance of this method in binding free energy calculations is already known (see for instance the original data set<sup>24</sup>).

Bearing this in mind, MK ESP charges from the B3LYP/6-311+G(d,p) model chemistry emerge with the highest recorded accuracy with respect to  $R^2$ , shown in Figure 12B,



**Figure 10.** RMSD values between average structures of the OPLS-AA and the different charge model simulations summed over poses, that is  $\sum_j \text{RMSD}(\langle \text{OPLSAA} \rangle_j, \langle \text{CM}_n \rangle_j)$ , accumulated for each ligand.



**Figure 11.** The docked pose for compound **68** in green, superpositioned with the corresponding OPLS-AA and Mulliken/6-311+G(d,p) average structures in blue and yellow, respectively. The Mulliken average structure is the most disparate with respect to OPLS-AA in terms of RMSD.

closely followed by CM1A and Vcharge in Figure 12C,D. However, taking the mean unsigned error into account the ESP method is superseded by CM1A, which is mainly due to the outlying compound **46** whose squared residual inflicts a heavy penalty on the coefficient of determination, as can be seen in Figure 13. From Figure 12 it seems fairly clear that the two bad binders **46** and **68**, whose experimental affinities are well separated from the rest, remain distin-

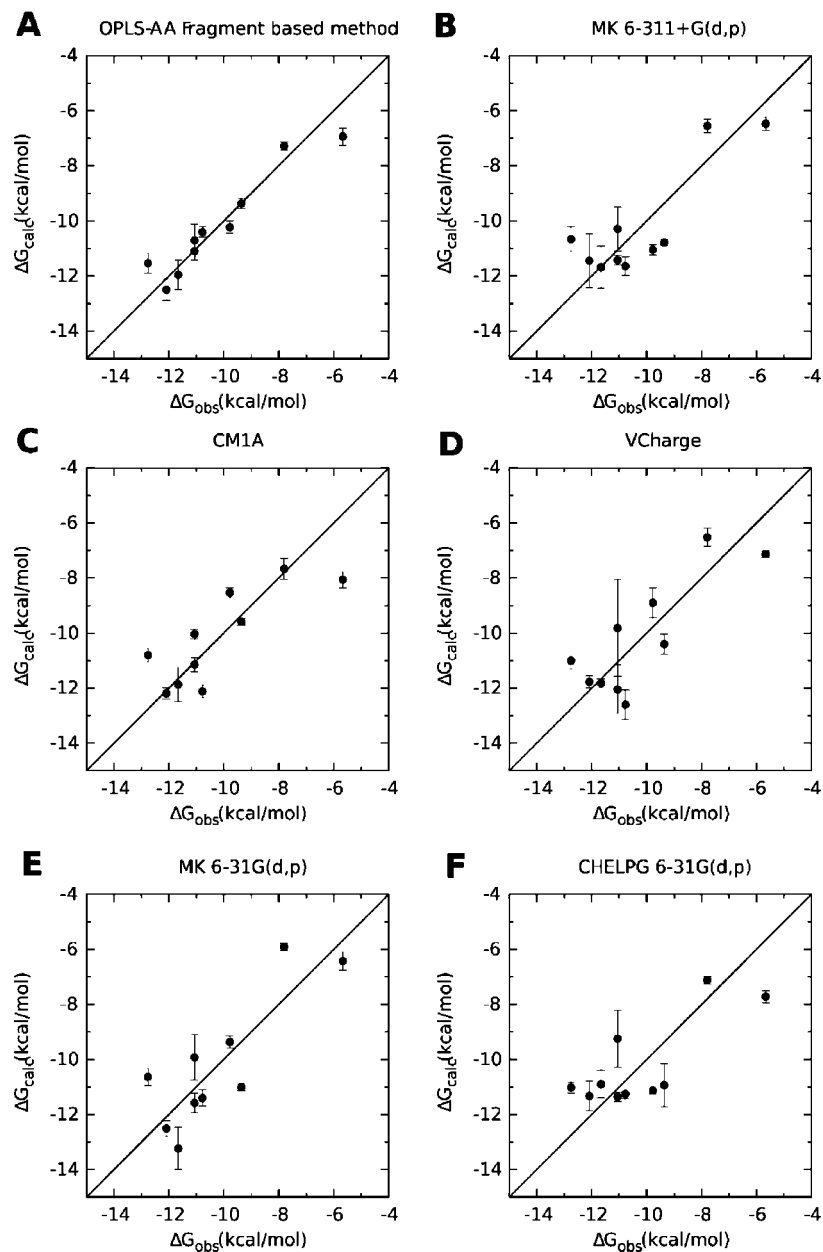
**Table 3.** Charge Deviations with Respect to the OPLS-AA Charges As Well as the Mean Unsigned Error,  $R^2$ ,  $Q^2$ , and Spearman's  $\rho$  from Binding Free Energy Calculations for the Different Charge Sets<sup>a</sup>

	charge RMSD <sup>b</sup>	(Err) <sup>c</sup>	$R^2$	$Q^2$	$\rho$	$\gamma$
OPLS-AA	N/A	0.49	0.90	0.88	0.96	-10.03
MK 6-311+G(d,p)	0.24	0.95	0.71	0.64	0.53	-8.97
CM1A	0.09	0.87	0.66	0.58	0.76	-10.35
Vcharge	0.09	1.09	0.64	0.56	0.61	-8.18
MK/6-31G(d,p)	0.20	1.11	0.61	0.52	0.67	-8.04
CHELPG/6-31G(d,p)	0.15	1.15	0.59	0.50	0.51	-7.48
Mulliken/6-31G(d,p)	0.14	1.20	0.53	0.42	0.61	-7.65
CM1A*1.14	0.10	1.34	0.40	0.26	0.70	-13.05
CHELPG/6-311+G(d,p)	0.18	1.41	0.36	0.21	0.35	-7.89
RESP/MK/HF 6-31G(d)	0.18	1.36	0.28	0.12	0.70	-8.93
MK/6-31G(d,p)/CPCM	0.21	1.56	-0.11	-0.37	0.67	-10.59
CHELPG/6-31G(d,p)/CPCM	0.16	1.90	-0.29	-0.59	0.44	-9.14

<sup>a</sup> The methods are presented in descending order with respect to  $R^2$ . <sup>b</sup> Given in [e]. <sup>c</sup> Given in [kcal/mol].

guished from the cluster of good binders for all these charge sets. However, Spearman's coefficient of rank correlation given in Table 3 seems to capture the apparent lack of agreement within the cluster of good binders for the MK/6-311+G(d,p) and CHELPG/6-31G(d,p) charges, giving a relatively low  $\rho$  of 0.53 and 0.51, respectively. This can be contrasted with the CM1A ranking where  $\rho$  is 0.76, followed by CM1A\*1.14 and the RESP charges with a  $\rho$  of 0.70. Hence, the rather low values of  $R^2$  and  $Q^2$  for the latter two methods seem not to be reflected in the ranking of the compounds.

As judged from their relative contribution to the SSE in Figure 13, it appears that the compounds **46** and **62** are the hardest to predict with LIE regardless of charge model,



**Figure 12.** Calculated versus experimentally observed binding free energies for the highest ranked charge models in this survey.

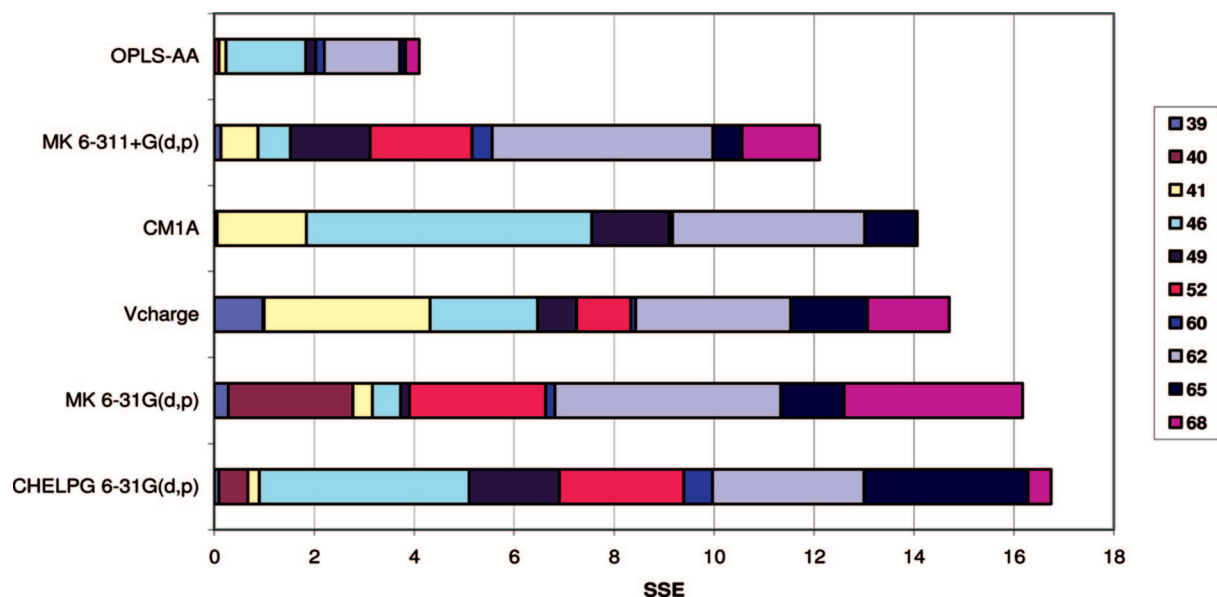
whereas **52** and **68** have relatively inaccurate predictions with the ESP and Vcharge methods but not with CM1A and OPLS-AA. Ligands **52** and **68** have a benzyl group in the  $R_3$  moiety in common, but the most striking differences in actual charge are rather found in the nonvariable part of the compounds.

Turning to the predictive ability, the six most accurate charge sets in Table 3 also have a  $Q^2 \geq 0.5$  within the LIE standard model, which is considered acceptable for these purposes. Although leave-many-out cross-validation will generally provide a better measure of this, the two aforementioned bad binders would inflict a heavier penalty than warranted if two or more compounds were excluded in this process.

Interestingly, whereas CM1A and Vcharge produce sets that are relatively close to the OPLS-AA charges, as judged by the low charge rms of  $0.09 e$  in Table 3, the MK ESP

charges are seen to differ with as much as  $0.24 e$ . This could be compared with the least accurate scheme with respect to  $R^2$  and the mean unsigned error given in this table, CHELPG from the solvent polarized B3LYP/6-31G(d,p) wave function, that has a significantly lower rms of  $0.16 e$  but is seen to produce rather poor predictions. For this reason, the relative success of the MK ESP charges indicates that partial charges can still perform well with the LIE method even when they are dissimilar from the fragment based charges, which are in turn derived to lie as close to the force field definition as possible. In contrast, the accuracy of CM1A and Vcharge presumably follows from their relative similarity with these charges.

Introducing solvent polarization, either by linear scaling or a continuum model, generally has a negative effect on accuracy both for the *ab initio* ESP schemes and CM1A. It is conceivable that these polarization models actually amplify



**Figure 13.** The contribution from each ligand to the sum of squared errors (SSE) with respect to experimental and calculated binding free energies for the highest ranked charge models.

errors in the charge assignment, such as the neglect of protein induced polarization. If this is the case, it is conceivable that ligand polarization by e.g. QM/MM methods<sup>76</sup> would provide better charge descriptions in this respect.

The fact that relatively minor changes in the rms charge when performing CM1A\*1.14 linear scaling has a significantly negative impact on estimation accuracy may be somewhat unexpected, especially since it has been shown to give excellent hydration free energy estimates,<sup>13,14</sup> but highlights the difference between solvation and binding free energies.

Taken together with the apparent success of ESP methods and simpler schemes adjusted to reproduce the overall electrostatic potential of the compounds, it appears that the LIE method is more sensitive toward observable changes in the electrostatic properties of the system than variations of the nonobservable and somewhat arbitrary partial charges of the constituent atoms. This suggests that these results should be readily transferrable to other protein systems and ligand classes as well. Furthermore, this could also be the reason why RESP fitting from the HF/6-31G(d) wave function is seen not to give nearly as precise predictions as the plain MK scheme applied to the higher level DFT density, although their charges are seen to be very similar (cf. Figure 4).

## Conclusions

After having taken measures of precision, accuracy, and internal statistical validation into account, the Merz–Kollman–Singh, CM1A, Vcharge, and CHELPG schemes are seen to provide ligand partial charges that perform well in binding free energy calculations with LIE. Among these, CM1A and Vcharge are also both computationally cheap and easily automated. Since CM1A has the additional advantage of providing a good ranking correlation with respect to experiment, this method emerges as an attractive choice for high-throughput LIE with the OPLS-AA force field. In view of

the close relationship between LIE and free energy perturbation/thermodynamic integration simulations, one would expect that the results obtained herein also should apply to those methods.

**Acknowledgment.** We wish to thank Professor William Jorgensen for insightful comments on charge scaling. Support from the Swedish Foundation for Strategic Research (SSF/Rapid) and the Swedish Research Council (VR) is gratefully acknowledged.

## References

- (1) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. *Adv. Protein Chem.* **2003**, *66*, 123–158.
- (2) Foloppe, N.; Hubbard, R. *Curr. Med. Chem.* **2006**, *13*, 3583–3608.
- (3) Gohlke, H.; Klebe, G. *Angew. Chem., Int. Ed.* **2002**, *41*, 2645–2676.
- (4) Åqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385–391.
- (5) Almlöf, M.; Brandsdal, B. O.; Åqvist, J. *J. Comput. Chem.* **2004**, *25*, 1242–1254.
- (6) Bjelic, S.; Nervall, M.; Gutierrez-de-Teran, H.; Ersmark, K.; Hallberg, A.; Åqvist, J. *Cell. Mol. Life Sci.* **2007**, *64*, 2285–2305.
- (7) Almlöf, M.; Andér, M.; Åqvist, J. *Biochemistry* **2007**, *46*, 200–209.
- (8) Andér, M.; Luzhkov, V. B.; Åqvist, J. *Biophys. J.* **2008**, *94*, 820–831.
- (9) Bortolato, A.; Moro, S. *J. Chem. Inf. Model.* **2007**, *47*, 572–582.
- (10) Carlsson, J.; Åqvist, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385–5395.
- (11) Huang, D. Z.; Luthi, U.; Kolb, P.; Cecchini, M.; Barberis, A.; Caffisch, A. *J. Am. Chem. Soc.* **2006**, *128*, 5436–5443.

- (12) Kolb, P.; Huang, D.; Dey, F.; Cafilisch, A. *J. Med. Chem.* **2008**, *51*, 1179–1188.
- (13) Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.
- (14) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- (15) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.
- (16) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.
- (17) Bieglerkonig, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *3*, 317–328.
- (18) Bieglerkonig, F. W.; Nguyendang, T. T.; Tal, Y.; Bader, R. F. W.; Duke, A. J. *J. Phys. B: At., Mol. Opt. Phys.* **1981**, *14*, 2739–2751.
- (19) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (20) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
- (21) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.
- (22) Gilson, M. K.; Gilson, H. S. R.; Potter, M. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.
- (23) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.
- (24) Carlsson, J.; Boukharta, L.; Åqvist, J. *J. Med. Chem.* **2008**, *51*, 2648–2656.
- (25) Benjihad, A.; Croisy, M.; Monneret, C.; Bisagni, E.; Mabire, D.; Coupa, S.; Poncelet, A.; Csoka, I.; Guillemont, J.; Meyer, C.; Andries, K.; Pauwels, R.; de Bethune, M. P.; Himmel, D. M.; Das, K.; Arnold, E.; Nguyen, C. H.; Grierson, D. S. *J. Med. Chem.* **2005**, *48*, 1948–1964.
- (26) Himmel, D. M.; Das, K.; Clark, A. D.; Hughes, S. H.; Benjihad, A.; Oumouch, S.; Guillemont, J.; Coupa, S.; Poncelet, A.; Csoka, I.; Meyer, C.; Andries, K.; Nguyen, C. H.; Grierson, D. S.; Arnold, E. *J. Med. Chem.* **2005**, *48*, 7582.
- (27) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–748.
- (28) Marelus, J.; Kolmodin, K.; Feierberg, I.; Åqvist, J. *J. Mol. Graphics Modell.* **1998**, *16*, 213–225.
- (29) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (30) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Rw, I.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (31) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (32) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647–3661.
- (33) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100–3107.
- (34) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512–9521.
- (35) Hansson, T.; Marelus, J.; Åqvist, J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27–35.
- (36) Luthi, H. P.; Ammeter, J. H.; Almlof, J.; Faegri, K. *J. Chem. Phys.* **1982**, *77*, 2002–2009.
- (37) Löwdin, P. O. *J. Chem. Phys.* **1950**, *18*, 365–375.
- (38) Momany, F. A. *J. Phys. Chem.* **1978**, *82*, 592–601.
- (39) Cox, S. R.; Williams, D. E. *J. Comput. Chem.* **1981**, *2*, 304–323.
- (40) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.
- (41) Connolly, M. L. *Science* **1983**, *221*, 709–713.
- (42) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.
- (43) Chirlian, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *8*, 894–905.
- (44) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (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.* **1996**, *118*, 2309–2309.
- (46) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (47) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (48) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.
- (49) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.
- (50) Hinze, J.; Whitehead, M. A.; Jaffe, H. H. *J. Am. Chem. Soc.* **1963**, *85*, 148.
- (51) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.
- (52) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 616–641.
- (53) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (54) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (55) Miehlisch, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.
- (56) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.
- (57) *Gaussian 03, revision C.02*; Gaussian Inc.: Wallingford, CT, 2004.
- (58) *AIMPAC, version 94*; Department of Chemistry, McMaster University: Hamilton, Ontario, Canada, 1994.
- (59) *AmberTools, 1.2*; University of California: San Francisco, CA, 2008.
- (60) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M. J.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.
- (61) *AMSO, version 7.0*; University of Minnesota: Minneapolis, MN, 2003.
- (62) *Vcharge, 1.0*; VeraChem LLC: Germantown, MD, 2004.
- (63) Pugh, D. Electric multipoles, polarizabilities and hyperpolarizabilities In *Chemical Modelling: Applications and Theory*; Hinchliffe, A., Ed.; RSC: Cambridge, U.K., 2000; Vol. 1, pp 1–37.
- (64) Herzberg G. Vibrational infrared and Raman spectra. In *Infrared and Raman spectra of polyatomic molecules*; 1st ed.; Van Nostrand: New York, NY 1945; pp 239–269.
- (65) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.
- (66) Miertus, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239–245.



- (67) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799, 805.
- (68) Klamt, A. *J. Phys. Chem.* **1995**, 99, 2224–2235.
- (69) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, 24, 669–681.
- (70) Ochterski, J. Vibrational analysis in Gaussian, 1999. Gaussian white papers. [http://www.gaussian.com/g\\_whitepap/vib.htm](http://www.gaussian.com/g_whitepap/vib.htm) (accessed Aug 11, 2007).
- (71) Kendall, M. G. The measurement of rank correlation. In *Rank Correlation Methods*; 1st ed.; Charles Griffin and Co.: London, U.K., 1948; pp 8–10.
- (72) Seifert, M. H. J.; Kraus, J.; Kramer, B. *Curr. Opin. Drug Discovery Dev.* **2007**, 10, 298–307.
- (73) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, 20, 269–276.
- (74) Martens, H. A.; Dardenne, P. *Chemom. Intell. Lab. Syst.* **1998**, 44, 99–121.
- (75) Johnson, R. D., NIST Computational Chemistry Comparison and Benchmark Database, 2005. <http://srdata.nist.gov/cccbdb> (accessed Aug 11, 2007).
- (76) Illingworth, C. J. R.; Gooding, S. R.; Winn, P. J.; Jones, G. A.; Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2006**, 110, 6487–6497.

CT800404F

# JCTC

Journal of Chemical Theory and Computation

## Lennard–Jones Parameters for B3LYP/CHARMM27 QM/MM Modeling of Nucleic Acid Bases

Ulla Pentikäinen,<sup>\*,†,‡</sup> Katherine E. Shaw,<sup>†</sup> Kittusamy Senthilkumar,<sup>†,§</sup>  
Christopher J. Woods,<sup>†</sup> and Adrian J. Mulholland<sup>\*,†</sup>

*Centre for Computational Chemistry, School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, United Kingdom, and Department of Biological and Environmental Science and NanoScience Center, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland*

Received April 20, 2008

**Abstract:** Combined quantum mechanics/molecular mechanics (QM/MM) methods allow computations on chemical events in large molecular systems. Here, we have tested the suitability of the standard CHARMM27 forcefield Lennard–Jones van der Waals (vdW) parameters for the treatment of nucleic acid bases in QM/MM calculations at the B3LYP/6–311+G(d,p)-CHARMM27 level. Alternative parameters were also tested by comparing the QM/MM hydrogen bond lengths and interaction energies with full QM [B3LYP/6–311+G(d,p)] results. The optimization of vdW parameters for nucleic acid bases is challenging because of the likelihood of multiple hydrogen bonds between the nucleic acid base and a water molecule. Two sets of optimized atomic vdW parameters for polar hydrogen, carbonyl carbon, and aromatic nitrogen atoms for nucleic acid bases are reported: base-dependent and base-independent. The results indicate that, for QM/MM investigations of nucleic acids, the standard forcefield vdW parameters may not be appropriate for atoms treated by QM. QM/MM interaction energies calculated with standard CHARMM27 parameters are found to be too large, by around 3 kcal/mol. This is because of overestimation of electrostatic interactions. Interaction energies closer to the full QM results are found using the optimized vdW parameters developed here. The optimized vdW parameters [developed by reference to B3LYP/6–311+G(d,p) results] were also tested at the B3LYP/6–31G(d) QM/MM level and were found to be transferable to the lower level. The optimized parameters also model the interaction energies of charged nucleic acid bases and deprotonation energies reasonably well.

### Introduction

Computational studies of chemical reactions in condensed phases ideally require a method that describes electronic changes in the region of interest. Current quantum mechanical (QM)–molecular electronic structure approaches can provide such descriptions, but the relatively high cost of these

methods limits the size of the systems that can be treated.<sup>1</sup> However, when for example enzyme-catalyzed reactions are studied, inclusion of the surrounding enzyme and water molecules can be crucial for the reliable treatment of the reaction energetics.<sup>2</sup> This, however, increases the size of the system significantly, making calculations too computationally expensive for pure QM methods. To investigate the effects of the environment on chemical events, an implicit or explicit representation of the environment is needed. For biological systems, combined QM and molecular mechanical (MM) methods are increasingly popular and important.<sup>3–9</sup> QM/MM methods enable computations on complex chemical events in large systems, by dividing the system into a

\* E-mail: ulla.m.pentikainen@jyu.fi; Adrian.Mulholland@bris.ac.uk.

† University of Bristol.

‡ University of Jyväskylä.

§ Present address: Department of Physics, Bharathiar University, Coimbatore 641 046 India.

quantum region and molecular mechanics region. QM/MM methods are well suited for studying enzyme-catalyzed reactions, which take place in a solvent and biomolecular (e.g., protein, nucleic acids, carbohydrates, or lipids) environment, involving thousands of atoms.<sup>3,6,7</sup>

The total Hamiltonian for the molecular system under consideration in the QM/MM framework can be written as:

$$H = H_{\text{QM}} + H_{\text{QM/MM}} + H_{\text{MM}} \quad (1)$$

where  $H_{\text{QM}}$  and  $H_{\text{MM}}$  are the normal QM and MM Hamiltonians that correspond to the atoms in the QM and MM regions, respectively. In eq 1 the QM/MM coupling term,  $H_{\text{QM/MM}}$ , typically contains terms for the electrostatic, van der Waals (vdW), and bonded interactions (eq 2).<sup>10–12</sup>

$$H_{\text{QM/MM}} = H_{\text{QM/MM(vdW)}} + H_{\text{QM/MM(elec)}} + H_{\text{QM/MM(bonded)}} \quad (2)$$

The  $H_{\text{QM/MM(bonded)}}$  term is required only where the partitioning into the QM and MM regions breaks covalent bonds. For such partitioning, the molecular mechanical bonding term is usually retained for interactions between covalently bonded QM and MM atoms, at the QM/MM boundary. The valency of the QM region is satisfied with the addition of link atoms<sup>10–14</sup> or by frozen orbital<sup>6,15</sup> or generalized hybrid orbital<sup>17–21</sup> approaches.

Several different methods have been used to describe the electrostatic interactions between the QM and MM regions, of which a so-called ‘electrostatic embedding’ scheme is the most common. In this model, interactions with the MM atomic point charges are included in the one-electron Hamiltonian of the QM region.<sup>14,16</sup> This model directly allows for the electronic polarization of the QM region by the MM environment. This is likely to be important to include in QM/MM studies of biological macromolecules because of their polar nature.

The vdW interaction between the QM and MM atoms in QM/MM calculations is typically included through a Lennard–Jones 12–6 potential,<sup>11</sup> as in standard biomolecular MM forcefields:<sup>22,23</sup>

$$V_{\text{vdW}}^{\text{QM/MM}} = \sum_A \sum_B 4\epsilon_{AB} \left[ \left( \frac{\sigma_{AB}}{R_{AB}} \right)^{12} - \left( \frac{\sigma_{AB}}{R_{AB}} \right)^6 \right] \quad (3)$$

where  $A$  and  $B$  are indices representing the QM and MM atoms, respectively,  $R_{AB}$  is the distance between the QM and MM atoms, and  $\epsilon_{AB}$  and  $\sigma_{AB}$  are calculated from vdW parameters for each atom in the forcefield, using standard combination rules. In the widely used CHARMM27 forcefield, the Lorentz–Berthelot combination rules are used.<sup>23</sup> This nonelectrostatic interaction term represents dispersion attractions, that fall off as  $R^{-6}$ , and also prevents molecular collapse at short distances between the QM and MM atoms (the  $R^{-12}$  term is used for computational convenience). The vdW interaction term is written in a slightly different form in CHARMM:

$$V_{\text{vdW}}^{\text{QM/MM}} = \sum_A \sum_B \epsilon_{AB} \left[ \left( \frac{R_{\text{min}}^{A,B}}{R_{AB}} \right)^{12} - 2 \left( \frac{R_{\text{min}}^{A,B}}{R_{AB}} \right)^6 \right] \quad (4)$$

where

$$R_{\text{min}} = \frac{1}{2} \sigma \quad (5)$$

and

$$\epsilon_{AB} = \sqrt{\epsilon_A \epsilon_B} \quad (6)$$

Hence in this paper we optimize  $R_{\text{min}}/2$  and  $\epsilon$ .

Typical MM forcefields used in calculations on biomolecules use vdW parameters that have been optimized to describe bonded and nonbonded interactions or to reproduce experimental thermodynamic data for small molecules.<sup>24</sup> Derivation of vdW parameters can be a time-consuming and laborious process. It would be convenient to be able to use existing MM vdW parameters in QM/MM modeling. It is, however, possible that this could lead to significant errors in QM/MM calculations, which employ a different theoretical basis. In addition, the fact that electronic polarization (of the QM region by the MM region) is included in the QM/MM calculations but is modeled only indirectly, in an average way, by MM, could lead to optimal MM and QM/MM parameters being quite different. The suitability of vdW parameters for normal forcefield calculations for use in QM/MM calculations for different complexes has been studied earlier by several groups.<sup>25–31</sup> A systematic study of Lennard–Jones parameters and QM/MM hydrogen bonding energies was reported by Gao and Xia in 1992.<sup>25</sup> They calculated hydrogen bonding energies and geometries of 53 water complexes, covering functional groups on amino acids and nucleotide bases, using Monte Carlo AM1/TIP3P simulations. Comparison of AM1/TIP3P and ab initio 6–31G(d) results showed that adjustment of the OPLS Lennard–Jones parameters for H, C, N, and O in the QM region was necessary to get the best agreement between QM/MM and pure QM hydrogen bonding energies. This, however, resulted in a reduction in the accuracy of the geometries. The necessity to optimize vdW parameters for atoms in the QM region was also observed later, in a study of 45 organic small molecule–water complexes, using the AM1/TIP3P model.<sup>26</sup> Refinement of vdW parameters has also been found to be necessary when ab initio QM/MM methods are used.<sup>27</sup> When optimized vdW parameters were used, a good agreement between ab initio 3–21G and the MM OPLS–TIP3P potential and ab initio 6–31G(d) results was observed, for over 80 hydrogen bonded complexes of organic compounds with water. In the parametrization process it was also found that it was necessary to use different Lennard–Jones parameters for oxygen and nitrogen atoms in ionic and neutral molecules.<sup>27</sup> However, for hydrogen and carbon atoms, the same parameters were found to be suitable for both ionic and neutral complexes.<sup>27</sup> In addition to hydrogen bonded complexes, refinement of vdW parameters has also found to be necessary when chemical reactions in the condensed phase are studied with QM/MM methods (AM1/CHARMM).<sup>28</sup> More recently, Riccardi et al.<sup>29</sup> compared their optimized vdW parameters for the SCC-DFTB/CHARMM method with results for vdW parameters selected from the CHARMM22 forcefield.<sup>23</sup> While the different parameter sets gave clear differences in results for gas-phase clusters and solvent structure around the solutes, it was observed that thermodynamic quantities (activation free

energies and reduction potentials) in the condensed phase were not very sensitive to the vdW parameters used in the QM/MM calculations. In contrast, Luque and co-workers, who studied the hydrogen bonded complexes between various functional groups and a water molecule at B3LYP, AM1, and PM3 levels, observed that the vdW parameters given in normal forcefields are not transferable from MM to QM/MM calculations.<sup>30</sup> In addition, they pointed out that vdW parameters are sensitive to the QM/MM formalism and parametrization details and warned against the direct transferability of vdW parameters between different QM/MM methods.<sup>30</sup> However, in several studies optimization of vdW parameters has been observed to be necessary, for example Lennard–Jones parameters in the B3LYP/6–31G(d)/AMBER potential were found to accurately reproduce B3LYP/6–31G(d) hydrogen bond energies and geometries when amino acid–water complexes are studied.<sup>31</sup>

In the current work, we have studied the geometries and interaction energies of nucleic acid base–water complexes at a full QM level [B3LYP/6–311+G(d,p), using the popular B3LYP hybrid density functional technique] and an equivalent QM/MM [B3LYP/6–311+G(d,p)-CHARMM27] level. We have evaluated the suitability of the standard CHARMM27 forcefield vdW parameters for nucleic acid bases<sup>23,24,32,33</sup> for the QM/MM calculations at the B3LYP/6–311+G(d,p) and B3LYP/6–31G(d) levels. With these relatively high levels of QM theory, it is not yet feasible to calculate thermodynamic properties by simulations. Previous studies of the suitability of normal (MM) forcefield vdW parameters for QM/MM calculations have generally concentrated on complexes containing other small molecules and not nucleic acid bases<sup>26–28,31</sup> except for example the early works of Gao et al., where the AM1/TIP3P model was used to study hydrogen bonding energies and geometries of nucleic acid base–water complexes.<sup>25,34</sup> In those studies, good agreement between ab initio Hartree–Fock 6–31G(d) and AM1/TIP3P interaction energies was observed. The interactions distances from AM1/TIP3P calculations were also observed to be in reasonable agreement with the ab initio results. Here we apply significantly higher levels of theory than these early studies. Reliable models of the reactions and energetics of nucleic acids are crucial because of their vital roles in all living systems. For example, the reactions of catalytic RNA molecules, i.e. ribozymes,<sup>35–38</sup> interstrand cross-linking of DNA bases<sup>39,40</sup> (which is believed to be responsible for the biological activity of a number of antitumor agents<sup>41</sup>), stacking of nucleic acid bases in nucleic acid structures,<sup>42,43</sup> drug binding, and also reactivity<sup>44–48</sup> have been studied. Here, we report optimized vdW parameters for polar hydrogen, carbonyl carbon, and aromatic nitrogen atoms of nucleic acid bases by using a set of hydrogen bonded complexes at the B3LYP/6–311+G(d,p) QM and the B3LYP/6–311+G(d,p)-CHARMM27 QM/MM levels. In addition, the transferability of the optimized parameters to the lower B3LYP/6–31G(d)-CHARMM27 level, which is commonly used in QM/MM calculations,<sup>7,49</sup> is also tested. The suitability of the optimized nucleic acid base parameters was also tested on complexes other than those used in the parameter optimization. Also the suitability of the standard

CHARMM27<sup>23,24,32,33</sup> and optimized vdW parameters was tested for modeling chemical changes for simple reactions of nucleic acid bases.

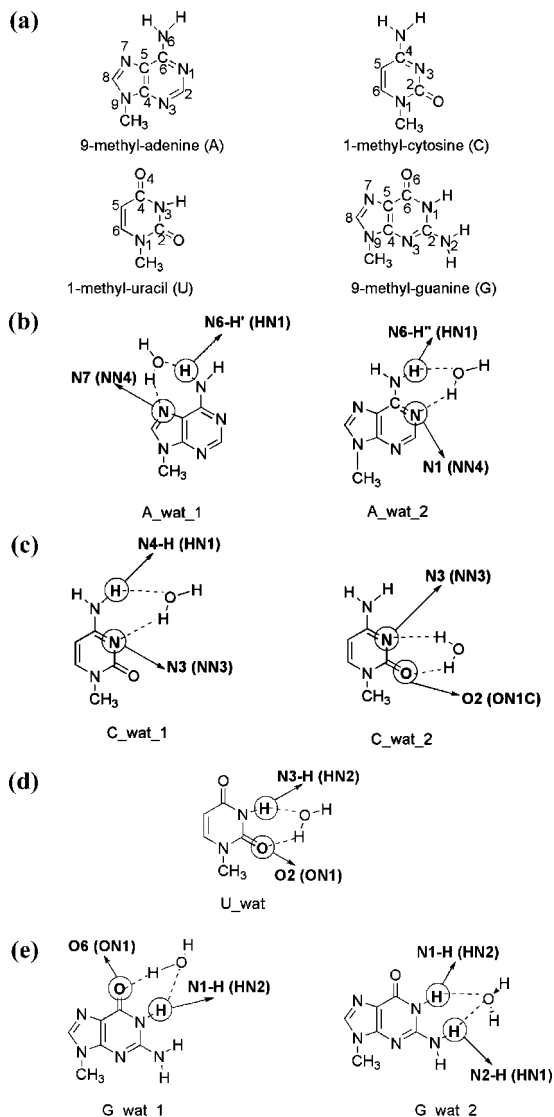
## Computational Details

QM calculations were performed using the Jaguar 6.0<sup>50</sup> and Gaussian03<sup>51</sup> programs. Full QM optimizations of nucleic acid base (Figure 1A)–water hydrogen bonded complexes shown in Figures 1B–E, 2, and 3 were performed at the B3LYP/6–311+G(d,p) level of theory. Work by Jorgensen and co-workers has shown that the B3LYP/6–311+G(d,p) level of theory (compared to electron correlated ab initio methods) treats hydrogen-bonded complexes well.<sup>52</sup> The counterpoise correction scheme of Boys and Bernardi<sup>53</sup> was used as an estimate to correct for basis set superposition error (BSSE) in the calculated QM interaction energies.

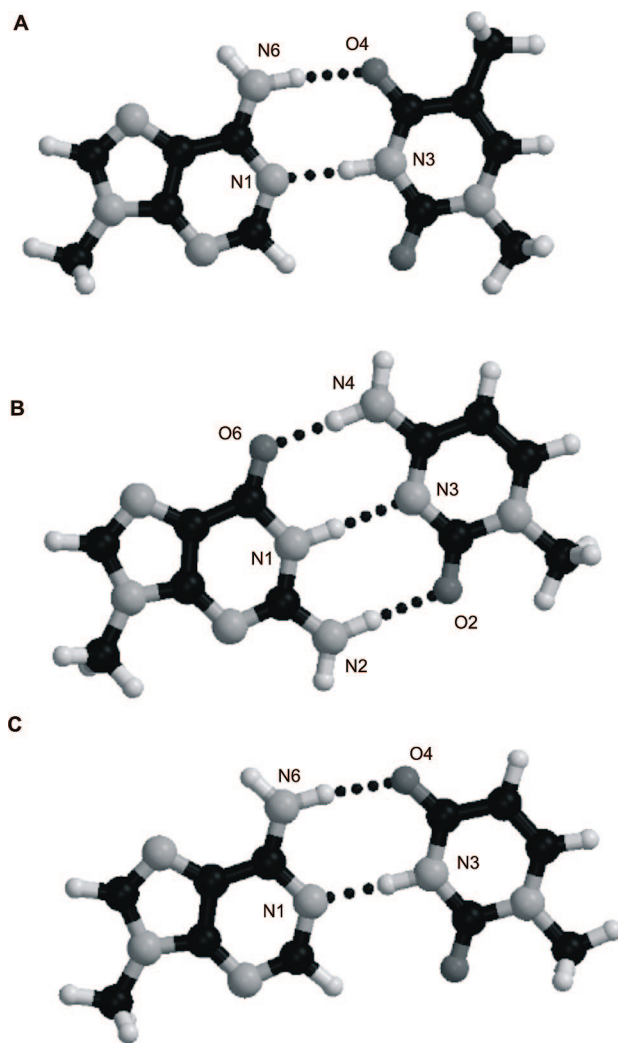
In the QM/MM calculations, the QM calculations were performed using Jaguar 5.0<sup>50</sup> at the B3LYP/6–311+G(d,p) or B3LYP/6–31G(d) level. The Tinker molecular mechanics program<sup>54</sup> with the CHARMM27 all-atom forcefield<sup>23,24,32,33</sup> was used to evaluate the MM terms. The results from the QM and the MM calculations were combined using the QM/MM interface program QoMMMa.<sup>55</sup> The QoMMMa program optimizes the QM geometry within the MM environment, and the positions of the MM atoms are fully relaxed at each QM step using the Tinker program.<sup>54</sup> In the nucleic acid–water complexes, the nucleic acid base was always defined as the QM region and the water molecule as the MM region. In nucleic acid base pair calculations, one base was defined as the QM region and the other as the MM region. MM calculations were performed with the CHARMM program.<sup>56,57</sup>

## Parameter Optimization

The vdW parameters of polar hydrogen, carbonyl carbon, and aromatic nitrogen atoms in four nucleic acid bases (Figure 1A) were optimized. The Lennard–Jones parameters (the minimum in the vdW interaction energy surface  $R_{\min}/2$  and the well depth,  $\epsilon$ ) for the QM atoms were adjusted systematically to obtain the best agreement for hydrogen bond energies and geometries between the QM/MM and the full QM calculations. At the beginning of the optimization, an increment of 0.1 Å for  $R_{\min}/2$  and 0.01 kcal/mol for  $\epsilon$  were used. As we got closer to hydrogen bond energies and geometries obtained from the full QM calculations, smaller increments were used. Both  $R_{\min}/2$  and  $\epsilon$  were adjusted simultaneously. The best parameters were determined by eye, and if it was difficult to get parameters which gave both optimal hydrogen bond energies and geometries, then parameters were chosen to give optimal hydrogen bond energy rather than geometry. Two different types of parameters, base-dependent and base-independent, have been developed here. By base-dependent parameters, we mean that unique parameters for each atom type were developed for each nucleic acid base. Base-independent parameters are, in contrast, the same for each nucleic acid base, i.e., dependent only on atom type. Parameters of atoms in the MM region, (i.e., the atomic charges and Lennard–Jones parameters in



**Figure 1.** A. The four nucleic acid bases used for parameter optimization. The structures also show atom numbering. B. The two hydrogen bonded complexes of adenine are shown here. The bold circled atoms are the two QM nucleic acid base atoms that form hydrogen bonds (see Figure 1A for atom numbering), and it is for these atoms that parameters are optimized. The CHARMM27 atom types for these hydrogen bonding atoms are given in parentheses. C. The two hydrogen bonded complexes of cytosine are shown here. The bold circled atoms are the two QM nucleic acid base atoms that form hydrogen bonds (see Figure 1A for atom numbering), and it is for these atoms that parameters are optimized. The CHARMM27 atom types for these hydrogen bonding atoms are given in parentheses. D. The hydrogen bonded complex of uracil is shown here (used as a model for thymine). The bold circled atoms are the two QM nucleic acid base atoms that form hydrogen bonds (see Figure 1A for atom numbering), and it is for these atoms that parameters are optimized. The CHARMM27 atom types for these hydrogen bonding atoms are given in parentheses. E. The two hydrogen bonded complexes of guanine are shown here. The bold circled atoms are the two QM nucleic acid base atoms that form hydrogen bonds (see Figure 1A for atom numbering), and it is for these atoms that parameters are optimized. The CHARMM27 atom types for these hydrogen bonding atoms are given in parentheses.

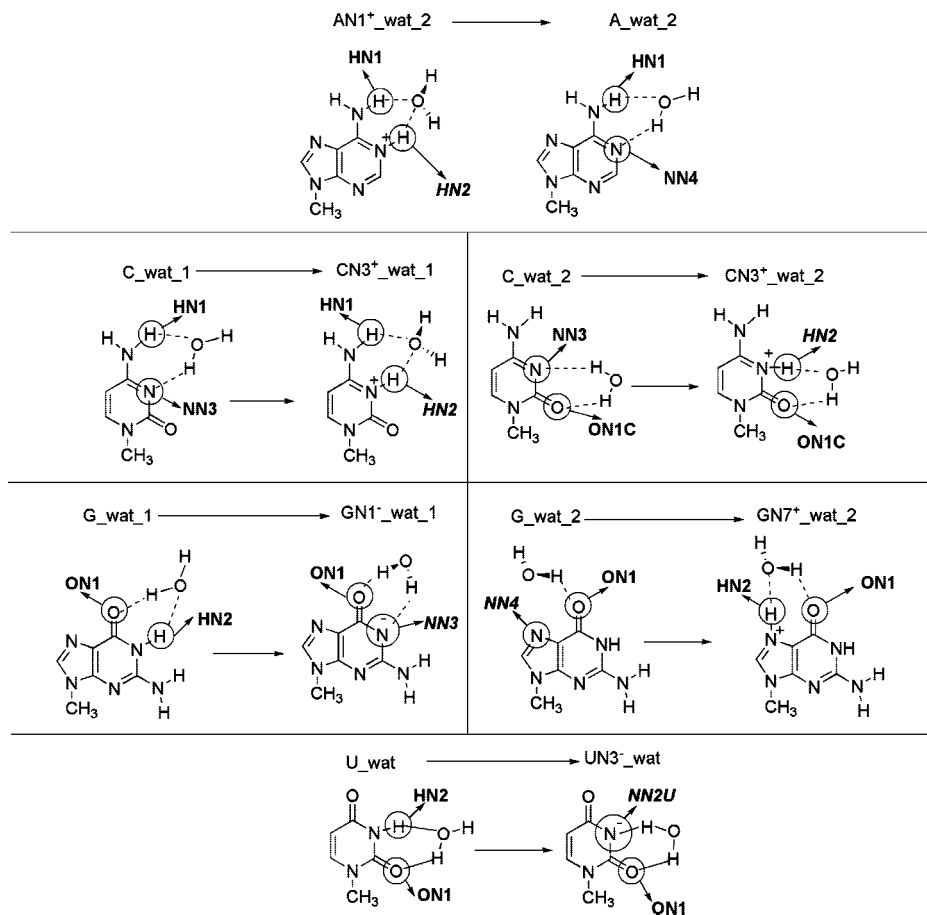


**Figure 2.** Structures of Watson Crick base pairs used in vdW parameter testing. A. Adenosine-thymine (AT) base pair. B. Guanosine-cytidine (GC) base pair. C. Adenosine-uracil (AU) base pair. Coloring of atoms: carbon atoms are in black, hydrogen atoms are in white, nitrogen atoms are in light gray, and oxygen atoms are in dark gray. The atom numbers of the atoms involved in the hydrogen bonding are shown on the structures (for atom numbering, see Figure 1A).

the CHARMM version of the TIP3P water model (TIP3P)<sup>58</sup> (which is modified to include vdW parameters for the water hydrogen atoms<sup>12,59</sup>) were kept unchanged, to ensure transferability and consistency with the CHARMM27 force-field. Geometries and interaction energies were also calculated using the optimized (base-dependent and independent) parameters at the lower B3LYP/6-31G(d) QM/MM level.

## Results and Discussion

**Complexes Used in vdW Parameter Optimization.** To model nucleosides, nucleic acid bases with a methyl group in place of the 1' carbon of the ribose ring were used (Figure 1A). The structures of the nucleic acid base–water complexes used in vdW parameter optimization are shown in Figures 1B–E. Uracil (U, Figure 1E) was used as a model for both uracil and thymine because uracil (found in RNA) is smaller than thymine (found in DNA). All the complexes



**Figure 3.** Protonation and deprotonation reactions studied with the newly developed vdW parameters. CHARMM27 atom types for nucleic acid base atoms involved in the reaction are shown (circled atoms). For atom types shown in italics, standard CHARMM27 parameters were used, as no base-dependent parameters were available. For atom numbering, see Figure 1A.

were constructed so that they contained two hydrogen bonds between the water molecule and the nucleic acid base, so that either the oxygen atom of the water molecule interacts with either the hydrogen atom of the exocyclic or endocyclic amino group, or the hydrogen atom of the water molecule interacts with either the carbonyl oxygen atom or the ring nitrogen atom of the nucleic acid base. For adenosine (A, Figure 1B), cytosine (C, Figure 1C), and guanosine (G, Figure 1D), two different water complexes were constructed (denoted A\_wat\_1, A\_wat\_2, etc.). In this way, vdW parameters for all different nitrogen [adenosine and guanosine N7 (CHARMM atom type NN4), adenosine N1 (NN4), cytosine N3 (NN3), see Figures 1B–D], hydrogen [adenosine N6-H (HN1), cytosine N4-H (HN1), uracil N3-H (HN2), guanosine N2-H (HN2), see Figures 1B–E], and oxygen [uracil O2 (ON1) and guanosine O6 (ON1)]; see Figures 1E and 1D] atom types found in nucleic acid bases could be studied.

**Comparison of QM and MM Geometries and Interaction Energies with QM/MM Results Using Standard CHARMM27 vdW Parameters.** The hydrogen bond distances for the nucleic acid base–water complexes obtained from the QM [B3LYP/6–311+G(d,p)] and QM/MM [B3LYP/6–311G+(d,p)-CHARMM27] optimizations are listed in Table 1. Generally, the hydrogen bond distances are shorter (by 0.02–0.22 Å, measured between the heavy atom and the hydrogen atom) in the QM/MM optimized

structures than in the full QM optimized structures, consistent with earlier studies.<sup>29,30,34</sup> However, in some cases, the opposite situation is observed: in the A\_wat\_2 complex (between N6-H'' and water), in the C\_wat\_2 complex (between N3 and water), and in the U\_wat complex (between O2 and water), the hydrogen bonds are clearly longer (by 0.3–0.5 Å) in the QM/MM optimized complexes than in the QM optimized complexes (Table 1; see Figures 1B–E for details of complexes). In the A\_wat\_1 complex, the hydrogen bond between N7 and water is also slightly longer (by 0.03 Å) in the QM/MM optimized complex than in the QM optimized complex (Table 1). The largest deviation between the QM and QM/MM optimized structures is seen for the C\_wat\_2 complex, where the distance between the nitrogen N3 atom and the hydrogen of the water molecule is 3.4 Å in the QM/MM optimized structure but only 2.9 Å in the QM optimized structure. However, this interaction is not a strong hydrogen bond, being too long. A similar observation was earlier made by Freindorf et al., who observed that the largest discrepancies in hydrogen bond distances optimized at QM [B3LYP/6–31+G(d)] and QM/MM [B3LYP/6–31+G(d)-AMBER] levels were observed when the A...H distance is longer than 2.5 Å, i.e., having more long-range electrostatic interaction character than hydrogen bond character.<sup>31</sup> In many (but not all) cases, the conformations of NH<sub>2</sub> groups are planar but slightly pyramidal in some QM (and QM/MM) calculations. As the

**Table 1.** Hydrogen Bond Distances (H-acceptor distances, Å) Calculated at the B3LYP/6-311+G(d,p) QM and B3LYP/6-311+G(d,p)-QM/MM Levels Using Either CHARMM27, Base-Dependent Parameters (Table 3), or Base-Independent Parameters (Table 4) for Complexes Shown in Figures 1B–E<sup>a</sup>

complex <sup>b</sup>	H-bond <sup>c</sup>	QM	QM/MM (CHARMM) <sup>d</sup>	QM/MM (OPT-BI) <sup>e</sup>	QM/MM (OPT-BD) <sup>f</sup>	MM (CHARMM) <sup>d</sup>
A_wat_1	N7–Hwat	1.89	1.92 (–0.03)	1.95 (–0.07)	1.95 (–0.07)	1.89 (0.00)
A_wat_1	N6–H'–Owat	1.95	1.80 (0.16)	2.00 (–0.05)	2.00 (–0.05)	1.88 (0.07)
A_wat_2	N6–H''–Owat	2.05	2.38 (–0.33)	2.39 (–0.34)	2.39 (–0.34)	2.01 (0.04)
A_wat_2	N1–Hwat	1.92	1.93 (–0.01)	1.93 (–0.01)	1.93 (–0.01)	1.90 (0.02)
C_wat_1	N4–H–Owat	2.01	1.93 (0.08)	2.11 (–0.10)	1.99 (0.01)	2.01 (0.00)
C_wat_1	N3–Hwat	1.92	1.90 (0.02)	1.91 (0.01)	1.94 (–0.02)	1.95 (–0.3)
C_wat_2	N3–Hwat	2.90	3.40 (–0.50)	3.53 (–0.63)	3.36 (–0.46)	1.94 (0.96)
C_wat_2	O2–Hwat	1.95	1.79 (0.16)	1.85 (0.10)	2.12 (–0.17)	2.39 (–0.44)
G_wat_1	O6–Hwat	1.86	1.80 (0.06)	1.87 (–0.01)	1.91 (–0.05)	1.87 (–0.01)
G_wat_1	N3–H–Owat	1.95	1.93 (0.01)	2.08 (–0.13)	2.19 (–0.25)	1.99 (–0.04)
G_wat_2	N1–H–Owat	2.10	1.93 (0.16)	2.07 (0.02)	2.17 (–0.07)	2.08 (0.02)
G_wat_2	N2–H–Owat	2.21	2.12 (0.09)	2.14 (0.08)	2.22 (–0.01)	2.00 (0.21)
U_wat	N3–H–Owat	2.00	1.79 (0.22)	1.98 (0.02)	1.97 (0.03)	1.99 (0.01)
U_wat	O2–Hwat	1.97	2.25 (–0.28)	2.23 (–0.27)	2.21 (–0.24)	1.94 (0.02)
STDEV <sup>g</sup>			0.21	0.22	0.15	0.31
MSE <sup>h</sup>			0.01	–0.01	0.12	–0.04
MUE <sup>i</sup>			0.15	0.13	0.13	0.15

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . The hydrogen bond distances are calculated between hydrogen atoms and hydrogen bond acceptors. <sup>b</sup> See Figures 1B–E for names of complexes. <sup>c</sup> See Figure 1A for labeling. <sup>d</sup> Standard CHARMM27 vdW parameters were used. <sup>e</sup> Optimized base-independent parameters were used. <sup>f</sup> Optimized base-dependent parameters were used. <sup>g</sup> Standard deviation of the error from the full QM calculated value. <sup>h</sup> Mean signed error (MSE) from the full QM calculated value,  $MSE = 1/n \sum_{i=1}^n (f_i - y_i)$ ,  $f_i$  = calculated value,  $y_i$  = QM or experimental value. <sup>i</sup> Mean unsigned error (MUE) from the full QM calculated value,  $MUE = 1/n \sum_{i=1}^n |f_i - y_i|$ .

**Table 2.** Interaction Energies (kcal/mol) Calculated at the B3LYP/6-311+G(d,p) QM and B3LYP/6-311+G(d,p)-QM/MM Levels Using Either CHARMM27, Base-Dependent Parameters (Table 3) or Base-Independent Parameters (Table 4) for Complexes Shown in Figures 1B–E<sup>a</sup>

complex <sup>b</sup>	QM	QM/MM (CHARMM) <sup>c</sup>	QM/MM (OPT-BI) <sup>d</sup>	QM/MM (OPT-BD) <sup>e</sup>	MM (CHARMM) <sup>c</sup>
A_wat_1	–9.9	–13.3 (3.4)	–11.9 (2.0)	–12.6 (2.7)	–10.4 (0.5)
A_wat_2	–8.9	–10.0 (1.1)	–10.2 (1.2)	–10.3 (1.4)	–9.5 (0.6)
C_wat_1	–10.5	–13.8 (3.2)	–12.9 (2.4)	–12.7 (2.3)	–11.1 (0.6)
C_wat_2	–7.0	–9.1 (2.1)	–8.7 (1.7)	–7.4 (0.4)	–10.4 (3.4)
G_wat_1	–10.8	–14.1 (3.3)	–12.8 (2.0)	–11.3 (0.6)	–11.5 (0.7)
G_wat_2	–7.7	–10.8 (3.1)	–10.6 (2.9)	–9.7 (2.1)	–10.9 (3.2)
U_wat	–7.5	–10.4 (2.9)	–9.3 (1.8)	–9.3 (1.8)	–9.6 (1.1)
STDEV <sup>f</sup>		0.8	0.5	0.9	1.3
MSE <sup>g</sup>		–2.7	–2.0	–1.6	–1.4
MUE <sup>h</sup>		2.7	2.0	1.6	1.4

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . <sup>b</sup> See Figure 1B–E for names of complexes. <sup>c</sup> Standard CHARMM27 vdW parameters were used. <sup>d</sup> Optimized base-independent parameters were used. <sup>e</sup> Optimized base-dependent parameters were used. <sup>f</sup> Standard deviation of the error from the full QM calculated value. <sup>g</sup> Mean signed error of from the full QM calculated value, see notes in Table 1. <sup>h</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

QM data are the target for optimization, and the pyramidity is not affected by the vdW parameters (and has little effect on hydrogen bond strength), this is not an important factor here.

The pure MM (CHARMM27) optimized hydrogen bond distances are very close to the full QM optimized results (Table 1). Only in the C\_wat\_2 complex is there a large difference between the MM and QM structures: the N3–Hwat hydrogen bond distance is 0.96 Å shorter in the MM optimized structure than in the QM optimized structure, and the O2–Hwat hydrogen bond is 0.44 Å longer in the MM optimized structure than in the QM optimized structure (Table 1, see Figures 1C for details of complex).

The hydrogen bond angles,  $\theta(D-H\cdots A)$ , calculated at the full QM and QM/MM levels, are usually within  $\sim 5^\circ$  of each other, although, in the U\_wat complex and in the C\_wat\_2 complex (hydrogen bond from water to O2), a larger deviation ( $\sim 12^\circ$ ) is seen (see Supporting Information, Table 12). Comparison of the MM optimized structures with

the QM optimized structures shows a somewhat larger deviation in the hydrogen bond angles than seen for the comparison of the QM/MM results with the pure QM results (see Supporting Information, Table 12).

The QM/MM interaction energies are overestimated in all cases when using standard CHARMM27 vdW parameters,<sup>23,24,32,33</sup> compared to those calculated at the pure QM level (Table 2). In most cases, the interaction energies are overestimated by around 3 kcal/mol (2.9 to 3.4 kcal/mol), but in the A\_wat\_2 complex, the difference is only 1.1 kcal/mol and in C\_wat\_2 it is 2.1 kcal/mol. It is interesting that in A\_wat\_2, the interaction energy is overestimated in the QM/MM calculations compared to the QM calculations, even though both hydrogen bond distances are longer in the QM/MM optimized structure than in the QM optimized structure. The complexes contain examples of QM groups as hydrogen bond acceptors and donors. There does not seem to be any clear difference in the deviations in hydrogen bond energies depending on whether the donor is QM or MM.

**Table 3.** Standard CHARMM27<sup>23, 27, 28</sup> and Optimized Base-Dependent vdW Parameters ( $R_{\min}/2$  and  $\epsilon$ ) for Nitrogen Polar, Hydrogen, and Carbonyl Oxygen Atoms for Nucleic Acid Bases

atom type <sup>a</sup>	CHARMM		OPT	
	$R_{\min}/2$ (Å)	$\epsilon$ (kcal/mol)	$R_{\min}/2$ (Å)	$\epsilon$ (kcal/mol <sup>-1</sup> )
Uracil/Thymine				
HN2	0.2245	-0.046	0.7	-0.046
ON1	1.7	0.120	1.7	0.120
Cytosine				
NN3	1.85	-0.200	1.95	-0.200
HN1	0.2245	-0.046	0.4245	-0.0460
ON1C	1.7	-0.120	1.9	-0.120
Guanine				
ON1	1.7	-0.120	1.8	-0.086
HN1	0.2245	-0.046	0.8245	-0.086
HN2	0.2245	-0.046	0.8245	-0.086
Adenine				
NN4	1.85	-0.200	1.9	-0.200
HN1	0.2245	-0.046	0.7	-0.046

<sup>a</sup> See Figures 1B–E for atom types.

The (CHARMM27) MM interaction energies are closer than the QM/MM interaction energies to the QM interaction energies (Table 2). In five complexes out of seven, the MM interaction energies are within 0.5–1.1 kcal/mol of the full QM interaction energies. This again demonstrates that the CHARMM27 parameters provide a good description of nucleic acid interactions in MM calculations. In two complexes, namely C\_wat\_2 and G\_wat\_2, the deviation from the QM interaction energies is larger, 3.4 and 3.2 kcal/mol, respectively.

**Optimized Base-Dependent vdW Parameters.** As described above, both hydrogen bond distances (Table 1) and interaction energies (Table 2) calculated at the QM/MM level with standard CHARMM27 Lennard–Jones parameters differed significantly from the full QM results. We therefore tested the effects of varying the vdW parameters for the nucleic acid bases treated with QM in the QM/MM calculations, in an attempt to improve the agreement, particularly for interaction energies. The vdW parameters of nitrogen (CHARMM atom types NN3 and NN4), polar hydrogen (atom types HN1 and HN2), and carbonyl oxygen (atom types ON1 and ON1C) atoms of the nucleic acid bases were changed systematically. The vdW parameters of the water molecule forming the MM part of the system were always kept unchanged (CHARMM TIPS3P<sup>58</sup>), for consistency with the CHARMM27 forcefield,<sup>23,24,32,33</sup> and because this represents likely interactions in typical QM/MM studies of nucleic acid bases (i.e., nucleic acid base treated by QM and water with MM).

To get the best agreement between the hydrogen bond distances and interaction energies calculated at the QM and QM/MM levels, base-dependent vdW parameters (i.e., different atomic parameters for different nucleic acid bases), were developed. These base-dependent parameters are listed in Table 3. Each nucleic acid base–water complex contained two hydrogen bonds between water and the nucleic acid base, which makes vdW parameter optimization challenging. The adenosine–water complexes, A\_wat\_1 and A\_wat\_2 (Figure 1B), were especially difficult. This was due to the fact that

when the original CHARMM27 vdW parameters were used, the hydrogen bond distance between HN1 and water in the complex A\_wat\_1 is shorter in the QM/MM optimized structure than in the QM optimized structure, whereas in the complex A\_wat\_2 the situation is reversed (Table 1). As the error in the interaction energy between the original QM and the QM/MM calculations is larger for the A\_wat\_1 complex than the A\_wat\_2 complex (Table 2), the hydrogen bond distance between HN1 and water in A\_wat\_1 complex was optimized rather than that in the A\_wat\_2 complex. A similar situation is also seen with C\_wat\_1 and C\_wat\_2. When the standard CHARMM27 vdW parameters were used, in the C\_wat\_1 complex the distance between NN3 and water is shorter in the QM/MM optimized structure than in the QM optimized structure, while in the C\_wat\_2 complex the situation is the opposite (Table 1).

As can be seen from Table 3, where the optimized base-dependent parameters are listed, the largest change was made for the  $R_{\min}/2$  value of hydrogen atoms. The original CHARMM27<sup>23,24,32,33</sup> value of 0.2245 Å is increased to 0.4245 Å on cytosine HN1, 0.7 Å on uracil/thymine HN2 and adenine HN1, and 0.8245 Å on guanine HN1 and HN2. The  $\epsilon$  values for guanine HN1 and HN2 were also changed from -0.046 to -0.086 kcal/mol. VdW parameters for nitrogen and oxygen atoms were changed only slightly: the  $R_{\min}/2$  value for cytosine nitrogen (NN3) was changed from 1.85 Å to 1.95 Å, for cytosine oxygen (ON1C) from 1.7 Å to 1.9 Å, guanine oxygen (ON1) from 1.7 Å to 1.8 Å, and adenine nitrogen (NN4) from 1.85 Å to 1.9 Å. In addition, the  $\epsilon$  value of the guanine oxygen was made slightly less negative, changed from -0.12 to -0.086 kcal/mol. Otherwise the change of the  $\epsilon$  value did not have any meaningful effect, either on the hydrogen bond lengths or the interaction energies, and so they are left at the original CHARMM27 values.

The hydrogen bond distances obtained with the optimized base-dependent vdW parameters, as well as the differences in hydrogen bond lengths between the QM optimized and QM/MM optimized structures using both standard CHARMM27 parameters and optimized base-dependent parameters, are shown in Table 1. In six cases, the hydrogen bond distances obtained were improved, compared to those obtained with the standard CHARMM27 vdW parameters. In five cases, the optimization of vdW parameters did not have any substantial effect and in three cases the hydrogen bond distances, compared to the full QM results, were worse than when the original parameters were used. This illustrates the problem of optimization of parameters for systems having two or more hydrogen bonds.

The interaction energies calculated using the optimized base-dependent parameters are generally overestimated by 0.4–2.7 kcal/mol (Table 2), compared to those calculated at the full QM level. This is a significant improvement, by 0.7 to 2.8 kcal/mol, compared to the interaction energies calculated using the standard CHARMM27 parameters (Table 2). The A\_wat\_2 complex was the only one where the difference in the interaction energies between the QM and the QM/MM calculations became worse with optimized parameters. This was because the hydrogen bond distance



**Table 4.** Standard CHARMM27<sup>23; 27; 28</sup> and Optimized Base-Independent vdW Parameters for Nitrogen, Polar Hydrogen, and Carbonyl Oxygen in Nucleic Acid Bases

atom type <sup>a</sup>	CHARMM		OPT	
	$R_{\min}/2$ (Å)	$\epsilon$ (kcal/mol)	$R_{\min}/2$ (Å)	$\epsilon$ (kcal/mol)
HN1	0.2245	-0.046	0.7	-0.046
HN2	0.2245	-0.046	0.7	-0.046
NN3	1.85	-0.2	1.9	-0.2
NN4	1.85	-0.2	1.9	-0.2
ON1C	1.7	-0.12	1.8	-0.12
ON1	1.7	-0.12	1.8	-0.12

<sup>a</sup> See Figures 1B–E for atom types.

between HN1–water was longer in the QM/MM optimized A\_wat\_2 structure and was shorter in the A\_wat\_1 complex, as explained above. However, in the A\_wat\_2 complex, the differences between the calculated QM and QM/MM interaction energies are very small, regardless of which parameters are used (Table 2).

The hydrogen bond angles (D–H···A) with the optimized base-dependent vdW parameters changed only slightly from those obtained with the standard CHARMM27 vdW parameters (see Supporting Information, Table 12). In some cases the difference between the hydrogen bond angles, calculated at the full QM level and at the QM/MM level, were decreased and in some cases increased with the optimized base-dependent vdW parameters. Given the relative lack of sensitivity of hydrogen bond angles to the parameters, hydrogen bond angles were not considered specifically during the vdW parameter optimization process.

**Optimized Base-Independent vdW Parameters.** A typical approach for the development of MM forcefield parameters is that the MM parameters, optimized for small molecules or fragments, are then used in large compounds, in different molecular environments. Accordingly, the same vdW parameters are used, for example, for carbonyl oxygen atoms in all nucleic acid bases. Therefore, starting from the base-dependent parameters listed in Table 3, base-independent parameters were developed (shown in Table 4). In this set of parameters,  $R_{\min}/2 = 0.7$  Å and  $\epsilon = -0.046$  kcal/mol were used for polar hydrogen atoms. For nitrogen and oxygen atoms, the optimized  $R_{\min}/2$  and  $\epsilon$  parameters are 1.85 Å and -0.2 kcal/mol, and 1.7 Å and -0.12 kcal/mol, respectively (Table 4).

The hydrogen bond distances and interaction energies calculated using the base-independent parameters are listed in Tables 1 and 2, respectively. Hydrogen bond angles are reported in Table 12, Supporting Information. With the base-independent parameters, the deviation of the QM/MM calculated interaction energies and hydrogen bond distances from the QM values is slightly larger than that seen with base-dependent parameters, as is to be expected. Hydrogen bond distances found with the base-independent parameters differ by 0.01 to 0.6 Å (Table 1) from the QM values, while with the base-dependent parameters (Table 4), the deviation is 0.01–0.5 Å (Table 1). The interaction energies deviate by 1.2–2.9 kcal/mol (Table 2) from those calculated at the full QM level, whereas with base-dependent parameters, the deviation is 0.4–2.7 kcal/mol (Table 2). However, with the base-independent parameters, the interaction energies are still

significantly closer to the full QM interaction energies than when standard CHARMM27 parameters are used, as clearly seen from Table 2. Interestingly, the base-independent parameters provide the best agreement with hydrogen bond angles between the QM and QM/MM calculated structures (see Supporting Information Table 11).

**Transferability of the Optimized Parameters to the B3LYP/6–31G(d) Level.** QM/MM calculations are often performed at the B3LYP/6–31G(d) level. This is partly because this level of theory gives reasonable descriptions of many (e.g., biomolecular) systems,<sup>6,7,31,49,55</sup> while being computationally less demanding than, for example, the B3LYP/6–311+G(d,p) level, but also because diffuse functions can cause QM/MM calculations to be unstable.<sup>60</sup> The transferability of the optimized vdW parameters [both base-dependent (Table 3) and independent (Table 4)] to the widely used B3LYP/6–31G(d) QM/MM level was tested. The results show that the vdW parameters optimized at the B3LYP/6–311+G(d,p) level can be transferred to the lower level. Comparison of the interaction energies calculated at the B3LYP/6–31G(d) QM level and B3LYP/6–31G(d)/CHARMM27 QM/MM levels shows that with both base-dependent and base-independent parameters, very similar values are obtained (Table 5). The deviation between the QM and QM/MM interaction energies is only 0.1–1.4 kcal/mol with base-dependent parameters (Table 3) and 0.1–3.3 kcal/mol with base-independent parameters (Table 4). Accordingly, especially with base-dependent parameters, the agreement between the QM and QM/MM interaction energies is very good. Table 5 also clearly shows that both sets of optimized parameters give relatively consistent QM/MM interaction energies that are closer to the QM results than when standard CHARMM27 parameters are used. Comparison of the hydrogen bond distances also shows that the QM and QM/MM geometries are very similar, as seen from Table 6. However, the differences between the hydrogen bond distances calculated at QM and QM/MM levels are larger when the 6–31G(d) basis is used instead of the 6–311+G(d,p) basis. In addition, although better agreement with the QM interaction energies is achieved with the optimized parameters than with the standard CHARMM27 parameters, no significant improvement is observed in QM/MM geometries with the optimized parameters when the 6–31G(d) basis is used (Table 6 for hydrogen bond distances and Supporting Information, Table 13 for hydrogen bond angles).

**Nucleic Acid Base Pairs.** The standard Watson–Crick base pairs [adenosine-uracil (AU), guanosine-cytosine (GC), and adenosine-thymine (AT); Figure 2] were used to test the transferability of the newly optimized parameters to different structures. For nucleic acid base pairs, it is only possible to easily use the base-independent parameters (Table 4), because in the standard Tinker format<sup>54</sup> (which QoM-MMa<sup>55</sup> uses to evaluate the MM terms) the same atom types are found in different bases (see Figures 1B–E for details of atom types in different nucleic acid bases). Calculations were performed for each base pair in both possible QM/MM combinations, i.e., first with one base treated QM and the second by MM, then separately with the other base QM (and the first MM). Hydrogen bond distances (measured

**Table 5.** Interaction Energies (kcal/mol) Calculated at the B3LYP/6-31G(d) QM and B3LYP/6-31G(d)-QM/MM Levels, Using Base-Dependent Parameters (Table 3), Base-Independent Parameters (Table 4), or the Original CHARMM27 Parameters for the Complexes Shown in Figures 1B–E<sup>a</sup>

complex <sup>b</sup>	QM	QM/MM (OPT-BI) <sup>c</sup>	QM/MM (OPT-BD) <sup>d</sup>	QM/MM (CHARMM) <sup>e</sup>
A_wat_1	-11.1	-11.6 (0.5)	-11.6 (0.5)	-13.4 (2.3)
A_wat_2	-9.6	-9.7 (0.1)	-9.7 (0.1)	-9.8 (0.2)
C_wat_1	-11.1	-12.1 (0.9)	-11.8 (0.7)	-12.9 (1.9)
C_wat_2	-6.9	-10.2 (3.3)	-6.7 (-0.3)	-8.6 (1.7)
G_wat_1	-11.4	-11.9 (0.5)	-11.9 (0.5)	-13.5 (2.1)
G_wat_2	-11.9	-9.8 (-2.1)	-9.3 (-2.6)	-10.5 (-1.4)
U_wat	-8.3	-8.3 (0.0)	-9.7 (1.4)	-9.4 (1.1)
STDEV <sup>g</sup>		1.6	1.3	1.3
MSE <sup>h</sup>		-0.5	-0.0	-1.1
MUE <sup>h</sup>		1.1	0.9	1.5

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[\text{QM} - (\text{QM/MM})]\}$ . <sup>b</sup> See Figures 1B–E for names of complexes. <sup>c</sup> Optimized base-independent parameters were used. <sup>d</sup> Optimized base-dependent parameters were used. <sup>e</sup> Original CHARMM parameters were used. <sup>f</sup> Standard deviation of the error from the full QM calculated value. <sup>g</sup> Mean signed error from the full QM calculated value, see notes in Table 1. <sup>h</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

**Table 6.** Hydrogen Bond Distances (Å) Calculated at the B3LYP/6-31G(d) QM and B3LYP/6-31G(d)-QM/MM Levels, Using Either Base-Dependent Parameters (Table 3), Base-Independent Parameters (Table 4), or the Original CHARMM27 Parameters, For the Complexes Shown in Figures 1B–E<sup>a</sup>

complex <sup>b</sup>	H-bond <sup>c</sup>	QM	QM/MM (OPT-BI) <sup>d</sup>	QM/MM (OPT-BD) <sup>e</sup>	QM/MM (CHARMM) <sup>f</sup>
A_wat_1	N7–Hwat	1.90	1.87 (0.03)	1.87 (0.03)	1.85 (0.05)
A_wat_1	N6–H'–Owat	1.89	2.01 (-0.12)	2.01 (-0.12)	1.80 (0.09)
A_wat_2	N6–H''–Owat	1.96	2.42 (-0.46)	2.42 (-0.46)	2.39 (-0.43)
A_wat_2	N1–Hwat	1.95	1.85 (0.10)	1.85 (0.10)	1.90 (0.05)
C_wat1	N4–H–Owat	1.92	2.13 (-0.21)	2.06 (-0.14)	1.98 (-0.06)
C_wat1	N3–Hwat	1.94	1.90 (0.03)	1.92 (0.02)	1.88 (0.06)
C_wat2	N3–Hwat	2.18	1.95 (0.22)	3.37 (-1.20)	3.46 (-1.28)
C_wat2	O2–Hwat	2.24	2.70 (-0.46)	2.12 (0.12)	1.74 (0.50)
G_wat_1	O6–Hwat	1.86	1.85 (0.01)	1.78 (0.07)	1.76 (0.10)
G_wat_1	N3–H–Owat	1.88	2.12 (-0.23)	2.18 (-0.30)	1.96 (-0.08)
G_wat_2	N1–H–Owat	1.90	2.13 (-0.23)	2.22 (-0.32)	2.05 (-0.15)
G_wat_2	N2–H–Owat	2.52	2.13 (0.39)	2.18 (0.34)	1.96 (0.56)
U_wat	N3–H–Owat	1.93	1.02 (0.90)	2.14 (-0.22)	1.81 (0.12)
U_wat	O2–Hwat	1.94	2.23 (-0.29)	1.83 (0.11)	2.24 (-0.30)
STDEV <sup>g</sup>			0.36	0.37	0.44
MSE <sup>h</sup>			0.02	0.14	0.06
MUE <sup>g</sup>			0.27	0.25	0.27

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[\text{QM} - (\text{QM/MM})]\}$ . The hydrogen bond distances are calculated between hydrogen atoms and hydrogen bond acceptor. <sup>b</sup> See Figures 1B–E for names of complexes. <sup>c</sup> See Figure 1A for labeling. <sup>d</sup> Optimized base-independent parameters were used. <sup>e</sup> Optimized base-dependent parameters were used. <sup>f</sup> Standard CHARMM parameters were used. <sup>g</sup> Standard deviation of the error from the full QM calculated value. <sup>h</sup> Mean signed error from the full QM calculated value, see notes in Table 1. <sup>g</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

between heavy atoms) between the base pairs AU and CG obtained from QM, QM/MM, and MM (CHARMM27) calculations are very similar to those observed in crystal structures<sup>61</sup> (Table 7). For the AT base pair, no crystallographic structure is currently available. The hydrogen bond distances in base pair structures optimized using pure QM [at the B3LYP/6-311+G(d,p) level of theory] deviate from experimental values<sup>61</sup> by 0.01–0.11 Å (Table 7). Similar differences from experimentally determined hydrogen bond distances are also observed in the pure MM optimized structures (Table 7). In the QM/MM optimized structures, hydrogen bond distances seem to deviate more from experimental values, compared to the pure QM and pure MM optimized structures. This is not affected by the vdW parameters or the level of QM/MM theory used (Table 7). The AT base pair geometry calculated at the fully QM level is very similar to that of the fully QM optimized AU base pair structure. The deviations of the MM and QM/MM optimized geometries from the fully QM optimized structures are very similar to those seen with AU and CG base pairs.

Comparison of the interaction energies (Table 8) shows that at the QM/MM level, our optimized base-independent parameters give better agreement with the full QM interaction energies than the standard CHARMM27 vdW parameters.<sup>23,24,32,33</sup> With the B3LYP/6-311+G(d,p) level of theory and optimized base-independent parameters, the QM/MM interaction energies in five complexes out of six are within 0.8–1.8 kcal/mol of the QM interaction energies (Table 8). The largest deviation is seen for the AT base pair, when thymine is treated with QM. In this case, the interaction energy is overestimated by 4.4 kcal/mol. With the lower [B3LYP/6-311G(d)] level of QM/MM theory, for the same base pair, some overestimation is also observed but to a smaller extent (only 2.3 kcal/mol). When standard CHARMM27 parameters<sup>23,24,32,33</sup> are used in QM/MM calculations, the deviation of QM/MM interaction energies from the QM interaction energies is 2.4–6.3 kcal/mol (Table 8). The MM interaction energies (with standard CHARMM27 parameters) are very close (within 1–2.5 kcal/mol) to the QM interaction

**Table 7.** Hydrogen Bond Distances (Å) for Nucleic Acid Base Pairs Shown in Figure 2<sup>a</sup>

complex <sup>b</sup>	H-bond <sup>c</sup>	exp. <sup>d</sup>	QM <sup>e</sup>	MM (CHARMM) <sup>f</sup>	QM/MM[OPT-BI, 6-311+G(d,p)] <sup>g</sup>	QM/MM[OPT-BI, 6-31G(d)] <sup>h</sup>	QM/MM[CHARMM, 6-311+G(d,p)] <sup>i</sup>
AU*	N1...H-N3	2.82	2.88 (-0.06)	2.87 (-0.05)	3.04 (-0.22)	3.05 (-0.23)	2.84 (-0.3)
AU*	N6...O4	2.95	2.94(0.01)	2.89 (0.06)	3.01 (-0.06)	3.00 (-0.06)	2.79 (0.16)
A*U	N1...H-N3	2.82	2.88 (-0.06)	2.87 (-0.05)	2.98 (-0.16)	2.99 (-0.17)	2.84 (-0.02)
A*U	N6...O4	2.95	2.94(0.01)	2.89 (0.06)	3.08 (-0.13)	3.10 (-0.15)	2.84 (0.11)
CG*	N4-H...O6	2.91	2.80 (0.11)	2.83 (0.08)	3.04 (-0.13)	3.07 (-0.16)	2.82 (0.09)
CG*	N3...H-N1	2.95	2.96 (-0.01)	2.92 (0.03)	3.10(-0.15)	3.09 (-0.14)	2.85(0.10)
CG*	O2...H-N2	2.86	2.91 (-0.06)	2.84 (0.02)	3.03 (-0.17)	3.03 (-0.17)	2.75 (0.11)
C*G	N4-H...O6	2.91	2.80(0.11)	2.83 (0.08)	2.98 (-0.07)	3.02 (-0.11)	2.74 (0.17)
C*G	N3...H-N1	2.95	2.96 (-0.01)	2.92 (0.03)	3.10 (-0.15)	3.07 (-0.13)	2.89 (0.06)
C*G	O2...H-N2	2.86	2.91 (-0.06)	2.84 (0.02)	3.07(-0.12)	3.03 (-0.08)	2.85(0.10)
AT*	N1...H-N3	-	2.89	2.87 (0.01)	2.98 (-0.10)	3.04 (-0.15)	2.82 (0.07)
AT*	N6-H...O4	-	2.94	2.90 (0.04)	3.09 (-0.15)	3.08 (-0.14)	2.77 (0.17)
A*T	N1...H-N3	-	2.89	2.87 (0.01)	2.99 (-0.10)	3.13 (-0.24)	2.84 (0.04)
A*T	N6-H...O4	-	2.94	2.90 (0.04)	3.09 (-0.14)	3.02 (-0.08)	2.85 (0.09)
STDEV - EXP <sup>j</sup>			0.06	0.05	0.05	0.09	0.14
STDEV -QM <sup>k</sup>				0.02	0.03	0.07	0.06
MSE -EXP <sup>l</sup>		0.00		-0.03	0.14	0.14	-0.06
MSE- QM <sup>m</sup>				-0.03	0.12	0.16	-0.09
MUE -EXP <sup>n</sup>		0.05		0.05	0.14	0.14	0.12
MUE - QM <sup>o</sup>				0.03	0.12	0.16	0.09

<sup>a</sup> Numbers in parentheses are the deviation from experimental hydrogen bond distances or in case of the AT base pair the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . The hydrogen bond distances are calculated between hydrogen bond acceptors and donors (i.e., heavy atoms). <sup>b</sup> See Figure 2 for names of complexes. \* denotes the base treated with QM. <sup>c</sup> See Figure 2 for labeling. <sup>d</sup> Taken from ref. 61. <sup>e</sup> Fully QM. The B3LYP/6-311+G(d,p) level of theory was used. <sup>f</sup> Fully MM. <sup>g</sup> Optimized base-independent parameters (Table 4) and the. <sup>h</sup> Optimized base-independent parameters (Table 4). <sup>i</sup> Standard CHARMM27 vdW parameters (Table 4). <sup>j</sup> Standard deviation of the error from the experimental value. <sup>k</sup> Standard deviation of the error from the full QM calculated value. <sup>l</sup> Mean signed error from the experimental value, see notes in Table 1. <sup>m</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1. <sup>n</sup> Mean unsigned error from the experimental value, see notes in Table 1. <sup>o</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

**Table 8.** Interaction Energies (kcal/mol) Calculated for Nucleic Acid Base Pairs Shown in Figure 2<sup>a</sup>

complex <sup>b</sup>	QM <sup>c</sup>	MM(CHARMM) <sup>d</sup>	QM/MM [OPT-BI, 6-311+G(d,p)] <sup>e</sup>	QM/MM [OPT-BI, 6-311G(d)] <sup>f</sup>	QM/MM [CHARMM, 6-311+G(d,p)] <sup>g</sup>
A*U	-12.1	-13.2 (1.1)	-13.8 (1.7)	-12.5 (0.4)	-16.9 (4.8)
AU*	-12.1	-13.2 (1.1)	-13.8 (1.7)	-12.5 (0.4)	-17.0 (4.9)
A*T	-11.8	-12.8 (1.0)	-13.3 (1.5)	-11.0 (-0.8)	-14.2 (2.4)
AT*	-11.8	-12.8 (1.0)	-16.2 (4.4)	-14.1 (2.3)	-18.1 (6.3)
C*G	-23.7	-26.4 (2.5)	-24.7 (0.8)	-22.5 (-1.4)	-29.06 (5.2)
CG*	-23.7	-26.4 (2.5)	-25.7 (1.8)	-24.6 (0.7)	-30.2 (6.3)
STDEV <sup>h</sup>		0.8	1.2	1.3	1.4
MSE <sup>i</sup>		-1.5	-2.0	-0.3	-5.0
MUE <sup>j</sup>		1.5	2.0	1	5.0

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . <sup>b</sup> See Figure 2 for names of complexes. \* denotes the base which is treated with QM. <sup>c</sup> Full QM, the B3LYP/6-311+G(d,p) level of theory was used. <sup>d</sup> Full MM, standard CHARMM27 vdW parameters were used. <sup>e</sup> Optimized base-independent parameters (Table 4) and the B3LYP/6-311+G(d,p) level of theory were used. <sup>f</sup> Optimized base-independent parameters (Table 4) and the B3LYP/6-311G(d) level of theory were used. <sup>g</sup> Standard CHARMM27 vdW parameters (Table 4) and the B3LYP/6-311+G(d,p) level of theory were used. <sup>h</sup> Standard deviation of the error from the full QM calculated value. <sup>i</sup> Mean signed error from the full QM calculated value, see Table 1. <sup>j</sup> Mean unsigned error from the full QM calculated value, see Table 1.

energies, providing an illustration of their known good quality for (MM) calculations on nucleic acid complexes.

**Suitability of the Optimized Parameters for Modeling Chemical Reactions.** When chemical reactions are studied by QM/MM methods, the atoms directly involved in the reaction must be treated quantum mechanically, while the surrounding environment can be treated with MM. vdW parameters optimized, for example, for the initial (reactant) state may not be suitable for the product or transition state. The fact that the same Lennard-Jones parameters are typically used for the QM atoms throughout a chemical reaction is a limitation of most current QM/MM methods. Therefore, we tested the suitability of our optimized vdW parameters for investigating changes associated with chemical reaction, using a test set consisting of protonated or

deprotonated nucleic acid bases (Figure 3). The most physiologically important protonation and deprotonation sites were selected, i.e., protonation of adenosine at N2, cytosine at N3, and guanosine at N7 and deprotonation of uracil at N1 and guanosine at N1 positions. Interaction and deprotonation energies were calculated at the [B3LYP/6-311+G(d,p)] QM/MM level using standard CHARMM27,<sup>23,24,32,33</sup> base-dependent (Table 3), and base-independent (Table 4) parameters. The interaction and deprotonation energies obtained were compared to those from fully QM calculations [B3LYP/6-311+G(d,p); Table 9]. Standard vdW parameters were not available for some of the atoms in the complexes shown in Figure 3 for the interaction and deprotonation energy calculations. This is because these calculations involved modified nucleic acid base chemical structures (e.g., proto-

**Table 9.** Interaction Energies (kcal/mol) Calculated at the B3LYP/6-311+G(d,p) QM and B3LYP/6-311+G(d,p)-QM/MM Levels Using Either CHARMM27, Base-Independent Parameters (Table 4), or Base-Dependent Parameters (Table 3) for Neutral, Protonated and Deprotonated Nucleic Acid Bases (Figure 3)<sup>a</sup>

complex <sup>b</sup>	QM	QM/MM(CHARMM) <sup>c</sup>	QM/MM (OPT-BI) <sup>d</sup>	QM/MM (OPT-BD) <sup>e</sup>
A-wat	-8.9	-13.3 (4.4)	-10.2 (1.3)	-10.3 (1.4)
AN1 <sup>+</sup> -wat	-16.1	-18.4 (2.3)	-17.0 (-1.4)	-18.4 (2.3)
U-wat	-7.5	-10.4 (2.9)	-9.3 (1.8)	-9.3 (1.8)
UN3 <sup>-</sup> -wat	-16.0	-19.7 (3.7)	-19.3 (3.3)	-19.7 (3.7)
G-wat1	-10.8	-10.7 (-0.1)	-12.8 (2.0)	-11.3 (0.5)
GN1 <sup>-</sup> -wat	-15.9	-20.0 (-4.1)	-19.4 (3.5)	-19.7 (3.8)
G-wat2	-5.4	-9.0 (-3.6)	-6.6 (0.6)	-6.7 (1.3)
GN7 <sup>+</sup> -wat	-16.3	-19.4 (3.1)	-16.4 (0.1)	-16.0 (-0.3)
C-wat1	-10.5	-13.8 (3.3)	-12.9 (2.4)	-12.9 (2.4)
CN3 <sup>+</sup> -wat1	-16.5	-19.9 (3.4)	-18.5 (2.0)	-19.0 (2.5)
C-wat2	-7.0	-9.1 (2.1)	-8.7 (1.7)	-7.4 (0.4)
CN3 <sup>+</sup> -wat2	-14.6	-18.5 (3.9)	-15.6 (1.0)	-16.7 (2.1)
STDEV <sup>f</sup>		2.9	1.4	1.3
MSE <sup>g</sup>		-1.8	-1.5	-1.8
MUE <sup>h</sup>		3.1	1.8	1.9

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . <sup>b</sup> See Figure 3 for names of complexes. <sup>c</sup> Standard CHARMM27 vdW parameters were used. <sup>d</sup> Optimized base-independent parameters (Table 4) were used. <sup>e</sup> Optimized base-dependent parameters (Table 3) were used. <sup>f</sup> Standard deviation of the error from the full QM calculated value. <sup>g</sup> Mean signed error from the full QM calculated value, see notes in Table 1. <sup>h</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

**Table 10.** Deprotonation Energies [(energy of the deprotonated complex) - (energy of the protonated complex) in kcal/mol] Calculated at the B3LYP/6-311+G(d,p) Level for the Reactions Shown in Figure 3<sup>a</sup>

reaction <sup>b</sup>	QM	QM/MM (CHARMM27) <sup>c</sup>	QM/MM (OPT-BI) <sup>d</sup>	QM/MM (OPT-BD) <sup>e</sup>
AN1 <sup>+</sup> → A	-242.6	-240.6(-2.0)	-242.3 (-0.2)	-243.7 (1.1)
1: CN3 <sup>+</sup> → C	-245.3	-246.0(0.2)	-244.9 (-0.4)	-245.4(0.1)
2: CN3 <sup>+</sup> → C	-246.9	-248.8 (1.9)	-246.2 (-0.7)	-246.3 (0.4)
U → UN3 <sup>-</sup>	-346.8	-346.0 (-0.8)	-345.2 (-1.6)	-344.8 (-1.9)
G → GN1 <sup>-</sup>	-341.9	-337.3 (-4.6)	-340.1 (-1.8)	-338.3 (-3.6)
GN7 <sup>+</sup> → G	-251.3	-252.3 (1.0)	-249.8(-1.5)	-249.2 (-2.1)
STDEV <sup>f</sup>		2.3	0.7	1.8
MSE <sup>g</sup>		0.7	1.0	1.0
MUE <sup>h</sup>		1.9	1.0	1.5

<sup>a</sup> Numbers in parentheses are the deviation from the fully QM calculation  $\{\Delta[QM - (QM/MM)]\}$ . <sup>b</sup> See Figure 3 for names of reactions. <sup>c</sup> Standard CHARMM27 vdW parameters were used. <sup>d</sup> Optimized base-independent parameters (Table 4) were used. <sup>e</sup> Optimized base-dependent parameters (Table 3) were used. <sup>f</sup> Standard deviation of the error from the full QM calculated value. <sup>g</sup> Mean signed error from the full QM calculated value, see notes in Table 1. <sup>h</sup> Mean unsigned error from the full QM calculated value, see notes in Table 1.

nated/deprotonated), not included in the CHARMM27 parameter set (nor in our optimized set). Base-independent vdW parameters were available for all atom types except NN2U (in the UN3<sup>-</sup>\_wat complex, Figure 3). For the UN3<sup>-</sup>\_wat complex, standard CHARMM27 vdW parameters were used in all calculations. In addition, for the following atom types, the base-dependent vdW parameters (Table 3) were not available (see Figure 3 for details of the complexes): HN2 (AN1<sup>+</sup>\_wat, CN3<sup>+</sup>\_wat1 and CN3<sup>+</sup>\_wat2), NN3 (GN1<sup>-</sup>\_wat), and NN4 (G\_wat2). For these atom types in the complexes mentioned, standard CHARMM27 vdW parameters were used in the base-dependent calculations.

The QM/MM interaction energies for the charged complexes are overestimated, compared to the QM interaction energies, in almost all cases (Table 9). The optimized vdW parameters [both base-dependent (Table 3) and base-independent (Table 4)] in the QM/MM calculations give generally better agreement with the QM interaction energies than the standard CHARMM27 parameters<sup>23,24,32,33</sup> (Table 9). Only in one case is a larger error observed with our optimized parameters than with standard CHARMM27 parameters (G\_wat\_1, Table 9). With the optimized parameters, the interaction energies differ from the QM results by 0.1–3.8 kcal/mol, while with the standard CHARMM27

parameters<sup>23,24,32,33</sup> all except G\_wat\_1 vary by 2.1–4.4 kcal/mol. G\_wat\_1 is the only complex where the standard CHARMM27 parameters<sup>23,24,32,33</sup> give very good agreement with the QM interaction energies (an error of only 0.1 kcal/mol, Table 9). With our optimized parameters, the largest deviations between the QM and QM/MM results are seen with deprotonated complexes, i.e., UN3<sup>-</sup>\_wat and GN1<sup>-</sup>\_wat\_1 complexes, regardless of which parameter set is used (Figure 3 and Table 9).

Comparison of deprotonation energies for nucleic acid bases in complex with one water molecule (see Figure 3 for details of reactions) shows that deprotonation energies calculated at the full QM [B3LYP/6-311+G(d,p)] level and at the QM/MM [B3LYP/6-311+G(d,p)] level are generally surprisingly similar, within 0.1–4.6 kcal/mol (Table 10). In 8 reactions out of 18, the QM/MM deprotonation energies are within 1.0 kcal/mol of the QM results, and in 7 reactions out of 18, they are within 1–2 kcal/mol. Only in three cases do the QM/MM deprotonation energies deviate by more than 2.0 kcal/mol from the QM deprotonation energies (Table 10). Comparison of the deprotonation energies obtained with different vdW parameters (i.e., the standard CHARMM27,<sup>23,24,32,33</sup> base-dependent and independent) shows that the optimized base-independent parameters seem to be little better than the

**Table 11.** Electrostatic Contribution to the Interaction Energy (kcal/mol) Calculated at the B3LYP/6–31G(d) QM/MM and CHARMM27 MM Levels for Some Sample Base–Water Complexes

complex <sup>a</sup>	structure	hydrogen bonds <sup>b</sup>	QM/MM energy <sup>c</sup>	MM energy <sup>d</sup>
G_wat_1	MM <sup>e</sup>	MM QM	-16.23	-13.50
	QM <sup>f</sup>	MM QM	-15.41	-12.55
	QM/MM (CHARMM27) <sup>g</sup>	MM QM	-16.63	-14.02
	QM/MM (OPT_BI) <sup>h</sup>	MM –	-14.06	-12.14
	QM/MM (OPT-BD) <sup>i</sup>	MM –	-14.32	-12.33
G_wat_2	MM <sup>e</sup>	– QM	-13.06	-12.38
	QM <sup>f</sup>	– –	-10.75	-10.44
	QM/MM (CHARMM27) <sup>g</sup>	QM –	-12.87	-12.88
	QM/MM (OPT_BI) <sup>h</sup>	– –	-11.08	-12.14
	QM/MM (OPT-BD) <sup>i</sup>	– –	-10.22	-12.33
A_wat_1	MM <sup>e</sup>	MM QM	-15.96	-13.15
	QM <sup>f</sup>	MM QM	-15.02	-12.29
	QM/MM (CHARMM27) <sup>g</sup>	MM QM	-17.90	-14.58
	QM/MM (OPT_BI) <sup>h</sup>	MM QM	-15.00	-12.37
	QM/MM (OPT-BD) <sup>i</sup>	MM QM	-15.00	-12.37
C_wat_2	MM <sup>e</sup>	MM –	-15.39	-13.23
	QM <sup>f</sup>	– MM	-10.79	-10.03
	QM/MM (CHARMM27) <sup>g</sup>	– MM	-10.36	-10.04
	QM/MM (OPT_BI) <sup>h</sup>	– MM	-11.35	-10.59
	QM/MM (OPT-BD) <sup>i</sup>	– –	-7.83	-7.96

<sup>a</sup> See Figures 1B–E for names of complexes. <sup>b</sup> Characterization of hydrogen bonds in the system. QM denotes a hydrogen bond with a QM hydrogen, MM a hydrogen bond with a MM hydrogen, and – no hydrogen bond present (a hydrogen bond was deemed to be present at acceptor to H distances  $\leq 2$  Å). The distances are shown in Table 1. <sup>c</sup> Electrostatic contribution to the interaction energy, calculated using QM/MM. <sup>d</sup> Electrostatic contribution to the interaction energy, calculated using pure MM. <sup>e</sup> Pure MM calculation using CHARMM27. <sup>f</sup> Pure QM calculation, using Gaussian. <sup>g</sup> Standard CHARMM27 parameters were used. <sup>h</sup> Optimized base-independent parameters (Table 4) were used. <sup>i</sup> Optimized base-dependent parameters (Table 3) were used.

optimized base-dependent or the standard CHARMM27 parameters.<sup>23,24,32,33</sup> This may be due, at least partly, to the fact that in the base-independent calculations, optimized vdW parameters were available for all atoms (except NN2U in the UN3<sup>-</sup>\_wat complex (as mentioned previously standard CHARMM27 parameters were used)). However, in the base-dependent calculations more optimized parameters were unavailable, and hence standard CHARMM27 vdW parameters<sup>23,24,32,33</sup> had to be used (see above). Therefore, the base-independent parameters developed here (Table 4) would appear to be a better choice than standard CHARMM27 vdW parameters for QM/MM modeling of such reactions in nucleic acid bases.

**Electrostatic Contribution to the Calculated QM/MM Interaction Energies.** While the parametrization here has been generally successful, the improvements have not been uniform. When the original parameters were used, we found that typically QM/MM hydrogen bond interaction energies were too strong and hydrogen bond lengths were too short (Tables 1 and 2).<sup>29,30,34</sup> We attempted to address this problem by optimizing the vdW parameters, however, as can be seen from Tables 3 and 4, only the parameters of the hydrogens were altered significantly. This has been seen previously.<sup>27,31</sup> To further investigate why hydrogen bond strengths are overestimated when the original parameters are used, we have calculated the QM/MM and MM electrostatic contributions to the interaction energy for the QM optimized, MM optimized, and QM/MM optimized geometries. The results are shown in Table 11. The QM/MM electrostatic interactions in the QM optimized and MM optimized structures are consistently stronger than the corresponding MM electrostatic interactions. This indicates that the electrostatic interactions are the source of the error. Given that the electrostatic interactions are overestimated, it is not

surprising that the QM/MM optimized geometries have shorter hydrogen bond lengths. Our optimizations increased the vdW radii of QM (hydrogen) atoms, and, as expected, this lengthens the hydrogen bonds, resulting in a reduction in the QM/MM electrostatic interaction energy for these complexes. The results vary for the different complexes. In three cases hydrogen bonds are broken, resulting in either significant improvements (G\_wat\_2) or in overcompensation (G\_wat\_1 and C\_wat\_2). For A\_wat\_1, there are significant reductions in the electrostatic contribution that cannot be readily explained. For the base-independent results for C\_wat\_2, the optimizations result in the electrostatics becoming even more favorable. It is apparent that our optimization of the vdW parameters does not consistently improve the results, suggesting that more refined QM/MM models (e.g., with different treatment of QM/MM electrostatic interactions<sup>62</sup>) may be required for some applications.

The optimization of the vdW parameters compensates for the overestimated QM/MM electrostatic interaction by increasing the vdW radii of the QM hydrogen atoms involved in hydrogen bonds (large increases of the hydrogen  $R_{\min}/2$  values are seen). While this has been reasonably successful, it has not addressed the root cause of the problem. Optimizing the vdW parameters alone provides an improved model, but ideally treatment of QM/MM electrostatic interactions should also be considered and improved. Further study of the balance of electrostatic interactions in QM/MM systems is required.

## Conclusions

QM/MM interaction energies and geometries for the hydrogen bonded systems are fairly sensitive to the values of the van der Waals parameters used to describe the nonbonded

dispersion interactions between the QM and MM atoms. In general, the use of the standard CHARMM27 parameters for the QM atoms in QM/MM calculations on nucleic acid base–water complexes gives shorter hydrogen bond lengths than those obtained from full QM calculations. QM/MM calculations with the standard CHARMM27 parameters overestimate the interaction energies by 1.1–3.4 kcal/mol, compared to the equivalent QM results (using B3LYP hybrid density functional theory). Adjusting the vdW parameters on the QM atoms gives improved results for interaction energies and in most cases also for hydrogen bond lengths. As expected, better agreement between the QM/MM and QM geometries and interaction energies is obtained with base-dependent parameters rather than with base-independent parameters. Using vdW parameters optimized at the B3LYP/6–311+G(d,p) level in QM/MM calculations at the B3LYP/6–31G(d) level reproduces the full QM [B3LYP/6–31G(d)] interaction energies (very well) and also geometries (reasonably well). Thus, vdW parameters can be transferred from the higher level to the lower level, when the same DFT model is used, and so should be useful in QM/MM calculations applying the popular B3LYP/6–31G(d) level of QM theory. Transferability to other levels of QM/MM treatment is not guaranteed and should be tested, particularly when different types of QM/MM approach (e.g., applying ab initio molecular orbital methods) are used. For example, Hartree–Fock level ab initio QM/MM methods tend to show more overestimation of intermolecular interaction energies.<sup>63</sup> Luque et al. have reported that vdW parameters are sensitive to the QM/MM formalism and cannot be transferred between different QM/MM methods.<sup>30</sup> Further work is required to analyze QM/MM interactions. In other work, we have examined QM/MM modeling of biomolecular hydrogen bonds. Examination of changes in total electron density and natural bond orbital atomic charges, due to hydrogen bond formation, in selected complexes showed that charge leakage from the QM atoms to MM atomic point charges close to the QM/MM boundary is not a serious problem, at least with limited basis sets.<sup>64</sup> Clearly, current QM/MM methods can provide good descriptions of many biomolecular systems.<sup>65</sup> Further analysis and development of more sophisticated treatment of QM/MM interactions (in particular of electrostatic interactions) should remain a central goal of research in this area, however. This is particularly true now that it is possible to model biomolecules (e.g., enzyme reactions) with QM/MM techniques employing correlated ab initio methods that are capable of high accuracy<sup>66,67</sup> and to carry out extensive high-level QM/MM free energy calculations.<sup>68</sup> QM/MM free energy calculations may provide a more general approach to optimizing parameters for QM/MM interactions.<sup>69,70</sup>

In summary, the results here show that, when appropriate parameters are used, hydrogen bonding interactions of nucleic acids can be modeled well with QM/MM methods. Obtaining parameters which reliably reproduce both interaction energies and geometries (hydrogen bonds and angles) can be difficult, especially if there are multiple hydrogen bonds in a system, as is the case for the interactions of nucleic acid bases with water. The parameters presented here give better agreement

with full QM calculations, especially for interaction energies. They have also successfully been used for complexes not used in the parameter optimization. Therefore, these parameters should be useful for QM/MM investigations of nucleic acid structure employing the B3LYP hybrid density functional QM method. It should, however, be noted that these parameters have been developed to represent hydrogen bonding interactions of nucleic acid bases, not base-stacking interactions. The modifications made here seem unlikely to affect the treatment of stacking significantly, but this should be checked in applications where base-stacking is important. The parameters developed here (especially the base-independent parameters) also appear to perform reasonably well for modeling chemical changes, based on results for deprotonation reactions. However, parameters specifically optimized for a given reaction may be preferable, for enhanced accuracy.<sup>71</sup> For the chemical changes modeled here, the optimized parameters performed well, but it should be remembered that in current QM/MM implementations a compromise must usually be made between representing QM/MM interactions of the substrate and product (and for example the transition state), as the same van der Waals parameters are typically used for all stages of a reaction. It may be that a single set of (MM) parameters can represent the interactions of all species in the reaction satisfactorily, but this cannot be guaranteed in advance. Accordingly, we believe that the best approach for QM/MM reaction modeling, in general, is to test the results for a given reaction and to optimize the parameters for that specific application as required. The parameters developed here should be a good basis for the optimization of reaction-specific Lennard–Jones parameters for QM atoms in QM/MM (CHARMM27) studies of reactions involving nucleic acids.

**Acknowledgment.** The Academy of Finland is gratefully acknowledged for fellowships for UP (grant numbers 109343 and 121393). A.J.M. thanks BBSRC (with K.E.S.), EPSRC (with K.S. and C.J.W.) and the IBM High Performance Computing Life Sciences Outreach Programme for support. CSC, the Finnish IT center for science (Espoo, Finland), is acknowledged for access to computational resources (project number jyy2516). We thank Dr. Jeremy Harvey for many useful discussions.

**Supporting Information Available:** Hydrogen bond angles for nucleic acid base and water complexes, shown in Figures 1B–E, calculated with pure QM, pure MM and combined QM/MM methods, using CHARMM27 and optimized parameters. This information is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Gogonea, V.; Suarez, D.; van der Vaart, A.; Merz, K. M. *Curr. Opin. Struct. Biol.* **2001**, *11*, 217.
- (2) Warshel, A. In *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons: New York, 1991.
- (3) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
- (4) Warshel, A.; Karplus, M. *J. Am. Chem. Soc.* **1972**, *94*, 5612.

- (5) Gao, J.; Thompson, M. *Combined quantum mechanical and molecular mechanical methods*. ACS Symposium Series 712; American Chemical Society: Washington, D.C., 1998; p 712.
- (6) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442.
- (7) Mulholland, A. J. *Drug Discovery Today* **2005**, *10*, 1393.
- (8) Monard, G.; Merz, K. M. *Acc. Chem. Res.* **1999**, *32*, 904.
- (9) Gao, J. Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Donald, B. B., Eds.; VCH Publishers Inc.: New York, 1996; Vol. 7, pp 119–185.
- (10) Reuter, N.; Dejaegere, A.; Maignet, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.
- (11) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959.
- (12) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.
- (13) Singh, U.; Kollman, P. J. *J. Comput. Chem.* **1986**, *7*, 718.
- (14) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
- (15) Thery, V.; Rinaldi, D.; Rivail, J.; Maignet, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269.
- (16) Antes, I.; Thiel, W. *J. Phys. Chem. A* **1999**, *103*, 9290.
- (17) Pu, J.; Gao, J.; Truhlar, D. *J. Phys. Chem. A* **2004**, *108*, 5454.
- (18) Pu, J.; Gao, J. L.; Truhlar, D. *ChemPhysChem* **2005**, *6*, 1853.
- (19) Garcia-Viloca, M.; Gao, J. *Theor. Chem. Acc.* **2004**, *111*, 280.
- (20) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.
- (21) Pu, J. Z.; Gao, J. L.; Truhlar, D. *J. Phys. Chem. A* **2004**, *108*, 632.
- (22) Mulholland, A. J. In *Chemical Modelling: Applications And Theory*; RSC Specialist Periodical Reports; Hinchliffe, A., Ed.; The Royal Society of Chemistry: London, 2006; Chapter 2, pp 23–68.
- (23) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (24) MacKerell, A. D. *Annu. Rep. Comput. Chem.* **2005**, *1*, 91.
- (25) Gao, J.; Xia, X. *Science* **1992**, *258*, 631.
- (26) Gao, J. Computation of Intermolecular Interactions with the Combined Quantum Mechanical and Classical Approach In *Modeling the Hydrogen Bond*; Smith, D. A., Eds.; ACS Symposium Series 569, 1994, Chapter 2, pp 8–21.
- (27) Freindorf, M.; Gao, J. *J. Comput. Chem.* **1996**, *17*, 386.
- (28) Bash, P. A.; Ho, L. L.; MacKerell, A. D., Jr.; Levine, D., Jr.; Hallstrom, P. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3698.
- (29) Riccardi, D.; Li, G.; Cui, Q. *J. Phys. Chem. B* **2004**, *108*, 6467.
- (30) Luque, F. J.; Reuter, N.; Cartier, A.; Ruiz-Lopez, M. F. *J. Phys. Chem. A* **2000**, *104*, 10923.
- (31) Freindorf, M.; Shao, Y.; Furlani, T. R.; Kong, J. *J. Comput. Chem.* **2005**, *26*, 1270.
- (32) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.
- (33) MacKerell, A. D., Jr.; Banvali, N. K. *J. Comput. Chem.* **2000**, *21*, 105.
- (34) Gao, J. *Biophys. Chem.* **1994**, *51*, 253.
- (35) Zhou, D. M.; Taira, K. *Chem. Rev.* **1998**, *98*, 991.
- (36) Takagi, Y.; Warashina, M.; Stec, W. J.; Yoshinari, K.; Taira, K. *Nucleic Acids Res.* **2001**, *29*, 1815.
- (37) Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. *Cell* **1983**, *35*, 849.
- (38) Cech, T. R.; Zaug, A. J.; Grabowski, P. J. *Cell* **1981**, *29*, 487.
- (39) Noll, D.; Mason, T. M.; Miller, P. *Chem. Rev.* **2006**, *106*, 277.
- (40) Elcock, A. H.; Lyne, P. D.; Mulholland, A. J.; Nandra, A.; Richards, W. G. *J. Am. Chem. Soc.* **1995**, *117*, 4706.
- (41) Hopkins, P. B.; Millard, J. T.; Woo, J.; Weidner, M. F.; Kirchner, J. J.; Sigurdsson, S. T.; Raucher, S. *Tetrahedron* **1991**, *47*, 2475.
- (42) Nordberg, J. L.; Nilsson, L. *Biopolymers* **1996**, *39*, 765.
- (43) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.
- (44) Li, Y.; Breaker, R. R. *J. Am. Chem. Soc.* **1999**, *121*, 5364.
- (45) Kaukinen, U.; Venäläinen, T.; Lönnberg, H.; Peräkylä, M. *Org. Biomol. Chem.* **2003**, *1*, 2439.
- (46) Kaukinen, U.; Lönnberg, H.; Peräkylä, M. *Org. Biomol. Chem.* **2004**, *2*, 66.
- (47) Bibillo, A. M.; Figlerowicz, M.; Kierzik, R. *Nucleic Acids Res.* **1999**, *27*, 3931.
- (48) Kierzek, R. *Nucleic Acids Res.* **1992**, *20*, 5079.
- (49) Claeysens, F.; Ranaghan, K. E.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. *Chem. Commun.* **2005**, 5086.
- (50) Jaguar 6.0, Schrödinger Inc., Portland, OR, 2002.
- (51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.
- (52) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, *102*, 3782.
- (53) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

- (54) Ponder, J. W. *Tinker - Software Tools for Molecular Design*. St. Louis, MO, 2003 (<http://dasher.wustl.edu/tinker/>); accessed 10/17/08.
- (55) Harvey, J. N. *J. Faraday Discuss.* **2004**, *127*, 165.
- (56) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (57) MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *The Encyclopedia of Computational Chemistry*; John Wiley & Sons: Chichester, 1998; pp 271–277.
- (58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (59) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902.
- (60) Mulholland, A. J.; Lyne, P. D.; Karplus, M. *J. Am. Chem. Soc.* **2000**, *122*, 534.
- (61) Saenger, W. In *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.
- (62) Biswas, P. K.; Gogone, V. *J. Chem. Phys.* **2005**, *123*, 1.
- (63) Ridder, L.; Harvey, J. N.; Rietjens, I. M. C. M.; Mulholland, A. J. *J. Phys. Chem. B.* **2003**, *107*, 2118.
- (64) Senthilkumar, K.; Mujika, J. I.; Ranaghan, K. E.; Manby, F. R.; Mulholland, A. J.; Harvey, J. N. *J. R. Soc. Interface* **2008**, *5*, S207.
- (65) van der Kamp, M. W.; Shaw, K. E.; Woods, C. J.; Mulholland, A. J. *J. R. Soc. Interface* **2008**, *5*, S173.
- (66) Claeysens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schutz, M.; Thiel, S.; Thiel, W.; Werner, H. J. *Angew. Chem.* **2006**, *45*, 6856.
- (67) Mulholland, A. J. *Chem. Cent. J.* **2007**, *1*, 19.
- (68) Woods, C. J.; Manby, F. R.; Mulholland, A. J. *J. Chem. Phys.* **2008**, *128*, 014109.
- (69) Cummins, P. L.; Gready, J. E. *J. Comput. Chem.* **1997**, *18*, 1496.
- (70) Ridder, L.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 9926.

CT800135K



## Study of the Conformational Dynamics of the Catalytic Loop of WT and G140A/G149A HIV-1 Integrase Core Domain Using Reversible Digitally Filtered Molecular Dynamics

Sarah L. Williams and Jonathan W. Essex\*

*School of Chemistry, University of Southampton, Highfield,  
Southampton SO17 1BJ, U.K.*

Received May 14, 2008

**Abstract:** The HIV-1 IN enzyme is one of three crucial virally encoded enzymes (HIV-1 IN, HIV-1 PR, and HIV-1 RT) involved in the life-cycle of the HIV-1 virus, making it an attractive target in the development of drugs against the AIDS virus. The structure and mechanism of the HIV-1 IN enzyme is the least understood of the three enzymes due to the lack of three-dimensional structural information. X-ray crystallographic studies have not yet been able to resolve the full-length structure, and studies have been mainly focused on the catalytic domain. This central domain possesses an important catalytic loop observed to overhang the active site, and experimental studies have shown that its dynamics affects the catalytic activity of mutant HIV-1 IN enzymes. In this study, the enhanced sampling technique, Reversible Digitally Filtered Molecular Dynamics (RDFMD), has been applied to the catalytic domain of the WT and G140A/G149A HIV-1 IN enzymes and has highlighted significant differences between the behavior of the catalytic loop which may explain the decrease of activity observed in experimental studies for this mutant.

### 1. Introduction

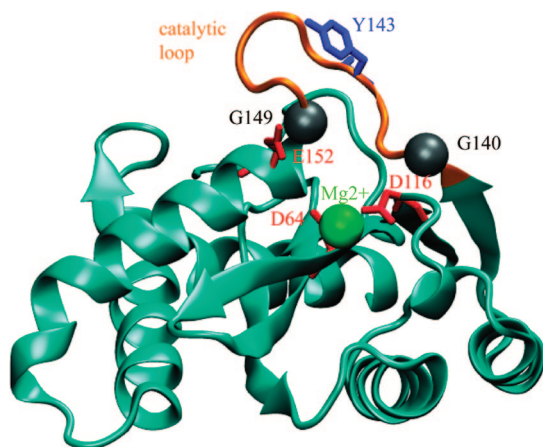
The HIV-1 integrase (HIV-1 IN) enzyme is one of three key enzymes involved in the life-cycle of the HIV-1 virus. The full length HIV-1 integrase enzyme comprises 288 residues which, based on partial proteolysis, functional, and structural studies, can be divided into three domains, the N- and C-terminal domains and the catalytic domain.<sup>1–3</sup> The N-terminal domain (residues 1–50) consists of three  $\alpha$ -helices and a zinc binding site where conserved histidine and cysteine residues coordinate a zinc ion, stabilizing the interaction between the helices. The binding of zinc is thought to promote the multimerisation of the enzyme which enhances the enzyme's activity.<sup>4–6</sup> The third, C-terminal domain (residues 212–288) is a nonspecific binding site of DNA<sup>7–11</sup> and additionally contributes to the multimerization, which is essential to the integration process.<sup>12</sup> In isolation, both the C- and N-terminal domains are dimeric in solution,

but the C-terminal is monomeric when attached to the catalytic domain.<sup>13</sup>

The central catalytic domain (residues 50–212) is composed of a mix of  $\alpha$ -helix and  $\beta$ -sheet structures (Figure 1). This domain is the most conserved of the three and contains the active motif, comprising residues Asp64, Asp116, and Glu152 (DDE). The domain possesses a catalytic loop (residues 140–149) which overhangs the active site and is postulated to play an important role in the catalytic activity of the enzyme through its role in positioning the substrates in the active site for processing.

The HIV-1 IN enzyme catalyzes two stages of the HIV-1 virus life-cycle which occur after the viral-RNA genome is reverse transcribed to produce the double-stranded DNA,<sup>14,15</sup> a process carried out by the HIV-1 reverse transcriptase (HIV-1 RT) enzyme. The first, 3'-processing stage involves the removal of 2 deoxynucleotides from each of the 3'-ends of the viral DNA, and the second, strand transfer stage then covalently ligates these processed 3'-ends to the host chromosomal DNA via transesterification reactions. The final

\* Corresponding author e-mail: J.W.Essex@soton.ac.uk.



**Figure 1.** Catalytic core domain of the HIV-1 integrase enzyme (PDB ID: 1BL3) with the following highlighted residues: active site residues, D64, D116, and E152 (red), catalytically important tyrosine residue, Y143 (blue), and  $\alpha$ -carbon atoms of hinge residues, G140 and G149 (silver),  $Mg^{2+}$  ion (green). The catalytic loop residues (140–149) are shown in orange.

product is the provirus, which comprises the intact double-stranded host DNA containing the inserted viral genetic information.

All three of the domains of the full-length HIV-1 IN are required to carry out the catalytic function of the enzyme.<sup>2</sup> However, in isolation, the catalytic domain has the ability to initiate the process of disintegration *in vitro*.<sup>16,17</sup> Disintegration is the reverse process of strand transfer, whereby the catalytic domain is able to release partially integrated viral DNA from the host DNA.

Owing to the same active site being used for the two distinct types of processes which utilize different substrates, it is likely that the active site undergoes a conformational change for each stage. Evidence for this hypothesis is given by the ability of diketo acid inhibitors to selectively inhibit the strand transfer while not affecting the 3'-processing process.<sup>18</sup>

The arrangement of the three domains in the full-length integrase enzyme still remains unknown due to the low solubility and the tendency of the enzyme to aggregate. Therefore, to gain structural information, the three domains have been determined and studied in isolation by crystallographic<sup>19–22</sup> and NMR methods.<sup>23–26</sup> The catalytic core was the first domain to be determined, and there are currently approximately 15 structures of this domain available in the Protein Data Bank (PDB),<sup>27</sup> each containing either the F185K or F185H solubility enhancing mutation. A few, more recent studies have published a two-domain structure comprising the catalytic core and C-terminal domains.<sup>28–30</sup> As with the determination of the core domain in isolation, not all residues have been resolved, with missing or poorly defined loop residues. It is believed that HIV-1 IN requires at least one divalent metal ion ( $Mg^{2+}$  or  $Mn^{2+}$ ), placed between the carboxylate groups of catalytic residues Asp64 and Asp116 in the active site (E152 is not involved),<sup>31–35</sup> and an activated water to behave as a nucleophile to be fully functional. The divalent ions are required for both the reactions and for the

formation of the HIV-1 IN complex with the viral DNA.<sup>12,32</sup> However, the first crystallographic studies of the catalytic domain were carried out in the presence of cacodylate (dimethylanionic acid),<sup>19</sup> which was critical in the enhancement of the solubility of the enzyme. It is now thought that the presence of this chemical provided a distorted representation of the active conformation where the active site aspartate residues are not close enough to be able to bind a metal ion. Experimental and MD studies have also found the absence of the metal ion to significantly increase the flexibility of the loop, presumably because of the absence of the interactions between the ion and catalytic residues. Disruption of the secondary structure has been noted in some MD studies.<sup>22,36–38</sup>

The location of a second metal ion in the presence of the DNA substrate has been suggested based on the similarities of the function with DNA polymerase, with the second ion being bound by D116 and E152. It is proposed that this ion is required for the stabilization of the active conformation in the presence of DNA.<sup>39</sup> However, in the absence of any structural data of integrase with the DNA substrate, there is currently no evidence to prove the existence of the second metal ion.

Studies of the activity of HIV-1 IN have been focused on the catalytic domain since residues 50–190 are sufficient to promote disintegration *in vitro*,<sup>16,17</sup> which has been shown to occur irrespective of whether or not cacodylate has been used in the crystallization process.<sup>40</sup> However, the structure of this domain has not yet been fully determined unambiguously by experimental techniques, especially the highly mobile loop region which shows a high degree of disorder, which is vital for the activity of the domain. As a consequence, only a few crystal structures exist with the entire loop resolved.<sup>35,22,41</sup> Bujacz et al.<sup>41</sup> determined a crystallographic structure said to contain the active loop in an extended conformation, with E152 shown to point away from the other two catalytic carboxylates. According to the assumed roles of these residues in the binding of metal ions, based on comparison with the related avian sarcoma virus (ASV) integrase, the authors surmise that the conformation of the loop they observe is not that seen during catalysis. Additionally, this structure does not contain the catalytically important divalent metal ion. In the structure resolved by Maignan et al.,<sup>35</sup> the location of a  $Mg^{2+}$  ion has been resolved, and the loop is positioned over the active site in a “closed” and more compact conformation, likely to correspond to an active conformation of the loop. The loop structure overhangs the active site, and although the active conformation adopted by the surface loop during the integration reaction is unknown, correlation has been found between the flexibility of the loop and enzymatic activity.<sup>42</sup> Mutagenesis experiments carried out by Greenwald et al.<sup>42</sup> replaced the loop hinge residues G140 and G149 (individually and as the double mutant) with alanine, resulting in reduced flexibility of the loop, demonstrated by the lower B values. The greatest reduction is seen in the double mutant, and they attribute the diminished catalytic HIV-1 IN activity observed in their experimental studies as a consequence of this decreased mobility. The authors suggest two possible mecha-

nisms causing the changes in enzyme activity, one being the alteration of the equilibrium between the different conformations required at different steps in the catalytic cycle compared with the WT. Alternatively, they suggest that the structure adopted by the constrained loop mutants may represent a nonfunctional conformation which is more stable. However, the crystallographic determination of these structures was carried out in the presence of cacodylate which is known to affect the conformational dynamics of the loop.<sup>36</sup>

**1.1. Theoretical Studies of the Conformational Dynamics of the WT and G140A/G149A Mutant HIV-1 IN Core Domain.** Lee et al.<sup>43</sup> performed a number of MD and locally enhanced sampling (LES)<sup>44,45</sup> simulations of the HIV-1 IN core domain of the WT and mutant containing loop hinge mutants (G140A, G149A, and G140A/G149A), using the AMBER2003 forcefield<sup>46</sup> and explicit solvent. The WT crystal structure (PDB code: 1QS4)<sup>47</sup> of the HIV-1 IN core domain used in this study contains the Mg<sup>2+</sup> ion coordinated by the D64 and D116 catalytic residues. The mutant HIV-1 IN structures were created through the mutation of the appropriate residues using the SwissPDB software.<sup>48</sup> This structure has two unresolved loop residues (I141 and P142), which have been modeled in, again using the SwissPDB software<sup>48</sup> from chain B of the 1BIS<sup>22</sup> (PDB code) crystal structure, which contains these missing residues. Initially, a number of short MD simulations were performed on the WT and mutant structures, followed by the extension (to 40 ns) of the most interesting trajectory. Their results report the WT loop to move from its initial open state to a closed state at ~8 ns, in which it remains stable for almost 30 ns. After applying LES to this closed state, the reopening of the loop is seen within 4 ns.

In the same study, MD and LES simulations involving three mutants (G140A, G149A, and the double mutant, G140A/G149A) produced results in agreement with experimental studies of Greenwald et al.,<sup>42</sup> demonstrating the hinge movement of the loop to be less prominent in the structures containing the single mutants, and completely eliminated in the structure possessing the double mutant. They suggest that the opening/closing ability of the loop is vital for catalytic activity, and the mutants studied hinder this loop mobility, thus affecting the activity of the enzyme.

**1.2. Role of Tyr143 in Catalysis.** The Tyr143 (Y143) residue is situated at the top of the loop of the core domain (Figure 1). Experimental studies have shown the presence of a conserved tyrosine residue in the catalytic domain near the active site in several retroviral integrase structures.<sup>1</sup> Its role in the mechanism of catalysis is based on the proposed structural arrangement of the active site of *E. coli* polymerase, where the residue is suggested to stabilize the activated water molecule.<sup>49</sup> Studies involving the mutation of this residue have resulted in alteration of the preference of the nucleophile during the 3'-processing reaction, from water (3'-processing) to alcohol (strand transfer)<sup>50,51</sup> thus demonstrating its importance. The side chain of the Y143 residue pointing toward the active site is generally assumed to be the active orientation, and simulations carried out by Lee et al.<sup>43</sup> and De Luca et al.<sup>52</sup> suggest that loop flexibility is required to position Y143 in close proximity to the substrate DNA when

the loop is closed, thus suggesting a correlation between the function of Y143 and loop dynamics.

In summary, this enzyme is relatively less well understood compared with the HIV-1 protease (HIV-1 PR) and HIV-1 reverse transcriptase (HIV-1 RT) enzymes of the HIV life cycle. Owing to difficulties in obtaining structural information, the catalytic core domain is the main focus of the majority of studies. The catalytic mechanism of this domain is not fully understood, but theoretical and experimental studies have demonstrated the importance of the side chain orientation of Y143 residue in the loop and the dynamics of the loop itself. However, the active form of the domain is still not accurately known, and there is some ambiguity concerning the highly mobile loop conformation and the number of metal ions present. The G140A/G149A mutant HIV-1 IN has been shown by an experimental<sup>42</sup> and theoretical<sup>43</sup> study to diminish the catalytic activity of the apoenzyme through the rigidifying of the loop. This reduction in loop mobility prevents the loop assuming active conformations and prevents the Y143 residue approaching the active site.

A frequent problem in the process of inhibitor design is the emergence of mutant variants affecting the binding ability of the inhibitor. Several studies of mutant forms of the catalytic domain have attributed a variation in enzyme activity and inhibition compared with the WT enzyme to a change in the dynamics of the important catalytic loop structure.<sup>42,43,53–55</sup> In this study, the conformational dynamics of the catalytic domain has been investigated for the WT and G140A/G149A HIV-1 IN enzymes to gain an increased understanding of the operation of this mutation which is reported to significantly decrease the catalytic activity of the enzyme.

## 2. Methodology

The starting structure, chain C of 1BL3,<sup>35</sup> an apo-form of the wild type catalytic domain of HIV-1 IN, was taken from the Protein Data Bank.<sup>27</sup> This structure was chosen as it was the only crystal structure available at the time possessing all the residues of the loop (residues 140–149) and included the catalytically important Mg<sup>2+</sup> ion. Three end residues, 210–212, are missing from this structure, but this was not considered significant as they are located away from the active loop, the focus of this study. The WHATIF<sup>56</sup> program was used to add polar hydrogens and to check the structure. The AMBER utility XLEAP<sup>57</sup> was used to add other hydrogen atoms and to solvate the system, with a minimum distance of 12 Å from the protein, in a box of 8693 TIPS3P<sup>58</sup> water molecules. One chloride counterion was added to neutralize the overall charge of the system.

This structure contains two solubility enhancing mutants, F185K and W131E, which were mutated back to their native forms using the SCAP<sup>59,60</sup> software, as it has been suggested that they may cause a deformation of the native structure. A previous experimental study has suggested that the F185K mutant resulted in the mutant protein being more active than the WT<sup>61</sup> and simulations carried out by Lee et al. suggest

the mutation to increase the flexibility of the catalytic loop through disruption of this region.<sup>43</sup>

All simulations, unless otherwise stated, have been carried out using the NAMD<sup>62</sup> molecular dynamics package and the CHARMM27 forcefield.<sup>63</sup>

Minimization was carried out in stages, starting with the protein only (5000 steps), followed by solvent (30,000 steps), ions (1000 steps), solvent and ions (20,000 steps), and finally the entire system (40,000 steps), giving a total of 96,000 steps. Two minimization algorithms were used, initially the steepest descent algorithm, followed by the conjugate gradient method. The minimized system was heated to 300 K in the NVT ensemble. The procedure employed a Langevin thermostat with a 10 ps<sup>-1</sup> damping parameter. The heating was carried out gradually in stages at 50 K intervals, each interval being 20,000 steps long. Following this, a further 50,000 steps in the NVT ensemble were carried out using a 5 ps<sup>-1</sup> thermostat damping parameter to control the temperature of the system.

Equilibration simulations, using a Nosé-Hoover barostat in the NPT ensemble were then carried out for 50,000 steps, with a target pressure of 1 atm. A decay parameter of 100 fs and a piston period of 200 fs were used. A further 50,000 steps were run, with a decay parameter of 300 fs and a piston period of 500 fs. The final equilibrated system had box dimensions of 66.57, 68.17, and 61.88 Å.

The apo mutant system, possessing the G140A/G149A double mutant, has also been studied. Since there are no complete crystal structures of this double mutant which possess the Mg<sup>2+</sup> ion, the equilibrated WT structure was taken, and the appropriate residues mutated using SCAP.<sup>59,60</sup> Careful minimization and equilibration of this mutated structure was carried out before use in simulations. Initially, the mutated residues were minimized for 1000 steps while restraining the rest of the system, followed by the minimization of the rest of the protein for 5000 steps, the solvent and counterion for 1000 steps, and, last, the entire system for 10,000 steps. Heating and equilibration of the system were then carried out as for the WT structure.

The final equilibrated system had box dimensions of 66.63, 68.27, and 61.91 Å.

For each system, one MD production simulation, 20 ns in length, has been carried out. All production MD simulations were run in the NVT ensemble using a 2 fs time step, a Langevin thermostat with a 5 ps<sup>-1</sup> damping parameter at a temperature of 300 K. Periodic boundary conditions were used, along with a particle mesh Ewald treatment of electrostatic interactions, using an interpolation order of 6, and switching function applied to the Lennard-Jones interactions between 9 Å and the 10.5 Å cutoff. PME gridsizes of 69 × 72 × 64 Å were used, similar values to those of the boxsizes. SHAKE<sup>64</sup> was applied to all bonds containing hydrogen, using a tolerance of 10<sup>-8</sup> Å.

**2.1. RDFMD Simulation Details.** Reversible Digitally Filtered Molecular Dynamics (RDFMD) enhances conformational change through amplification of the low frequency motions of specific structural regions of a protein. Prior to this study, the method has been successfully applied to a number of protein systems, including *E. coli* dihydrofolate

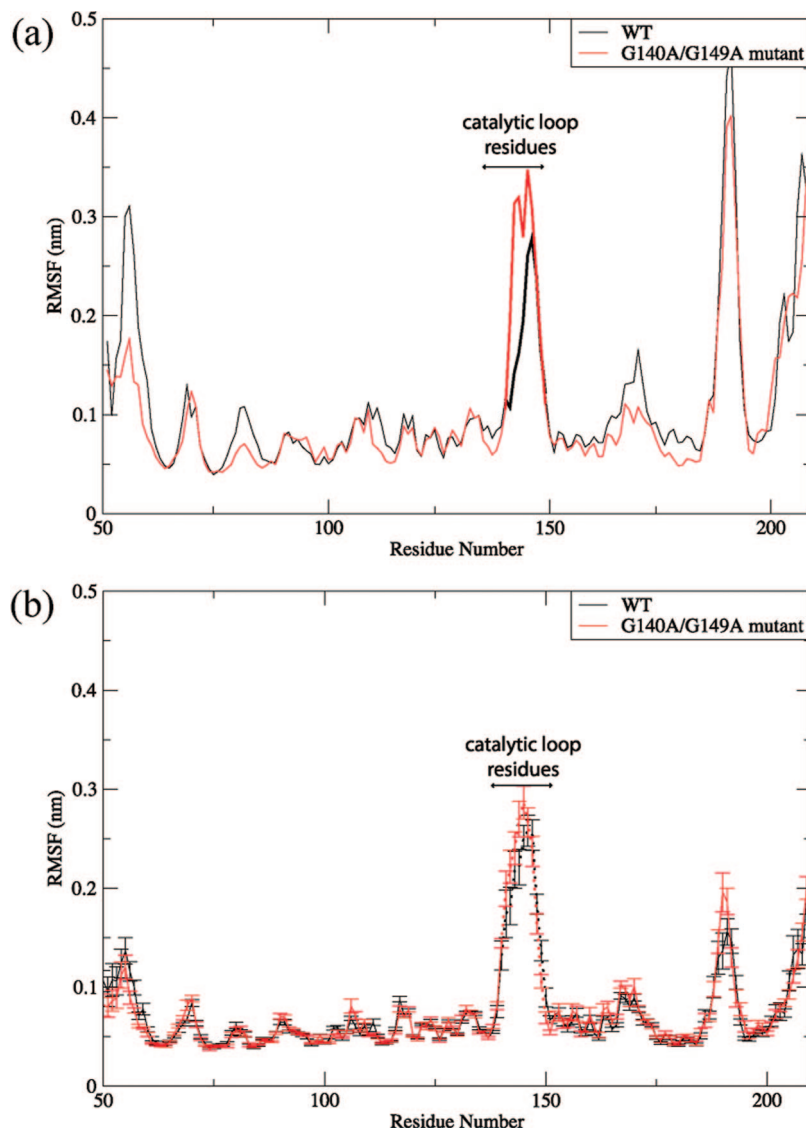
reductase<sup>65</sup> and apo WT HIV-1 PR.<sup>66</sup> The details of the method have been described previously in the literature,<sup>65,67,68,66</sup> and a protocol of parameters for use on regions of interest in proteins has been developed previously.<sup>65</sup> The parameters are heavily interrelated and some are system dependent, and, therefore, a suitable set of parameters has been optimized through trial and error for the study of the WT and mutant HIV-1 IN systems simulated here. These include the use of a digital filter designed to amplify frequencies between 0–100 cm<sup>-1</sup> using 201 coefficients, an amplification factor of 2, a temperature cap of 700 K, and a delay between filter applications of either 50 or 100 steps. The filter sequences were separated by 4 ps of molecular dynamics simulation in the NVT ensemble. This is sufficient time for the system temperature to return to 300 K, and it is during this period of time that conformations for analysis are generated. The final results are taken from piecing together the individual 4 ps MD runs. Each RDFMD simulation produces 100 4 ps MD sections, totalling 400 ps. Since the dynamics of the catalytic loop is thought to play a fundamental role in the activity of the enzyme, all of the atoms of residues in this region (140–149) have been selected as the target region of the filter in the RDFMD simulations. Simulations were run in the NVT ensemble using the Langevin thermostat with a 5 ps<sup>-1</sup> damping parameter.

A total of 12 RDFMD simulations have been carried out using six different starting structures with different velocities. The last stage of the equilibration process was extended for a further 60,000 steps, taking the velocities and starting coordinates after every 10,000 steps, to generate each of the six different starting structures.

### 3. Results

**3.1. Flexibility of the Core Domain.** Analysis of the root mean squared fluctuations (RMSF) of the α-carbon atoms of the residues of the catalytic domain over the length of the MD and RDFMD simulations shows two main regions of flexibility in both the WT and mutant HIV-1 integrase enzyme. Both of these regions are loop structures (residues 140–149 and 186–194), with residues 140–149 comprising the catalytic loop. The residues toward the ends of the catalytic domain, which would be connected to the N- and C-terminal domains in the full length HIV-1 IN, are also shown to possess flexibility. The overall relative higher flexibility of these regions is consistent with the profile of B-factors calculated from MD simulations by Lee et al.<sup>43</sup>

Figure 2 compares the RMSF of the core domain of WT and G140A/G149A HIV-1 IN over the length of the MD and RDFMD simulations. In both the MD and RDFMD simulations, the catalytic loop is shown to be slightly more flexible for the mutant HIV-1 IN compared with the WT, although this difference is perhaps negligible in the case of the RDFMD simulations. These results are in contrast to those reported in the theoretical studies of Lee et al.<sup>43</sup> and an X-ray crystallography study by Greenwald et al.<sup>42</sup> which concluded that the double mutant possesses significantly reduced flexibility.



**Figure 2.** RMSF of the  $\alpha$ -carbon atoms of all residues of the core domain for the WT and G140A/G149A mutant: (a) MD simulations and (b) RDFMD simulations. The standard error for each residue for each of the 12 RDFMD simulations has been calculated.

**3.2. Conformational Dynamics of WT and G140A/G149A HIV-1 IN Catalytic Loop.** Principal component analysis (PCA) was used to identify the major motions of the catalytic loop and highlight any differences in the conformational dynamics which may occur as a result of the mutation from glycine to alanine. In addition, this method has proved useful in the evaluation of the conformational sampling as a result of the RDFMD methodology used. All PCA analysis has been carried out using the utilities provided in tools of the gromacs simulation package.<sup>69,70</sup>

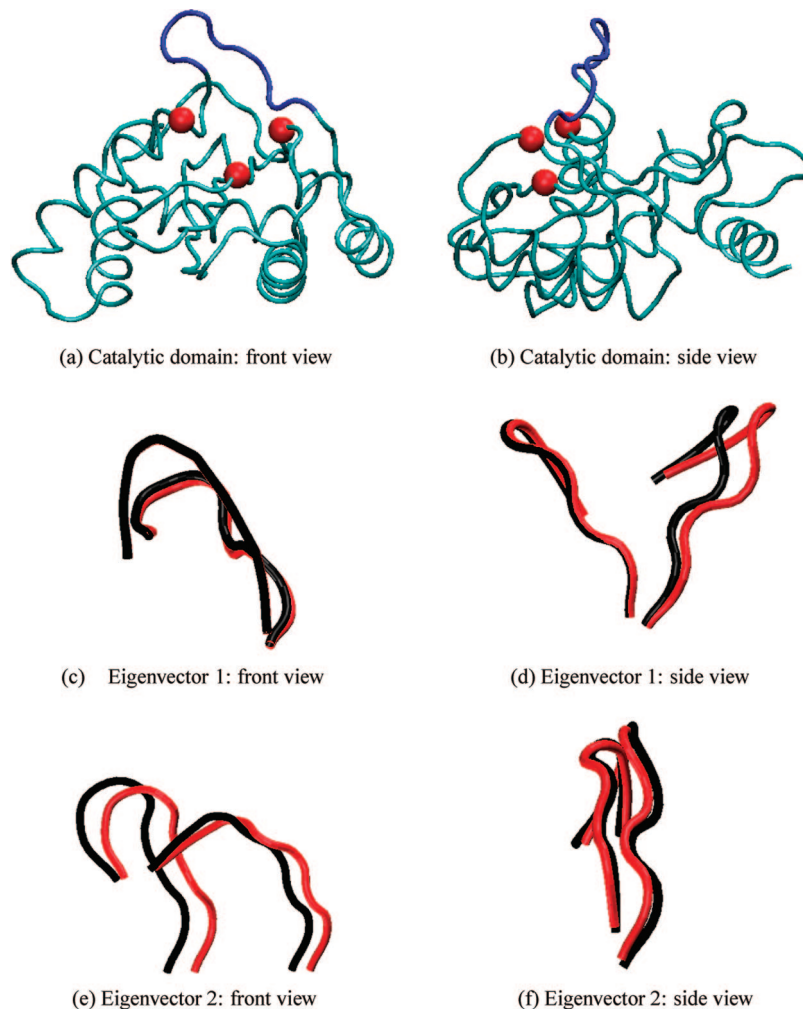
The trajectories of the  $\alpha$ -carbon atoms of the catalytic loop (residues 140–149) of each of the HIV-1 IN systems have been used as the data set in the calculation of the covariance matrix for PCA, with fitting carried out using the nonloop residues. The noncatalytic loop residues have been disregarded for this analysis to avoid the dynamics of the region of interest being obscured by the dynamics of the rest of the protein. Additionally, to compare the eigenvectors of the different MD and RDFMD trajectories, the same reference

structure has been used; the  $\alpha$ -carbon atoms of the first equilibrated WT HIV-1 IN system.

For each of the RDFMD and MD simulations, the number of eigenvectors chosen for study was based on the proportion of the total motion captured and by visual inspection of the motions they represent. As a result, the first two eigenvectors were selected (represent >70% of total motion in MD and RDFMD simulations), and further eigenvectors have been disregarded as they represent higher frequency motions and were harder to define.

**3.2.1. Cross-Correlation Analysis.** Comparison of the inner-products between two eigenvectors has been used to indicate the level of correlation between them, and correlation coefficients provide quantitative information to describe this correlation. A correlation coefficient value of 1 demonstrates the two eigenvectors being compared to be identical, and a value of 0 means the eigenvectors are orthogonal.

Initially, the twelve RDFMD trajectories of the WT and mutant HIV-1 IN systems were concatenated to form two



**Figure 3.** Extreme conformations sampled by the concatenated WT and G140A/G149A HIV-1 IN RDFMD trajectories projected on the first two eigenvectors (black: WT, red: G140A/G149A HIV-1 IN). Eigenvectors generated from the concatenation of the WT and G140A/G149A HIV-1 IN RDFMD trajectories.

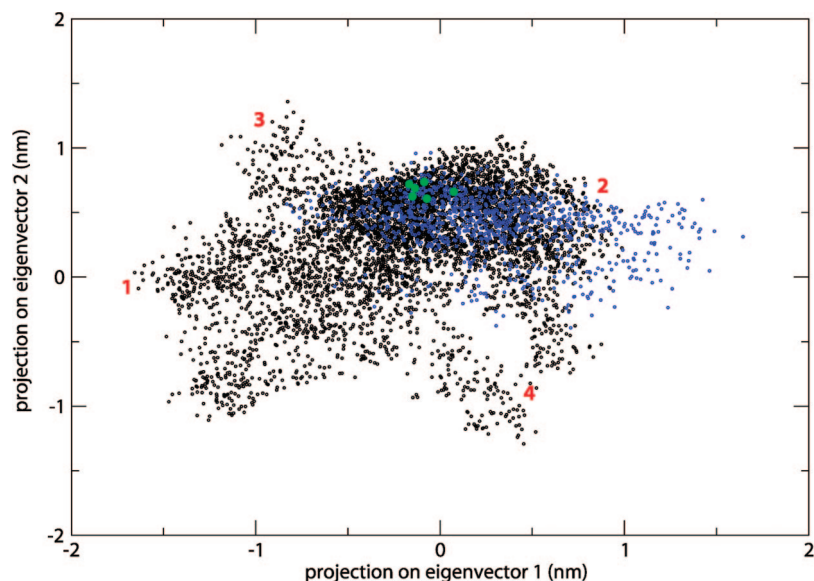
single trajectories (WT and mutant), and PCA was performed on the catalytic loop as described earlier. Comparison of the first few eigenvectors between the WT and mutant are shown to be highly comparable (correlation between inner-products of eigenvectors 1: 0.814, eigenvectors 2: 0.686—see the Supporting Information for cross-correlation plot), and so further PCA analysis was carried out on a single trajectory incorporating all the G140A/G149A and WT trajectories together (i.e., 24 RDFMD trajectories concatenated into one). This allows for direct conformational sampling comparisons to be made between the WT and mutant systems (details given later).

Analysis of the eigenvectors generated by MD simulations shows high diagonal correlation with those of the RDFMD simulations for both the WT (correlation coefficient between inner-products of the first eigenvectors: 0.870, second eigenvectors: 0.766—see the Supporting Information for cross-correlation plot) and mutant simulations (correlation coefficient between inner-products of the first eigenvectors: 0.754, second eigenvectors: 0.600). This demonstrates that the catalytic loop undergoes the same types of movement in RDFMD simulations as it does in MD simulations, thus confirming the RDFMD methodology to sample reasonable conformations of the catalytic loop. Owing to the similarity

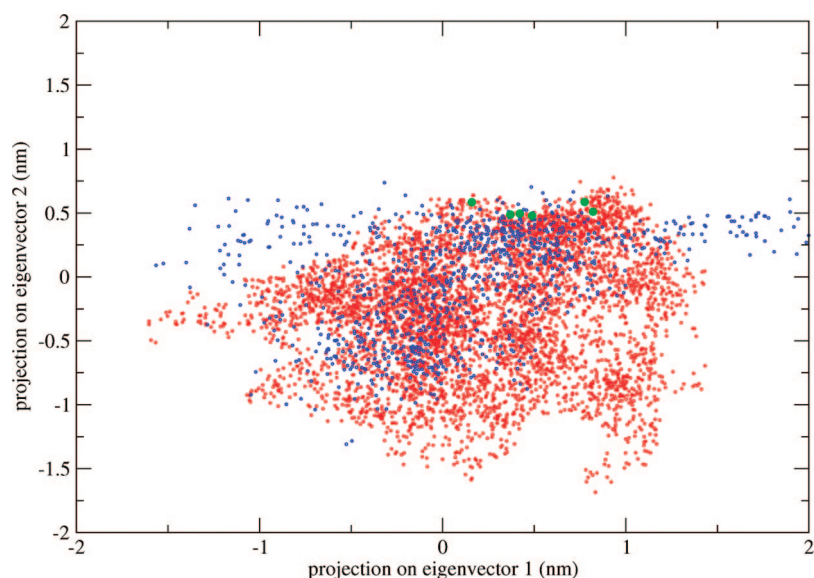
of the eigenvectors generated by MD and RDFMD simulations and to evaluate the sampling of the MD and RDFMD simulations rigorously, the trajectories were projected onto the same eigenvectors. For this purpose, the first two eigenvectors generated by the concatenated WT and mutant RDFMD trajectories were chosen. Although the first two eigenvectors of the MD and RDFMD simulations are similar, the eigenvectors of the RDFMD simulations have been generated through the concatenation of several simulations using several different starting structures, rather than based on just a single simulation, as with the MD simulations.

It is noted that an increase in motion is expected to be observed for the RDFMD simulations compared with the MD simulations as both sets of trajectories are projected onto the eigenvectors of the RDFMD simulations.

Figure 3 shows the extremes of loop motion of the WT and mutant HIV-1 IN loop as a result of the projection of the trajectories onto the first two eigenvectors. The first eigenvector shows the opening/closing gating motion of the catalytic loop toward and away from the active site (Figure 3(c),(d)), a motion thought to be associated with access to the active site, with the loop predicted to overhang the active site in the closed conformation. The second eigenvector shows the loop moving from a more compact structure which



**Figure 4.** Sampling of the first two eigenvectors, generated from the projection of the WT RDFMD (black) and the MD (blue) simulations onto the eigenvectors of the catalytic loop created from the concatenation of all RDFMD trajectories. (The six different starting structures are represented by green circles). The numbers show the extremes of sampling. Regions marked 1 and 2 on a plot are represented by Figure 3(c),(d) (black), and regions 3 and 4 marked on a plot are represented by Figure 3(e),(f) (black).



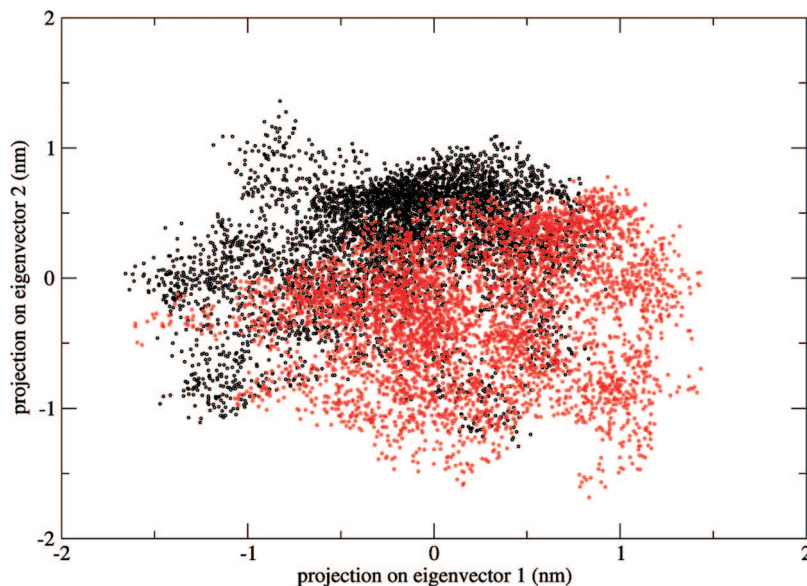
**Figure 5.** Sampling of the first two eigenvectors, generated from the projection of the G140A/G149A mutant RDFMD (red) and the MD (blue) simulations on the eigenvectors of the catalytic loop created from the concatenation of all RDFMD trajectories. (The six different starting structures are represented by green circles.)

leans over to one side, to a more extended conformation of the loop which spans a larger area (Figure 3(e),(f)).

**3.3. Comparison of the Conformations of Catalytic Loop Sampled by the WT and G140A/G149A HIV-1 IN MD and RDFMD Simulations.** The WT and G140A/G149A MD and RDFMD trajectories have been projected onto the first two eigenvectors generated from the concatenation of the 24 WT and G140A/G149A HIV-1 IN RDFMD trajectories, resulting in the two-dimensional plots shown in Figures 4 and 5. The sampling of these plots will demonstrate any correlation between these two eigenvectors and will also highlight differences in the conformations sampled by the MD and RDFMD simulations. The numbered

regions marked on the two-dimensional plot (Figures 4 and 5) show the extremes of sampling of the first two eigenvectors for the WT and G140A/G149A HIV-1 IN, and the corresponding conformations are visualized in Figure 3.

Sampling along the  $x$ -axis of the two-dimensional plots demonstrates the sampling of eigenvector 1, with negative values corresponding to the loop moving over the active site (closed conformation) and positive values corresponding to the loop moving backward, away from the active site (open conformation, Figure 3(c),(d)). Values along the  $y$ -axis demonstrate the sampling of the second eigenvector, with positive values corresponding to a compact loop conformation which leans to one side, and negative values cor-



**Figure 6.** Projection of concatenated WT (black) and G14A/G149A (red) RDFMD simulation trajectories of the loop onto the first and second eigenvectors.

responding to the more extended catalytic loop structure which spans a larger area (Figure 3(e),(f)).

The plots clearly show the RDFMD simulations to sample a greater range of conformations of the catalytic loop compared with the MD simulations, with the sampling by the RDFMD simulations shown to largely encompass the area of sampling of the MD simulations. Owing to the larger amount of conformational space sampled by the RDFMD simulations, comparisons between the loop conformations sampled by the WT and G140A/G149A HIV-1 IN were based on the RDFMD simulations.

Figure 6 overlays the sampling of the WT and mutant RDFMD simulations, demonstrating a difference in the sampling of these two eigenvectors for the WT and G140A/G149A HIV-1 IN RDFMD simulations.

The WT RDFMD simulations show a clear main sampling area where the sampling is shown to be denser on the two-dimensional plot. This is not the case in the mutant RDFMD simulations, whose sampling distribution appears to be more diffuse with no obvious main area of sampling.

Comparison of the sampling of the WT and G140A/G149A HIV-1 IN simulations along eigenvector 1 clearly shows the mutant HIV-1 IN simulations to preferentially sample a larger proportion of open-type conformations (positive values) and fewer closed conformations (negative values) and is able to open further compared to the WT HIV-1 IN simulations (Figure 3(d)). In contrast, the loop of the WT HIV-1 IN is shown to sample significantly more conformations where the loop is closed, overhanging the active site compared with the mutant HIV-1 IN, demonstrated by the greater sampling of the negative values of eigenvector 1.

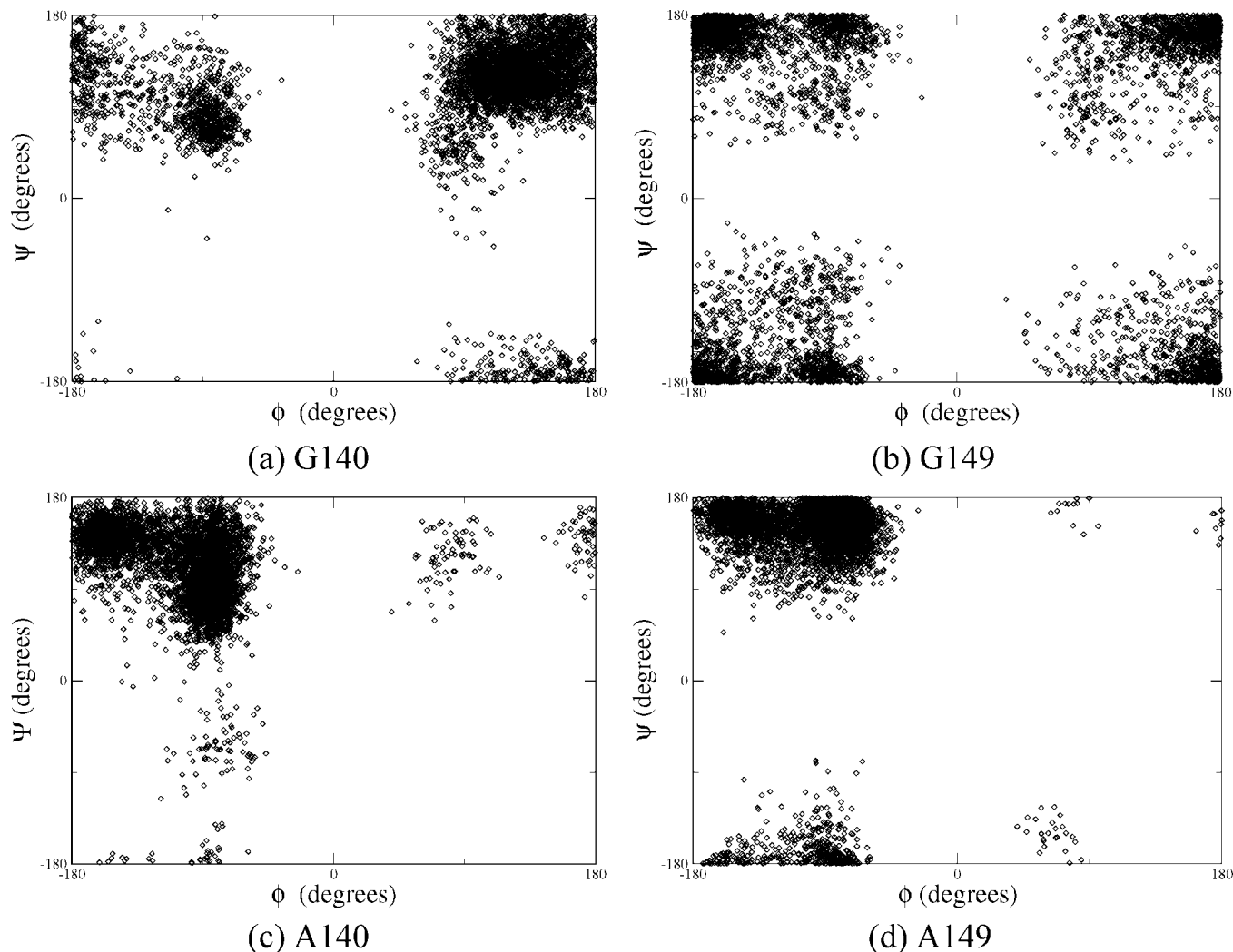
Assessment of the sampling of the second eigenvector also reveals differences between the loop conformations of the WT and G140A/G149A RDFMD simulations. The WT HIV-1 IN is shown clearly to preferentially sample conformations where the loop is in its compact conformation where it leans to one side (more positive values of eigenvector 2) and to be able to achieve conformations which are more compact compared with

those sampled by the RDFMD simulations of the mutant HIV-1 IN loop (see Figure 3(e)). In contrast, the loop of the G140A/G149A HIV-1 IN demonstrates a significantly greater amount of sampling of the extended conformation (negative values of eigenvector 2).

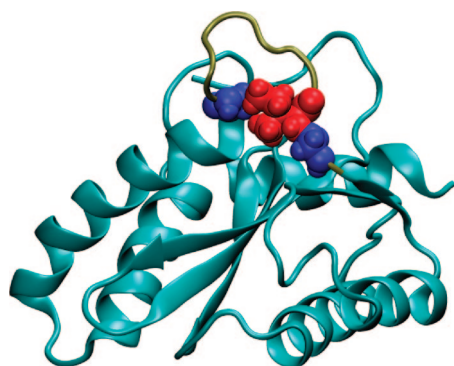
Generally, where the WT RDFMD simulations sample the more compact conformations of the catalytic loop, where it leans to one side, the loop is more likely to be in a more upright position, intermediate between fully open or closed (i.e., the most positive/negative values of eigenvector 1 are not sampled). At these times, the mutant RDFMD HIV-1 IN loop simulations are unable to sample the compact conformations seen in the WT RDFMD simulations and are shown to sample an increased range of more extended loop conformations. Additionally, where the loop is seen to overhang the active site in the closed position (negative values of eigenvector 1), the loop is shown to more likely sample the second eigenvector corresponding to slightly extended loop conformations (Figure 3(c),(d)). This correlation of sampling indicates that, in order for the loop to move into the most closed conformation, which is postulated to be an active conformation, in closer proximity to the active site residues (D116, D64, and E152), the loop cannot exist in the very compact conformations, which are associated with the loop leaning to one side; instead the loop must be at least slightly extended.

Analysis of the  $\phi$  and  $\psi$  torsion angles (Figure 7) of the loop hinge residues of the WT (G140/G149) and mutant (A140/A149) reveal the expected restricted dihedral sampling of the larger alanine residue, with virtually no sampling of positive  $\phi$  values for the A140 and A149 residues. This does not appear to impact on the overall flexibility of the mutant loop, as demonstrated by RMSF (Figure 2), with the mutant enzyme demonstrating similar loop flexibility to the WT. Instead, the results indicate a change in the equilibrium of open/closed conformations sampled. Analysis of the trajectories shows the additional methyl group of the alanine compared with the glycine residue to result in steric





**Figure 7.** Sampling of  $\phi$  -  $\psi$  dihedral angles of the two hinge residues by WT and G140A/G149A HIV-1 IN RDFMD simulations.



**Figure 8.** Compact conformation of catalytic loop sampled by WT RDFMD simulations. G140 and G149 residues highlighted in blue van der Waals representation. Ile141 represented in red van der Waals representation.

hindrance with other residues of the loop. An example is shown in Figure 8 where Gly149 and Ile141 are in close contact in the WT, allowing the loop to be in a compact conformation. In the case of the mutant catalytic loop, which is unable to sample such compact loop conformations, the residues would not be able to approach as closely owing to steric repulsion and resulting in a less compact conformation.

In summary, the loop of the G140A/G149A RDFMD simulations can open further than the WT (as also noted by Lee et al.)<sup>43</sup> and preferentially samples the open conformation, whereas the loop of the WT HIV-1 IN RDFMD simulations is able to close over the active site to a greater extent. The variation in the sampling of open/closed conformations seen between the WT and mutant HIV-1 IN loop may be due to the glycine to alanine mutation limiting the formation of the more closed conformations owing to steric repulsion with other residues of the loop.

**3.4. Role of Tyrosine 143 (Y143).** It has been predicted that the Y143 residue plays an important role in the catalytic activity of this core domain. Based on the activity of 3'-5' exonuclease of *E. coli* polymerase I,<sup>49</sup> it has been suggested that this residue may be involved in the stabilization of, and in guiding, the nucleophile through a hydrogen-bond interaction, thus assisting in the catalysis of the hydrolytic and phosphoryl transfer reactions. The assumed active side chain conformation of this residue points downward toward the active site, and it has been suggested that the dynamics of the catalytic loop may play a role in the positioning of the Y143 residue. Therefore, a change in the dynamics of the loop relative to the WT HIV-1 IN would affect the catalytic

activity of the domain. The loop of the G140A/G149A HIV-1 IN is shown to be able to open to a greater extent and unable to overhang the active site as was observed in the RDFMD simulations of the WT HIV-1 IN. Therefore, the Y143 would reside further from the active site for a greater amount of time compared with the WT and reduce catalytic activity, whereas the WT samples more closed conformations, thus providing more opportunity for the catalytically important Y143 to be near the active site residues.

#### 4. Conclusions

In this study, the RDFMD technique has been shown to efficiently enhance the range of conformations sampled for the HIV-1 IN enzyme compared with MD simulations. The results of RDFMD simulations highlight differences in the conformational dynamics of the catalytic loop between the WT and G140A/G149A HIV-1 IN which may explain the diminished disintegration activity observed in the presence of the G140A/G149A mutant. The results indicate agreement with previous suggestions proposing the importance of the Y143 residue in the catalytic mechanism and the function of the loop to position this residue in the correct orientation for the functional form.<sup>43,50–52</sup> However, unlike previous studies,<sup>42,43</sup> the mobility of the mutant loop is not reduced compared to the WT; when comparing data from several RDMFD simulations, the results indicate the mobility to be largely similar. It must be remembered, however, that the previous experimental studies<sup>42</sup> were carried out in the presence of cacodylate which is known to affect the conformational dynamics of the loop by preventing the binding of the essential metal ion.

The results of this study indicate the mechanism for the diminished catalytic activity could be due to a difference in the equilibrium between the open/closed conformations of the WT and G140A/G149A HIV-1 IN catalytic loops. As mentioned, the active conformation of the enzyme is presumed to involve the Y143 side chain being positioned close to the active site, with the hydroxyl group pointing downward. The PCA results show the G140A/G149A HIV-1 IN system to sample a larger number of conformations where the catalytic loop is open. Additionally, the loop is able to open further and is not able to close to the same extent as seen in the RDFMD simulations of the WT HIV-1 IN. The cause of the difference in sampling is postulated to be due to increased steric hindrance between loop residues in the mutant HIV-1 IN domain, as a result of the larger size of the alanine residue. These differences in conformational sampling of the loop may result in the decreased likelihood of the Y143 side chain being in a suitable location and conformation for the catalytic mechanism to take place in the mutant HIV-1 IN.

Owing to the sampling limitations presented by using conventional MD simulations, the RDFMD methodology has played a crucial role in this study, identifying the proposed differences in dynamics between the WT and G140A/G149A mutant HIV-1 IN.

**Acknowledgment.** We would like to acknowledge the contributions of S. Phillips, M. Swain, C. Edge, R. Gledhill, C. Woods, and A. Wiley for the development and improve-

ment of the RDFMD method. This work was supported by grants from the ESPRC and BBSRC.

**Supporting Information Available:** Cross-correlation plots comparing the eigenvectors of WT and G140A/G149A RDFMD simulations and comparing MD and RDFMD simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### References

- (1) Engelman, A.; Craigie, R. *J. Virol.* **1992**, *66*, 6361–6369.
- (2) Engelman, A.; Bushman, F. D.; Craigie, R. *EMBO J.* **1993**, *12*, 3269–3275.
- (3) van Gent, D. C.; Vink, C.; Groeneger, A. A.; Plasterk, R. H. *EMBO J.* **1993**, *12*, 3261–3267.
- (4) Lee, S. P.; Xiao, J.; Knutson, J. R.; Lewis, M. S.; Han, M. K. *Biochemistry* **1997**, *36*, 173–180.
- (5) Deprez, E.; Tauc, P.; Leh, H.; Mouscadet, J. F.; Auclair, C.; Brochon, J. C. *Biochemistry* **2000**, *39*, 9275–9284.
- (6) Zheng, R. J.; Jenkins, T. M.; Craigie, R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13659–13664.
- (7) Kahn, E.; Mack, J. P. G.; Katz, R. A.; Kulkosky, J.; Skalka, A. M. *Nucleic Acids Res.* **1991**, *19*, 851–860.
- (8) Vink, C.; Groeneger, A. M.; Plasterk, R. H. *Nucleic Acids Res.* **1993**, *21*, 1419–1425.
- (9) Woerner, A. M.; Marchis-Sekura, C. J. *Nucleic Acids Res.* **1993**, *21*, 3507–3511.
- (10) Engelman, A.; Hickman, A. B.; Craigie, R. *J. Virol.* **1994**, *68*, 5911–5917.
- (11) Puras-Lutzke, R. A.; Vink, C.; Plasterk, R. H. *Nucleic Acids Res.* **1994**, *22*, 4125–4131.
- (12) Brown, P. O. In *Retroviruses*; Coffin, J. M., Hughes, H. H., Varmus, H. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1997; pp 161–203.
- (13) Chen, J. C. H.; Krucinski, J.; Miercke, L. J. W.; Finer-Moore, J. S.; Tang, A. H.; Leavitt, A. D.; Stroud, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8233–8238.
- (14) Brown, P. O.; Bowerman, B.; Varmus, H. E.; Bishop, J. M. *Cell* **1987**, *49*, 347–356.
- (15) Lobel, L. I.; Murphy, J. E.; Goff, S. P. *J. Virol.* **1989**, *63*, 2629–2637.
- (16) Chow, S. A.; Vincent, K. A.; Eliason, V.; Brown, P. O. *Science* **1992**, *255*, 723–726.
- (17) Bushman, F. D.; Engelman, A.; Palmer, I.; Wingfield, P.; Craigie, R. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3428–3432.
- (18) Espeseth, A. S.; Felock, P.; Wolfe, A.; Witmer, M.; Grobler, J.; Anthony, N.; Egbertson, M.; Melamed, J. Y.; Young, S.; Hamill, T.; Cole, J. L.; Hazuda, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11244–11249.
- (19) Dyda, F.; Hickman, A. B.; Jenkins, T. M.; Engelman, A.; Craigie, R.; Davies, D. R. *Science* **1994**, *266*, 1981–1986.
- (20) Bujacz, G.; Jaskolski, M.; Alexandratos, J.; Wlodawer, A.; Merkel, G.; Katz, R. A.; Skalka, A. M. *J. Mol. Biol.* **1995**, *253*, 333–346.
- (21) Bujacz, G.; Jaskolski, M.; Alexandratos, J.; Wlodawer, A.; Merkel, G.; Katz, R. A.; Skalka, A. M. *Structure* **1996**, *4*, 89–96.

- (22) Goldgur, Y.; Dyda, F.; Hickman, A. B.; Jenkins, T. M.; Craigie, R.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9150–9154.
- (23) Eijkelenboom, A. P.; van den Ent, F. M.; Vos, A.; Doreleijers, J. F.; Hard, K.; Tullius, T. D.; Plasterk, R. H.; Kaptein, R.; Boelens, R. *Curr. Biol.* **1997**, *7*, 739–746.
- (24) Eijkelenboom, A. P.; Lutzke, R. A.; Boelens, R.; Plasterk, R. H.; Kaptein, R.; Hard, K. *Nat. Struct. Mol. Biol.* **1995**, *2*, 807–810.
- (25) Lodi, P. J.; Ernst, J. A.; Kuszewski, J.; Hickman, A. B.; Engelman, A.; Craigie, R.; Clore, G. M.; Groenenborn, A. M. *Biochemistry* **1995**, *34*, 9826–9833.
- (26) Cai, M.; Zheng, R.; Caffrey, M.; Craigie, R.; Clore, G. M.; Gronenborn, A. M. *Nat. Struct. Biol.* **1997**, *4*, 567–577.
- (27) Berman, H. M.; Westbrook, Z.; Feng, J.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res* **2000**, *28*, 235–242.
- (28) Chen, Z. G. *J. Mol. Biol.* **2000**, *296*, 521–533.
- (29) Chen, J. C.; Krucinski, J.; Miercke, L. J.; Finer-Moore, J. S.; Tang, A. H.; Leavitt, A. D.; Stroud, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8233–8238.
- (30) Yang, Z. N.; Muserm, T. C.; Bushman, F. D.; Hyde, C. C. *J. Mol. Biol.* **2000**, *296*, 535–548.
- (31) Wolfe, A. L.; Felock, P. J.; Hastings, J. C.; Blau, C.; Hazuda, D. J. *J. Virol.* **1996**, *70*, 1424–1432.
- (32) Ellison, V.; Brown, P. O. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 7316–7320.
- (33) Vink, C.; Lutzke, R. A.; Plasterk, R. H. *Nucleic Acids Res.* **1994**, *22*, 4103–4110.
- (34) Hazuda, D. J.; Felock, P. J.; Hastings, J. C.; Pramanik, B.; Wolfe, A. L. *J. Virol.* **1997**, *71*, 7005–7011.
- (35) Maignan, S.; Guilloteau, J. P.; Zhou-Liu, Q.; Clement-Mella, C.; Mikol, V. *J. Mol. Biol.* **1998**, *282*, 359–368.
- (36) Laboulais, C.; Deprez, E.; Leh, H.; Mouscadet, J.; Brochon, J.; Le Bret, M. *Biophys. J.* **2001**, *81*, 473–489.
- (37) Lins, R. D.; Briggs, J. M.; Straatsma, T. P.; Carlson, H. A.; Greenwald, J.; Choe, S.; McCammon, J. A. *Biophys. J.* **1999**, *76*, 2999–3011.
- (38) Wijitkosoom, A.; Tonmunphean, S.; Truong, T. N.; Hannongbua, S. *J. Biomol. Struct. Dyn.* **2006**, *23*, 613–624.
- (39) Lins, R. D.; Adesokan, A.; Soares, T. A.; Briggs, J. M. *Pharmacol. Ther.* **2000**, *85*, 123–131.
- (40) Molteni, V.; Greenwald, J.; Rhodes, D.; Hwang, Y.; Kwiatkowski, W.; Bushman, F. D.; Siegel, J. S.; Choe, S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2001**, *57*, 536–544.
- (41) Bujacz, G.; Alexandratos, J.; Qing, Z. L.; Clement-Mella, C.; Wlodawer, A. *FEBS Lett.* **1996**, *398*, 175–178.
- (42) Greenwald, J.; Le, V.; Butler, S. L.; Bushman, F. D.; Choe, S. *Biochemistry* **1999**, *38*, 8892–8898.
- (43) Lee, M. C.; Deng, J.; Briggs, J. M.; Duan, Y. *Biophys. J.* **2005**, *88*, 3133–3146.
- (44) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.
- (45) Roitberg, A.; Elber, R. *J. Chem. Phys.* **1991**, *95*, 9277–9287.
- (46) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiang, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (47) Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13040–13043.
- (48) Guex, N.; Peitsch, M. C. *Electrophoresis* **1997**, *18*, 2714–2723.
- (49) Beese, L. S.; Steitz, T. A. *EMBO J.* **1991**, *10*, 25–33.
- (50) van Gent, D. C.; Groeneger, A. A.; Plasterk, R. H. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9598–9602.
- (51) van Gent, D. C.; Groeneger, A. A.; Plasterk, R. H. *Nucleic Acids Res.* **1993**, *21*, 3373–3377.
- (52) De Luca, L.; Vistoli, G.; Pedretti, A.; Barreca, M. L.; Chimirri, A. *Biochem. Biophys. Res. Commun.* **2005**, *336*, 1010–1016.
- (53) Barreca, M.; Lee, W.; Chimirri, A.; Briggs, J. M. *Biophys. J.* **2003**, *84*, 1450–1463.
- (54) Brigo, A.; Lee, W.; Mustata, G. I.; Briggs, J. M. *Biophys. J.* **2005**, *88*, 3072–3082.
- (55) Brigo, A.; Lee, W.; Fogolari, F.; Mustata, G. I.; Briggs, J. M. *Proteins* **2005**, *59*, 723–741.
- (56) Vriend, G. *J. Mol. Graphics* **1990**, *8*, 52–56.
- (57) 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*, 1–41.
- (58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (59) Xiang, Z.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421–430.
- (60) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597–608.
- (61) Jenkins, T. M.; Hickman, A. B.; Dyda, F.; Ghirlando, R.; Davies, D. R.; Craigie, R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6057–6061.
- (62) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.
- (63) MacKerell, A. D.; Bashford, D.; Bellott, D.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joesph-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.
- (64) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (65) Wiley, A. P.; Swain, M. T.; Phillips, S. C.; Edge, C. M.; Essex, J. W. *J. Chem. Theory. Comput.* **2005**, *1*, 24–35.
- (66) Wiley, A. P.; Williams, S. L.; Essex, J. W. Unpublished.
- (67) Phillips, S. C.; Swain, M. T.; Wiley, A. P.; Essex, J. W.; Edge, C. M. *J. Phys. Chem. B* **2003**, *107*, 2098–2110.
- (68) Phillips, S. C.; Essex, J. W.; Edge, C. M. *J. Chem. Phys.* **2000**, *112*, 2586–2597.
- (69) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
- (70) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

# JCTC

Journal of Chemical Theory and Computation

## MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein–Protein Interface

Sergio Wong,<sup>\*,†,‡</sup> Rommie E. Amaro,<sup>\*,†,‡</sup> and J. Andrew McCammon<sup>‡,§,||</sup>

*Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics,  
Department of Pharmacology, and Howard Hughes Medical Institute, University of  
California at San Diego, La Jolla, California 92093-0365*

Received September 8, 2008

**Abstract:** The calculation of protein interaction energetics is of fundamental interest, yet accurate quantities are difficult to obtain due to the complex and dynamic nature of protein interfaces. This is further complicated by the presence of water molecules, which can exhibit transient interactions of variable duration and strength with the protein surface. The T-cell receptor (TCR) and its staphylococcal enterotoxin 3 (SEC3) binding partner are well-characterized examples of a protein–protein interaction system exhibiting interfacial plasticity, cooperativity, and additivity among mutants. Specifically engineered mutants induce intercalating interfacial water molecules, which subsequently enhance protein–protein binding affinity. In this work, we perform a set of molecular mechanics (MM) Poisson–Boltzmann (PB) surface area (SA) calculations on the wild type and two mutant TCR–SEC3 systems and show that the method is able to discriminate between weak and strong binders only when key explicit water molecules are included in the analysis. The results presented here point to the promise of MM-PBSA toward rationalizing molecular recognition at protein–protein interfaces, while establishing a general approach to handle explicit interfacial water molecules in such calculations.

### Introduction

Methods to calculate relative binding free energies vary in computational expense and accuracy. The more computationally expensive methods, i.e. free energy perturbation or thermodynamic integration,<sup>1</sup> can calculate relative binding free energies to within a few kcal/mol of experimental values or better. Absolute estimates of binding free energy remain difficult; however, for applications in drug and protein design, it can be useful to differentiate strong from weak binders.

Srivinasan et al.<sup>2</sup> proposed an intermediate method. It calculates average free energy differences between bound and unbound states via examination of a molecular dynamics simulation. A molecular mechanics (MM) force field is used to calculate the internal energy, while a Poisson–Boltzmann (PB) calculation yields the polar component of the solvation free energy. The nonpolar contribution correlates with the surface area (SA). The method is known as MM-PBSA.

Previous applications of MM-PBSA included binding to nucleic acids<sup>2,3</sup> and small molecule binding to enzymes.<sup>4,5</sup> Applications of MM-PBSA to protein–protein interactions are relatively new and far less common. An example is the work by Gohlke and Case<sup>6</sup> on the Ras-Raf system. Of particular interest is to gain insight into molecular recognition. The ability to design protein surfaces that bind a given target protein or molecule has great potential for therapeutic treatment.<sup>7</sup> This is challenging because it is necessary to capture small effects on binding affinity due to mutations or other perturbations at the protein surface. Furthermore, the

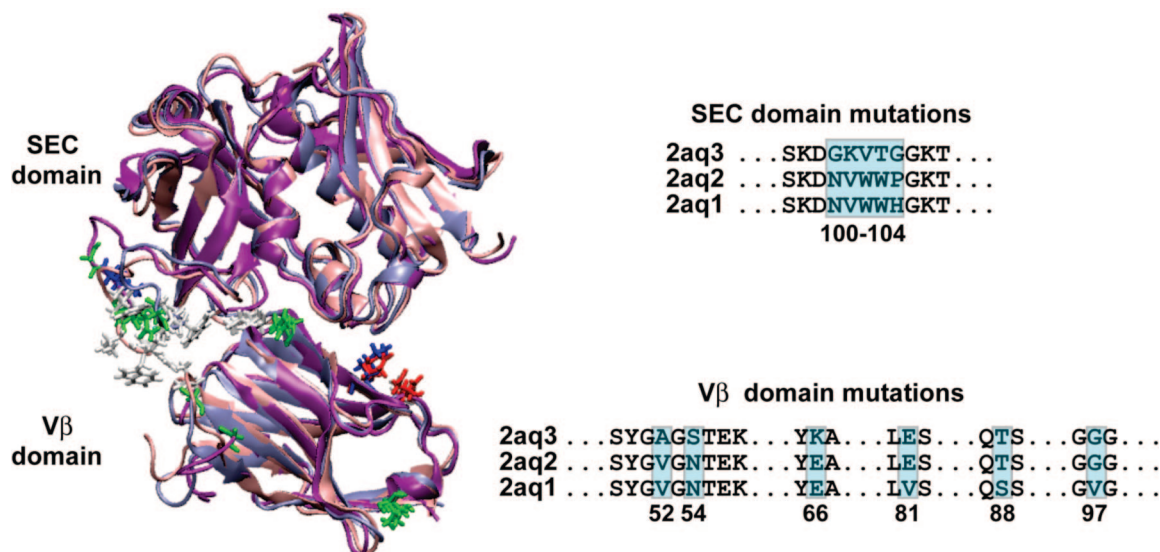
\* Corresponding author phone: (858)822-1469; fax: (858)534-4974; e-mail: swong@mccammon.ucsd.edu (S.W.), ramaro@mccammon.ucsd.edu (R.E.A.). Corresponding author address: Department of Chemistry & Biochemistry, University of California San Diego, 9500 Gilman Drive, Mail Code 0365, La Jolla, CA 92093-0365.

<sup>†</sup> These authors contributed equally to this work.

<sup>‡</sup> Department of Chemistry and Biochemistry and Center for Theoretical Biological Physics.

<sup>§</sup> Department of Pharmacology.

<sup>||</sup> Howard Hughes Medical Institute.



**Figure 1.** The three simulated systems are structurally aligned for comparison. The SEC domain and V $\beta$  domain are shown in cartoon representation, with the mutated positions shown in licorice (hydrophobic residues in white, polar in green, negatively charged in red, positively charged in blue). An excerpt of the full sequence alignment is shown with mutated positions highlighted and numbered.

effects may be subtle and in some cases involve intercalating water molecules.

An example of how mutations can induce intercalating water molecules and improve binding affinity is the engineering of a T-cell receptor mutant that binds staphylococcal enterotoxin 3 (SEC3) 1000 times more strongly than wild type<sup>8</sup> (Figure 1). These systems are exceptionally well characterized in terms of their binding, thermodynamics, and structures and are examples of protein–protein systems that exhibit interfacial plasticity, cooperativity, and additivity among mutants. The effect of each TCR mutation (G17E, A52V, S54N, K66E, E80V, L81S, T87S, G96V) was analyzed via extensive kinetic and structural studies.<sup>9,10</sup> In some cases, the affinity was additive, whereas in others it was cooperative.

The role of water at the interface of biomolecular complexes remains an open and intriguing question.<sup>11,12</sup> In the case of the barnase/barstar and the D1.3/lysozyme complexes, it was found that crystallographically resolved water molecules accounted for 25% of the total interaction energy.<sup>13</sup> There is evidence that removing water mediated contacts, via introduction of functional groups that replace the water, can diminish binding in some cases,<sup>14–17</sup> while it can be favorable in others.<sup>18–20</sup> Moreover, the environment surrounding the water molecule(s) seems to play an important role. Olano and Rick<sup>21</sup> found that transferring a water molecule from the bulk solvent to a hydrophilic cavity is favorable (−4.7 kcal/mol), whereas transferring it to a hydrophobic cavity will be unfavorable (4.7 kcal/mol). Thus a protein–protein interface, which may contain variable interaction types, may present a combination of favorable and unfavorable water mediated contacts.

In this work, we perform three separate explicitly solvated molecular dynamics (MD) simulations using the available high-resolution crystal structures of the TCR/SEC3 complexes and perform MM-PBSA analyses on the resulting

**Table 1.** Summary of Each of the Simulated Systems

system	mutants	number of total simulation atoms	time
2a1	H72Q-r:SEC3–1A4	54,541	16 ns
2a2	A52V/S54N/K66E: SEC3–1D3	55,435	16 ns
2a3	mTCR15-SEC3	54,722	16 ns

trajectories in order to capture their experimentally known binding affinities. The systems include the wild type and two strongly binding mutant systems. Our results show that the MM-PBSA method is able to discriminate between the strongly binding mutants and the weaker-binding wild type complex and suggest that including explicit water molecules in the binding energy calculations was crucial to obtaining the correct energetic trends with statistical significance.

## Methods

**Molecular Dynamics Simulations.** The crystal structures used in this study had PDB codes 2A1 ( $K_D = 5.50 \times 10^{-9}$  M), 2A2 ( $K_D = 1.14 \times 10^{-8}$  M), and 2A3 ( $K_D = 7.55 \times 10^{-6}$  M), which span 3 orders of magnitude in terms of their binding affinities. The protonation states of the histidines and other titratable groups was determined with the WHATIF program.<sup>22</sup> All crystallographically resolved water molecules were retained in the systems; however, the ions (zinc and sulfate) were removed. No additional water molecules were added at the interfaces of any of the complexes. The Amber99 force field<sup>23</sup> was used with xLeap in Amber<sup>9</sup><sup>24</sup> for system setup. A box of TIP3P water molecules<sup>25</sup> was added to solvate to each system. The composite systems each contain approximately 55,000 atoms (Table 1).

The systems were energy minimized for 50,000 steps with NAM2.6<sup>26</sup> and then equilibrated at 298.15 K in the isobaric–isothermal (NPT) ensemble for 2 ns. Periodic boundary conditions and the hybrid Nose-Hoover Langevin

piston method<sup>27</sup> were used to control pressure at 1 atm. After 2 ns, dynamics were continued in the canonical (NVT) ensemble for an additional 16 ns. All hydrogen bond lengths were constrained with the RATTLE algorithm, thus allowing a 2 fs time step. A multiple time-stepping algorithm was utilized, where bonded interactions were evaluated at every time step, and short-range nonbonded interactions were evaluated every 2 timesteps, and long-range electrostatic interactions were evaluated every 4 timesteps.<sup>28,29</sup> Particle mesh Ewald was employed to efficiently treat electrostatics.<sup>30</sup> Simulations were performed on the San Diego Supercomputer Center's Datastar platform with 64 processors, and each nanosecond of dynamics took approximately 0.18 days. The hydrogen bonding and salt bridge interaction analyses were performed with VMD<sup>31</sup> and Matlab.

**MM-PBSA Calculations.** MM-PBSA is a well-established method to calculate binding free energies. It requires dynamical sampling of the complexed system, usually in explicit water, and postprocessing of the trajectory structures. The binding free energy may be calculated by comparison of the complexed trajectory with separate trajectories of the unbound monomers or, as is more typically the case, from a single trajectory of the complex. The binding free energy is calculated using a simple thermodynamic cycle from the energy difference between the complex and the two unbound binding partners. The free energy of each species is calculated as follows

$$G_{\text{tot}} = H_{\text{MM}} + G_{\text{solv}} - T\Delta S_{\text{conf}} \quad (1)$$

where  $H_{\text{MM}}$  corresponds to the molecular mechanics energy, or enthalpic, contribution and is given by

$$H_{\text{MM}} = \sum_{\text{bonds}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}(1 + \cos[n\varphi - \gamma]) + \sum_{i < j}^{\text{atoms}} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \sum_{i < j}^{\text{atoms}} \frac{q_i q_j}{\epsilon R_{ij}} \quad (2)$$

where, per the Amber9 manual, the first sum is over all the chemical bonds, the second term sums over all the angles, the third addresses the dihedral angle potential, and the last two terms explicitly describe the van der Waals and electrostatics contributions, respectively. The indices  $i$  and  $j$  denote individual atoms.  $\epsilon$  is the dielectric constant.

$G_{\text{solv}}$  denotes the solvation free energy. There are two parts to this term. First there is the nonpolar contribution, *i.e.* the cost of opening a cavity in the condensed phase. The product of the surface area and an effective surface tension term often approximates the nonpolar contribution. There are, however, further corrections based on attractive and repulsive solvent–solute interactions that improve the estimate of the nonpolar contribution.<sup>32</sup> Second, the surrounding dielectric, water, responds to protein atomic charges inside the cavity. The work involved is the polar contribution to solvation. Unlike the molecular mechanics contribution, it implicitly includes the solvent entropy.

The entropy term should, in theory, account for the conformational entropy change of the two binding partners

upon complexation. However, due to the complicated and computationally intensive nature of calculating entropy, only an approximate quantity is computed. Here, we perform a normal-mode analysis, using Nmode in Amber9, to compute the vibrational, rotational, and translational entropy.

For each complex snapshot, free energy calculations for the structure of each binding partner are carried out separately (in the absence of the other binding partner). The binding free energy is approximated by the difference

$$\Delta G_{\text{bind}} = G_{\text{tot}}(\text{complex}) - G_{\text{tot}}(\text{monomer A}) - G_{\text{tot}}(\text{monomer B}) \quad (3)$$

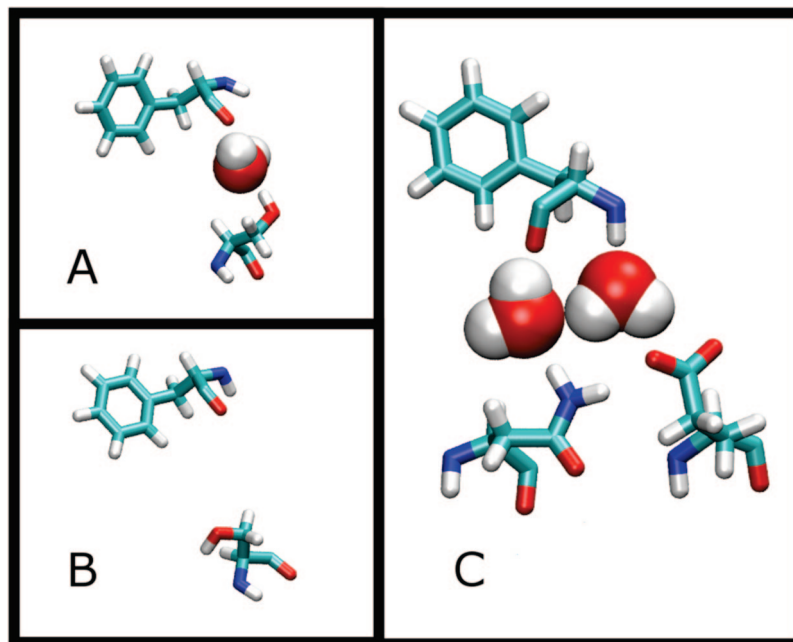
When comparing protein–protein binding of single residue mutants, Massova and Kollman<sup>33</sup> found the entropic contribution difference nearly canceled. Because of the high computational cost of this calculation and its approximate nature, it is often omitted from the overall binding free energy estimate.

Here, the MM-PBSA analysis was performed using the Amber parm99 force field for the MM contribution and APBS<sup>34</sup> for the Poisson–Boltzmann contribution. In order to achieve this, the iAPBS<sup>35</sup> patch was used to call APBS from Sander, the MD engine in Amber 9.

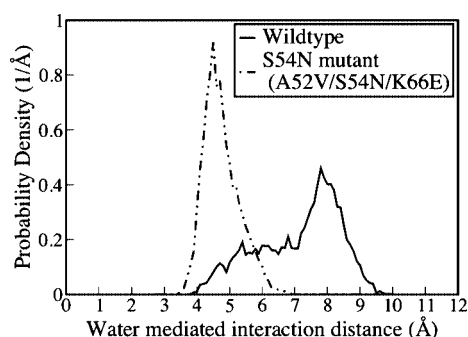
For the MM contributions, the dielectric constant was set to 1.0, and the interaction cutoff distance equaled 999 Å. Other parameters were default. For the APBS calculation, the grid spacing was 0.5 Å in each dimension, the solvent dielectric was set to 80.0, the protein dielectric was set to 1.0, the solvent radius was set to 1.4 Å, the boundary condition (bcfl) was set to 2, there were no counterions, the cubic spline window was set to 0.3, and the rest of the parameters were default values.

**Selection of Interface Water Molecules.** In this study, we chose to consider the effect of interfacial water molecules between SEC3 and its binding partners, *e.g.*, the wild type and mutant TCRs, on the MM-PBSA calculations. In one case, we included all the interface water molecules; more specifically, the closest 200 water molecules to the protein interface were selected and included in the end point free energy calculations. The closest distances of each water molecule to the SEC3 domain and the TCR domain were computed. At every trajectory snapshot and for each water molecule, these two distances were summed, and this sum was used as a metric for the selection process. The 200 water molecules with the smallest sum of squared distances were chosen as interface water molecules for each snapshot. The interface water molecules were considered part of SEC3 for the purposes of the MM-PBSA calculations.

In the second case, we focus on fewer, specific water molecules that were suggested by the crystal structures to mediate contacts between residues Asn54 and Glu56 of TCR with the backbone amide of SEC3 Phe206 (Figure 2). To do that, at each trajectory snapshot, the shortest distance from each water molecule to TCR residues 54, 56, and SEC 206 was computed. For each water molecule, the square of the minimum distance to each of the residues was summed. This sum was used as a metric of how close any given water molecule was to the site of interaction. This list was sorted, and the two water molecules with the smallest sum of squared



**Figure 2.** The S54N mutation stabilizes a water-mediated contact. A) Ser54 makes a water mediated contact with the backbone carbonyl of SEC3 Phe206. B) Ser54 in a conformation where the contact is broken. C) Asn54 making the water-mediated contact and also a hydrogen bond interaction with Asp56.



**Figure 3.** Distance probability density for a water-mediated interaction across the interface. The interaction involves the backbone carbonyl of SEC3 Phe206 and either 1) the OH group of Ser54 (wild type) or 2) the amide group of the S54N mutant side chain. It is clear the Asn54 side chain makes this water mediated contact nearly all of the time.

distances were chosen as the bridging water molecules for each snapshot. The same procedure was followed for the wild type and mutant trajectories; the two interface waters closest to TCR residues 54 and 56 and SEC3 residue 206 were included in the wild type calculation. The intercalating water molecules were considered part of SEC3 for the purposes of the MM-PBSA calculations.

**Block Averaging.** Each trajectory was divided in 5 equal sections. MM-PBSA results for each of these segments were averaged. The standard error of the mean for these five data points was reported.

## Results and Discussion

Kieke et al.<sup>8</sup> produced TCR mutants and selected the best binders against SEC3. Two iterations of this process yielded

a mutant system that binds 1000 times stronger than wild type. The strongly binding complex involved nine mutations: G17E, A52V, S54N, K66E, Q72H, E80V, L81S, T87S, and G96V.<sup>10</sup> Four of the mutations are located at the binding interface (A52V, S54N, K66E, and Q72H); the other five are distal to the interface and did not show a significant contribution to the binding affinity.<sup>10</sup>

To explore the structural effects of the mutations, Cho et al.<sup>9</sup> resolved crystal structures of two mutant complexes. The H72Q-r system (PDB code: 2aq1) has eight of the total nine mutations and a dissociation constant of  $5.5E(-9)$  M; Q72H was reversed, but it has a minor effect on the binding affinity ( $K_d = 5.5E(-9)$  M versus  $5.3E(-9)$  M). The A52V/S54N/K66E system (PDB code: 2aq2) involves three mutations (A52V, S54N, and K66E) and has a  $1.1E(-8)$  M dissociation constant. Molecular dynamics simulations of these two mutants and the wild type complex were performed (Table 1), and the resulting trajectories were analyzed via MM-PBSA.

**MD Simulations.** MD simulations of the two mutants and wild type complexes spanned 16 ns. These were explicit solvent simulations, under periodic boundary conditions, with neutralizing counterions and where electrostatics were treated via PME.<sup>30,36</sup> Except for minor fluctuations, the  $C_\alpha$  rmsd of the trajectories is below  $2.5 \text{ \AA}$  for the wild type system and below  $2.0 \text{ \AA}$  for the mutants (Supporting Information).

**Effect of Interfacial Water.** Cho et al.<sup>9</sup> noted that the structure of the Ser54Asn mutant introduces several new bridging water molecules across the interface that were not present in the wild type system. These ordered interfacial water molecules, herein called intercalating water molecules, persisted in their original location at the interface throughout the 16 ns simulation (Figure 2). The wild type Ser54 system is also able to order water across the interface, but it easily

**Table 2.** MM-PBSA Results Using iAPBS<sup>a</sup>

iAPBS/MM-PBSA results	H72Q-r	std error of mean	A52V S54N K66E	std error of mean	wild type	std error of mean
experimental $K_d$	5.5 E(-9) M		1.1 E(-8) M		7.6 E(-6) M	
no water molecules included	-47.7	3.1	-47.2	0.8	-44.1	1.2
interfacial water molecules (200)	-146.3	4.7	-143.8	14.7	-148.2	3.0
intercalating water molecules	-55.3	2.8	-54.3	0.8	-46.2	1.4

<sup>a</sup> All results are in kcal/mol, except for the experimental  $K_d$  data. The internal entropy contributions are not included in these estimates since it was nearly the same for the three cases.

adopts a second conformation that disrupts the hydrogen bond network. When Asn substitutes Ser, it forms a hydrogen bond with Asp56, which better positions it for ordering the interactions with the intercalating water molecules. Both the Ser and Asn contact the carbonyl oxygen of Phe206 on SEC3 via this water-mediated interaction. To quantify the difference in behavior between these two residues, the distance distribution between the F206 carbonyl oxygen and either the alcohol hydrogen of Ser54 or amide hydrogen in the case of the S54N mutant was calculated (Figure 3). As this distance is mediated by two water molecules, it is longer than the usual 3.5 Å between hydrogen bond donor and acceptor atoms. This result shows a dramatic shift in the distance distribution toward a shorter distance for the Asn mutant.

**End Point Free Energy Calculations.** In the first and simplest case, we performed a MM-PBSA calculation on only the protein domains, without including any explicit water molecules. Such calculations have been shown in the literature to work for other protein-protein systems, such as the Ras-Raf complex.<sup>6</sup> The results of this calculation were indeed able to predict the correct binding affinity trend for the three systems, but their statistical uncertainties overlap (Table 2). The absolute free energy binding estimates were not correct, nor would we expect them to be given the approximations used in this study, such as neglecting the internal (strain) energy of the systems and entropic contributions. The standard error of the mean, however, indicates that when only the protein domains are included in the calculations, the three systems yield binding energy results that are all within statistical error.

Entropy contribution estimates, using a harmonic approximation, yielded essentially the same results for the three complexes ( $T\Delta S = -39.8, -40.8, -40.0$  with a standard deviation of  $\sim 10$  kcal/mol). Due to their similar values and large standard deviations, these values were not included in the analysis. The similarity of these entropy values is not surprising. In their work on computational alanine scanning, Massova and Kollman<sup>33</sup> found entropy contributions to be nearly the same for alanine mutants of a protein-peptide complex. It is also important to keep in mind that it is difficult to converge these estimates,<sup>44</sup> which may explain the large standard deviations. Fortunately, even without including these contributions, the effect of alanine mutations can be captured, to some extent, via MM-PB(GB)SA analysis.<sup>6</sup>

As a second case, all interfacial water molecules were included in the free energy calculations. More specifically, the 200 closest water molecules to both subunits were selected in each frame and included as part of the SEC3 domain in the MM-PBSA calculations. This approach is similar to hybrid solvent models where the first hydration

shell is explicitly included, and a continuum solvent model approximates the bulk solvation beyond that point.<sup>37,38</sup> The results including all interface water molecules yielded an incorrect energetic trend as compared to experimental binding affinities and high absolute values for the binding affinities in all three cases (Table 2). In addition, the statistical error is significantly higher than the other two scenarios considered. The larger number of explicit electrostatic and van der Waals contributions to the energetic terms causes the perceived higher binding affinity. In principle, this effect should be the same for the three cases and the relative difference between them should be unaffected. However, the statistical error in this case increases beyond the binding affinity differences; therefore, the ability to discern among the mutant and wild type systems is further reduced.

In a third case, the MM-PBSA analysis explicitly included two interfacial water molecules, which were first identified in the Ser54Asn mutant crystal structure (Figure 2). The two water molecules were considered as part of SEC for these calculations. The actual water molecule coordinates were taken from each individual MD snapshot so that they were the closest two water molecules to TCR residues 54 and 56 and SEC3 residue 206. When these two intercalating water molecules were included, the correct energetic trends were reproduced, and it was possible to discern between the strongly binding mutants and the weaker binding wild type complex with statistical significance (Table 2). This result underscores the importance of including *specific* interface water molecules, and not necessarily *all* water molecules, for the computational prediction of the binding energetics. In this case, the Ser54Asn mutation introduced an important water-mediated contact between the two protein subunits. Without accounting for explicit water, the contact is lost, along with the higher binding affinity that accompanies this mutation. These results are further substantiated by a previous study that employed end point free energy calculations for a nucleic acid system, which also showed better performance when a key explicit water molecule was included in the analysis.<sup>3</sup>

Considered more broadly, these results are not surprising considering that water is well-known to play an important role in protein dynamics and function.<sup>39</sup> The water molecules in the first hydration shell, which make direct contacts to protein residues, adapt to the topology and physicochemical character of the protein surface.<sup>40</sup> The subsequent dynamics of these water molecules is affected by the hydrophobicity and curvature of the protein surface.<sup>41</sup> Longer residence times of water molecules that make ordered interactions with exposed protein groups are frequently exhibited at protein-protein interfaces, and these longer residence times are typically related to stronger interaction energies.<sup>42</sup>



**Limitations of the Method.** The failure to differentiate between the H72Q-r and A52V/S54N/K66E mutants may be attributed, at least in part, to two reasons. First, the 10-fold difference in dissociation constants between the complexes corresponds to  $\sim 1.6$  kcal/mol binding free energy. This difference is very close to the error margin ( $\sim 1.35$  kcal/mol) for the more rigorous and computationally intensive free energy of perturbation or thermodynamic integration calculations.<sup>43</sup> Therefore, it would be very surprising if this method could reliably rank complexes so similar in affinity. Second, the two mutant complexes are identical at the interface. The differences between the two complexes are located away from the interface.<sup>10</sup>

In addition, the entropic cost of fixing a water molecule at the interface was neglected. This value is particularly difficult to converge,<sup>45</sup> but it may be up to  $\sim 2.1$  kcal/mol.<sup>46</sup> Although we do not attempt to calculate it here, we note that accounting for this entropic penalty may bring our binding free energy estimates closer to the experimentally determined values. As a first order approximation, one may assume that the only difference in solvent entropy among the wild type and mutant systems is the ordering of the two intercalating water molecules. Given the value provided by Dunitz, one would estimate an entropic penalty of approximately 4.2 kcal/mol. Such an assumption would reduce, but not eliminate, the statistically significant difference between the wild type and mutant systems.

The main finding of this work is that including key intercalating water molecules in MM-PBSA calculations can help discriminate between strong- and weak-binding complexes. In the case of the TCR and SEC3 systems, the importance of particular intercalating water molecules was established experimentally, wherein the crystallographic structure of the mutant complexes showed that these ordered water molecules mediate interfacial contacts of the mutated residues.<sup>9</sup> No direct interface contacts were introduced by the mutations. In this work we show, by comparison to other scenarios where the interface water molecules are either completely excluded or included, that explicitly including select water molecules improves the predictive ability of the MM-PBSA calculations. Although we concede that ignoring the entropy loss of these water molecules will introduce some error that may overestimate the stability of the complex, calculations at this level of approximation may be sufficiently accurate to achieve the goal of discriminating between strong- and weak-binding protein-protein complexes.

## Conclusions

The results presented here highlight the crucial role that intercalating water molecules play in protein-protein interaction energetics. The results also point to the limitations of using a completely continuum solvent model, such as PBSA. However, we show that such errors may be rescued if key water molecules, such as those present in the first solvent shell or as suggested from crystallographic data, are included explicitly in the calculations. More broadly, the ability to computationally discern between strong- and weak-binding complexes can be particularly useful in the study of

molecular recognition and in the prediction and design of new or mutant protein systems. This work shows MM-PBSA may be of use in that effort.

**Acknowledgment.** We thank Dr. Robert Konecny for kindly providing the iAPBS wrapper for the calculations and Drs. David Cerutti and Tushar Jain and Profs. Roy Mariuzza and Mike Gilson for helpful discussions. R.E.A. is funded by NIH F32-GM077729 and NSF MRAC CHE060073N. Funding by NIH GM31749, NSF MCB-0506593, and MCA93S013 to J.A.M. also supported this work. Additional support from the Howard Hughes Medical Institute, San Diego Supercomputing Center, Accelrys, the W.M. Keck Foundation, the National Biomedical Computational Resource, and the Center for Theoretical Biological Physics is gratefully acknowledged.

**Supporting Information Available:**  $C_\alpha$  rmsd trajectories for wild type and mutant system simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Simonson, T.; Archontis, G.; Karplus, M. Free energy simulations come of age: Protein-ligand recognition. *Acc. Chem. Res.* **2002**, *35* (6), 430–437.
- (2) Srinivasan, J.; Cheatham, T. E., III.; Kollman, P.; Case, D. Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (3) Spackova, N.; Cheatham, T. E., III.; Ryjacek, F.; Lankas, F.; van Meervelt, L.; Hobza, P.; Sponer, J. Molecular dynamics simulations and thermodynamics analysis of DNA-drug complexes. Minor groove binding between 4',6-diamidino-2-phenylindole and DNA duplexes in solution. *J. Am. Chem. Soc.* **2003**, *125* (7), 1759–1769.
- (4) El-Barghouthi, M. I.; Jaime, C.; Al-Sakhen, N. A.; Issa, A. A.; Abdoh, A. A.; Al Omari, M. M.; Badwan, A. A.; Zughul, M. B. Molecular dynamics simulations and MM-PBSA calculations of the cyclodextrin inclusion complexes with 1-alkanols, para-substituted phenols and substituted imidazoles. *J. Mol. Struct.: THEOCHEM* **2008**, *853* (1–3), 45–52.
- (5) Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophys. J.* **2004**, *86* (1), 67–74.
- (6) Gohlke, H.; Kiel, C.; Case, D. Insights into Protein-Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras-Raf and Ras-RalGDS Complexes. *J. Mol. Biol.* **2003**, *330* (4), 891–913.
- (7) Babine, R.; Bender, S. Molecular Recognition of Protein-Ligand Complexes: Applications to Drug Design. *Chem. Rev.* **1997**, *97* (5), 1359–1472.
- (8) Kieke, M. C.; Sundberg, E.; Shusta, E. V.; Mariuzza, R. A.; Witttrup, K. D.; Kranz, D. M. High affinity T cell receptors from yeast display libraries block T cell activation by superantigens. *J. Mol. Biol.* **2001**, *307* (5), 1305–1315.
- (9) Cho, S.; Swaminathan, C. P.; Yang, J.; Kerzic, M. C.; Guan, R.; Kieke, M. C.; Kranz, D. M.; Mariuzza, R. A.; Sundberg, E. J. Structural basis of affinity maturation and intramolecular

- cooperativity in a protein-protein interaction. *Structure* **2005**, *13* (12), 1775–1787.
- (10) Yang, J.; Swaminathan, C. P.; Huang, Y.; Guan, R.; Cho, S.; Kieke, M. C.; Kranz, D. M.; Mariuzza, R. A.; Sundberg, E. J. Dissecting cooperative and additive binding energetics in the affinity maturation pathway of a protein-protein interface. *J. Biol. Chem.* **2003**, *278* (50), 50412–50421.
- (11) Li, Z.; Lazaridis, T. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.* **2007**, *9*, 573–581.
- (12) Papoian, G.; Ulander, J.; Wolynes, P. Role of Water Mediated Interactions in Protein-Protein Recognition Landscapes. *J. Am. Chem. Soc.* **2003**, *125*, 9170–9178.
- (13) Covell, D.; Wallqvist, A. Analysis of protein-protein interactions and the effects of amino acid mutations on their energetics. The importance of water molecules in the binding epitope. *J. Mol. Biol.* **1997**, *269* (2), 281–297.
- (14) Clarke, C.; Woods, R.; Gluska, J.; Cooper, A.; Nutley, M.; Boons, G. Involvement of Water in Carbohydrate-Protein Binding. *J. Am. Chem. Soc.* **2001**, *123*, 12238–12247.
- (15) Mikol, V.; Papageorgiou, C.; Borer, X. The role of water molecules in the structure-based design of (5-hydroxynorvaline)-2-cyclosporin: synthesis, biological activity and crystallographic analysis with cyclophilin A. *J. Med. Chem.* **1995**, *38* (17), 3361–3367.
- (16) Sharrow, S.; Edmonds, K.; Goodman, M.; Novotny, M.; Stone, M. Thermodynamic consequences of disrupting a water-mediated hydrogen bond network in a protein: pheromone complex. *Protein Sci.* **2005**, *14*, 249–256.
- (17) Branden, B.; Goldman, E.; Mariuzza, R.; Poljack, R. Anatomy of an antibody molecule: structure, kinetics, thermodynamics and mutational studies of the antilysozyme antibody D1.3. *Immunol. Rev.* **1998**, *163*, 45–57.
- (18) Chen, J.; Wawrzak, Z.; Basarab, G.; Jordan, D. Structure-based design of protein inhibitors of scytalon dehydratase: displacement of a water molecule from the active site. *Biochemistry* **1998**, *37* (51), 17735–17744.
- (19) Weber, P.; Pantoliano, M.; Simons, D.; Salemme, J. Structure-Based Design of Synthetic Azobenzene Ligands for Streptavidin. *J. Am. Chem. Soc.* **1994**, *116*, 2717–2724.
- (20) Connelly, P.; Aldape, R.; Bruzzese, F.; Chambers, S.; Fitzgibbon, M.; Fleming, M.; Itoh, S.; Livingston, D.; Navia, M.; Thomson, J.; Wilson, K. Enthalpy of hydrogen bond formation in a protein-ligand binding reaction. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91* (5), 1964–1968.
- (21) Olano, R.; Rick, S. Hydration free energies and entropies for water in protein interiors. *J. Am. Chem. Soc.* **2004**, *126* (25), 7991–8000.
- (22) Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graphics* **1990**, *8* (1), 52–56.
- (23) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21* (12), 1049–1074.
- (24) Case, D. A.; Darden, T.; Cheatham, T. E., III.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Merz, K.; Pearlman, D.; Crowley, M.; Walker, R.; Zhang, W.; Wang, B.; Hayik, A.; Roiberger, A.; Seabra, G.; Wong, K.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Morgan, J.; Hornak, V.; Cui, G.; Beroza, P.; Matthews, D.; Schfmeister, C.; Ross, W.; Kollman, P. AMBER 9; 2006.
- (25) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
- (26) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
- (27) Feller, S.; Zhang, Y.; Pastor, R.; Brooks, B. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613.
- (28) Schlick, T.; Skeel, R. D.; Brunger, A. T.; Kalé, L. V.; Board, J. A.; Hermans, J.; Schulten, K. Algorithmic Challenges in Computational Molecular Biophysics. *J. Comput. Phys.* **1999**, *151* (1), 9–48.
- (29) Grubmuller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Mol. Simul.* **1991**, *6* (1), 121–142.
- (30) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An  $N \log(N)$  Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.
- (31) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33–38.
- (32) Wagoner, J. A.; Baker, N. A. Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (22), 8331–8336.
- (33) Massova, I.; Kollman, P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **1999**, *121* (36), 8133–8143.
- (34) Baker, N.; Sept, D.; Joseph, S.; Holst, M.; McCammon, J. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.
- (35) Konecny, R. iAPBS interface on the Web. <http://mccammon.ucsd.edu/iapbs> (accessed April 4, 2008).
- (36) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.
- (37) Lee, M. S.; Olson, M. A. Evaluation of Poisson solvation models using a hybrid explicit/implicit solvent method. *J. Phys. Chem. B* **2005**, *109* (11), 5223–5236.
- (38) Beglov, D.; Roux, B. Dominant solvation effects from the primary shell of hydration: Approximation for molecular dynamics simulations. *Biopolymers* **1995**, *35* (2), 171–178.
- (39) Helms, V. Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem* **2007**, *8* (1), 23–33.
- (40) Pizzitutti, F.; Marchi, M.; Sterpone, F.; Rossky, P. J. How protein surfaces induce anomalous dynamics of hydration water. *J. Phys. Chem. B* **2007**, *111* (26), 7584–7590.
- (41) Hua, L.; Huang, X.; Zhou, R.; Berne, B. J. Dynamics of water confined in the interdomain region of a multidomain protein. *J. Phys. Chem. B* **2006**, *110* (8), 3704–3711.
- (42) Schroder, C.; Rudas, T.; Boresch, S.; Steinhauser, O. Simulation studies of the protein-water interface. I. Properties at the molecular resolution. *J. Chem. Phys.* **2006**, *124* (23), 234907.

- (43) Shirts, M.; Pitera, J.; Swope, W.; Pande, V. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- (44) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2003**, *25* (2), 238–250.
- (45) Hamelberg, D.; McCammon, J. A. Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- (46) Dunitz, J. D. Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, *264*, 670.

CT8003707

## Conformational Studies of Methyl $\beta$ -D-Arabinofuranoside Using the AMBER/GLYCAM Approach

Hashem A. Taha,<sup>†</sup> Norberto Castillo,<sup>†</sup> Pierre-Nicholas Roy,<sup>\*,‡</sup> and Todd L. Lowary<sup>\*,†</sup>

*Department of Chemistry and Alberta Ingenuity Centre for Carbohydrate Science, Gunning-Lemieux Chemistry Centre, University of Alberta, Edmonton, Alberta, Canada T6G 2G2, and Department of Chemistry, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*

Received September 12, 2008

**Abstract:** Furanosides are important constituents of a number of glycoconjugates from many microorganisms. The highly flexible nature of these furanosyl moieties is believed to contribute significantly to their role in biological processes. Therefore, an understanding of the conformational preferences of these molecules is an important area of research. As part of a larger program involved in the conformational analysis of mycobacterial oligofuranosides, molecular dynamics simulations on methyl  $\beta$ -D-arabinofuranoside (**3**) have been carried out using the AMBER forcefield and the GLYCAM carbohydrate parameter set. This approach was used to predict the rotamer population distribution about the hydroxymethyl group (C4–C5 bond) as well as the ring puckering of this flexible ring system. Comparison of the conformer distributions obtained during the simulation of **3** using the TIP3P water model with those obtained by analysis of  $^1\text{H}$ – $^1\text{H}$  coupling constant data indicated that this water model was insufficient to describe the solvation of this system. However, the use of the TIP4P and TIP5P models led to improved agreement with conformer populations obtained from NMR data.

### Introduction

Tuberculosis (TB) is a significant world health concern that affects one-third of the world's population and kills nearly three million people annually.<sup>1</sup> The dramatic increase in drug-resistant strains of the bacterium that causes TB, *Mycobacterium tuberculosis*, has heightened interest in the development of novel vaccines and antibiotics for the prevention and treatment of the disease.<sup>2–4</sup> The treatment of bacterial disease often involves the use of antibiotics that act by inhibiting the biosynthesis of the bacterial cell wall.<sup>5–7</sup> Some of the clinically used anti-TB drugs act in this fashion,<sup>8</sup> and the mycobacterial cell wall has attracted significant attention in the development of new drugs for the treatment of this disease.<sup>2,9,10</sup>

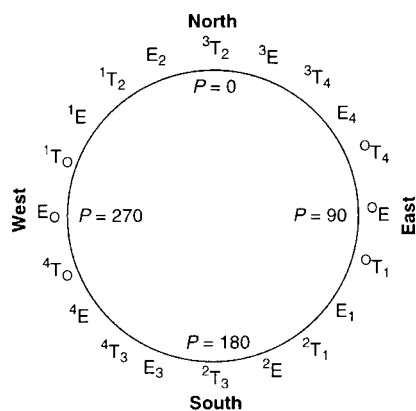
The cell wall in mycobacteria, a complex array of polysaccharides, proteins, lipids, and glycolipids,<sup>11</sup> consists

of two major carbohydrate components, an arabinogalactan (AG) and a lipoarabinomannan (LAM). These polymers contain arabinose and galactose residues present exclusively in the furanose ring form. Compared to their pyranose counterparts, these five-membered rings are the least thermodynamically stable forms of most monosaccharides. Although pyranosides generally exist in well-defined energetically favorable chair conformations that have little ring strain, furanosides have increased ring strain. Consequently, furanoside rings are flexible species that can adopt several conformational states, typically separated by low energy barriers.<sup>12</sup> The inherent flexibility of these five-membered ring carbohydrates is postulated<sup>13</sup> to play a role in the protection of the organism against its environment by providing a malleable scaffold that promotes the ideal packing of the mycolic acid residues attached to the terminal ends of the AG.<sup>11</sup> These densely packed lipids form a lipophilic barrier at the outer part of the cell wall that protects the organism against both the passage of antibiotics and the immune system of the infected host.<sup>14</sup>

\* Corresponding author e-mail: pnroy@uwaterloo.ca (P.-N.R.), todd.lowary@ualberta.ca (T.L.L.).

<sup>†</sup> University of Alberta.

<sup>‡</sup> University of Waterloo.

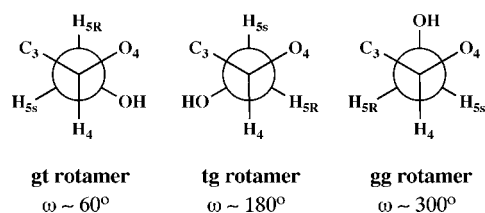


**Figure 1.** Pseudorotational itinerary for a D-aldofuranose ring.

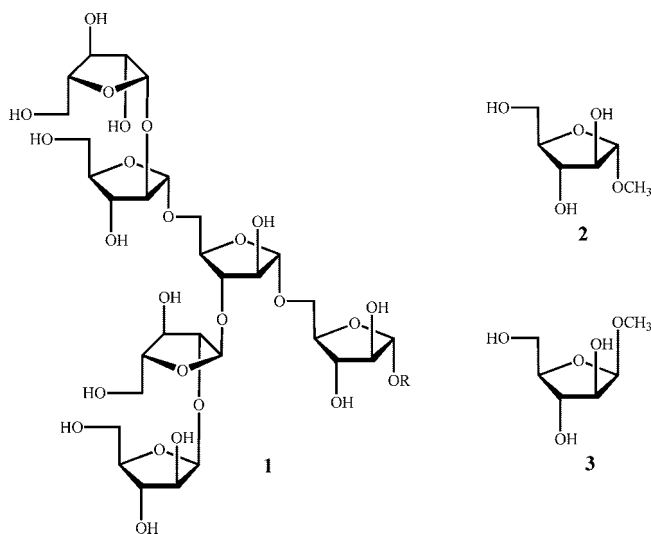
The critical role that the arabinofuranose ring system plays in the cell wall structure of mycobacteria has prompted our interest in understanding the conformational preferences of oligosaccharides containing arabinofuranose rings.<sup>12,15–19</sup> This work has been driven by hope that this knowledge will aid in the design of potential inhibitors of the enzymes involved in biosynthesis of the arabinofuranose-containing portions of the mycobacterial cell wall.<sup>9</sup>

The conformational preferences of furanose rings are difficult to determine due to their great flexibility; these rings can adopt various envelope and twist conformations as depicted on the pseudorotational itinerary<sup>20</sup> (Figure 1). The PSEUROT program<sup>21,22</sup> has been an important tool in the determination of the solution conformers of these furanose rings by analysis of the vicinal  $^1\text{H}$ – $^1\text{H}$  coupling constants of the ring hydrogen atoms. This program assumes a two-state approximation, where a North (N) and a South (S) conformer equilibrate via pseudorotation, rather than inversion through a high-energy eclipsed planar ring form. Each conformer can be described by its pseudorotational phase angle,  $P$ , and puckering amplitude,  $\Phi_{\text{max}}$ . From a set of vicinal  $^1\text{H}$ – $^1\text{H}$  coupling constants obtained from a given five-membered ring,  $P$  can be calculated for both conformers along with the N/S population ratio. The identity and relative populations of the N and S conformers depend on several factors, including steric and stereoelectronic effects.<sup>23,24</sup> These factors also influence the distribution of rotamers about the C4–C5 bond of these ring systems. In the past, the hydroxymethyl group rotamer populations of several furanose-containing mono- and oligosaccharides have been calculated.<sup>25–28</sup> This was done by analysis of the coupling constants between the two pro-chiral C5 protons and the C4 proton, measured from the  $^1\text{H}$  NMR spectra of these compounds. A generalized Karplus equation<sup>29,30</sup> was then used to determine the relative populations of the three rotamers. The three minimal staggered rotamers depicted in Figure 2 describe the C4–C5 bond conformation.

Although the majority of the D-arabinofuranose residues present in AG and LAM are in the  $\alpha$ -configuration,  $\beta$ -D-arabinofuranosyl residues are also present. These moieties are typically found at the periphery of the polysaccharides, and their hydroxymethyl group is usually substituted with other groups that play key roles in the survival and pathogenicity of the organism.<sup>11</sup> For example, in the AG,



**Figure 2.** Definition of *gt*, *tg*, and *gg* rotamers about the C4–C5 bond. The quantity  $\omega$  describes the dihedral angle between the endocyclic oxygen O4 and the hydroxyl group OH-5.



**Figure 3.** Hexasaccharide motif (1) found at the nonreducing termini of mycobacterial AG and LAM, methyl  $\alpha$ -D-arabinofuranoside (2), and methyl  $\beta$ -D-arabinofuranoside (3).

this group is the site to which the mycolic acids are esterified,<sup>11</sup> while in the LAM a range of “capping” motifs are found attached to this position.<sup>31</sup>

Several structural investigations on mycobacterial arabinofuranosides have been carried out,<sup>15–19</sup> including various high-level *ab initio* and density functional theory (DFT) calculations. Our current interests lie in the study of larger oligomers of D-arabinofuranose, for which we have NMR data.<sup>25,32</sup> Of particular interest is a hexasaccharide motif found at the nonreducing end of AG and LAM (1, Figure 3), comprised of both  $\alpha$  and  $\beta$  arabinofuranose residues. However, the size of these molecules renders their treatment with *ab initio* or DFT methods impractical. As a result, the use of force field models to probe the conformation of these oligosaccharides is the only practical way to model these systems. In previous conformational studies of methyl  $\alpha$ -D-arabinofuranoside<sup>33</sup> (2, Figure 3) as well as similar investigations on pyranosides,<sup>34–36</sup> it was found that the use of the AMBER<sup>37</sup> force field in conjunction with the GLYCAM<sup>38</sup> parameter set was an effective method for modeling carbohydrates. Herein, we report our investigations on the use of this computational approach to study the conformation of methyl  $\beta$ -D-arabinofuranoside (3). In particular, we examine the ability of the method to reproduce the distribution of rotamers about the C4–C5 bond and the ring conformers as determined by NMR spectroscopy. In the course of these studies we demonstrated that the water model used had an

important influence on the ability of this method to reproduce experimentally determined conformer populations.

## Methodology

The AMBER forcefield with the GLYCAM parameter set was employed for the description of methyl  $\beta$ -D-arabinofuranoside **3**. All molecular dynamics (MD) simulations were carried out using the *Sander* module in the AMBER 9.0 suite of programs,<sup>37</sup> and the electronic structure calculations were performed with Gaussian 03.<sup>39</sup>

**Ring Averaged Charge Calculations.** To carry out the MD simulations on **3**, partial atomic charges were required. Due to the flexibility of furanose rings, a novel ring-averaged approach developed by us,<sup>33</sup> which incorporates the effects of ring flexibility, was implemented to obtain partial charges for **3**. This method is a modification of the usual GLYCAM approach developed by Woods and workers.<sup>36</sup> The input geometry of **3** was obtained from crystallographic data,<sup>40</sup> and an *ab initio* geometry optimization was performed at the HF/6–31G\* level of theory. An initial set of restrained partial atomic charges was obtained using the RESP approach.<sup>41</sup> After a 50 ns MD simulation, 200 random conformations were selected from the resulting trajectory, and a constrained *ab initio* geometry optimization (HF/6–31G\*) was performed for each conformation, restricting the dihedral angles involving the hydroxyl protons to the values obtained from the simulations. For the 200 new conformations, each with unique ring geometry and torsion angles, a single point HF/6–31G\* calculation was performed to obtain the RESP fit, and the final charge of each atom was calculated as an average. The value of the RESP restraint weight was set to 0.01, and fitting was only performed to the nonaliphatic hydrogen atoms.<sup>42</sup> The charges obtained from this procedure were ensemble averaged over several exocyclic torsions and ring conformations. The charges used in the simulations are provided in the Supporting Information.

**Coupling Constant Analysis.** An NMR spectrum of methyl  $\beta$ -D-arabinofuranoside **3** in D<sub>2</sub>O solution was obtained using a Varian Inova 400 MHz spectrometer. Rotamer populations were calculated using a generalized Karplus relationship<sup>30</sup> represented by eq 1

$${}^3J_{HH} = 14.63\cos^2\varphi - 0.78\cos\varphi + 0.60 + \sum_i \lambda_i \{0.34 - 2.31\cos^2[\xi_i\varphi + 18.4|\lambda_i|]\} \quad (1)$$

where  $\lambda_i$  is the difference in electronegativities of non-hydrogen substituents along the coupling path, and  $\xi_i$  is +1 or –1 depending on the relative orientation of substituents. The angle  $\varphi$  is the dihedral angle between the two coupled protons. A number of different calculations were carried out. In one, the angles used the values for ideally staggered rotamers (60°, 180°, and –60°). In another set of calculations, this angle was equal to a value corresponding to the most probable dihedral angle obtained from the respective MD simulations (e.g., 61°, 174°, and –61° for the TIP3P H<sub>4</sub>–H<sub>5R</sub> dihedral). This data are provided in the Supporting Information. Equation 1 produces values for limiting coupling constants of the individual rotamers. Due to rapid intercon-

version of rotamers, the experimentally observed vicinal couplings are the weighted sums of these values. This is represented by eqs 2–4, where  $X_i$  represents the mole fractions of each rotamer, and the coefficients are the limiting coupling constants calculated based on idealized staggered geometry:

$${}^3J_{H_4-H_{5R}} = 0.9X_{gg} + 11.1X_{gt} + 4.5X_{tg} \quad (2)$$

$${}^3J_{H_4-H_{5R}} = 2.7X_{gg} + 2.7X_{gt} + 11.1X_{tg} \quad (3)$$

$$1 = X_{gg} + X_{gt} + X_{tg} \quad (4)$$

Once a set of experimental couplings is obtained, eqs 2–4 can be simultaneously solved to produce the desired rotameric distribution.

**Solution Simulations.** A 200 ns molecular dynamics simulation of **3** was performed in a box of 264 TIP3P<sup>43</sup> water molecules under NPT conditions. The total box size was 22.771 × 25.372 × 25.544 Å. The temperature of the system was set to 300 K and the pressure to 1 atm using a constant temperature thermostat with the weak coupling algorithm (*ntt* = 1) and a constant pressure barostat with isotropic position scaling (*ntp* = 1). A cutoff of 8 Å was used for nonbonded interactions. Scaling parameters (SCNB and SCEE) were set to 1.0 in accordance with the GLYCAM approach. All simulations were carried out under NPT conditions using the SHAKE<sup>44</sup> algorithm to constrain all hydrogen-containing bonds. Prior to production dynamics, minimization of the water molecules was performed, followed by minimization of the whole system, 100 ps of annealing and 150 ps of equilibration. Ewald summation was used to handle long-range electrostatics.

**Hydrogen Bonding Analysis.** Hydrogen bond analysis of the 200 ns trajectory was performed using the *ptraj* module in the AMBER suite. All oxygen atoms in **3** were assigned as potential hydrogen bond donors, and all hydroxyl hydrogen atoms were assigned as potential acceptors. The criteria used for a hydrogen bond was a cutoff distance of 3.5 Å between the two heavy atoms and an angle cutoff of 120.0°. All intramolecular interactions were also specified in the analysis using the INCLUDESELF keyword.

**Gas-Phase Simulations.** A simulation of **3** was performed in the gas phase to highlight the importance of the effects of explicit solvation. The same simulation parameters were employed as those for the solution simulations. However, periodic boundary conditions and Ewald summation were not used. A cutoff of 18.0 Å was utilized for the long-range interactions.

**TIP4P and TIP5P Simulations.** Simulations of **3** using the TIP4P<sup>43</sup> and TIP5P<sup>45</sup> water models were also performed. Using TIP4P, a water box of 711 molecules with dimensions of 30.086 × 32.853 × 32.457 Å was used. For TIP5P, a box of 568 molecules was used with dimensions of 29.803 × 33.253 × 34.560 Å. All other parameters were identical to the TIP3P simulations.

## Results and Discussion

**Atomic Charges.** Our modification of the usual GLYCAM approach for charge derivation, which incorporates the effects

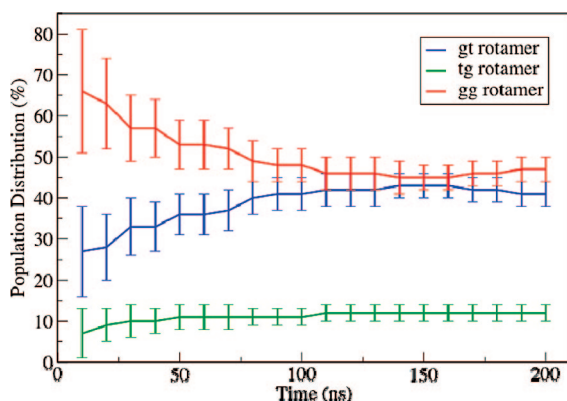
**Table 1.** Experimental Rotameric Distributions about the C4–C5 Bond in **3**

	ideal geometry	TIP3P	GAS	TIP4P	TIP5P
$X_{gt}$ (%)	57	55	58	56	54
$X_{tg}$ (%)	8	10	6	8	12
$X_{gg}$ (%)	35	35	36	36	34

of the ring flexibility, was described previously.<sup>33</sup> The charges obtained from this procedure are presented in the Supporting Information. An average rmsd of the carbon atoms of the ring based on the 200 conformations used in the ring averaging was calculated, and a value of 0.17 with a fluctuation of 0.07 was obtained. This parameter provides a measure of the ring flexibility of the system. To quantify the magnitude of the rmsd in terms of puckering, a correlation study was carried out that indicates what change in ring puckering corresponds to a particular value of rmsd. A large number of conformations (namely 100000) were selected for this correlation study from the simulation based on our ring-averaged atomic charges (discussed below). This number of conformations was chosen to obtain a statistically meaningful estimate.

**C4–C5 Rotamer Populations.** Using eqs 2–4 for ideally staggered rotamers, a distribution of 57:8:35 *gt:tg:gg* was obtained from the NMR data (Table 1). Moreover, experimental distributions using the most probable dihedral angle values ( $H_4-C_4-C_5-H_{5S}$  and  $H_4-C_4-C_5-H_{5R}$ , see the Supporting Information) obtained from the MD simulations were also calculated; these results are reported in Table 1.

In our previous studies of methyl  $\alpha$ -D-arabinofuranoside (**2**),<sup>33</sup> the length of the MD simulations required to obtain convergence of C4–C5 rotamer populations was established to be 200 ns. A convergence study of the rotamer populations in **3** as a function of simulation time is presented in Figure 4. As was observed for **2**, a simulation time of 200 ns was optimal for the convergence of rotamer populations in **3** to reasonable uncertainties (see the Supporting Information). Figure 5 shows a time-dependence study of the C4–C5 torsion angle and its associated distribution. Integration of the peaks in the histogram results in a distribution of 40:12:48 for the *gt:tg:gg* rotamers. The *gg* rotamer is found to be the most populated, a result that contradicts the experimentally observed rotamer population distributions (Table 1).

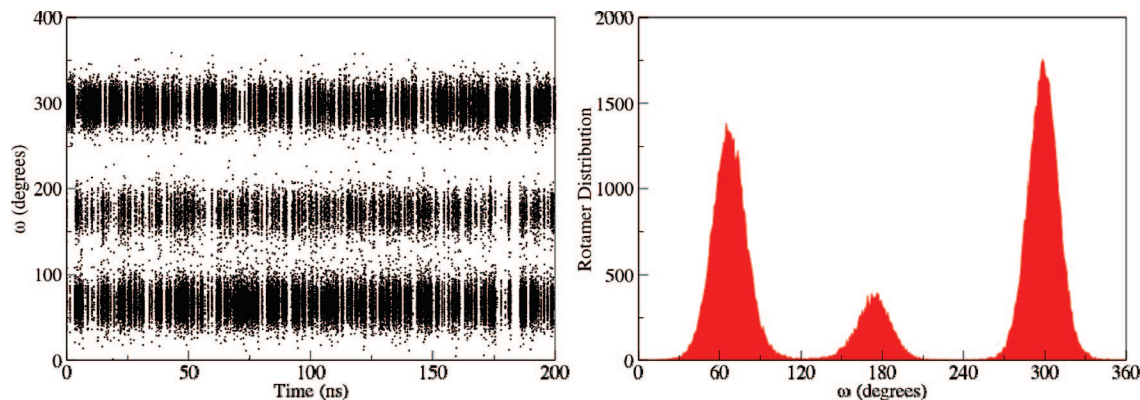
**Figure 4.** Convergence of the rotamer distribution of **1** as a function of simulation time in TIP3P water.

The discrepancy between the results from the MD simulation and experiment may be rationalized by solvation effects. In aqueous solution, it could be expected that the C5 hydroxyl group is heavily solvated by water molecules, rendering it too sterically demanding to be favorably oriented in the *gg* rotamer. In addition, solvation of this hydroxyl group would also diminish intramolecular hydrogen bonds present within the sugar that may stabilize the *gg* rotamer.

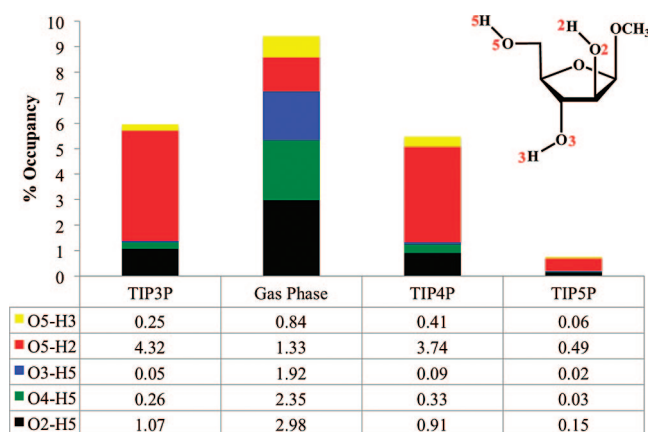
To assess intramolecular hydrogen bond patterns in the conformer ensemble of **3**, these interactions were analyzed as described in the Methodology section. The results of this analysis suggest that there exists intramolecular hydrogen bonding interactions within **3** that may be responsible for the discrepancy in the results between experiment and simulation. In the 200 ns trajectory, a hydrogen bond exists between O5 and H2 in 4.3% occupancy and between O2 and H5 in 1.1% occupancy (Figure 6, TIP3P). This occupancy of hydrogen bonds is most likely to occur in the *gg* rotamer where these particular hydrogen donors (H2 and H5) are in close proximity to the oxygen acceptors (O5 and O2, respectively). We hypothesize that this collective occupancy of 5.4% should be reduced or eliminated by competing *intermolecular* hydrogen bonds with water and therefore affect the difference in *gg* and *gt* populations. In other words, if the *gt* population is increased by 5.4% or more, and the *gg* population is decreased by the same amount, the populations will near experimental values. Other intramolecular H-bonds (e.g., O5...H3 and O4...H5) do exist but in minimal occupancy.

To probe further intramolecular hydrogen bonding in **3**, a gas-phase simulation was performed. Analysis of the resulting trajectory showed that there was, as expected, an overall increase in the intramolecular hydrogen bonding interactions compared to those observed in the TIP3P simulations (Figure 6, gas phase). Concomitantly, the rotamer distribution diverged further from the experimental data. The C4–C5 bond rotamer distributions from these gas-phase MD simulations are presented in Figure 7 and compared to results from TIP3P simulations and experiment. In the gas phase, the *gt* rotamer (23%) significantly decreased in population, while the *gg* (55%) and *tg* (22%) rotamer populations increased. This suggests a higher propensity for the *gg* and *tg* rotamers to form intramolecular hydrogen bonds compared with the *gt* rotamer. This deviation from experiment led to the implication that the TIP3P water model used in the simulations may be insufficient to provide an accurate representation of the aqueous solvation of **3**.

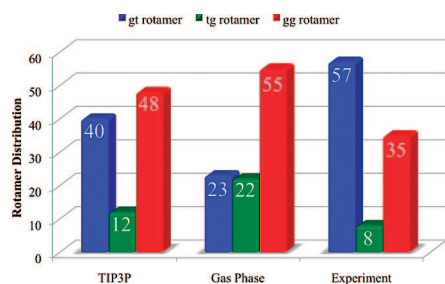
This result led to the examination of alternative water models for use in the simulations. A number of different potential functions for liquid water have been developed over the past several years.<sup>46</sup> Although many bear little resemblance to “real water”, there exist models that are more sophisticated than TIP3P, which would generally give better agreement. TIP3P can be depicted as an effective rigid pair potential composed of Lennard-Jones (LJ) and coulombic interaction terms, where the LJ site is on the oxygen atom and the charge sites are on the hydrogen atoms; it uses atom-centered point charges to represent the electrostatic interactions. As an alternative, TIP4P<sup>43</sup> is a branched and rigid water



**Figure 5.** Time dependence of the C4–C5 torsion angle (left) and its associated distribution (right) in TIP3P water.



**Figure 6.** Percentage occupancy of intramolecular hydrogen bonds.



**Figure 7.** Rotamer population distribution of **1** in solution (TIP3P) and gas phase and their comparison to experiment.

model that has a LJ site on the oxygen and bare charge sites on the hydrogen atoms and along the bisector between the hydrogen atoms. Because this model utilizes more interaction sites, it provides an improved description of liquid water and thus may aid in obtaining a satisfactory computational outcome. Moreover, TIP5P<sup>45</sup> is a branched rigid model that utilizes an additional two charge sites that represent the electron lone pairs in “real” water. Although these models offer a better representation of water, their use in MD simulations would typically result in a significantly increased computational cost, as the number of interaction sites increases.

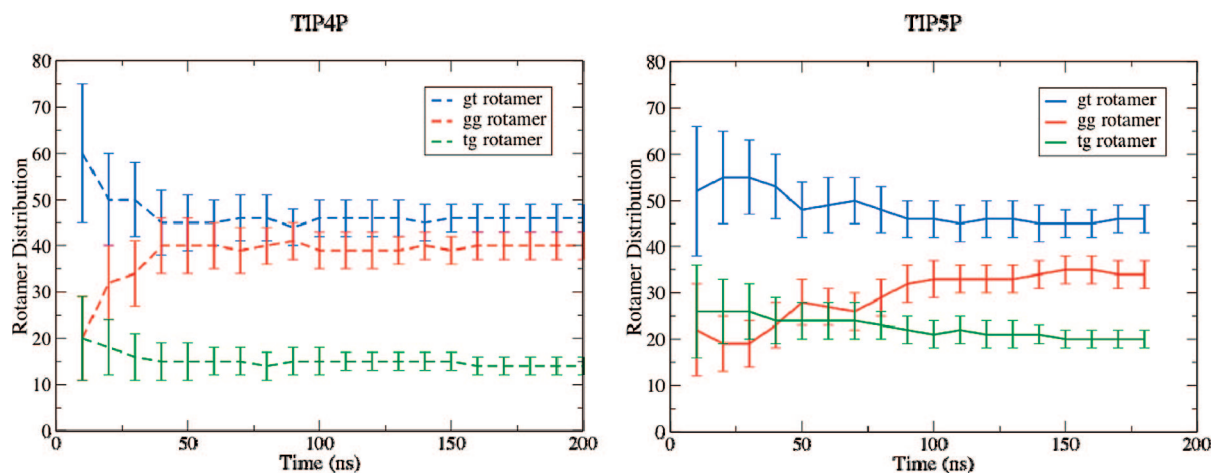
To study the effect of the water model on the conformational ensemble, MD simulations of **3** were performed using the TIP4P and TIP5P water models, at simulation times of

200 and 185 ns, respectively; the rotamer populations converged at these times (Figure 8). The distributions of the C4–C5 bond torsion angles of **3** in these two water models are presented in Figure 9; a comparison of the rotamer distributions obtained from all simulations is shown in Figure 10.

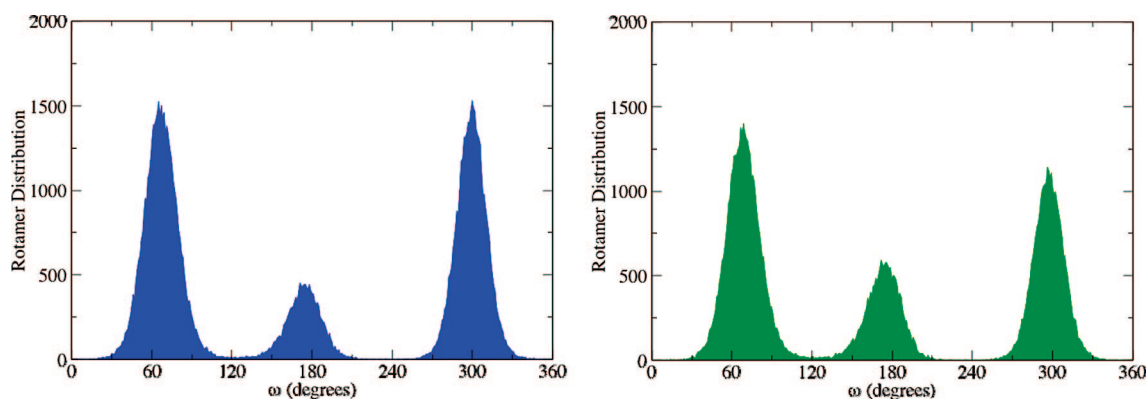
Gratifyingly, and in contrast to the results using TIP3P, the general trend in rotamer populations obtained from the TIP4P simulations parallels that found in experiment ( $gt > gg > tg$ ); the calculated populations of the  $gt$  and  $gg$  rotamers are similar within the error bars. The  $gt$  rotamer is the most populated of the three, whereas the  $tg$  rotamer is the least populated. A similar trend was observed using the TIP5P water model. From these results we conclude that the use of the more refined water models in the simulations provides a conformational ensemble of C4–C5 bond rotamers in **3** that are in better agreement with experiment. Moreover, a comparison of the MD rotameric distributions with the experimental values obtained by using the most probable dihedral angle values from the respective simulations (Table 1) shows a slightly better agreement in each case when compared to the idealized geometry.

Hydrogen bond analysis of the resulting trajectories of the TIP4P and TIP5P simulations shows a decreased percentage in the occupancy of the intramolecular hydrogen bonds compared to that observed in the TIP3P simulations (Figure 6, TIP4P and TIP5P). With the TIP4P model, the O5···H2 hydrogen bond is 3.7% populated, whereas the O2···H5 hydrogen bond is 0.9% populated, levels that are reduced compared to the TIP3P simulations (4.3% and 1.1%, respectively). Although these differences are small, the better agreement with experiment demonstrates the superiority of the TIP4P water model for simulations of **3**. The formation of intramolecular hydrogen bonds is even more reduced in the TIP5P simulation – the overall percentage of these intramolecular interactions is extremely minor (a total of 0.75% occupancy) – suggesting that with this model intermolecular interactions with solvent molecules compete well with the formation of intramolecular hydrogen bonds. Therefore, this implies that the more sophisticated water models may provide a more accurate representation of the solvent effects involved in this system. Although this is not unexpected when comparing TIP4P and TIP5P to real water, the simulation results do demonstrate that the use of less

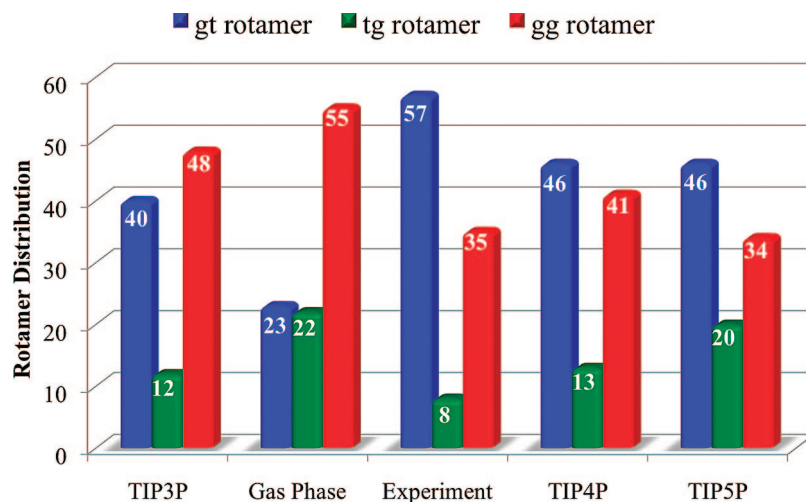




**Figure 8.** Convergence of the rotamer distribution of **1** as a function of simulation time in TIP4P (right) and TIP5P (left) water.



**Figure 9.** Distribution of the C4–C5 torsion angle in TIP4P (blue, left) and TIP5P (green, right) water.

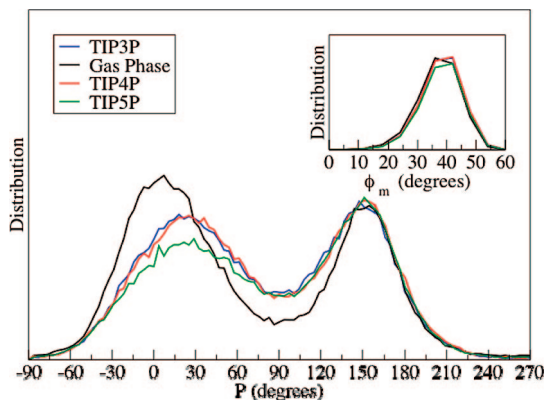


**Figure 10.** A summary of the distribution of rotamer populations about the C4–C5 bond of **3** in TIP3P water, gas phase, TIP4P water, and TIP5P water compared to experiment.

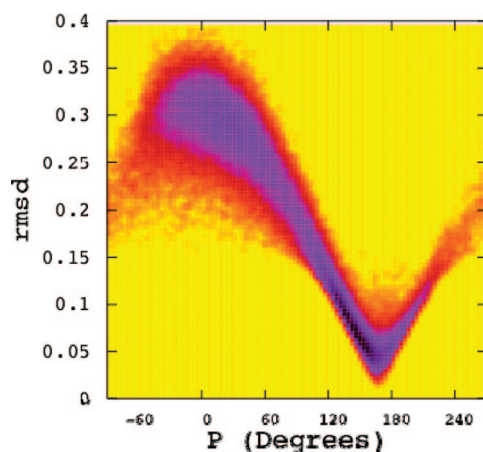
sophisticated water models is insufficient to represent the solvation of this particular system. Moreover, it was surprising to observe that this outcome is unique when compared to the  $\alpha$  system **2**<sup>33</sup> where TIP3P was sufficient for good agreement with experiment.

**Ring Conformer Populations.** Having successfully identified simulation parameters that provided C4–C5 rotamer populations in good agreement with experiment, our attention

turned to ring parameters in **3**,  $P$  and  $\Phi_{\max}$ . Figure 11 presents the variation in  $P$ , which describes the ring puckering, in all simulations; the inset shows the variation in puckering amplitude,  $\Phi_{\max}$ . The MD simulations correctly predict the two-state model generally used to model the solution conformation for furanoside ring systems. From the 200 ns trajectory, 55% of the conformations are present in the northern hemisphere of the pseudorotational itinerary (see



**Figure 11.** The distribution of the pseudorotational phase angle ( $P$ ) for **3** in the gas phase, TIP3P water, TIP4P, and TIP5P water. The distribution of puckering amplitude,  $\Phi_{\max}$ , is given in the inset.



**Figure 12.** Joint probability distribution of the puckering angle (in degrees),  $P$ , and the rmsd (in Å) of the ring carbon atoms.

Figure 1), adopting pseudorotational phase angles of  $-90^\circ$  to  $60^\circ$ . The remaining 45% of conformations exist in a range of  $P$  values from 120 to  $240^\circ$ . These magnitudes of the populations of these conformers do not correspond well with the experimental values ( $P_N = -7^\circ$ , 86%;  $P_S = 162^\circ$ , 14%) determined by PSEUROT.<sup>12</sup> However, the area of conformational space in the southern hemisphere is centered about  $P = 160^\circ$  and in the northern hemisphere about  $P \approx 10^\circ$ , which corresponds well respectively to the S and N conformer determined for **3** experimentally. The distribution in  $\Phi_{\max}$  is centered about  $39^\circ$  in all simulations, and this corresponds well to earlier calculations on **3** as well as to the puckering parameters of this molecule in the crystal structure.<sup>40</sup>

The discrepancy in the pseudorotational phase angles could possibly arise from our modification of the standard GLYCAM approach to derive the partial atomic charges. Furanosides are flexible molecules, and thus, a single conformation cannot solely be used for the charge derivation. Rather, charges must be averaged over a number of rings to better represent all the conformations accessible to the system.<sup>33</sup> Figure 12 illustrates the correlation between the rmsd of the ring atoms and the distribution of the puckering angle,  $P$ . The graph indicates that an rmsd of 0.17 Å as obtained in the

ring-averaged charge derivation procedure corresponds to a change of more than  $100^\circ$  in  $P$ . Fluctuations in the rmsd result in even greater changes in the ring puckering. This result is not unexpected, as previous simulations on **2** also yielded similar variations in  $P$ . Unlike our previous results, however, simulations on **3** suggested that the two-state conformational model for assessing ring conformation using PSEUROT is valid. Therefore, another possible source of discrepancy in  $P$  compared to experiment is that the generalized Karplus equation used in the PSEUROT analysis of **3** may not be accurate for this ring system. This could be circumvented by the calculation of Karplus curves specific for each coupling pathway in  $\beta$ -arabinofuranoside rings, using theoretical methods.<sup>47,48</sup> Moreover, access to accurate Karplus curves tailored to **3** would allow a direct comparison of ring  $^3J_{\text{H,H}}$  obtained from NMR spectroscopy, with those obtained from the conformational ensemble generated by AMBER/GLYCAM MD simulations. Such an approach may circumvent possible sources of error entailed by the PSEUROT approach. It appears that Karplus relationships more specific to **3** are essential. An attempt to calculate  $^3J_{\text{H,H}}$  of **3** from the conformer distribution provided by the MD simulations using the generalized Karplus equations<sup>30</sup> led to poor agreement with experiment (data not shown).

## Conclusions

In this paper, the combined AMBER/GLYCAM forcefield model was applied to simulations of methyl  $\beta$ -D-arabinofuranoside (**3**). A recently developed method for the calculation of atomic charges was used to take into account the flexibility of the furanose ring in **3**. Initial simulations with TIP3P water model yielded C4–C5 rotamer populations that were not in good agreement with experimental data, and hydrogen bond analysis as well as gas-phase simulations suggested that more sophisticated water models were required for the proper representation of the solvation of **3**. The TIP4P and TIP5P water models were then employed, and both demonstrated good agreement of the C4–C5 rotamer distribution with the experimental results. Analysis of the ring puckering showed that although the magnitude of populations of the two low energy conformations are different than experiment, good agreement was obtained with respect to the identity of each conformer as well as the puckering amplitude ( $\Phi_{\max}$ ). The discrepancy in puckering results may be attributed to the charge derivation procedure or an inaccurate Karplus relationship that was used in the analysis of experimental  $^3J_{\text{H,H}}$  data. Having successfully applied the AMBER/GLYCAM approach to investigate the conformation of **3**, this method will be used in the conformational studies of larger furanose containing oligosaccharides, such as **1**. In addition, novel Karplus equations for each  $^1\text{H}$ – $^1\text{H}$  coupling fragment in **3** are currently being developed using theoretical methods.

**Acknowledgment.** This work was supported by the Alberta Ingenuity Centre for Carbohydrate Science and the Natural Sciences and Engineering Research Council of Canada.

**Supporting Information Available:** Calculated partial atomic charges, with standard deviations; MD simulations convergence criteria, with errors in rotamer populations; coupling constants for **3**, compared to previously reported values; values for the most probable dihedral angles in each simulation; and comparison of MD rotamers and experimental rotamers in each case. This material is available free of charge via the Internet at <http://pubs.acs.org>.

### References

- Bloom, B. B.; Murray, C. L. *J. Science* **1992**, *257*, 1055–1064.
- Zhang, Y. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 529–564.
- Hamasur, B.; Källenius, G.; Svenson, S. B. *Vaccine* **1999**, *17*, 2853–2861.
- Dietrich, J.; Lundberg, C. V.; Andersen, P. *Tuberculosis* **2006**, *86*, 163–168.
- Williams, D. H. *Nat. Prod. Rep.* **1996**, *13*, 469–477.
- Gadebusch, H. H.; Stapley, E. O.; Zimmerman, S. B. *Crit. Rev. Biotechnol.* **1992**, *12*, 225–243.
- Woodin, K. A.; Morrison, S. H. *Pediatr. Rev.* **1994**, *15*, 440–447.
- Janin, Y. L. *Bioorg. Med. Chem.* **2007**, *15*, 2479–2513.
- Lowary, T. L. *Mini Rev. Med. Chem.* **2003**, *3*, 689–702.
- Barry, C. E. *Biochem. Pharmacol.* **1997**, *54*, 1165–1172.
- Brennan, P. J.; Nikaido, H. *Annu. Rev. Biochem.* **1995**, *64*, 29–63.
- Houseknecht, J. B.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2003**, *107*, 5763–5777.
- Connell, N. D.; Nikaido, H. In *Tuberculosis: Pathogenesis, Protection and Control*; Bloom, B. R., Ed.; American Society for Microbiology: Washington, DC, 1994; pp 333–352.
- Draper, P., Daffé, M. The Cell Envelope of *Mycobacterium tuberculosis* with special reference to the capsule and outer permeability barrier. In *Tuberculosis and the Tubercle Bacillus*; Cole, S. T., Eisenach, K. D., McMurray, D. N., Jacobs, W. R., Jr., Eds.; American Society for Microbiology: Washington, DC, 2005; pp 261–273.
- Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Am. Chem. Soc.* **1999**, *121*, 9682–9692.
- McCarren, P. R.; Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2001**, *105*, 5911–5922.
- Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Org. Chem.* **2000**, *65*, 4954–4963.
- Houseknecht, J. B.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2003**, *107*, 372–378.
- Houseknecht, J. B.; McCarren, P. R.; Lowary, T. L.; Hadad, C. M. *J. Am. Chem. Soc.* **2001**, *123*, 8811–8824.
- Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* **1972**, *94*, 8205–8212.
- PSEUROT, version 6.2*; Gorlaeus Laboratories, University of Leiden: Leiden, NL, 1995.
- Deleeuw, F.; Altona, C. *J. Comput. Chem.* **1983**, *4*, 428–437.
- Plavec, J.; Tong, W.; Chattopadhyaya, J. *J. Am. Chem. Soc.* **1993**, *115*, 9734–9746.
- Plavec, J.; Thibaudeau, C.; Chattopadhyaya, J. *Pure Appl. Chem.* **1996**, *11*, 2137–2144.
- D'Souza, F. W.; Ayers, J. D.; McCarren, P. R.; Lowary, T. L. *J. Am. Chem. Soc.* **2000**, *122*, 1251–1260.
- Callam, C. S.; Lowary, T. L. *J. Org. Chem.* **2001**, *66*, 8961–8972.
- Joe, M.; Sun, D.; Taha, H.; Completo, G. C.; Croudace, J. E.; Lammas, D. A.; Besra, G. S.; Lowary, T. L. *J. Am. Chem. Soc.* **2006**, *128*, 5059–5072.
- Wu, G. D.; Serianni, A. S.; Barker, R. *J. Org. Chem.* **1983**, *48*, 1750–1757.
- Altona, C.; Ippel, J. H.; Hoekzema, A. J. A. W.; Erkelens, C.; Groesbeek, M.; Donders, L. A. *Magn. Reson. Chem.* **1989**, *27*, 564–576.
- Altona, C.; Francke, R.; de Haan, R.; Ippel, J. H.; Daalman, G. J.; Hoekzema, J. A. W.; van Wijk, J. *Magn. Reson. Chem.* **1994**, *32*, 670–678.
- Nigou, M.; Gilleron, M.; Puzo, G. *Biochimie* **2003**, *85*, 153–166.
- Rademacher, C.; Shoemaker, G. K.; Kim, H. S.; Zheng, R. B.; Taha, H.; Liu, C.; Nacario, R. C.; Schriemer, D. C.; Klassen, J. S.; Peters, T.; Lowary, T. L. *J. Am. Chem. Soc.* **2007**, *129*, 10489–10502.
- Seo, M.; Castillo, N.; Ganzynkowicz, R.; Daniels, C. R.; Woods, R. J.; Lowary, T. L.; Roy, P.-N. *J. Chem. Theory Comput.* **2008**, *4*, 184–191, and appropriate references within.
- Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541–10545.
- Gonzalez-Outeiriño, J.; Kirschner, K. N.; Thobhani, S.; Woods, R. J. *Can. J. Chem.* **2006**, *84*, 569–579.
- Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22*, 1125–1137.
- 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*, 1668–1688.
- Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraserreid, B. *J. Phys. Chem.* **1995**, *99*, 3832–3846.
- Gaussian 03, Revision C.02*; 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, Inc.*: Wallingford, CT, 2004.
- Evdokimov, A.; Gilboa, A. J.; Koetzle, T. F.; Klooster, W. T.; Schultz, A. J.; Mason, S. A.; Albinati, A.; Frolow, F. *Acta Crystallogr., Sect. B: Struct. Sci.* **2001**, *57*, 213–220.

- (41) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (42) Woods, R. J.; Chappelle, R. *J. Mol. Struct.* **2000**, *527*, 149–156.
- (43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (44) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (45) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.
- (46) Wallqvist, A.; Mountain, R. D. *Rev. Comput. Chem.* **1999**, *13*, 183–247.
- (47) Stenutz, R.; Carmichael, I.; Widmalm, G.; Serianni, A. S. *J. Org. Chem.* **2002**, *67*, 949–958.
- (48) Zhao, H.; Pan, Q.; Zhang, W.; Carmichael, I.; Serianni, A. S. *J. Org. Chem.* **2007**, *72*, 7071–7082.

CT800384H